IJCLA **ISSN 0976-0962**

International Journal of Computational Linguistics and Applications

Vol. 1 No. 1-2 Jan-Dec 2010

Editor-in-Chief
Alexander Gelbukh

ISSN 0976-0962

International Journal of Computational Linguistics and Applications

Vol. 1 No. 1-2 Jan-Dec 2010

International Journal of Computational Linguistics and Applications – IJCLA (started in 2010) is a peer-reviewed international journal published twice a year, in March and September. It publishes original research papers related to computational linguistics, natural language processing, human language technologies and their applications.

The views expressed herein are those of the authors. *International Journal of Computational Linguistics and Applications* reserves the right to edit the material.

Subscription: India: Rs. 500
Rest of the world: US$ 120

Payments can be made by Cheques/Bank Drafts/International Money Orders drawn in the name of BAHRI PUBLICATIONS, NEW DELHI and sent to:

BAHRI PUBLICATIONS
1749A/5, 1st Floor, Gobindpuri Extension,
P. O. Box 4453, Kalkaji, New Delhi 110019
Telephones: 011-65810766, (0) 9811204673, (0) 9212794543
E-mail: bahrius@vsnl.com; bahripublications@yahoo.com
Website: http://www.bahripublications.com

Printed & Published by Deepinder Singh Bahri, for and on behalf of **BAHRI PUBLICATIONS**, New Delhi.

International Journal of Computational Linguistics and Applications

Vol. 1 **No. 1-2** **Jan-Dec 2010**

CONTENTS

Computational Semantics

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 9-26
RECEIVED 23/11/09 ACCEPTED 16/01/10 FINAL 11/03/10

Exploring the Lexical Semantics of Dialogue Acts

NICOLE NOVIELLI[1] AND CARLO STRAPPARAVA[2]

[1] *Università degli Studi di Bari, Italy*
[2] *FBK-irst, Italy*

ABSTRACT

People proceed in their conversations through a series of dialogue acts to yield some specific communicative intention. In this paper, we study the task of automatic labeling dialogues with the proper dialogue acts, relying on empirical methods and simply exploiting lexical semantics of the utterances. In particular, we present some experiments in both a supervised and an unsupervised framework on an English and an Italian corpus of dialogue transcriptions. In the experiments we consider the settings of dealing with or without additional information from the dialogue structure. The evaluation displays good results, regardless of the used language. We conclude the paper exploring the relation between the communicative goal of an utterance and its affective content.

1 INTRODUCTION

When engaged in dialogues, people ask for information, agree with their partner, state some facts and express opinions. They proceed in their conversations through a series of dialogue acts to yield some particular communicative intention.

Dialogue Acts (DA) have been well studied in linguistics [1,2] and attracted computational linguistics research for a long time [3,4]. There is a large number of application domains that can benefit from the automatic extraction of the underlying structure of dialogues: dialogue systems for human-computer interaction, conversational agents for monitoring and supporting human-human conversations forums and chat logs analysis

for opinion mining, affective state recognition by mean of dialogue pattern analysis, automatic meeting summarization and so on. This kind of applications requires a deep understanding of the conversational structure and dynamic evolution of the dialogue: at every step of the interaction the system should be able to understand who is telling what to whom. With the advent of the Web, a large amount of material about natural language interactions (e.g. blogs, chats, conversation transcripts) has become available, raising the attractiveness of empirical methods of analyses on this field.

In this paper, we study the task of automatic labeling dialogues with the proper speech acts. We define a method for DA recognition by relying on empirical methods that simply exploit lexical semantics of the sentences. Even if prosody and intonation surely play a role (e.g. [5,6]), nonetheless language and words are what the speaker uses to convey the communicative message and are just what we have at disposal when we consider texts found on the Web.

We present some experiments in a supervised and unsupervised framework on both an English and an Italian corpus of dialogue transcriptions. In particular we consider the classification of dialogue acts with and without taking into account dialogue contextual features. We achieved good results in all settings, independently from the used language. Finally, we explore the relation between the communicative goal of an utterance and its affective content, using a technique [7] for checking the emotional load in a text.

The paper is organized as follows. Section 2 gives a brief sketch of the NLP background on Dialogue Act recognition. In Section 3 we introduce the English and Italian corpora of dialogues, their characteristics, DA labeling and preprocessing. Then, Section 4 explains the supervised and unsupervised settings, showing the experimental results obtained on the two corpora and providing detailed results and error analysis. In Section 5 we presents the results considering also dialogue contextual features. Section 6 describes the preliminary results of a qualitative study about the relation between the dialogue acts and their affective load. Finally, in Section 7 we conclude the paper with a brief discussion and some directions for future work.

2 BACKGROUND

A DA can be identified with the communicative goal of a given utterance [1]. Researchers use different labels and definitions to address the com-

Table 1. An excerpt from the Switchboard corpus

Speaker	Dialogue Act	Utterance
A	OPENING	*Hello Ann.*
B	OPENING	Hello Chuck.
A	STATEMENT	*Uh, the other day, I attended a conference here at Utah State University on recycling*
A	STATEMENT	*and, uh, I was kind of interested to hear cause they had some people from the EPA and lots of different places, and, uh, there is going to be a real problem on solid waste.*
B	OPINION	Uh, I didn't think that was a new revelation.
A	AGREE /ACCEPT	*Well, it's not too new.*
B	INFO-REQUEST	So what is the EPA recommending now?

municative goal of a sentence: Searle [2] talks about *speech act*; Schegloff [8] and Sacks [9] refer to the concept of *adjacency pair part*; Power [10] adopts the definition of *game move*; Cohen and Levesque [11] focus more on the role speech acts play in interagent communication.

Traditionally, the NLP community has employed DA definitions with the drawback of being domain or application oriented. In the recent years some efforts have been made towards unifying the DA annotation [4]. In the present study we refer to a domain-independent framework for DA annotation, the DAMSL architecture (Dialogue Act Markup in Several Layers) by Core and Allen [3].

Recently, the problem of DA recognition has been addressed with promising results. Stolcke et al. [5] achieve an accuracy of around 70% and 65% respectively on transcribed and recognized words by combining a discourse grammar, formalized in terms of Hidden Markov Models, with evidences about lexicon and prosody. Reithinger and Klesen's approach [12] employs a bayesian approach achieving 74.7% of correctly classified labels. A partially supervised framework by Venkataraman et al. [13] has also been explored, using five broad classes of DA and obtaining an accuracy of about 79%. Regardless of the model they use (discourse grammars, models based on word sequences or on the acoustic features or a combination of all these) the mentioned studies are developed in a supervised framework. Rather than improving the performance

of supervised frameworks, our main goal is to explore the use of an unsupervised methodology.

3 DATA SETS

Table 2. The set of labels employed for Dialogue Act

Label	**Description and Examples**	**Italian**	**English**
INFO-REQUEST	Utterances that are pragmatically, semantically, and syntactically questions - *'What did you do when your kids were growing up?'*	34%	7%
STATEMENT	Descriptive, narrative, personal statements - *'I usually eat a lot of fruit'*	37%	57%
S-OPINION	Directed opinion statements - *'I think he deserves it.'*	6%	20%
AGREE-ACCEPT	Acceptance of a proposal, plan or opinion - *'That's right'*	5%	9%
REJECT	Disagreement with a proposal, plan, or opinion - *'I'm sorry no'*	7%	.3%
OPENING	Dialogue opening or self-introduction - *'Hello, my name is Imma'*	2%	.2%
CLOSING	Dialogue closing (e.g. farewell and wishes) - *'It's been nice talking to you.'*	2%	2%
KIND-ATT	Kind attitude (e.g. thanking and apology) - *'Thank you.'*	9%	.1%
GEN-ANS	Generic answers to an Info-Request - *'Yes'*, *'No'*, *'I don't know'*	4%	4%
total cases		1448	131,265

In the experiments described in this paper we exploit two corpora, both annotated with Dialogue Acts labels. We aim at developing a recognition methodology as much general as possible, so we selected corpora that differ in the content and in the used language: the Switchboard corpus [14] of English telephone conversations about general interest topics, and an Italian corpus of dialogues in the healthy-eating domain [15].

The Switchboard corpus is a collection of transcripts of English human-human telephone conversations [14] involving couples of randomly se-

lected strangers: they were asked to select one general interest topic and to talk informally about it. Full transcripts of these dialogues are distributed by the Linguistic Data Consortium. A part of this corpus is annotated [16] with DA labels (overall 1155 conversations, for a total of 205,000 utterances and 1.4 million words) [3]. Tables 1 shows a short sample fragment of dialogue from this corpus.

The Italian corpus had been collected in the scope of some previous research about Human-ECA (Embodied Conversational Agent) interaction: to collect these data a Wizard of Oz tool was employed [15] in which the application domain and the ECA's appearance may be settled at the beginning of simulation. During the interaction, the ECA played the role of an artificial therapist and the users were free to interact with it in natural language, without any particular constraint. This corpus is about healthy eating and contains overall 60 dialogues, 1448 users' utterances and 15,500 words.

Labelling. The two corpora are annotated in order to capture the communicative intention of each dialogue move. Defining a DA markup language is out of the scope of the present study, hence we employed the original annotation of the two corpora [17,16], which is consistent, in both cases, with the Dialogue Act Markup in Several Layers (DAMSL) scheme [3]. In particular the Switchboard corpus employs the SWBD-DAMSL revision [16].[4]

Table 2 shows the set of labels employed for the purpose of this study, with definitions and examples: it maintains the DAMSL main characteristic of being domain-independent and it is also consistent with the original semantics of the SWBD-DAMSL markup language employed in the Switchboard annotation. As shown in Table 3, the SWBD-DAMSL had been automatically converted into the categories included in our markup language. Also we did not consider the utterances formed only by non-verbal material (e.g. laughter). The DA label distribution and the total number of cases (utterances) considered in the two data sets are reported in Table 2.

[3] `ftp.ldc.upenn.edu/pub/ldc/public_data/swb1_dialogact_annot.tar.gz`

[4] The SWBD-DAMSL modifies the original DAMSL framework by further specifying some categories or by adding extra (mainly prosodic) features, which were not originally included in the scheme.

Table 3. The Dialogue Act set of labels with their mapping with the SWBD-DAMSL correspondent categories

Label	**SWBD-DAMSL**
INFO-REQ	*Yes-No question* (*qy*), *Wh-Question* (qw), *Declarative Yes-No-Question* (qyˆd), *Declarative Wh-Question* (qwˆd), *Alternative ('or') question* (qr) *and OR-clause* (qrr) , *Open-Question* (qo), *Declarative* (ˆd) *and Tag questions* (ˆg)
STATEMENT	*Statement-non-opinion* (sd)
S-OPINION	*Statement-opinion* (sv)
AGREE-ACC	*Agreement /accept* (aa)
REJECT	*Agreeement /reject* (ar)
OPENING	*Conventional-opening* (fp)
CLOSING	*Conventional-closing* (fc)
KIND-ATT	*Thanking* (ft) *and Apology* (fa)
GEN-ANS	*Yes answers* (ny), *No answers* (nn), *Affirmative non-yes answers* (na) *Negative non-no answers* (ng)

Data preprocessing. To reduce the data sparseness, we used a POS-tagger and morphological analyzer [18] for preprocessing the corpora and we used lemmata instead of tokens. No feature selection was performed, keeping also stopwords. In addition, we augment the features of each sentence with a set of linguistic markers, defined according to the semantics of the DA categories. We hypothesize, in fact, these features could play an important role in defining the linguistic profile of each DA. The addition of these markers is performed automatically, by just exploiting the output of the POS-tagger and of the morphological analyzer, according to the following rules:

- **WH-QTN**, used whenever an interrogative determiner is found, according to the output of the POS-tagger (e.g. 'when' does not play an interrogative role when tagged as conjunction);
- **ASK-IF**, used whenever an utterance presents some cues of the pattern 'Yes/No' question. ASK-IF and WH-QTN markers are supposed to be relevant for the recognition of the INFO-REQUEST category;
- **I-PERS**, used for all declarative utterance whenever a verb is in the first person form, singular or plural (relevant for the STATEMENT);
- **COND**, used when a conditional form is detected.
- **SUPER**, used for superlative adjectives;

– **AGR-EX**, used whenever an agreement expression (e.g. 'You are right', 'I agree') is detected (relevant for AGREE-ACCEPT);
– **NAME**, used whenever a proper name follows a self-introduction expression (e.g. 'My name is') (relevant for the OPENING);
– **OR-CLAUSE**, used when the utterance is an or-clause, i.e. it starts with the conjunction 'or' (should be helpful for the characterization of the INFO-REQUEST);
– **VB**, used only for the Italian, it is when a dialectal form of agreement is detected.

4 MINIMALLY SUPERVISED DIALOGUE ACT RECOGNITION

It is not always easy to have large training material at disposal, partly because of manual labeling effort and moreover because often it is not possible to find it. Schematically, our unsupervised methodology consists of the following steps: (i) building a semantic similarity space in which words, set of words, text fragments can be represented homogeneously, (ii) finding seeds that properly represent dialogue acts and considering their representations in the similarity space, and (iii) checking the similarity of the utterances.

To get a similarity space with the required characteristics, we used Latent Semantic Analysis (LSA). LSA is a corpus-based measure of semantic similarity proposed by Landauer [19]. In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a singular value decomposition (SVD) on the term-by-document matrix $\mathbf{T}$ representing the corpus.

SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. In our case, SVD decomposes the term-by-document matrix $\mathbf{T}$ into three matrices $\mathbf{T} = \mathbf{U}\mathbf{\Sigma_k}\mathbf{V}^T$ where $\mathbf{\Sigma_k}$ is the diagonal $k \times k$ matrix containing the k singular values of $\mathbf{T}$, $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_k$, and $\mathbf{U}$ and $\mathbf{V}$ are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is recomposed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\mathbf{\Sigma_{k'}}\mathbf{V}^T$.

LSA can be viewed as a way to overcome some of the drawbacks of the standard vector space model (sparseness and high dimensionality). In fact, the LSA similarity is computed in a lower dimensional space, in which second-order relations among terms and texts are exploited. The

similarity in the resulting vector space is then measured with the standard cosine similarity. Note also that LSA yields a vector space model that allows for a *homogeneous* representation (and hence comparison) of words, sentences, and texts. For representing a word set or a sentence in the LSA space we use the *pseudo-document* representation technique, as described by Berry [20]. In practice, each text segment is represented in the LSA space by summing up the normalized LSA vectors of all the constituent words, using also a *tf.idf* weighting scheme [21].

Table 4. The complete sets of seeds for the unsupervised experiment

Label	**Seeds**
INFO-REQ	WH-QTN, '?', ASK-IF
STATEMENT	I-PERS, I
S-OPINION	Verbs which directly express opinion or evaluation (guess, think, suppose)
AGREE-ACC	AGR-EX, yep, yeah, absolutely, correct
REJECT	Verbs which directly express disagreement (disagree, refute)
OPENING	Expressions of greetings (hi, hello), words and markers related to self-introduction formula (name, NAME)
CLOSING	Interjections/exclamations ending discourse (alright, okey, '!'), Expressions of thanking (thank) and farewell (bye, bye-bye, goodnight)
KIND-ATT	Lexicon which directly expresses wishes (wish), apologies (apologize), thanking (thank) and sorry-for (sorry, excuse)
GEN-ANS	no, yes, uh-huh, nope

The methodology is unsupervised[5] as we do not exploit any 'labeled' training material. For the experiments reported in this paper, we run the SVD using 400 dimensions (i.e. k') respectively on the English and Italian corpus, without any DA label information. Starting from a set of seeds (words) representing the communicative acts, we build the corresponding vectors in the LSA space and then we compare the utterances to find the communicative act with the highest similarity.

[5] Or minimally supervised, since providing hand-specified seeds can be regarded as a minimal sort of supervision.

Table 5. Evaluation of the supervised and unsupervised methods on the two corpora

	Italian						English					
	SVM			LSA			SVM			LSA		
Label	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1
INFO-REQ	.92	.99	.95	.96	.88	.92	.92	.84	.88	.93	.70	.80
STATEMENT	.85	.68	.69	.76	.66	.71	.79	.92	.85	.70	.95	.81
S-OPINION	.28	.42	.33	.24	.42	.30	.66	.44	.53	.41	.07	.12
AGREE-ACC	.50	.80	.62	.56	.50	.53	.69	.74	.71	.68	.63	.65
REJECT	-	-	-	.09	.25	.13	-	-	-	.01	.01	.01
OPENING	.60	1.00	.75	.55	1.00	.71	.96	.55	.70	.20	.43	.27
CLOSING	.67	.40	.50	.25	.40	.31	.83	.59	.69	.76	.34	.47
KIND-ATT	.82	.53	.64	.43	.18	.25	.85	.34	.49	.09	.47	.15
GEN-ANS	.20	.63	.30	.27	.38	.32	.56	.25	.35	.54	.33	.41
micro	.71	.71	.71	.66	.66	.66	.77	.77	.77	.68	.68	.68

Table 4 shows the complete sets of seeds used for building the vector of each DA. We defined seeds by only considering the communicative goal and the specific semantics of every single DA, just avoiding the overlapping between seed groups as much as possible. We wanted to design an approach which is as general as possible, so we did not consider domain words that would have made easier the classification in the specific corpora. The seeds are the same for both languages, which is coherent with our goal of defining a language-independent method. There are only a few exceptions: in Italian it is not necessary to specify the pronoun when formulating a sentence so we did not include the 'I' equivalent pronoun in the seeds for the STATEMENT label; the VB linguistic marker is used only for the Italian and is included in the seeds for the S-OPINION vector.

An upper-bound performance is provided by running experiment in a supervised framework. We used Support Vector Machines [22], in particular SVM-light package [23] under its standard configuration. We randomly split the two corpora into 80/20 training/test partitions. SVMs have been used in a large range of problems, including text classification, image recognition tasks, bioinformatics and medical applications, and they are regarded as the state-of-the-art in supervised learning. To allow comparison, the performance is measured on the same test set partition for both the unsupervised and supervised experiments.

4.1 *Experimental Results and Discussion*

We evaluated the performance of our method in terms of precision, recall and F1-measure (see Table 5) according to the DA labels given by annotators in the datasets. As baselines we can consider (i) most-frequent label assignment (respectively 37% for Italian, 57% for English) for the supervised setting, and (ii) random DA selection (11%) for the unsupervised one.

We got .71 and .77 of F1 respectively for the Italian and the English corpus in the supervised condition, and .66 and .68 respecitvely in the unsupervised one. The performance is quite satisfying and is comparable to the state of the art in the domain. In particular, the unsupervised technique is significantly above the baseline, for both the Italian and the English corpus experiments. We note that the methodology is independent from the language and the domain: the Italian corpus is a collection of dialogue about a very restricted domain (advice-giving dialogue about healthy-eating) while in the Switchboard corpus the conversations revolve around general topics chosen by the two interlocutors. Moreover, in the unsupervised setting we use the same seed definitions. Secondly, it is independent on the differences in the linguistic style due to the specific interaction scenario and input modality. Finally, the performance is not affected by the difference in size of the two data sets.

Error analysis. After conducting an error analysis, we noted that many utterances are misclassified as STATEMENT. One possible reason is that statements usually are quite long and there is a high chance that some linguistic markers that characterize other dialogue acts are present in those sentences too. On the other hand, looking at the corpora we observed that many utterances that appear to be linguistically consistent with the typical structure of statements have been annotated differently, according to the actual communicative role they play. The following is an example of a statement-like utterance (by speaker B) that has been annotated differently because of its context (speaker A's move):

A: 'In fact, it's easier for me to say, uh, the types of music that I don't like are opera and, uh, screaming heavy metal.' STATEMENT
B: 'The opera, yeah, it's right on track.' AGREE-ACCEPT

For similar reasons, we observed some misclassification of S-OPINION as STATEMENT. The only significative difference between the two labels seems to be the wider usage of 'slanted' and affectively loaded lexicon

when conveying an opinion. Another source of confounding is the misclassification of the OPENING as INFO-REQUEST. The reason is not clear yet, since the misclassified openings are not question-like. Eventually, there is some confusion among the backchannel labels (GEN-ANS, AGREE-ACC and REJECT) due to the inherent ambiguity of common words like *yes*, *no*, *yeah*, *ok*.

Recognition of such cases could be improved (i) by enabling the classifiers to consider not only the lexical semantics of the given utterance but also the knowledge about a wider context window (e.g. the previous *n* utterances), (ii) by enriching the data preprocessing (e.g. by exploiting information about lexicon polarity and subjectivity parameters).

5 Exploiting Contextual Features

The findings in Section 4.1 highlight the role played by the context in determining the actual communicative goal of a given dialogue turn: manual annotation of utterances is shown to depend not only on the linguistic realization itself. On the contrary, the knowledge about the dialogue history constitutes a bias for human annotators.

This is consistent with Levinson's theory of conversational analysis. Both local and global contextual information contribute in defining the communicative intention of a dialogue turn [24]. In this perspective, top-down expectation about the next likely dialogue act and bottom-up information (i.e. the actual words used in the utterance or its acoustic and prosodic parameters) should be combined to achieve better performance in automatic DA prediction.

Stolcke et al. [5] propose an approach that combines HMM discourse modeling with consideration of linguistic and acoustic features extracted from the dialogue turn. Poesio and Mikheev [25] exploit the hierarchical structure of discourse, described in terms of game structure, to improve DA classification in spoken interaction. Reithinger and Klesen [12] employ a Bayesian approach to build a probabilistic dialogue act classifier based on textual input.

In this section we present some experiments that exploit knowledge about dialogue history. In our approach, each utterance is enriched with contextual information (i.e. the preceding DA labels) in form of either 'bag_of_words' or 'n-grams'. We explore the supervised learning framework, using SVM, under five different experimental settings. Then, we propose a bootstrap approach for the unsupervised setting. In order to al-

low comparison with the results in Section 4 we refer, for both languages, to the same train/test partitions employed in our previous experiments.

Supervised. We have tested the role played by the context in DA recognition, experimenting with: (i) the number of turn (one vs. two turns) considered in extracting contextual features (i.e. DA labels) based on the dialogue history of a given turn and (ii) the approach used for representing the knowledge about the context, i.e. Bag_of_Words style (BoW) vs. n-grams.

Data preprocessing involves enriching both, the train and test sets, with contextual information, as shown in Table 6. When building the context for a given utterance we only consider the label included in our DA annotation language (see Table 2). In fact, our markup language does not allow mapping of SWBD-DAMSL labels such as 'non verbal turn' or 'abandoned turn'. According to our goal of defining a method which simply exploits textual information, we consider all cases originally annotated with such labels as a lack of knowledge about the context.

Table 6. Enriching the data set with contextual features

natural language input:		
(a1)	STATEMENT	'I don't feel comfortable about leaving my kids in a big day care center'
(b1)	INFO-REQ	'Worried that they're not going to get enough attention?'
(a2)	GEN-ANS	'Yeah'
correspondent dataset item for the utterance a2:		
BoW	STATEMENT:1 INFO-REQUEST:1 yeah:1	
Bigram	STATEMENT&INFO-REQUEST:1 yeah:1	

Table 7 (a) shows the results in terms of precision, recall and F1-measure. As comparison, we also report the global performance when no context features are used in the supervised setting. For both the Italian and English corpora, bigrams seem to best capture the dialogue structure. In particular, using a BoW style seems to even lower the performance with respect to the setting in which no information about the context is exploited. Neither combining bigrams with Bag_of_Words nor using higher-order n-gram improve the performance.

Table 7. Overall performance of the different approaches for exploiting contextual information in the supervised setting (a) and bootstrap on the unsupervised method (b)

	English		
Experimental Setting	prec	rec	F1
no context	*.77*	*.77*	*.77*
1 turn of context	.49	.49	.49
BoW (2 turns)	.76	.76	.76
Bigrams (2 turns)	.83	.83	.83
BoW + Bigrams (2 turns)	.83	.83	.83
	Italian		
no context	*.71*	*.71*	*.71*
Bigrams (2 turns)	.82	.82	.82

(a)

	English		
Experimental Setting	prec	rec	F1
no context	*.68*	*.68*	*.68*
Bigrams (2 turns)	.70	.70	.70
	Italian		
no context	*.66*	*.66*	*.66*
Bigrams (2 turns)	.72	.72	.72

(b)

Unsupervised. According to the results in the previous section, we decided to investigate the use of bigrams in the unsupervised learning condition using a bootstrap approach. Our bootstrap procedure is composed by the following steps: (i) annotating the English and Italian corpora using the unsupervised approach described in Section 4; (ii) using the result of this unsupervised annotation for extracting knowledge about contextual information for each utterance: each item in the data sets is then enriched with the appropriate bigram, as shown in Table 6; (iii) training an SVM classifier on the bootstrap data enriched with bigrams. Then performance is evaluated on the test sets (see Table 7 (b)) according to the actual label given by human annotators.

6 Affective Load of Dialogue Acts

Sensing emotions from text is a particularly appealing task of natural language processing [26,27]: the automatic recognition of affective states is becoming a fundamental issue in several domains such as human-computer interaction or sentiment analysis for opinion mining. Recently there have been several attempts to integrate emotional intelligence into user interfaces [28,29,15]. A first attempt to exploit affective information in dialogue act disambiguation has been made by Bosma and André [30], with promising results. In their study, the recognition of emotions is based on sensory inputs which evaluate physiological user input.

In this section we present some preliminary results of a qualitative study aimed at investigating the affective load of DAs. To the best of our knowledge, this is the first attempt to study the relation between the communicative act of an utterance and its affective load by applying lexical similarity techniques to textual input.

We calculate the affective load of each DA label using the methodology described in [7]. The idea underlying the method is the distinction between *direct* and *indirect* affective words. For direct affective words, authors refer to the WordNet Affect [31] lexicon, an extension of the WordNet database [32] which employs six basic emotion labels (anger, digust, fear, joy, sadness, surprise) to annotate WordNet synsets. LSA is then used to learn, in an unsupervised setting, a vector space from the British National Corpus[6]. As said before, LSA has the advantage of allowing homogeneous representation and comparison of words, text fragments or entire documents, using the pseudo-document technique exploited in Section 4. In the LSA space, each emotion label can be represented in various way. In particular, we employ the 'LSA Emotion Synset' setting, in which the synsets of direct emotion words are considered. The affective load of a given utterance is calculated in terms its lexical similarity with respect to one of the six emotion labels. The overall affective load of a sentence is then calculated as the average of its similarity with each emotion label.

Results are shown in Table 8 (a) and confirm our preliminary hypothesis (see error analysis in Section 4.1) about the use of slanted lexicon in opinions. In fact, S-OPINION is the DA category with the highest affective load. Opinions are immediately followed by KIND-ATT due to the high frequency of politeness formulas in such utterances (see Table 8 (b)).

7 CONCLUSIONS AND FUTURE WORK

The long-term goal of our research is to define an unsupervised method for Dialogue Acts recognition. The techniques employed have to be independent from some important features of the corpus used such as domain, language, size, interaction scenario.

In this study we propose a method that simply exploits the lexical semantics of dialogue turns. In particular we consider DA classification with and without considering contextual features. The methodology starts

[6] `http://www.hcu.ox.ac.uk/bnc/`

Table 8. Affective load of DA labels (a) and examples of slanted lexicon (b)

Label	Affective Load
S-OPINION	.1439
KIND-ATT	.1411
STATEMENT	.1300
INFO-REQ	.1142
CLOSING	.0671
REJECT	.0644
OPENING	.0439
AGREE-ACC	.0408
GEN-ANS	.0331

(a)

S-OPINION
Gosh uh, it's getting pathetic now, absolutely pathetic.
They're just horrid, you'll have nightmares, you know.
That's no way to make a decision on some terrible problem.
They are just gems of shows. Really, fabulous in every way.
And, oh, that is so good. Delicious.
KIND-ATTITUDE
I'm sorry, I really feel strongly about this.
Sorry, now I'm probably going to upset you.
I hate to do it on this call.

(b)

with automatically enriching the corpus with additional features (linguistic markers). Then the unsupervised case consists of defining a very simple and intuitive set of seeds that profiles the specific dialogue acts, and subsequently performing a similarity analysis in a latent semantic space. The performance of the unsupervised experiment has been compared with a supervised state-of-art technique such as Support Vector Machines.

Results are quite encouraging and show that lexical knowledge plays a fundamental role in distinguishing among DA labels. Though, the analysis of misclassified cases suggested us to (i) include the consideration of knowledge about context (e.g. the previous *n* utterances) and (ii) to check the possibility of enriching the preprocessing techniques by introducing new linguistic markers (e.g. features related to the use of slanted lexicon, which seems to be relevant in distinguishing between objective statements and expressions of opinion).

Regarding the consideration of knowledge about the dialogue history, we have tested first of all the role played by contextual features in different experimental settings, achieving promising results. In particular bigrams are shown to cause a significant improvement in the DA recognition performance especially in the supervised framework. The improve-

ment is less significant in the unsupervised learning condition, in which a bootstrap based approach is implemented. Improving the bootstrap approach for including contextual information in our unsupervised framework will be object of further investigation in our future research.

We also performed a qualitative study about the affective load of utterances. The experimental results are preliminary but show that a relation exists between the affective load and the DA of a given utterance. According to these experimental evidences, we decided to further investigate, in the next future, the possibility of considering the affective load of utterances in disambiguating DA recognition. In particular, it would be interesting to exploit the role of slanted or affective-loaded lexicon to deal with the misclassification of opinions as statements. Along this perspective, DA recognition could serve also as a basis for conversational analysis aimed at improving a fine-grained opinion mining in dialogues.

REFERENCES

1. Austin, J.: How to do Things with Words. Oxford University Press, New York (1962)
2. Searle, J.: Speech Acts: An Essay in the Philosophy of Language. Cambridge University Press, Cambridge, London (1969)
3. Core, M., Allen, J.: Coding dialogs with the DAMSL annotation scheme. In: Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines, Cambridge, MA (1997) 28–35
4. Traum, D.: 20 questions for dialogue act taxonomies. Journal of Semantics **17** (2000) 7–30
5. Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Ess-Dykema, C.V., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M.: Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics **26** (2000) 339–373
6. Warnke, V., Kompe, R., Niemann, H., Nöth, E.: Integrated dialog act segmentation and classification using prosodic features and language models. In: Proceedings of 5th European Conference on Speech Communication and Technology. Volume 1., Rhodes, Greece (1997) 207–210
7. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: SAC '08: Proceedings of the 2008 ACM symposium on Applied computing, New York, NY, USA, ACM (2008) 1556–1560
8. Schegloff, E.: Sequencing in conversational openings. American Anthropologist **70** (1968) 1075–1095
9. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50** (1974) 696–735

10. Power, R.: The organisation of purposeful dialogues. Linguistics **17** (1979) 107–152
11. Cohen, P.R., Levesque, H.J.: Communicative actions for artificial agents. In: in Proceedings of the First International Conference on Multi-Agent Systems, AAAI Press (1995) 65–72
12. Reithinger, N., Klesen, M.: Dialogue act classification using language models. In: In Proceedings of EuroSpeech-97. (1997) 2235–2238
13. Venkataraman, A., Liu, Y., Shriberg, E., Stolcke, A.: Does active learning help automatic dialog act tagging in meeting data? In: Proceedings of EUROSPEECH-05, Lisbon, Portugal (2005)
14. Godfrey, J., Holliman, E., McDaniel, J.: SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), San Francisco, CA, IEEE (1992) 517–520
15. Clarizio, G., Mazzotta, I., Novielli, N., deRosis, F.: Social attitude towards a conversational character. In: Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication, Hatfield, UK (2006) 2–7
16. Jurafsky, D., Shriberg, E., Biasca, D.: Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-01, University of Colorado Institute of Cognitive Science (1997)
17. Novielli, N.: Hmm modeling of user engagement in advice-giving dialogues. Journal on Multimodal User Interfaces (2009)
18. Pianta, E., Girardi, C., Zanoli, R.: The TextPro tool suite. In: Proceedings of LREC-08, Marrakech, Morocco (2008)
19. Landauer, T.K., Foltz, P., Laham, D.: Introduction to latent semantic analysis. Discourse Processes **25** (1998)
20. Berry, M.: Large-scale sparse singular value computations. International Journal of Supercomputer Applications **6** (1992)
21. Gliozzo, A., Strapparava, C.: Domains kernels for text categorization. In: Proc. of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005), University of Michigan, Ann Arbor (2005) 56–63
22. Vapnik, V.: The Nature of Statistical Learning Theory. Springer-Verlag (1995)
23. Joachims, T.: Text categorization with Support Vector Machines: learning with many relevant features. In: Proceedings of the European Conference on Machine Learning. (1998)
24. Levinson, S.C.: Pragmatics. Cambridge University Press, Cambridge; New York (1983)
25. Poesio, M., Mikheev, A.: The predictive power of game structure in dialogue act recognition: Experimental results using maximum entropy estimation. In: Proceedings of ICSLP-98, Sydney (1998)
26. Strapparava, C., Mihalcea, R.: SemEval-2007 task 14: Affective Text. In: Proceedings of the 4^{th} International Workshop on Semantic Evaluations (SemEval 2007), Prague (2007) 70–74

27. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2** (2008) 1–135
28. Conati, C.: Probabilistic assessment of user's emotions in educational games. Applied Artificial Intelligence **16** (2002) 555–575
29. Picard, R.W., Klein, J.: Computers that recognise and respond to user emotion: Theoretical and practical implications. Technical report, MIT Media Lab (2001)
30. Bosma, W., André, E.: Exploiting emotions to disambiguate dialogue acts. In: IUI '04: Proceedings of the 9th international conference on Intelligent user interfaces, New York, NY, USA, ACM (2004) 85–92
31. Strapparava, C., Valitutti, A.: WordNet-Affect: an affective extension of WordNet. In: Proceedings of LREC. Volume 4. (2004) 1083–1086
32. Fellbaum, C., ed.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998)

NICOLE NOVIELLI
UNIVERSITÀ DEGLI STUDI DI BARI,
DIPARTIMENTO DI INFORMATICA
VIA ORABONA, 4 - 70125 BARI, ITALY
E-MAIL: <NOVIELLI@DI.UNIBA.IT>

CARLO STRAPPARAVA
FBK-IRST, ISTITUTO PER LA RICERCA
SCIENTIFICA E TECNOLOGICA,
VIA SOMMARIVE, 18 - I-38050 POVO TRENTO, ITALY
E-MAIL: <STRAPPA@FBK.EU>

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 27-43
RECEIVED 27/11/09 ACCEPTED 16/01/10 FINAL 10/03/10

Learning Event Semantics From Online News

HRISTO TANEV, MIJAIL KABADJOV, AND MONICA GEMO

Joint Research Centre, Italy

ABSTRACT

In this paper we present a multilingual algorithm for automatic extension of an event extraction grammar by unsupervised learning of semantic clusters of terms. In particular, we tested our algorithm to learn terms which are relevant for detection of displacement and evacuation events. Such events constitute an important part in the process of development of humanitarian crises, conflicts and natural and man made disasters. Apart from the grammar extension we consider our learning algorithm and the obtained semantic classes as a first step towards the semi-automatic building of a domain-specific ontology of disaster events. We carried out experiments both for English and Spanish languages and obtained promising results.

1 INTRODUCTION

Automatic event extraction is a relatively new sub-branch of information extraction, whose ultimate goal is the automatic extraction of structured information about events described in text sources, such as news. We look at the events as complex processes including interactions among several entities. Each of these participating entities has an event-specific semantic role, which defines the way in which the entity participates in the event and interacts with the other entities. The event-specific semantic roles are related to the nature of the entities to which they are assigned, however this relation is not straightforward. For example, in the context of evacuation events, an entity which belongs to the category *buildings*

can be assigned the event-specific role *evacuated-place*, *target-place-of-evacuation* (a place where people are evacuated) or it can be related only loosely to the dynamics of the event.

In the context of our preceding and current work, this relation is modeled through event extraction grammars which connect linguistic expressions, such as "were evacuated" with semantic classes, such as *person-group*, *facility*, etc. and event-specific semantic roles, such as *evacuated-people*, *evacuated-place*, etc. For example, the following sample rule detects evacuation events and extracts descriptions of evacuated people and evacuated places:

person-group :evacuated-people *were evacuated from* **NP(**head-noun: *place*): evacuated-place

This rule will match a text like: "Five women were evacuated from a hotel." and will extract "five women" as *evacuated-people* and "a hotel" as *evacuated-place*.

In order to automatize partially the process of creation of such rules, we propose a semi-automatic approach for extending event extraction grammars. The core of our approach is an unsupervised algorithm for learning of semantically consistent term clusters. In particular, we tested our algorithm to acquire terms, which are relevant for detection of displacement and evacuation events. Such events constitute an important part in the process of evolution over time of humanitarian crises, conflicts and natural and man made disasters. Apart from the grammar extension, we consider our learning algorithm and the obtained semantic clusters as a first step towards the semi-automatic building of domain-specific ontology of disaster events.

The starting point for us is an existing event extraction grammar for detection of evacuations and displacements from online news reports. The grammar is an integral part of NEXUS [1], an automatic system for event extraction from online news, which is profiled in the domain of security and crises-management.NEXUS makes use of over 90 event-specific patterns and a noun-phrase recognition grammar to detect boundaries of phrases which refer to groups of people.Using these two resources the system can identify text fragments such as "about 200000 people have abandoned their homes", where the phrase "about 200000 people" will be labeled with the event-specific semantic category *displaced people*.

Similarly, NEXUS can identify phrases about evacuations, such as "five women were evacuated", where "five women" will be labeled as *evacuated people*.

However, further, crucial information about these event scenarios is often encoded by the subcategorization frames in which the main verb phrases of the patterns may occur or by some verb adjuncts. In both cases, they consist of highly productive prepositional phrases (with selectional restrictions), where the noun-phrase head typically belongs to a specific semantic category. For example, in the text fragment "more than 1000 people were evacuated after a chemical leak" the prepositional phrase contains the crucial information about the event which caused the evacuation. In a similar way, the phrase "20000 people displaced to Beddawi camp" reports both the number of displaced people as well as the place where they were moved.

We developed an algorithm which expands automatically the event extraction grammar by learning a subset of the scenario-related subcategorization frames of verb phrases from unannotated news corpus.

The main part of our learning algorithm is an unsupervised term extraction and clustering approach which is a new way of combining several state-of-the-art term acquisition and classification techniques.

Clearly, there is far more structure within the subcategorization frames of the domain-specific verbs than standard surface level patterns of NEXUS can detect. Consequently, more work will be necessary to obtain a better picture about the different syntactic positions in which the semantic clusters can be introduced with respect to the main verbs. At this stage, we regard our experiments just as a first step towards automatic or semi-automatic learning of syntactico-semantic rules.

The rest of the paper proceeds as follows: Section 2 makes a review of the related work. Section 3 introduces the event extraction grammar, currently exploited by NEXUS. Section 4 explains our approach for learning of semantic classes and extending the event extraction grammar. Section 5 describes our experiments and the evaluation we did. Finally, section 6 presents our conclusions and discusses future research directions.

2 RELATED WORK

Relevant to our work are approaches for learning of verb subcategorization frames. In particular the work of [2] share similarities with our method, as far as it is based on automatic term clustering to acquire semantic clusters in unsupervised manner. However, they rely on manual

attachment of the obtained semantic clusters to the prepositions. Moreover, their semantic clusters are of limited size and contain only nouns which appear with specific predicates. Apart from application on very specific domains, such an approach will require a large training corpus to compensate for potential data-sparseness problems.

Another group of approaches in this field were introduced by the work of [3]: They use thesauri or taxonomies, such as WordNet to find the right level of semantic generalization in the subcategorization frames. The problem is that these methods are hardly applicable for languages other than English, due to the extensive use of semantic resources

Clustering and classification of words based on the distributional similarity of their contexts is not new: [4] proposed this approach for automatically clustering of nouns. Later, [5] used different syntactic features to cluster semantically similar words. Recently, the interest to distributional-similarity approaches was revived in the context of Ontology Learning and Population - [6] introduced unsupervised approach for ontology population, based on context distributional similarity between named entities, such as “Trento” and semantic categories, such as “city”; based on this work, [7] introduced some limited-scale supervision in the form of semi-automatically acquired seed sets of named entities thus improving the performance. The approach, presented by [8] uses contextual based similarity to cluster words into concept clusters; a particular feature of this work is that it explores deeper the usage of concept attributes such as contextual features. The problem with these approaches is that they begin with a predefined set of terms, taken from ontologies or other sources, which are next clustered or classified. It is not clear what will be the performance when combined with term extraction from free texts.

Another type of approaches for semantic classification follow the pioneering work of Marti Hearst [9]. It puts forward a small set of hypernym-hyponym extraction patterns, which relate a concept word, such as “city” with its possible hypernyms, e.g. “place”. A similar pattern-based approach was used by [8] to extract concept attributes. However, such pattern-based approaches are strongly affected by the data sparseness problem (see [7]), some authors promote the use of the Web [10] via a search engine, which however brings under consideration problems such as efficiency, maximal number of allowed queries, access policies of the search engines, etc.

3 EVENT EXTRACTION GRAMMAR

The grammar currently used by NEXUS is a finite state cascade grammar. The first grammar level recognizes references to groups of people, such as "100 women", "fifty Chinese workers", etc. The first level works on the top of a tokenizer and a dictionary with person referring nouns, such as "women", "workers", etc., nations, such as "Chinese", "Russian", etc. As an example, consider the following grammar rule which can parse phrases like "5 Canadian soldiers":

[person-group] → **(digit-number** | **word-number+) nation? person-noun-plural**

The second grammar cascade combines the recognized person phrases from the first level with the linear patterns, listed in a dictionary. As an example consider the following patterns for recognition of displacement events:

[person-group] *were forced out of their homes*

[person-group] *were displaced*

[person-group] *were uprooted*

These patterns and others of their type are encoded at the second grammar level through one rule:

[person-group] right-context-displacement-pattern

In this rule *right-context-displacement-pattern* refers to a class of string patterns, listed in the pattern dictionary, such as "were displaced", "were uprooted", etc. These strings, when appearing on the right from a description of a person group, designate a description of displacement event, in which the person group phrase refers to the displaced people in this event

4 EXTENDING THE GRAMMAR

The goal of the grammar expanding algorithm is the learning of syntactic adjuncts which are introduced in the description of the events usually

through prepositional phrases. More concretely, we would like to recognize phrases like "many people were evacuated to temporary shelters". In order to do this, our system has to recognize patterns like

[person-group]*were displaced to* **NP**(*facility*)

where *NP(facility)* refers to a noun phrase whose head noun belongs to the category *facility*, which should be described through a list of nouns. The grammar should also assign the event-specific semantic label *place-of-displacement* to this noun phrase. In the context of our experiments, we learn grammar extensions in the form of triples *(preposition, semantic cluster, event-specific role)*. For example, *(to; F; place-of-displacement)*, where F is a cluster of words, which can be considered as belonging to the category $facility$ in our event specific context.

We do not specify which triple to which pattern can be attached. This was not done, since many patterns are based on the same verbs or at least on verbs which share the same or similar sub-categorization frames. Therefore, the sample triple , *(to; F; place-of-displacement)* will be encoded in the extended grammar as

[person-group] right-context-displacement-pattern *to* (**NP**(F)):**place-of-displacement**

left-context-displacement-pattern [person-group] *to* (**NP**(F)):**place-of-displacement**

where F refers to a cluster, represented via dictionary which contains words which are likely to be $facilities$, e.g. "school", "hospital", "refugee camp", etc. Such rules can recognize text fragments, such as "1000 people were displaced to government shelters", provided that "shelters" is a member of the cluster F. Moreover, "government shelters" will be tagged with the event-specific semantic labels *place-of-displacement*.

4.1 *Algorithm overview*

In order to learn such grammar extensions, we propose the following multilingual machine learning algorithm, on which we elaborate in the following subsections:

1. Create a superficial seed terminology extraction grammar which recognizes preposition phrases which appear after displacement/evacuation patterns. We obtained this grammar via extending the multilingual

event extraction grammar of NEXUS. Note, that this is NOT the final extended grammar which was discussed in the beginning of this section, although its structure is very similar. It is rather a grammar for extraction of seed terminology.
2. We run the term extraction grammar on a news corpus and we extract all the pairs of a preposition and a head noun which appear after the event extraction templates. These pairs are grouped by preposition. In such a way, we obtain for each preposition a list of nouns which appear after it.
3. For each preposition, we cluster the corresponding nouns, using distributional similarity of their contexts.
4. We extend the clusters, using a multilingual term extraction based on context distribution similarity.
5. Clusters are cleaned using Hearst hypernym-hyponym templates applied on the Web.
6. Manually, we link each learned pair of a preposition and a semantic cluster to an event-specific semantic role, such as *cause-of-displacement*.
7. Extend the event extraction grammar by adding the learned adjuncts

4.2 *Seed terminology extraction grammar*

We construct the term extraction grammar by extending the second level event extraction grammar, described in the previous section. We created simple noun phrase recognition rules which utilize the output of a morphological processor. These rules constitute an intermediate grammar level between the first and the second one. The output of this level is the structure $NP(head : N)$, which denotes a noun phrase with head N. The second level rules for displacement and evacuation are modified by adding an adjunct introduced by a preposition. For example,

[person-group] right-context-displacement-pattern

will become

[person-group] right-context-displacement-pattern Prep NP

where $Prep$ can match any preposition and *NP* matches any noun phrase. Similarly, the preposition and the NP are attached to left context rules and the same we do for evacuation patterns. This grammar is used during the learning phase to extract a list of terms which are next used to form seed semantic classes.

4.3 *Learning semantic classes*

We run the term extraction grammar on a news corpus and we extract all the pairs of a preposition and a head noun which matches the construction *Prep NP(head-noun:n)*. If the NP has a main noun modified by another noun, e.g. "rain fall", then the whole bi-gram is taken as a head noun. As an example, from the text "five people were evacuated from a burning hotel", the term extraction grammar will extract the pair ("from", "hotel")

All the extracted pairs are grouped together with respect to the preposition. For each preposition, we keep only these nouns which appear with it at least a certain number of times. In such a way we obtain a list of prepositions, and for each preposition we have a list of associated nouns (or noun bi-grams, as explained before). For example, let's assume that for the preposition "after" we obtain the list: "day", "fire", "forest fire", "dam break", "flood", "rainstorm". Then, the following cluster learning algorithm is applied:

1. For each word in a cluster we obtain a list of contextual features. They are uni-grams, bi-grams and tri-grams which co-occur with the word in a news corpus. Weighting is carried out using an algorithm similar to the one described in [7], however we use superficial features, similar to the ones used by [11]. The feature weighting is described in more details in the next subsection.
2. The nouns corresponding to one preposition are clustered based on their contextual features extracted in the previous step. This step is necessary, since the same preposition can be followed by nouns from different semantic classes, which introduce different event-specific semantic roles. For example, the nouns occurring after the preposition "after" are clustered in three seed clusters:
   - day
   - fire, forest fire
   - dam break, flood, rainstorm
3. We ignore seed clusters with less than 3 elements as unreliable, therefore only the third cluster will remain in the previous example. Since clusters are formed based on contextual features, then words in a cluster will tend to appear in similar contexts. According to Harris' distributional hypothesis words which appear in similar contexts have similar semantics.
4. We use each seed cluster as a seed set to learn new terms which have similar contextual features and therefore are semantically similar. We used our in-house term extraction system, opulis, to perform this task. The system is based on a weakly supervised ontology

population approach introduced in [7] and modified for the use with superficial features. From an initial seed set of terms, Ontopopulis learns a list of terms with similar contextual distribution. The list is ordered by similarity with the seed set. We use the first 300 most similar elements from it to form our extended cluster. As an example, consider the highest scored members of the extended cluster obtained from the seed cluster "dam break", "flood", "rainstorm"; the top-scored members of the extended cluster, obtained from it, are: "flood", "quake", "floods", "fire", "tsunami", "disaster", "flooding", "earthquake", "storm", "cyclone", "hurricane", etc.

5. The extended cluster generated by the top 300 elements returned by Ontopopulis have significant amount of noise due to the big number of accepted terms. On the other hand, we found that some correct terms can have low similarity score due to data sparseness, semantic ambiguity, etc. Therefore, reducing the number of the accepted terms would result in low coverage. In order to improve the semantic consistency of our clusters without discarding many appropriate terms, we propose a semantic validation approach, based on superficial hypernym-hyponym patterns, similar to the ones introduced by Marti Hearst in [9]; we used the Web as a corpus. The approach has three main steps:
   - First, for the seed cluster, for example ("dam break", "flood", "rainstorm"), it forms the plural forms of the words: "dam breaks", "floods", "rainstorms" and queries the Web, using Yahoo API, with the pattern "such * as W", where W is substituted with the plural form of each word in the seed cluster, e.g. "such '* as dam breaks". For Spanish we used the pattern "W y otros * ".The assumption is that what appears at the position of the asterisk will be mostly a word X, such that there is an *is-a* relation between the word W and X. That is, X can be considered a hypernym of the word, at least in certain contexts.
   - Next, we learn one hypernym word H which co-occurs with most of the words from the seed cluster. (We use a simple co-occurrence measure based on frequencies). For the example seed cluster we obtain the hypernym word: "disasters"
   - For each word W from the extended cluster the algorithm forms its plural form WP and queries Yahoo API with the check pattern "H such as WP' (for Spanish it becomes "WP y otros H")', e.g. "disasters such as hurricanes". If seven or more pages are found on the Web which contain the pattern, then the word W

is accepted, otherwise it is filtered out from the extended cluster. In such a way we leave in the clusters mostly words which are likely to have an *is a* relation with one and the same concept. This improves semantic consistency of the final cluster. Note, that for English the check pattern is a slightly modified version of the Hearst pattern used to learn a hypernym in the previous step; the motivation for using two patterns is empirical - the first one is more precise and therefore better for learning of hypernyms, however we found it to be too restrictive as a check pattern.

At the end, we link each semantic cluster with the prepositions from which its seed cluster co-occurs. Therefore, at the end of this learning phase we have a list of word clusters $C_1, C_2, ...C_n$, which are mostly semantically consistent and a list of pairs $(Prep, C_i)$, where $Prep$ denotes a preposition and C_i denotes a cluster.

CONTEXTUAL FEATURES The basis of the semantic cluster learning are the contextual features. In our work a contextual feature of a word w is defined to be any lowercase word, bi-gram or a tri-gram which co-occurs in a corpus immediately on the left or on the right from w, it is not a stop-word, and co-occurs at least certain number of times. The co-occurrence feature specifies also the position of the n-gram (left or right) with respect to the words. For example, the word "hurricane" has a feature "X destroyed", where X shows the position of the word (in this case "hurricane") with respect to the feature. Every contextual feature is weighted, based on its co-occurrence with the word. Co-occurrence is measured using the Pointwise Mutual Information. These contextual features were used both for intital word clustering for obtaining the seed clusters, as well as for their expansion with Ontopopulis. When measuring the contextual similarity of two words, the dot product of their feature vectors is calculated.

4.4 *Extending the event-extraction grammar*

As it was pointed out before, at the end of the previous step we obtain a list of semantic clusters $C_1, C_2, ..., C_n$ and a list of pairs $(Prep, C_i)$, where $Prep$ denotes a preposition and C_i denotes a cluster. We manually link each pair to an event-specific semantic role, such as *cause-for-displacement* (e.g. "forest fire"), *target-place-of-displacement, means-of-evacuation, psychological-state-of-evacuated* (e.g., "left the building in

panic"), *evacuated-place* (e.g. "were evacuated from a skyscraper"), etc. In such away, we transform each pair *(Prep, Ci)* into a triple *(Prep, Ci, event-role)*, where *event-role* is a manually-assigned event-specific semantic role, such as *cause-for-displacement*. Each triple $(Prep, Ci, event-role)$ is used to transform each domain specific rule, such as:

[person-group] right-context-displacement-pattern

into

[person-group] right-context-displacement-pattern Token? Token? Token? Prep NP(head-noun:Ci): event-role

The term $NP(head-noun : C_i)$ will match each noun phrase, whose head noun belongs to the cluster Ci. (We used the noun-phrase extraction grammar layer, described in the second subsection.) In order to augment the coverage of the extended rules, we allow for several optional tokens to appear between the original pattern and the prepositional phrase. The *event-role* shows what event specific role will be assigned to the noun phrase, which matches $NP(head-noun : C_i)$. After several experiments with this grammar we reached the conclusion that some semantic clusters nearly always introduce the same event specific role, when appearing closely to the pattern, and this does not depend on the preposition. For example, the cluster with disasters always shows the reason of displacement. In such cases we omit from the rules the specification of the preposition and allow for more optional tokens, which can appear between the pattern and the noun phrase matched by $NP(head-noun : C_i)$. In such a way we increased the generality of our rules and obtained higher coverage for the extended grammar.

5 EXPERIMENTS AND EVALUATION

We applied our algorithm on English language online news, we obtained several semantic clusters, which we used to extend our event extraction grammar and extract three new types of event-specific roles, namely *cause for displacement/evacuation*, *evacuated place* and *target place of the evacuation*. We carried out also experiments with Spanish-language online news. Since we did not have enough time, we did not run the whole learning algorithm for the Spanish. We learned two semantic clusters for

this language and added one new semantic role to the Spanish event extraction grammar, namely *cause for displacement/evacuation*.

5.1 *Experiments for English*

We extended the English language grammar to obtain a term extraction grammar, as explained in section 4.2. The grammar uses 103 patterns for displacement and evacuation events.

Then, we run the algorithm for learning of the semantic classes, described in section 4.3 : We run the event extraction grammar on a $6GB$ corpus of news articles excerpts and extracted 11 prepositions which tend to appear after patterns for evacuation and displacement. For each preposition the event extraction grammar extracted also a list of nouns which tend to appear frequently after it. For our experiments we chose 5 of them for which there were sufficient number of nouns. These were the prepositions: “after”, “to”, “into”, “from” and “in”. For each of them we took the list of nouns and performed agglomerative clustering, based on contextual features, which were extracted from a news corpus. We chose in random a couple of clusters from each preposition; we expanded and cleaned them using the learning algorithm described in section 4.3. In such a way, we obtained 8 semantic clusters. We manually labeled with events-specific semantic roles all, but one of the combinations of preposition-cluster pairs. We used three types of event-specific semantic roles: *cause for displacement/evaluation*, *source (evacuated place)* and *target place of evacuation/displacement*. Then, we expanded the event extraction grammar, as described in section 4.4. In table 1 we list the clusters together with the main general and specific semantic category which they mostly represent, the corresponding prepositions with which these clusters were obtained and the event-specific semantic role, they were assigned.

5.2 *Experiments for Spanish*

For the Spanish language, we applied partially the learning algorithm described in section 4.3. We did not have time to collect necessary data for running the entire procedure. Instead of applying the whole algorithm, we translated two English-languages seed clusters into Spanish. More concretely, the first cluster contained four words, all designating different types of buildings and the second one consisted of three words, all designating disasters. Then, we applied the learning algorithm from step 4. That is, we performed cluster expansion using Ontopopulis and cluster

Table 1. Evaluation of the semantic consistency of the clusters (SF stands for Settlement and facility)

	size	gen. category	purity	sub-category	purity	prep.	ev.role
English							
c1	67	Calamity	85%	Natural disaster	67%	after	Cause
c2	48	Calamity	85%	Natural disaster	68%	after	Cause
c3	70	SF	70%	Facility	39%	to	Target
c4	95	SF	75%	Facility	58%	to	Target
c5	87	SF	71%	Facility	62%	into	Target
c6	69	Calamity	80%	Manmade disaster	56%	from	Cause
c7	111	SF	86%	Facility	84%	from	Source
c8	21	Situation	62%	Threat	33%	in	-
Spanish							
c9	72	Calamity	75%	Natural disaster	71%	-	Cause
c10	112	SF	55%	Facility	49%	-	-

Table 2. Accuracy of assigning event-specific roles using an extraction grammar

	Cause	Target place	Source (evacuated place)
English	86%	58%	100%
Spanish	32%	-	-

cleaning using Hearst patterns on the Web. In such a way, we obtained two extended semantic clusters for Spanish. In our experiments we used the extended cluster with the disasters to expand the Spanish event extraction grammar with one additional semantic role, namely *cause for displacement/evaluation*. We did not use the first stage of our algorithm which extracts seed terms (instead, the seed set was obtained as a translation of the English seed sets), therefore the clusters were not attached to specific prepositions. Regarding cluster $c10$, we did not include it in our grammar and therefore, we did not attach to it an event-specific role, however it can be used to find target or source places of evacuation and displacement events.

5.3 *Evaluation*

We carried out two types of evaluation: First, we evaluate the semantic consistency of each cluster and second, we run the extended event extraction grammar on a corpus of online news and extracted the entities,

which were assigned the newly added semantic roles; then we calculated the accuracy of assigning event-specific roles.

Semantic consistency was calculated by asking one English-speaking and one Spanish-speaking judge to define which is the semantic category which is predominant in each cluster. At first, the judge suggested quite generic categories, then they were asked to choose one more specific sub-category and to mark the cluster members which belong to the general category and to the more specific sub-category. Then, we calculated the *purity* of the cluster with respect to the generic and to the more specific categories as a ratio of the words which belong to the category and the cluster size. Results are presented in table 1.

The average purity of the English-language clusters with respect to the general category is 77%; it is 58% with respect to the more specific category. The corresponding purity values for the Spanish clusters is 65% and 60%. The purity of the Spanish-language clusters is comparable to the English ones. This is a good indicator for the multilingual nature of our algorithm. It is also important that cluster members, which we considered irrelevant for our evaluation, can still be considered relevant for the domain of displacements and evacuations: For example, our system learned words referring to vehicles and people, but they were mixed with other categories in the same cluster.

Regarding the extraction of event specific semantic roles, we run the extended grammar on an English and Spanish online news corpora, consisting of news clusters (each news cluster is a set of news articles, which refer to the same topic). For English we used a corpus of about 22,000 clusters and for Spanish we used a corpus of about 33,500 clusters. We calculated the accuracy of extraction for each of the event-specific semantic roles. We did not calculate recall, since at this stage our extended grammar was created mostly for experimental purposes and did not encode all the possible syntactic variations via which adjuncts can be connected to the event describing phrase. The results are presented in table 2.

The tangible result of our experiments was that new event specific roles were added to the event extraction grammar. In particular, the new slot $Cause$ was important, since it captured the events which lead to the displacement events.

The importance of detecting new semantic roles goes beyond extracting additional information. Detecting a semantic role, such $Cause$ together with some event-specific predicate, such as “flee” can be used to detect reliably an event of interest. For example, our grammar correctly

extracts "conflict" as a cause for displacement from the text: "...tens of thousands of civilians trying to $flee$ the $conflict$". Extracting such information allows us to detect reliably a report about a displacement event.

On the other hand, a word, such as "flee" alone does not provide enough evidence that an event of interest took place. For example, in the the following text "Meanwhile, the leopard injured several persons while making repeated attempts to $flee$.", no displacement event is described, still the word "flee" appears.

Most of the errors in our experiments were due to poorly clustered frequent words. For example, for Spanish we had the frequent words "pais" (country) wrongly clustered together with disasters. Similarly, the English word "killing" was clustered together with disasters and calamities, which lead to incorrect detection of a displacement event. With a little bit of manual cleaning, the accuracy of the obtained grammars could significantly be improved. Interestingly, some not very well classified words lead to grammar performance, which we considered correct. For example, the word "bomb" was clustered together with "war", "conflict" and other disastrous events. However, "bomb" is not an event. Nevertheless, the system extracts "bomb" as a cause for evacuation from the text: "Thousands of residents fled $bomb$-blasted parts of northern Mogadishu on Tuesday". This can be considered as a nearly correct match which lead to correct detection of an evacuation event, although strictly speaking the cause for evacuation was bombing and not "bomb". Similarly, "volcano" was clustered as a disaser and subsequently was extracted as cause for evacuation from the followiing text: "Thousands of people have been evacuated after a $volcano$ erupted" Such examples show that semantic similarities between words which belong to different categories (e.g. between "bomb" and "conflict") can be useful for practical purposes. Such kind of similarities cannot be found in semantic dictionaries, such as WordNet, however distributional word clustering successfully finds them. Clearly, distributional clustering is never 100% correct, however we think it is much easier to clean the errors from an already acquired dictionary, rather than creating one from scratch.

6 Conclusions and Future Work

In this paper we presented a multilingual algorithm for extending event extraction grammars by unsupervised learning of semantic classes. Although the results can be improved further, they show the viability of our approach.

The method we presented here can be used to automatize partially building of domain-specific grammars, which is quite a laborious task. As we demonstrated, the method can easily be adapted between languages.

Since our approach obtains word clusters, which model semantic concepts, it can also be used in the process of ontology building.

REFERENCES

1. Tanev, H., Piskorski, J., Atkinson, M.: Real-time news event extraction for global crisis monitoring. In: Proceedings of 13th International Conference on Applications of Natural Language to Information Systems, LNCS. (2008)
2. Faure, D., Nedellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In: LREC workshop on Adapting lexical and corpus resources to sublanguages and applications. (1998)
3. Li, H., Abe, N.: Generalizing case frames using a thesaurus and the mdl principle. Computational Linguistics **24** (1998) 214–244
4. Hindle, D.: Noun classification from predicate-argument structures. In: Proceedings of the Meeting of the Association for Computational Linguistics. (1990)
5. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of the ACL'98. (1998)
6. Cimiano, P., Völker, J.: Towards large-scale, open-domain and ontology-based named entity classification. In: Proceedings of Recent Advances in Natural Language Processing (RANLP), Borovets, Bulgaria (2005)
7. Tanev, H., Magnini, B.: Weakly supervised approaches for ontology population. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press (2008)
8. Almuhareb, A., Poesio, M.: Extracting concept descriptions from the web: the importance of attributes and values. In: Ontology Learning and Population: Bridging the Gap between Text and Knowledge, IOS Press (2008)
9. Hearst, M.: Automatic discovery of wordnet relations. In: WordNet: An Electronic Lexical Database, MIT Press (1998) 131–152
10. Markert, K., Malvina, N., Modjeska, N.: Using the web for nominal anaphora resolution. In: EACL Workshop on the Computational Treatment of Anaphora. (2003) 39–46
11. Lian, S., Sun, J., Che, H.: Populating crab ontology using context-profile based approaches. In: Knowledge Science, Engineering and Management. (2007)

HRISTO TANEV
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <HRISTO.TANEV@EXT.JRC.EC.EUROPA.EU>

MIJAIL KABADJOV
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <MIJAIL.KABADJOV@JRC.EC.EUROPA.EU>

MONICA GEMO
JOINT RESEARCH CENTRE, EUROPEAN COMMISSION,
VIA E. FERMI 2749, I-21027,ISPRA, ITALY
E-MAIL: <MONICA.GEMO@JRC.EC.EUROPA.EU>

LANGUAGE FORUM

Language Forum – LF (started in 1975) is a peer-reviewed international journal that has been publishing original research papers devoted to the studies of LANGUAGE & LITERATURE for more than the past three decades. It is published in March and September every year.

The scope of the journal has over time, varied from pure Linguistics to pure Literature and sometimes both. Over the years, while it has dealt brilliantly with issues in modern Indo-English fiction, studies in the "comparative literature" domain, the issue of "tolerance" and "identity" in Indian poetry and fiction, the writings of and for the marginal sect of Indian society i.e. Dalits, the Indian and foreign English writers such as Anita Desai and T. S. Eliot etc, it has also thrown significant insight towards what is happening not only in Kashmiri linguistics but also in the vocabulary teaching methodologies, computational linguistics, stylistics, E.L.T, E.F.L, second language acquisition etc. The scope has been deliberately kept very broad to include the following disciplines and sub-disciplines of Language & Literature both:

- Linguistics in general
- Natural language processing
- Applied literary criticism
- Indo-English Literature
- Modern language teaching methods
- E. L. T, E. F. L. & E. S. L.
- Language curriculum planning
- Linguistic analysis of Indian languages
- Language development
- Indo-English fiction
- Structuralism and post structuralism
- Communicative competence
- First and second language acquisition
- Comparative literature
- Pedagogy
- Language problems
- Language disorder
- Semantics
- Stylistics
- Bilingualism
- Narratives
- Globalization & cultures
- Art and aesthetics
- Discourse analysis
- Folklore and media

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.com>.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 45-59
RECEIVED 28/11/09 ACCEPTED 16/01/10 FINAL 08/03/10

Exploiting Higher-level Semantic Information for the Opinion-oriented Summarization of Blogs

ALEXANDRA BALAHUR[1], MIJAIL KABADJOV[2], JOSEF STEINBERGER[2]

[1] *University of Alicante, Spain*
[2] *European Commission – Joint Research Centre, Italy*

ABSTRACT

Together with the growth of the Web 2.0, people have started more and more to communicate, share ideas and comment in blogs, social networks, forums and review sites. Within this context, new and suitable techniques must be developed for the automatic treatment of the large volume of subjective data, to appropriately summarize the arguments presented therein (e.g. as "in favor" and "against"). This article assesses the impact of exploiting higher-level semantic information such as named entities and IS-A relationships for the automatic summarization of positive and negative opinions in blog threads. We first run a sentiment analyzer (with and without topic detection) and subsequently a summarizer based on a framework drawing on Latent Semantic Analysis. Further on, we employ an annotated corpus and the standard ROUGE scorer to automatically evaluate our approach. We compare the results obtained using different system configurations and discuss the issues involved, proposing a suitable method for tackling this scenario.

Keywords: opinion mining, sentiment analysis, text summarization, social media.

1 INTRODUCTION

The recent growth in access to technology and the Internet, together with the development of the Web 2.0 (Social Web), has led to the birth of new and interesting social phenomena. On the one hand, the possibility to express opinion "by anyone, anywhere, on anything", in blogs, forums, review sites has made it possible for people all around the world to take better and more informed decisions at the time of buying products and contracting services. On the other hand, the companies and public persons are more informed on the impact they have on people, because the large amount of opinions expressed on them offers a direct and unbiased, global feedback. Moreover, people all over the world can express their opinion on the issues that affect their lives – events in politics, economics, the social sphere – or simply discuss on their hobbies and everyday lives. Thus, the past few years, due to the growing access to the Internet and the development of such Web 2.0 phenomena, have lead to the creation on the web of extensive quantities of subjective and opinionated data. Such information cannot be manually processed, although their analysis (discovery of opinions, their classification into positive and negative), could be useful to a high diversity of entities (potential customers, companies, public figures and institutions etc.), for a large variety of tasks (opinion analysis for marketing, sociological or political studies, decision support etc.). Therefore, automatic systems must be built, with the aim of processing the subjective data available and extracting the information that is relevant to the users.

For example, when a potential customer is interested in buying a new digital camera, they would like to know what others think about the features of the different models available on the market, within a price range, and whether others recommend the product or not. An automatic system assisting such a user would have to retrieve all the opinionated texts on the customer's products of interest, extract the product features and the opinions expressed on them, classify the opinions as positive or negative and present the user with percentages of positive and negative opinions on each of the product features. One step further could be that of summarizing the positive and negative opinions, so that the users can read for themselves the reasons for liking or disliking the product.

Another example involving the treatment of subjective data is that of a public person constantly monitoring his/her public image. Such a person would require the daily or weekly analysis of all the opinions expressed on them and their actions. An automatic system

implementing this task would have to gather all the opinions expressed on the person every day, analyze them to determine whether they are positive or negative and present the user with an overview of the general opinion (in percentages, organized depending on the opinion source, or in the form of an extractive summary).

Finally, an example of a system analyzing subjective data to respond to the needs of different users is one that is capable of extracting, from discussion threads, such as those present in blogs, the arguments "in favor" and "against" a topic, be it the economic crisis or cooking recipes. Such a system can extract the relevant opinions expressed on the topic and eliminating the redundant information, presenting the user with a clear list of arguments explaining the general view on the matter.

This article presents and compares different methods implemented with the aim of creating a system of the latter type. We show how the subjective content can be analyzed from the pure opinion and combined topic-opinion point of view and how the relevant parts can subsequently be summarized, based on the polarity of the opinions expressed. In what follows, Section 2 presents the related work and previous experiments in related tasks. Further on, Section 3 motivates the approaches proposed and indicates the contribution of this article to the task. In Section 4, we present the data we employ in our experiments and in Section 5, we depict the preliminary experiments conducted on it. Section 6 presents an in-depth description of the experiments performed and the results of the different evaluations. Finally, we conclude in Section 7, by discussing our findings and proposing the lines for future work.

2 RELATED WORK

While the task of summarization has been tackled for a longer period of time within the field of Natural Language Processing (NLP), literature in sentiment analysis has only flourished in the past few years, due to the massive growth in the quantity of subjective data available on the web. Thus, whilst there is abundant literature on text summarization [1, 2, 3, 4, 5] and sentiment analysis [6, 7, 8, 9, 10], there is still limited work at the intersection of these two areas [11, 12, 13]. This is easily explainable by: a) the fact that both systems performing opinion mining, as well as those automatically summarizing must have a certain level of maturity, so that errors do not propagate along the processing

pipeline; b) the task of summarization within the opinion context may be different from the traditional view on text summarization [14].

The 2008 edition of the Text Analysis Conference (TAC 2008), organized by the US National Institute of Standards and Technology (NIST), contained a pilot task, within the summarization track – i.e. Summarization Opinion Pilot. Being a pilot task within the summarization track, most of the techniques employed by the participants were based on the already existing summarization systems. New characteristics were added to these systems to account for the assessment of opinions present in the text (sentiment, positive/negative sentiment, positive/negative opinion). Examples of such systems are: CLASSY [15]; CCNU [16]; LIPN [17]; IIITSum08 [18]. Other participants, outside de summarization track, focused more on the opinion mining part of the task, thus doing the retrieval and filtering based on polarity - DLSIUAES [19]- or on separating information rich clauses – italica [20]. The results of the competition showed that, on the one hand, systems concentrating on the summarization part lost on the opinion content, and, on the other hand, systems lacking proper summarization components lose as far as the linguistic quality of the results is concerned and introduce much noise due to not being able to filter out redundant or marginal information.

Zhou and Hovy [21] and [22] present approaches to summarizing threads in blogs and online discussions, but focusing on the factual content. They demonstrate why this type of summarization is more difficult than traditional summarization in newswire and model subtopics and topic drifts.

Recently, [12] propose an approach to summarize threads in blogs using a combination of an opinion mining and a summarization system. They analyze the output as far as linguistic quality is concerned, to assess the difficulty of the task in the context of blogs, demonstrating that the difficulty in performing opinion summarization of blog threads resides in the language used, the topic inconsistency and the high redundancy of information. [13] claim that topic detection is crucial to the summarization of blog threads, but no experiments are done in this sense.

[14] assess the difference between the traditional task of summarization and opinion summarization in blogs, showing that through the nature of blog texts and the high subjectivity they contain, opinion summarization differs to a large degree from the traditional task. They experiment with the hypothesis of whether, in this context, the intensity of polarity is a good summary indicator.

3 MOTIVATION AND CONTRIBUTION

As demonstrated by the body of research that has tackled this issue, summarizing opinion is a difficult task, especially when pursued in the context of blogs.

Even if the behavior of bloggers has changed in the past few years, as shown by the Technorati "State of the blogosphere" reports[1] in 2008[2] and 2009[3], one of the main difficulties when addressing opinion expression in blogs is that it contains many references to outside sources, as well as "copy+paste"s from newspaper articles, photos, videos and other types of multimodal information that supports the argument that is made. While in 2006, Zhou and Hovy (2006) were writing that the predominance in blogs is given by the original blog message of the blog author, in 2009, we find that the vast majority of the thread body is given by comments written by other bloggers. This fact is supported by the Technorati report on the state of the blogosphere in 2009, where commenting in other blogs is found to be one of the strategies employed for attracting audience to one's blog, along with the tagging of content, regular updating of content and others.

Contrary to the general belief, blogs are mainly written by highly educated people and they can constitute a manner to consult expert opinion on different subjects. That is why, our first motivation in our experiments to search for and summarize opinions on different topics in blogs is given by the possibility blogs give to acquire useful and timely information.

Secondly, the research done so far in this area has not taken into consideration the use of methods to detect sentiment that is directly related to the topic. In the experiments we have performed, we detect sentences where the topic is mentioned, by using Latent Semantic Analysis.

Thirdly, most summarization systems do not take into consideration semantic information or include Named Entity variants and co-references. In our approach, also employed in the TAC 2009

[1] The Technorati reports on the state of the blogosphere have been published online since 2004, and are available at http://technorati.com/. They present statistics and overviews on the number of blogs, their topics, the social background and motivation of bloggers, as well as results of questionnaires enquiring on the behavior of bloggers.

[2] http://technorati.com/blogging/feature/state-of-the-blogosphere-2008/

[3] http://technorati.com/blogging/feature/state-of-the-blogosphere-2009/

summarization, we employ these methods and show how we can obtain better results through their use.

4 DATA

The data used in our experiments is described in [12] and it has also been used in [13] and [14]. It consists of 51 blog entries with their corresponding comments (threads) in English, summing up to a total of 1829 posts with 299.568 words. This corpus was selected, on the one hand, because it gives us the possibility to compare the results obtained with the ones reported in the related studies and, on the other hand, because it contains the annotations of the topics discussed in the posts and labeling of the topic-relevant sentences as far as source (the author of the text snippet), target (the topic it addresses), polarity (positive and negative) and intensity of the polarity (low, medium, high) are concerned. Although the threads are centered mostly on economy, science and technology, cooking, society and sport, their annotation contains a finer-grained identification of subtopics – e.g. the economic crisis, idols, VIPs and so on.

The gold standard for the summarization process is marked by the annotations on this corpus. We consider that the correct sentences that should appear in the final summaries (separately considering the positive and negative arguments on a topic) are the ones that are relevant for the topic, have the required polarity and score high on intensity.

5 PRELIMINARY EXPERIMENTS

Before reaching the present configuration of the system, we have performed several experiments on the presented blog data, as well as on quotations (reported speech)- shorter pieces of text representing a direct statement of opinion, from a source to a target. From these preliminary experiments, we could extract several useful conclusions, which influenced the final setting of the experiments presented.

5.1 Preliminary sentiment analysis approach

The first and easiest approach that we carried out was based on two processing phases: the first one identified the subjective sentences -

using the Subjectivity Indicators in [23] - and, in the second phase, the polarity of the sentences classified as subjective was computed as sum of the opinion words found in them - using different combinations of affect and opinion lexicons: MicroWordNet Opinion [24], SentiWordNet [25], WordNet Affect [26] and a list of in-house terms denominated the JRC List. In order to perform these two steps on the data, the blog threads were split into files containing the initial post and, individually, the comments given by other bloggers on this post and subsequently the posts were split into sentences using LingPipe[4]. The best results on the blog data presently used were obtained when a combination of all resources was employed, leading to a precision of classification for positive opinion of 0.67, with a recall of 0.22 and a precision of classification for negative opinion of 0.53, with a recall of 0.89. The low results were attributed mostly to the lack of topic determination; the analysis of the accuracy for sentence classification revealed that many of the sentences had been correctly classified from the opinion polarity point of view, but they were not on the topics identified in the blogs. The summarization process, based on Latent Semantic Analysis [27] had a performance, given by the ROUGE scores, of 0.21 and 0.22 (R_1 for positive and negative, respectively) and 0.05 and 0.09 (for R_2 and R_{SU4} for positive and negative, respectively).

5.2 Opinion classification around Named Entities

Filtering sentences according to their topic, when the latter is a wide concept, such as economics or politics, is not a trivial task. However, when the topic is a Named Entity – its mentions, under its name or title (e.g. Gordon Brown, mentioned as such, or as Gordon, or "the British prime-minister") – the task becomes easier. Thus, in a parallel experiment, we tested, under the same conditions, the possibility to classify opinion on different public persons, by assessing the context surrounding their mentions in newspaper quotations. The results of these experiments showed significant improvements over the previous results, with an accuracy of 83% in classifying opinion among positive, negative and neutral (objective), using a combination of MicroWNOp, the JRC List and the General Inquirer (Stone et al. 1966). We employ this same strategy in order to compute the opinion on the topic of interest, using the topic words discovered with LSA as anchors around which opinion words are sought.

[4] http://alias-i.com/lingpipe/

6 EXPERIMENTS AND EVALUATION

As seen in the preliminary experiments, the performance of the opinion summarization, as it was tackled so far (without taking into consideration the topic) was rather low. From the human evaluation of the obtained summaries, we could see that the sentiment analysis system classified the sentences correctly as far as opinion, polarity and intensity are concerned. However, many topic irrelevant sentences were introduced in the summaries, leaving aside the relevant ones. On the other hand, we could notice that in the experiments taking into consideration the presence of the opinion target and its co-references and computing the opinion polarity around the mentions of the target reaches a higher level of performance. Therefore, it became clear that a system performing opinion summarization in blogs must include a topic component.

6.1 Sentiment analysis system

In the first stage, we employ the same technique as in the preliminary approach, but using only the resources that best scored together (MicroWordNet Opion, JRC Lists and General Inquirer). We map each of these resources into four classes (of positive, negative, high positive and high negative, and assign each of the words in the classes a value, of 1, -1, 4 and -4, respectively. We score each of the blog sentences as sum of the values of the opinion words identified in it (Fig.1).

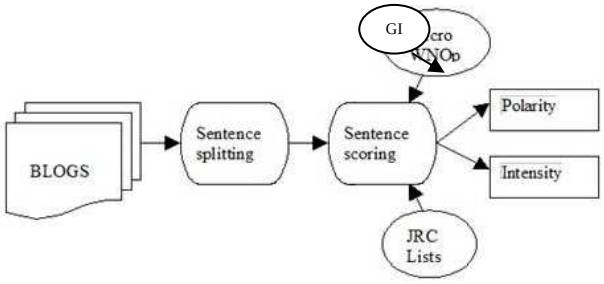


Fig. 1. Sentiment analysis system

In the second stage, we first filter out the sentences that are associated to the topic discussed, using LSA. Further on, we score the sentences identified as relating to the topic of the blog post, in the same manner as in the previous approach (Fig. 2).

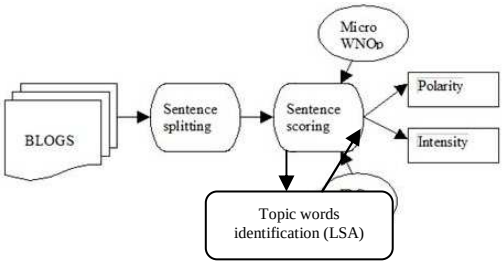


Fig. 2. Sentiment analysis system with topic words identification through LSA

Topic words identification using LSA. In order to filter for processing only the sentences containing opinions on the post topic, we first create a small corpus of blog posts on each of the topics included in our collection. These small corpora (30 posts for each of the five topics) are gathered using the search on topic words on http://www.blogniscient.com/. For each of these 5 corpora, we apply LSA, using the Infomap NLP Software[5]. Subsequently, we compute the 100 most associated words with 2 of the terms that are most associated with each of the 5 topics and the 100 most associated words with the topic word. For example, for the term "bank", which is associated to "economy", we obtain (the first 20 terms):

```
bank:1.000000;money:0.799950;pump:0.683452;
switched:0.682389;interest:0.674177;easing:0.661366
;
authorised:0.660222;coaster:0.656544;roller:0.65654
4;
maintained:0.656216;projected:0.656026;apf:0.655364
;
requirements:0.650757;tbills:0.650515;ordering:0.64
8081;
eligible:0.645723;ferguson's:0.644950;proportionall
y:0.63358;
integrate:0.625096;rates:0.624235
```

[5] http://infomap-nlp.sourceforge.net/

6.2 Summarization system

The summarization process is based on LSA, which is enriched with semantic information coming from two sources: the Medical Subject Headings (MeSH) taxonomy[6] and a Named Entity recognizer and disambiguator [28].

The LSA approach to summarization entails a two-fold process: firstly, a term-by-sentence matrix from the source is built and secondly, Singular Value Decomposition (SVD) is applied to the initial matrix. The decomposition is then used to select the most informative sentences. The enrichment of semantic information takes place during the step of building the term-by-sentence matrix. Full details of the approach can be found in [29].

6.3 Evaluation

We include the usual ROUGE metrics: R_1 is the maximum number of co-occurring unigrams, R_2 is the maximum number of co-occurring bigrams, R_{su4} is the skip bigram measure with the addition of unigrams as counting unit, and finally, R_L is the longest common subsequence measure (Lin, 2004). In the cases of the baseline systems we present the average *F1* score for the given metric and within parenthesis the 95% confidence intervals.

Table 1. Summarization performance.

System	R_1	R_2	R_{SU4}	R_L
Sent+BLSumm$_{neg}$	0.22 (0.18-0.26)	0.09 (0.06-0.11)	0.09 (0.06-0.11)	0.21 (0.17-0.24)
Sent+Summ$_{neg}$	**0.268**	0.087	0.087	**0.253**
Sent+BLSumm$_{neg}$	0.21 (0.17-0.26)	0.05 (0.02-0.09)	0.05 (0.02-0.09)	0.19 (0.16-0.23)
Sent+Summ$_{neg}$	**0.275**	0.076	0.076	**0.249**

[6] The MeSH thesaurus is prepared by the US National Library of Medicine for indexing, cataloguing, and searching for biomedical and health-related information and documents. Although, it was initially meant for biomedical and health-related documents, since it represents a large IS-A taxonomy it can be used in more general tasks (http://www.nlm.nih.gov/mesh/meshhome.html). Additionally, thanks to NGO Health-on-the-Net (HON, http://www.hon.ch/), a tool for recognizing terms in free text and grounding them to the MeSH taxonomy was available to us.

There are four rows in table 1: the first one, *Sent+BLSumm*$_{neg}$, is the performance of the baseline LSA summarizer on the negative posts (i.e., using only words), the second one, *Sent+Summ*$_{neg}$, is the enhanced LSA summarizer exploiting entities and IS-A relationships as given by the MeSH taxonomy, the third one, *Sent+BLSumm*$_{pos}$, presents the performance of the baseline LSA summarizer on the positive posts and the fourth one, *Sent+Summ*$_{pos}$, is the enhanced LSA summarizer for the positive posts.

Based on table 1 we can say that the results obtained with the enhanced LSA summarizer are overall better than the baseline summarizer. The numbers in bold show statistically significant improvement over the baseline system (note they are outside of the confidence intervals of the baseline system). The one exception where there is a slight drop in performance of the enhanced summarizer with respect to the baseline system is in the case of the negative posts for the metrics R_2 and R_{su4}, however, the *F1* is still within the confidence intervals of the baseline system, meaning the difference is not statistically significant.

We note that the main improvement in the performance of the enhanced summarizer comes from better precision and either no loss or minimal loss in recall with respect to the baseline system. The improved precision can be attributed, on one hand, to the incorporation of entities and IS-A relationships, but also, on the other hand, to the use of a better sentiment analyzer than the one used to produce the results of the baseline system.

We conclude that exploiting higher-level semantic information such as entities and IS-A relationships does bring a tangible improvement for the opinion-oriented summarization of blogs.

7 CONCLUSIONS

In this paper we measured the impact of exploiting higher-level semantic information such as named entities and IS-A relationships for the automatic summarization of positive and negative opinions in blog threads. We ran in tandem a sentiment analyzer and an LSA-based summarizer in two configurations: one using only words which we set as our baseline system, and another one making use in addition of entities and IS-A relations which we called the enhanced LSA summarizer. We used an annotated corpus and the standard ROUGE scorer to automatically evaluate the performance of our system. We

conclude that making use of higher-level semantic information as given by named entities and IS-A relationships does bring a tangible improvement for the opinion-oriented summarization of blogs.

In future work, we intend to analyze in more detail the cases where our system fails as well as the cases where a standard framework for evaluating summarization system falls short in providing adequate results for the task of producing opinion-oriented summaries.

REFERENCES

1. Kabadjov, M., Steinberger, J., Pouliquen, B., Steinberger, R., and Poesio, M. Multilingual statistical news summarisation: Preliminary experiments with English. In Proceedings of theWorkshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), (2009)
2. Steinberger, J., Poesio, M., Kabadjov, M., and Jezek, K. Two uses of anaphora resolution in summarization. Information Processing and Management, 43(6):1663–1680. Special Issue on Text Summarization (Donna Harman, ed.), (2007)
3. Hovy, E. H. Automated text summarization. In Mitkov, R., editor, The Oxford Handbook of Computational Linguistics, pages 583–598. Oxford University Press, Oxford, UK, (2005)
4. Erkan, G. and Radev, D. R. LexRank: Graph-based centrality as salience in text summarization. Journal of Artificial Intelligence Research (JAIR), (2004)
5. Gong, Y. and Liu, X. Generic text summarization using relevance measure and latent semantic analysis. In Proceedings of ACM SIGIR, New Orleans, US, (2002)
6. Pang, B. and Lee, L. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Annual Meeting of the Association for Computational Linguistics, pages 271–278, Barcelona, Spain, (2004)
7. Riloff, E. and Wiebe, J. Learning extraction patterns for subjective expressions. In Proceeding of the Conference on Empirical Methods in Natural Language Processing, (2003)
8. Kim, S. and Hovy, E. (2004). Determining the sentiment of opinions. In Proceedings of the International Conference on Computational Linguistics (COLING), (2004)
9. Turney, P. and Littman, M. Measuring praise and criticism: Inference of semantic orientation from association. ACM Transactions on Information Systems, 21, (2003)

10. Balahur, A., Steinberger, R., van der Goot, E., Pouliquen, B. and Kabadjov, M. Opinion mining from newspaper quotations. In Proceedings of the Workshop on Intelligent Analysis and Processing of Web News Content at the IEEE / WIC / ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT), (2009)
11. Stoyanov, V. and Cardie, C. Toward opinion summarization: Linking the sources. In Proceedings of the COLING-ACL Workshop on Sentiment and Subjectivity in Text, Sydney, Australia. Association for Computational Linguistics, (2006)
12. Balahur, A., Lloret, E., Boldrini, E., Montoyo, A., Palomar, M., and Martínez-Barco, P. Summarizing threads in blogs using opinion polarity. In Proceedings of the Workshop on Emerging Text Types (eETTs 2009), satellite workshop to RANLP, (2009)
13. Balahur, A., Kabadjov, M., Steinberger, J., Steinberger, R. and Montoyo, A. Summarizing opinion in blog threads. In Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation (PACLIC 2009).
14. Kabadjov, M., Balahur, A. and Boldrini, E. (2009b). Sentiment Intensity: Is It a Good Summary Indicator? In Proceedings of the Language and Technology Conference (LTC 2009)
15. Conroy, J. and Schlesinger, S. Classy at TAC 2008 metrics. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
16. He, T., Chen, J., Gui, Z., and Li, F. CCNU at TAC 2008: Proceeding on using semantic method for automated summarization yield. . In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
17. Bossard, A., Généreux, M., and Poibeau, T. Description of the LIPN systems at TAC 2008: Summarizing information and opinions. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
18. Varma, V., Pingali, P., Katragadda, R., Krisha, S., Ganesh, S., Sarvabhotla, K., Garapati, H., Gopisetty, H., Reddy, V., Bysani, P., and Bharadwaj, R. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
19. Balahur, A., Lloret, E., Ferrández, O., Montoyo, A., Palomar, M., and Muñoz, R. The DLSIUAES team's participation in the TAC 2008 tracks. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
20. Cruz, F., Troyano, J., Ortega, J., and Enríquez, F. (2008). The Italica system at TAC 2008 opinion summarization task. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology, (2008)
21. Zhou, L. and Hovy, E. Digesting Virtual "Geek" Culture: The Summarization of Technical Internet Relay Chats. In Proceedings of ACL 2005, (2005)

22. Zhou, L. and Hovy, E. On the summarization of dynamically introduced information: online discussions and blogs. In Proceedings of the AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), (2006)
23. Wilson, T., Wiebe, J., Hoffman, P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of Proceeding of the Conference on Empirical Methods in Natural Language Processing, (2005)
24. Cerini, S., Compagnoni, V., Demontis, A., Formentelli, M., and Gandini, G. Language resources and linguistic theory: Typology, second language acquisition, English linguistics, chapter Micro-WNOp: A gold standard for the evaluation of automatically compiled lexical resources for opinion mining. Franco Angeli Editore, Italy, (2007)
25. Esuli, A., Sebastiani, F.. SentiWordNet: A Publicly Available Resource for Opinion Mining. In Proceedings of 4th International Conference on Language Resources and Evaluation, LREC (2006)
26. Strapparava, C. Valitutti, A. WordNet-Affect: an affective extension of WordNet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, LREC (2004)
27. Deerwester, S., Dumais, S., Furnas, G. W., Landauer,T. K., and Harshman, R. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 3(41), (1990)
28. Pouliquen, B. and Steinberger, R. Automatic construction of multilingual name dictionaries. In Cyril Goutte, Nicola Cancedda, Marc Dymetman, and George Foster, editors, Learning Machine Translation. MIT Press, NIPS series (2009)
29. Steinberger, J., Kabadjov, M., Pouliquen, B., Steinberger, R., Poesio, M. (2009). WB-JRC-UT's participation in TAC 2009: Update Summarization and AESOP tasks. In Proceedings of the Text Analysis Conference, organized by the National Institute of Standards and Technology (2009)

ALEXANDRA BALAHUR
UNIVERSITY OF ALICANTE,
DEPARTMENT OF SOFTWARE AND COMPUTING SYSTEMS,
AP. DE CORREOS 99, E-03080 ALICANTE, SPAIN
E-MAIL: <ABALAHUR@DLSI.UA.ES>

MIJAIL KABADJOV
EUROPEAN COMMISSION – JOINT RESEARCH CENTRE
IPSC - GLOBESEC - OPTIMA
(OPENSOURCE TEXT INFORMATION MINING AND ANALYSIS)
T.P. 267, VIA FERMI 2749. 21027 ISPRA (VA), ITALY
E-MAIL: <MIJAIL.KABADJOV@JRC.EC.EUROPA.EU>

JOSEF STEINBERGER
EUROPEAN COMMISSION – JOINT RESEARCH CENTRE
IPSC - GLOBESEC - OPTIMA
(OPENSOURCE TEXT INFORMATION MINING AND ANALYSIS)
T.P. 267, VIA FERMI 2749. 21027 ISPRA (VA), ITALY
E-MAIL: <JOSEF.STEINBERGER@JRC.EC.EUROPA.EU>

Indian Journal of Applied Linguistics

Indian Journal of Applied Linguistics – IJOAL (started in 1975) is a peer-reviewed international journal that has been publishing original research papers devoted to LANGUAGE AND LINGUISTICS, for more than past three decades. It is published in March and September every year.

It provides a forum for the discussion of language related problems faced by L1 and L2 learners, the various language teaching methodologies adopted, apart from dealing with the general linguistic theories, and the branches and sub branches of linguistics.

It also attempts to place before its reader's new theoretical and methodological ideas and research from the several disciplines engaged in applied linguistics. The Linguist, the Anthropologist, the Psychologist, the Applied Linguist, and the Language Teacher may find it a useful forum, for both descriptive and experimental studies.

Some of the key areas of focus are:

- First and second language acquisition
- Modern language teaching methods
- Communicative competence
- Neurolinguistics
- Ethnolinguistics
- Sociolinguistics
- Natural language processing
- Language problems & language planning
- E. L. T, E. F. L. & E. S. L.
- Pragmatics
- Pedagogy
- Bilingualism
- Psycholinguistics
- Computational linguistics
- Applied literary criticism
- Stylistics

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.com>.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 61-83
RECEIVED 12/12/09 ACCEPTED 16/01/10 FINAL 08/03/10

Thai Rhetorical Structure Tree Construction

SOMNUK SINTHUPOUN[1] AND OHM SORNIL[2]

[1]Maejo University, Thailand
[2] National Institute of Development Administration, Thailand 10240

ABSTRACT

A rhetorical structure tree (RS tree) is a representation of elementary discourse units (EDUs) and discourse relations among them. An RS tree is very useful to many text processing tasks utilizing relations among EDUs such as text understanding, summarization, and question-answering. Thai language with its distinctive linguistic characteristics requires a unique RS tree construction technique. This article proposes an approach to Thai RS tree construction; it consists of two major steps: EDU segmentation and RS tree construction. Two hidden Markov models constructed from grammatical rules are employed to segment EDUs, and a clustering technique with its similarity measure derived from Thai semantic rules is used to construct a Thai RS tree. The proposed technique is evaluated using three Thai corpora. The results show the Thai RS tree construction effectiveness of 94.90%.

Keywords: Thai Language, Elementary Discourse Unit, Rhetorical Structure Tree.

1 INTRODUCTION

A rhetorical tree (RS tree) is a tree-like representation of elementary discourse units (EDUs) and discourse relations (DRs) among them. It can be defined as: RS tree = (status, DR, promotion, left, right) where status is a set of EDUs; DR is a set of discourse relations; promotion is a subset

of EDUs; and left and right can either be NULL or recursively defined objects of type RS tree [14, 16].

Definition of EDU may vary. Some researchers consider an EDU to be a clause or a clause-like [16] excerpt while others consider them to be a sentence [18] in discourse parsing. A number of techniques are proposed to determine EDU boundaries for English language such as those using discourse cues [1, 6, 15], punctuation marks [6, 16], and syntactic information [16, 18, 19].

Many discourse relations can be used in writings. Some have a single nucleus such as elaboration and condition while others have multiple nucleuses such as contrast [13]. A number of techniques for determining relations between EDUs are proposed, such as those using verb semantics [20] to build verb-based events, using cue phrases/discourse markers (e.g., “because”, “however”) [15], and using machine learning techniques [16].

Chaniak [5] constructs RS trees by using statistical techniques, taking into account part-of-speech tagging on syntax, and using a corpus like the Penn tree-bank [20] to produce statistical RS trees. Statistical RS Trees work by assigning probabilities to possible RS trees of sentences. The probability of an entire RS tree is the product of the probabilities for each of the rules used therein.

Ito, *et.al.* [10] construct RS trees by using linguistic clues and rules to identify relation types, i.e., clausal-sequence, conjunction, means and circumstance, and using features of subject and verb in the clauses to predicate adjacent child units of the relations.

For Thai language, Sukvaree, *et.al.* [21] purpose a technique to construct an RS tree by using global and local spanning trees which makes decisions by discourse markers.

This article proposes a new approach to Thai RS Tree construction which consists of two major steps: EDU segmentation and RS tree construction. Two Hidden Markov models constructed from syntactic properties of Thai language are used to segment EDUs, and a clustering technique with its similarity measure derived from semantic properties of Thai language is then used to construct a Thai RS tree.

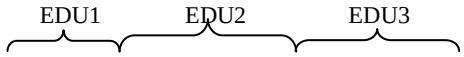
2 ISSUES IN THAI RS TREE CONSTRUCTION

Thai language has unique characteristics both syntactically and semantically. This makes techniques proposed for other languages not

directly applicable to Thai language. A number of important issues with respect to constructions of Thai RS trees are discussed in this section.

2.1 No Explicit EDU Boundaries

Unlike English, Thai language has no punctuation marks (e.g., comma, full stop, semi-colon, and blank) to determine the boundaries of EDUs. Therefore, EDU segmentation in Thai language becomes a nontrivial issue.



Thai : [w1w2…wmwm+1wm+2…wnwn+1wn+2…wo]
English : [w1 w2 … wm],[wm+1 wm+2 … wn];[wn+1 wn+2 … wo].
Where w_i is a word in text.

2.2 EDU Constituent Omissions

Given two EDUs, an absence of subject, object or conjunction in the anaphoric EDU may happen, such as a situation where an anaphoric EDU omits the subject that refers back to the object of the cataphoric EDU. Accordingly, EDU boundaries are ambiguous.

Thai text :	“เพื่อนจะขอยืมหนังสือ เพราะหาซื้อไม่ได้” (A friend’s going to borrow this book because she hasn’t been able to find it.)
Three possibilities :	1) [S(เพื่อน)V(จะขอยืม)O(หนังสือ)]$_{EDU1}$ [because S(Φ) V(หาซื้อไม่ได้)]$_{EDU2}$
	2) [S(เพื่อน)V(จะขอยืม)O(หนังสือ)]$_{EDU1}$ [because(Φ)S(Φ)V(หาซื้อไม่ได้)]$_{EDU2}$
	3) [S(เพื่อน)V(จะขอยืม)O(Φ)]$_{EDU1}$ [because(Φ)S(หนังสือ)V(หาซื้อไม่ได้)]$_{EDU2}$

In addition, the absence of subject, object or preposition which is a modifier nucleus of VP especially in the anaphoric EDU makes the use of word co-occurrence alone not sufficient to determine the relation between EDU1 and EDU2. For example,

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage properties.)

EDU2: Φ1 จะสั่งยกเลิกการแยก Φ2 ได้ (Φ1 can cancel the partition of Φ2.)

In the example, EDU2 omits subject “ศาล” (court) and object ‘สินสมรส’ (marriage properties). Therefore, word co-occurrence alone is not sufficient to determine this relation.

2.3 Implicit Markers

The absences of discourse markers in Thai language are often occurred. In the example below, “แต่” (but) is a discourse marker which is omitted, but the relation between EDU1 and EDU2 is still able to determine.

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage property.)
EDU2: Φ ภริยาหรือสามีคัดค้าน (Φ a wife or a husband may contest.)

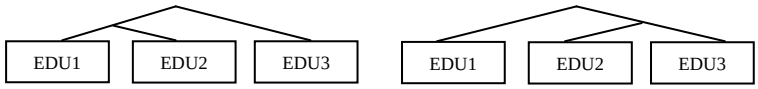
Therefore, considering markers or cue phrases alone is not sufficient to determine the relation between EDUs.

2.4 Adjacent Markers

Given three EDUs with two markers, as shown in the example below, two RS Trees are possible.

EDU1: ศาลได้มีคำสั่งให้แยกสินสมรส (A court has ordered partition of marriage properties.)
EDU2: แต่ถ้าภริยาหรือสามีคัดค้าน (but if a wife or a husband contests,)
EDU3: ศาลจะสั่งยกเลิกการแยกได้ (the court can cancel the partition.)

The first possibility, EDU1 and EDU2 relate first by a discourse marker “แต่” (but), next (EDU1, EDU2) and EDU3 relate by a marker “ถ้า” (if). For the other possibility, EDU2 and EDU3 relate first by a marker “ถ้า” (if), next that between (EDU2, EDU3) and EDU1 relate by a marker “แต่” (but).



a) The RS tree with “but” applied first b) The RS tree with “if” applied first

Fig. 1. Adjacent markers issue

3 STRUCTURES OF THAI EDUS

A Thai EDU consists of infrastructure and adjunct constituents. The twelve possible arrangements of Thai EDUs [17] are shown in Table 1. The structure of an EDU "A teacher usually doesn't drink alcohol" is shown in Fig. 2.

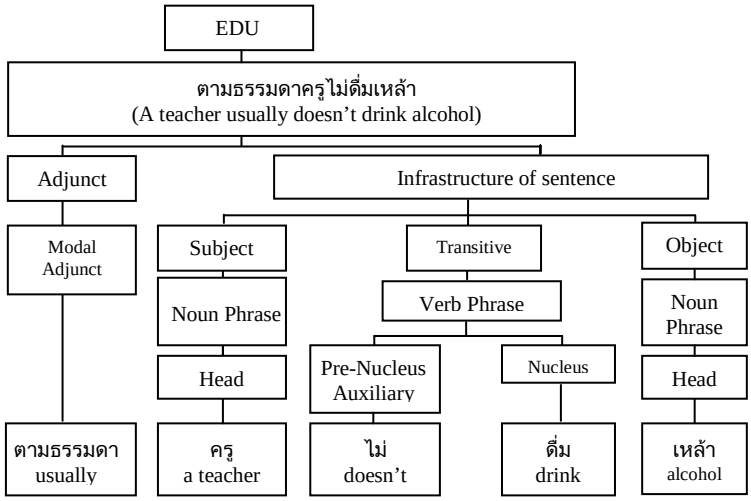


Fig. 2. Structure of the EDU "A teacher usually doesn't drink alcohol."

4 EDU SEGMENTATION

This section describes the EDU segmentation technique proposed in this research. To reduce the segmentation ambiguities caused from omissions of words or discourse markers, and the appearances of modifiers, noun phrases and verb phrases which are constituents of EDUs are first determined, according to the syntactic properties of Thai language. These phrases are then used to identify boundaries of EDUs.

Table 1: The possible arrangements of Thai EDUs.

EDUs	**Examples**	**Rules**
Vi	หิว (I'm hungry.)	NP_S-*Vi*-NP_S
S-Vi	ฝน-ตก (It's rain.)	

Vi-S	เจ็บไหม-คุณ (Are you pain?)	
Vt-O	หิว- น้ำ (I'm hungry.)	NP_O-NP_S-*Vt*-NP_O
S-Vt-O	รถ-ชน-เด็ก (The car hit the boy.)	
O-S-Vt	รูปนี้-ฉัน-ดูแล้วละจ๊ะ (I've already seen this photograph.)	
Vtt-O-I	ยังไม่ได้ให้-ยา-คนไข้ (I haven't given the patient the medicine.)	NP_S-*Vtt*-NP_O-NP_I
S-Vtt-O-I	ใคร-ให้-ลูกกวาด-หนู (Who gave you the sweet?)	
O-S-Vtt-I	ความลับ-ใครละ-จะกล้าถาม- คุณ (Who would dare to ask you the secret?)	NP_O-NP_S-*Vtt*-NP_I
I-S-Vtt-O	หนู-ป้า-จะให้-บ้านนี้ (Niece, I am going to give you this house.)	NP_I-NP_S-*Vtt*-NP_O
N	ป้า (Auntie)	NP_N-NP_N
N-N	นี่ปากกา-ใคร (Whose pen is this?)	
N-N	นี่ปากกา-ใคร (Whose pen is this?)	

A noun phrase (NP) is a noun or a pronoun and its expansions which may function as one of the four Thai EDU constituents, namely subject (S), object (O), indirect object (Oi) and nomen (N). The general structure of a noun phrase consists of five constituents which are: head (H), intransitive modifier (Mi), adjunctive modifier (Ma), quantifier (Q), and determinative (D).

A verb phrase (VP) is a verb and its expansions which may function as one of the three Thai EDU constituents, namely intransitive verb (Vi), transitive verb (Vt) and double transitive verb (Vtt). The general structure of a verb phrase consists of four constituents which are: nucleus (Nuc), pre-nuclear auxiliary (Aux1), post-nuclear auxiliary (Aux2), and modifier (M).

There are twenty five possible arrangements of noun phrase and ten arrangements of verb phrases [17], which are shown in Table 2.

4.1 Phrase Identification

To perform phrase identification, word segmentation and part of speech (POS) tagging are performed using SWATH [7] which extracts words and classifies them into 44 types such as common noun (NCMN), active verb (VACT), personal pronoun (PPRS), definite determiner

(DDAC), unit classifier (CNIT) and negate (NEG). A hidden Markov model (HMM) [11] employs these POS tag categories to determine phrases. The model assumes that at time step *t* the system is in a hidden state *PC(t)* which has a probability b_{jk} of emitting a particular visible state of POS tag *tag(t)*, and a transition probability between hidden states a_{ij}:

$$a_{ij} = p(PC_j(t+1)|PC_i(t)). \quad (1)$$

$$b_{jk} = p(tag_k(t)|PC_j(t)). \quad (2)$$

where *PC(t)* is the phrase constituent at time step *t*, and *tag(t)* is POS tag at time step *t*.

Table 2: The possible arrangements of Thai NPs and VPs.

Noun Phrases	**Noun Phrases (cont.)**	**Verb Phrases**
H-Ma	H	Nuc
H-Mi-Ma	H-Mi	Nuc-Aux2
H-Q-Ma	H-Q	Nuc-M
H-Ma-Q	H-D	Nuc-Aux2-M
H-D-Ma	H-Mi-Q	Nuc-M-Aux2
H-Mi-Q-Ma	H-Q-Mi	Aux1-Nuc
H-Q-Mi-Ma	H-Mi-D	Aux1-Nuc-Aux2
H-Mi-D-Ma	H-Q-D	Aux1-Nuc-M
H-Q-D-Ma	H-D-Q	Aux1-Nuc-Aux2-M
H-D-Q-Ma	H-Mi-Q-D	Aux1-Nuc-M-Aux2
H-Mi-Q-D-Ma	H-Mi-D-Q	
H-Mi-D-Q-Ma	H-Q-Mi-D	
H-Q-Mi-D-Ma		
H-Q-Mi-D-Ma		

The probability of a sequence of *T* hidden states PC^T = *{PC(1), PC(2), …, PC(T) }* can be written as:

$$p(PC^T) = \prod_{t=1}^{T} p(PC(t) \mid PC(t-1)) \quad (3)$$

The probability that the model produces the corresponding sequence of POS tag tag^T, given a sequence of PCs PC^T can be written as:

$$p(tag^T \mid PC^T) = \prod_{t=1}^{T} p(tag(t) \mid PC(t)) \quad (4)$$

Then, the probability that the model produces a sequence tag^T of visible POS tag states is:

$$p(tag^T) = \arg\max_{PC_{1,n}} \prod_{t}^{T} p(tag(t) \mid PC(t))\, p(PC(t) \mid PC(t-1)) \quad (5)$$

The Baum-Welch [11] learning algorithm is applied to determine model parameters, i.e., a_{ij} and b_{jk}, from an ensemble of training samples.

Given a sequence of visible state tag^T, the Viterbi algorithm [11] is used to find the most probable sequence of hidden states by recursively calculating $p(tag^T)$ of visible POS states. Each term $p(tag(t)|PC(t))$ $p(PC(t)|PC(t-1))$ involve only $tag(t)$, $PC(t)$, and $PC(t-1)$ by the following definition:

$$\delta_t(j) = \begin{cases} 0, & t = 0 \text{ and } j \neq \text{initial state} \\ 1, & t = 0 \text{ and } j = \text{initial state} \\ \arg\max_i \delta_{t-1}(i) a_{ij} b_{jkt}, & \text{otherwise} \end{cases} \quad (6)$$

Figure 3 shows a phrase identification model of string “เพื่อนจะขอยืมหนังสือเล่มนี้ เพราะΦ_1ซื้อไม่ได้Φ_2 ดังนั้นΦ3จึงต้องยืมหนังสือฉัน” (A friend’s going to borrow this book. Because she (Φ_1) hasn’t been able to buy it (Φ_2). Therefore she (Φ_3) must borrow it from me.) POS tags of the string is “เพื่อน (A friend-NCMN) จะขอ (is going to-XVMM) ยืม (borrow-VACT) หนังสือ (book-NCMN) เล่ม (numerative-CNIT) นี้ (this-DDAC) เพราะ (Because-CONJ) เธอ (she(Φ_1)-PPRS) ไม่ (hasn’t been-NEG) สามารถ (able to-XVMM) ซื้อ (buy-VACT) มัน (it(Φ_2)) ดังนั้น (Therefore-CONJ) เธอ (she(Φ_3)-PPRS) จึงต้อง (must-XVMM) ยืม (borrow-VACT) หนังสือ(book-NCMM) ฉัน (me-PPRS)”.

The hidden state of a phrase model consists of H(NCMN-book (2/4), -friend (1/4); PPRS-me (1/4)), D(CNIT-numerative (1/2); DDAC-this (1/2)), Discourse-marker(CONJ-because (1/2), -therefore (1/2)), Aux1(XVMM-is going to (1/4), -must (1/4), -able to (1/4); NEG-hasn’t been (1/4)) and Nuc(VACT-borrow (2/3), -buy (1/3)).

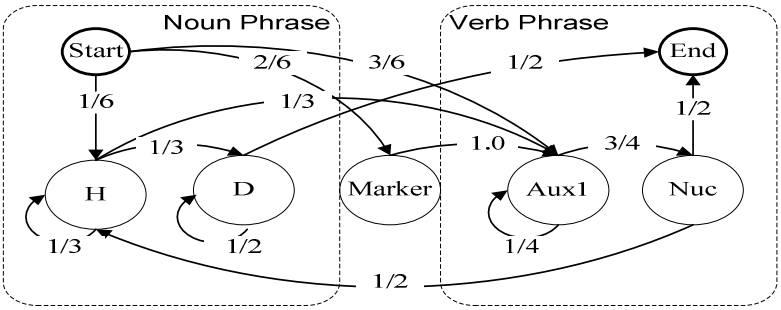


Fig. 3. A phrase identification model.

		เพื่อน	จะขอ	ยืม	หนังสือ	เล่ม	นี้	
	Start	NCMN	XVMM	VACT	NCMN	CNIT	DDAC	END
Start	1	0	0	0	0	0	0	0
H	0	1/6*3/4	0	0	$8*10^{-3}$	0	0	0
D	0	0	0	0	0	$1*10^{-3}$	$3*10^{-4}$	0
Marker	0	2/6*0	0	0	0	0	0	0
Aux1	0	3/6*0	$3*10^{-2}$	0	0	0	0	0
Nuc	0	0	0	$2*10^{-2}$	0	0	0	0
End	0	0	0	0	0	0	0	$1*10^{-4}$
T =	0	1	2	3	4	5	6	7
Output	Start	H	Aux1	Nuc	H	D	D	End

Fig.4. The results of Viterbi tagging on the phrase identification model in Fig 3.

4.2 EDU Boundary Determination

After we determine NPs and VPs, another HMM on EDU constituents (shown in Fig. 5.) is then created to determine the boundaries of EDUs. This model can handle the subject and object omission problems, discussed earlier.

Fig. 5 shows an example of the EDU segmentation model for an EDU "เพื่อน-จะขอ-ยืม-หนังสือ-เล่ม-นี้" (A friend's going to borrow this book.)

The EDU segmentation model can be expressed as:

$$p(tag^T) = \arg \max_{EDUC_{1,n}} \prod_t^T p(tag(t) \mid EDUC(t)) p(EDUC(t) \mid EDUC(t-1)) \quad (7)$$

where *EDUC(t)* is EDU constituent at time step *t*, and *tag(t)* is the phrase tag at time step *t*.

The expression, *p(EDUC(t)|EDUC(t-1))* is the probability of EDU constituent (*EDUC*) at time *t* given the previous *EDUC(t-1)*, and *p(tag(t)|EDUC(t))* is the probability of phrase tag *tag(t)* given *EDUC(t)*.

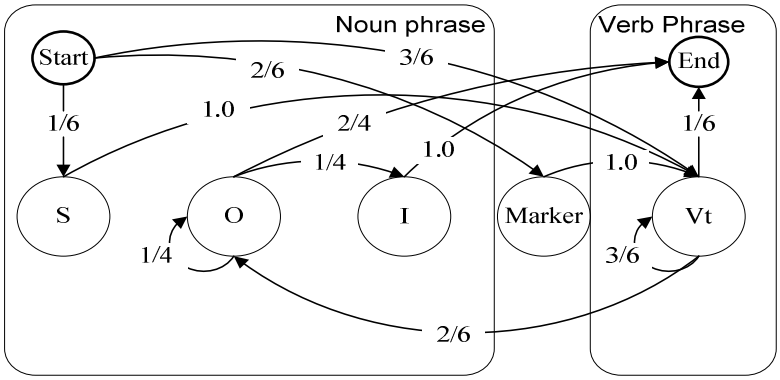


Fig.5. An example of a Thai EDU segmentation model.

		เพื่อน	จะขอ	ยืม	หนังสือ	เล่ม	นี้	
	Start	H	Aux1	Nuc	H	D	D	END
Start	1	0	0	0	0	0	0	0
S	0	1[1/6*1]	0	0	0	0	0	0
O	0	0	0	0	$3*10^{-3}$	$6*10^{-4}$	$1*10^{-4}$	$5*10^{-5}$
I	0	0	0	0	0	0	0	0
Marker	0	1[2/6*0]	0	0	0	0	0	0
Vt	0	1[3/6*0]	$9*10^{-2}$	$2*10^{-2}$	0	0	0	0
End	0	0	0	0	0	0	0	0
t =	0	1	2	3	4	5	6	7
Output	Start	S	Vt	Vt	O	O	O	End

Fig.6. The results of Viterbi tagging
on the Thai EDU segmentation model in Fig.5.

4.3 EDU Constituent Grouping

Once EDU boundaries are determined, syntactic rules in Table 1 are then applied to group EDU constituents into a larger unit that will be used to match the semantic rules in further steps, For example a string "เพื่อน-จะขอ-ยืม-หนังสือ-เล่ม-นี้" (A friend's going to borrow this book.), the result from the Viterbi tagging on the EDU segmentation model is S, Vt, Vt, O, O, O. The matched rule of "NP_O-NP_S-Vt-NP_O" is applied, and the result becomes: "NP_S – (V, V)$_t$ – (NP, NP, NP)$_O$."

5 THE REFERENCES SECTION

In this section, we describe our proposed technique based on semantic rules derived from Thai linguistic characteristics to construct an RS tree from a corpus. The rules are classified into three types which are Absence, Repetition, and Addition rules [2, 3, 4, 12, 17]. Given a pair of EDUs, an author may write by using any combination of the rules. A similarity measure is calculated from these rules, and a hierarchical clustering algorithm employing this measure is used to construct an RS tree.

5.1 Semantic Rules for EDU Relations

Absence Rules

In Thai language, it has been observed that frequently in writings some constituents of an EDU may be absent while its meaning remains the same. In the example below, the NP (object) "ขนม" (dessert) is absent from the anaphoric EDU, according to rule Φ (O, O).

Cataphoric EDU (Vt-O) : อยากจะทำขนมไหม (Would you like to make a dessert?)
Anaphoric EDU (Vt) : อยากจะทำ (Yes, I do.)

Repetition Rules

It has been observed that frequently an anaphoric EDU relates to its cataphoric EDU by a repetition of NP (subject, object) or a preposition phrase (PP) functioning as a modifier of a nucleus or a verb phrase (VP). In the following example, two EDUs relate by a repetition of an object (NP) "บ้าน" (house), according to the rule ฯ (O, O).

Cataphoric EDU (Vtt-O-I) : ผมกำลังจะขายบ้านให้เขา (I'm going to sell him a house.)
Anaphoric EDU (Vt-O) : จะขายบ้านหลังไหน (Which house are you going to sell?)

Addition Rules

It has been observed that frequently an anaphoric EDU relates to its cataphoric EDU by an addition of a discourse marker, and possibly accompanied by Absence and/or Repetition rules. In the example below a discourse marker "เพราะ" (because) is added in front of the anaphoric EDU, according to the rule Д (Marker, Before).

Cataphoric EDU (Vtt-O-I) : ฉันอยากจะยืมหนัง (I want to borrow films.)

Anaphoric EDU (Vt-O) : เพราะหาซื้อไม่ได้ (because I have not been able to buy it.)

Table 3 lists Repetition, Absence, and Addition rules, for example, я (S, S) means that the subject of the cataphoric EDU is repeated in the anaphoric EDU; Φ(S, S) means that the subject is present in the cataphoric EDU but absent from the anaphoric EDU; and Д (Marker, Before) means that a discourse marker is added in front of this particular EDU.

5.2 EDU Similarity

Similarity between two EDUs can be calculated from the semantic rules in Table 3, as follows:

5.2.1 Feature Calculations

Given a pair of EDUs, for each rule, an EDU calculates a feature vector which consists of the following elements: Subject, Absence of Subject, Object, Absence of Object, Preposition, Absence of Preposition, Nucleus, Modifier Nucleus, Head, Absence of Head, Modifier Head, Absence of Modifier Head, Marker Before, and Marker After elements. The value of an element is dependent upon the type of rule, as follows:

Table 3: Repetition, Absence, and Addition rules.

Repetition (Я)	Absence (Ф)	Addition (Д)
я (S, S)	Ф (S, S)	Д (Marker, After)
я (O, S)	Ф (O, S)	Д (Marker, Before)
я (S, O)	Ф (S, O)	Д (Key Phrase, After)
я (O, O)	Ф (O, O)	Д (Key Phrase, Before)
я (S, Prep)	Ф (Only H, H)	
я (O, Prep)	Ф ((H, M), H)	
я (Prep, S)	Ф ((H, M), M)	
я (Prep, O)	Ф (S, Prep)	
я ((S, Prep), (S, Prep))	Ф (O, Prep)	
я ((O, Prep), (S, Prep))	Ф (Prep, S)	
я ((Prep, Prep), (S, Prep))	Ф (Prep, O)	
я ((S, Prep), (O, Prep))		
я ((O, Prep), (O, Prep))		
я((Prep, Prep), (O, Prep))		
я (Only H, Only H)		
я (H, M)		
я (Only M, Only Nuc)		
я (Only M, Only M)		
я ((Nuc, M), (Nuc, M))		

The following example is used to illustrate calculations related to semantic rules:

EDU1: ชาวบ้าน (Subject) ประกอบ (Nucleus) อุตสาหกรรมในครอบครัว (Object) (The villagers
perform the family-industry.)

EDU2: และ (Before) Ф (Absence of Subject) หวงแหน (Nucleus) สมบัติของชาติ (Object) (and
protect properties of the nation.)

EDU3: อุตสาหกรรมในครอบครัว (Subject) จึงเป็น (Nucleus) สมบัติของชาติ (Object) (Therefore, the
family-industry is a property of the nation.)

To describe the calculations related to semantic rules, the following notations will be used. C_{Cat} is a constituent of the cataphoric EDU, C_{Ana} is a constituent of the anaphoric EDU, Pos_{Cat} is the position of cataphoric EDU, and Pos_{Ana} is the position of anaphoric EDU. *X:Y* where *X* can be either Cataphoric or Anaphoric, and *Y* is an element in the vector of X, e.g., *Cataphoric:Subject* is the Subject element in the vector of the cataphoric EDU. *X:rule* is an Addition rule applied to *X* (i.e., a cataphoric or an anaphoric EDU).

Features based on an Absence rules:

Feature vectors of the cataphoric and anaphoric EDUs are filled for a

matched Absence rule, as follows:

$$If\ \Phi(C_{Cat}, C_{Ana})\ is\ true\ then$$
$$Cataphoric C_{Cat} = Anaphoric(Absence\ of\ C_{Ana}) = 1 - \frac{|Pos_{Cat} - Pos_{Ana}|}{Total\ \#\ of\ sentences} \quad (8)$$

In this example, the properties of EDU1 and EDU2 match with the rule *Φ(S, S)* with the absence of subject "ชาวบ้าน" (villager) in the anaphoric EDU, thus:

$$Cataphoric: Subject = Anaphoric: Absence\ of\ Subject = 1 - \frac{|1-2|}{3} \quad (9)$$

Features based on Repetition rules:

Feature vectors of the cataphoric and anaphoric EDUs is filled for a matched Repetition rule, as follows:

$$If\ \Re(C_{Cat}, C_{Ana})\ is\ true\ then$$
$$Cataphoric: C_{Cat} = Anaphoric: C_{Ana}$$
$$= \frac{|Pos_{Cat} - Pos_{Ana}|}{Total\ \#\ of\ sentences} * \frac{Total\ \#\ of\ repeating\ words}{Total\ \#\ of\ words\ in\ sentences} \quad (10)$$

In the example, the properties of EDU1 and EDU3 match with the rule я (O, S) with a repetition of an object "อุตสาหกรรมในครอบครัว" (family-industries) in the cataphoric EDU as a subject in the anaphoric EDU, thus:

$$Cataphoric: Object = Anaphoric: Subject = (1 - \frac{1-3}{3}) * (\frac{1}{3} * \frac{1}{3}) \quad (11)$$

Features based on Addition rules:

Feature vectors of the cataphoric and anaphoric EDUs is filled for a matched Addition rule, as follows:

$$\begin{array}{c} If\ Cataphoric{:}Д\ (Marker, After)\ is\ true\ then \\ Cataphoric{:}Marker\ After = Anaphoric{:}Marker\ Before = 1 \\ else\ if\ Anaphoric{:}Д\ (Marker, Before)\ is\ true\ then \\ Anaphoric{:}Marker\ Before = Cataphoric{:}Marker\ After = 1 \end{array} \quad (12)$$

In this example, the properties of EDU1 and EDU2 match with the rule Д (Marker, Before) at EDU2, thus:

$$Anaphoric{:}Marker\ Before = Cataphoric{:}Marker\ After = 1 \quad (13)$$

5.2.2 Rule Scoring

After for each rule, the two vectors of the EDU pair are calculated, the vectors are then combined into a rule score which depends on the type of rule and the distance between the two EDUs, as follows:

Absence and Repetition Rules:
These rules consist of two parts (cataphoric and anaphoric). If both parts of an Absence or a Repetition rule are true, then the rule is true. But if a part of an Absence or a Repetition rule is false, then the rule is false, thus:

$$if \mid Pos_{Cat} - Pos_{Ana} \mid < MD \ \ then$$

$$RS_{\substack{Absense \\ or \\ Repetition}} = [Magnitude\ of\ \ EDU_{Cataphoric} \ * \ Magnitude\ of\ \ EDU_{Anaphoric}] \quad (14)$$

where Pos_{Cat} and Pos_{Ana} are the positions of cataphoric and anaphoric EDUs, and MD is the maximum distance between the EDUs (from experiments $MD = 4$ in this research)

Addition Rules:
In this type of rules, if one part of the rule is true, then the rule is true, thus:

$$if \mid Pos_{Cat} - Pos_{Ana} \mid < MD \ then$$

$$RS_{Addition} = [Magnitude\ of\ \ EDU_{Cataphoric} + Magnitude\ of\ \ EDU_{Anaphoric}] \quad (15)$$

5.2.3 Rule Scoring
Once rule scores are available, similarity between two EDUs (cataphoric and anaphoric) can be calculated as a sum of all the rule scores (each normalized into a range from 0 to 1) according to the CombSum method [8].

6 RHETORICAL TREE CONSTRUCTION

A hierarchical clustering algorithm is applied to create an RS tree where each sample (an EDU in this case) begins in a cluster of its own; and while there is more than one cluster left, two closest clusters are combined into a new cluster, and the distance between the newly formed cluster and each other cluster is calculated. Hierarchical clustering algorithms studied in this research are shown in Table 4, and two example RS trees created from two different algorithms are shown in Fig. 7.

Table 4. Hierarchical clustering algorithms studied in this research.

Algorithms	Distance Between Two Clusters
Single Linkage	The smallest distance between a sample in cluster A and a sample in cluster B.
Unweighted Arithmetic Average	The average distance between a sample in cluster A and a sample in cluster B.
Neighbor Joining	A sample in cluster A and a sample in cluster B are the nearest. Therefore, define them as neighbors.
Weighted Arithmetic Average	The weighted average distance between a sample in cluster A and a sample in cluster B.
Minimum Variance	The increase in the mean squared deviation that would occur if clusters A and B were fused.

7 EXPERIMENTAL EVALUATION

7.1 Rule Scoring

In order to evaluate the effectiveness of the EDU segmentation process, a consensus of five linguists, manually segmenting EDUs of Thai family law, is used. The dataset consists of 10,568 EDUs in total.

The EDU segmentation model is trained with 8,000 random EDUs, and the rest are used to measure performance.

The training continues until the estimated transition probability changes no more than a predetermined value of 0.02, or the accuracy achieves 98%.

The performances of both phrase identification and EDU segmentation are evaluated using recall (Eq. 16) and precision (Eq. 17) measures, which are widely used to measure performance.

$$\mathrm{Re}\,call = \frac{\#\,correct\ \ (phrases\ \ or\ \ EDUs)\ \ identified\ \ by\ \ HMM}{\#\,(phrase\ \ or\ \ EDUs)\ \ identified\ \ by\ \ linguists} \tag{16}$$

$$\mathrm{Pr}\,ecision = \frac{\#\,correct\ \ (phrases\ \ or\ \ EDUs)\ identified\ \ by\ \ HMM}{total\ \#\,(phrases\ \ or\ \ EDUs)\ identified\ \ by\ \ HMM} \tag{17}$$

The results show that the proposed method achieves the recall values of 84.8% and 85.3%; and the precision values of 93.5% and 94.2% for phrase identification and EDU segmentation, respectively.

7.2 Evaluation of EDU Constituent Grouping

In order to evaluate the effectiveness of the EDU constituent grouping, three corpuses are used which consist of Absence data (84 EDUs), Repetition data (117 EDUs) and a subset of the Family law with 367 EDUs). The Absence data contains EDUs mostly those following the Absence rules while the Repetition data contains mostly those following the Repetition rules. Five linguists create training and testing data sets by manually grouping EDU constituents.

Table 5 shows the results of grouping EDU constituents (subject (S), object (O), indirect object (I) and nomen (N)) by using rules based on NPs, assuming the positions of verb phrases (Vi, Vt and Vtt) are known. From the results, in general all rules, except NP_O-NP_S-Vtt-NP_I and NP_I-NP_S-Vtt-NP_O, perform well.

Table 5: Performance of grouping EDU constituents

Rules	Absence Data	Repetition Data	Family Law
NP_S-Vi-NP_S	NP_S (100%)	NP_S (100%)	NP_S (100%)
NP_O-NP_S-Vt-NP_O	NP_S & NP_O (100%)	NP_S &NP_O (100%)	NP_S &NP_O (100%)
NPS-Vtt-NPO-NPI	NP_S &NP_O&NP_I (100%)	NP_S &NP_O&NP_I (100%)	NP_S &NP_O&NP_I (100%)
NP_O-NP_S-Vtt-NP_I NP_I-NP_S-Vtt-NP_O	NP_S (100%), NP_O&NP_I (91.37%)	NP_S (100%), NPO&NP_I (79.59%)	NP_S (100%), NP_O&NP_I (90.21%)
N-N	NP_N (100%)	NP_N (100%)	NP_N (100%)

To further resolve ambiguities with respect to these two rules, a probability table of terms in positions of NP_I and NP_O following Vtt ($P(Vtt| NP_I, NP_O)$) is used. The results of determining functions of EDU constituents by using the rules based on NPs together with the probability table show higher performance for Absence data (92.24%), Repetition data (85.78%), and Family law (93.71%).

7.3 Evaluation of Thai RS Tree Construction

In order to evaluate the effectiveness of the proposed Thai RS tree construction process, linguists manually construct the rhetorical structure trees of three texts used above with a total of 568 EDUs. The

algorithms are evaluated by using recall (Eq. 18) and precision (Eq. 19) measures. Recall and precision are calculated with respect to how close an RS tree constructed from the proposed technique to that created by a consensus of the linguists.

$$\text{Re } call = \frac{\#\, correct \;\; internal \;\; nodes \;\; identified \;\; by \;\; RS \;\; Tree}{\#\, internal \;\; nodes \;\; identifi \;\; ed \;\; by \;\; linguists} \quad (18)$$

$$\text{Pr } ecision = \frac{\#\, correct \;\; \text{internal nodes} \;\; identified \;\; by \;\; \text{RS Tree}}{\text{Total \# of internal nodes} \;\; identified \;\; by \;\; \text{RS Tree}} \quad (19)$$

For the Absence and Repetition data sets, though relations between EDUs follow mostly Absence rules and Repetition rules, respectively, in reality when examined in details, many types of rules are used together in writing. For example,

Anaphoric EDU (S-Vt-O) : บุรุษไปรษณีย์ (S) จะคัดเลือก (Vt) จดหมาย (ย O)
(A Postman will sort letters)

Cataphoric EDU ((S)-Vt-O) : และ (Д) (Φ S) รับส่ง (Vt) จดหมาย (ย O)
(And will deliver letters)

Table 6 shows calculations of recall and precision of RS trees created by the Minimum Variance and Unweighted Arithmetic Average algorithms, in Fig. 7.

Table 7 shows the results of evaluating Thai RS Tree construction on the three data sets. The performance on the Family law dataset which combines many kinds of rules in its content is 94.90% recall and 95.21% precision. The results also show that Unweighted Arithmetic Average clustering algorithm gives the best performance for Thai RS Tree construction.

8 CONCLUSIONS

Thai rhetorical structure tree (RST) construction is an important task for many textual analysis applications such as automatic text summarization and question-answering. This article proposes a novel two-step technique to construct Thai RS tree combining machine learning techniques with linguistic properties of the language.

Table 6: RS tree construction performance of two clustering algorithms

The correct RS tree	Minimum Variance	Unweighted Arithmetic Average
3’	3’	3’
4’	4’	4’
1’	1’	1’
9’	9’	6’
2’	2’	2’
5’	5’	5’
6’	6’	
7’	7’	
8’	8’	
		7’
		8’
		9’
		10’
	Precision = 9/9	Precision = 6/10
	Recall = 9/9	Recall = 6/9

Table 7: Performance of the RS tree construction

Data	Num EDUs	Clustering Method	Recall	Precision
Absence	84	Neighbor Joining	87.23	89.13
		Single Linkage	82.97	84.78
		Un weighted Arithmetic Average	87.23	89.13
		Minimum Variance	89.40	91.30
		Weighted Arithmetic Average	87.23	89.13
Repetition	117	Neighbor Joining	89.70	91.04
		Single Linkage	83.82	85.07
		Unweighted Arithmetic Average	89.70	91.04
		Minimum Variance	77.94	79.10
		Weighted Arithmetic Average	89.70	91.04
Family-Law	367	Neighbor Joining	85.98	86.26
		Single Linkage	64.01	64.21
		Unweighted Arithmetic Average	94.90	95.21
		Minimum Variance	63.37	63.57
		Weighted Arithmetic Average	90.44	90.73

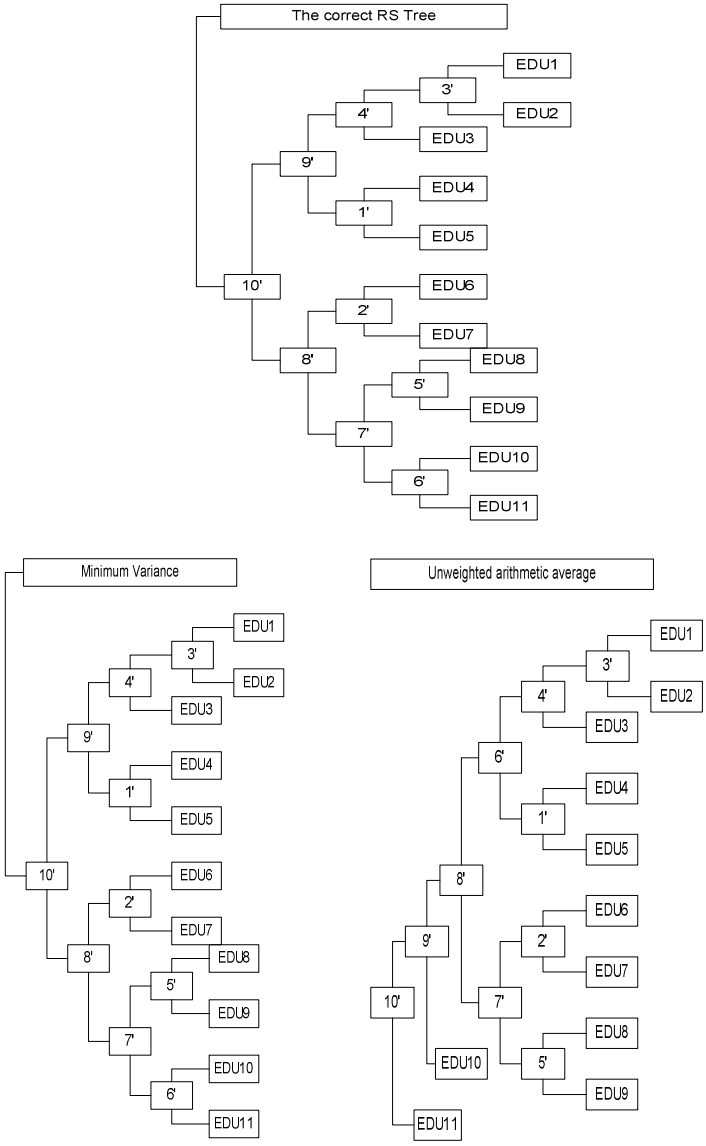


Fig. 7. RS trees from two hierarchical clustering algorithms

First, phrases are determined and then are used to segment elementary discourse units (EDUs). The phrase segmentation model is a hidden Markov model constructed from the possible arrangements of Thai phrases based on part-of-speech of words, and the EDU segmentation model is another hidden Markov model constructed from the possible phrase-level arrangements of Thai EDUs. Linguistic rules are applied after EDU segmentation to group related constituents into a large unit. Experiments show the EDU segmentation effectiveness of 85.3% and 94.2% in recall and precision, respectively.

A hierarchical clustering algorithm whose similarity measure derived from semantic rules of Thai language is then used to construct an RS tree. The technique is experimentally evaluated, and the effectiveness achieved is 94.90% and 95.21% in recall and precision, respectively.

REFERENCES

1. Alonso, L. and Castellon, I.: Towards a delimitation of discursive segment for Natural Language Processing applications. International Workshop on Semantics, Pragmatics and Rhetorics, San Sebastian (2001)
2. Aroonmanakun, W.: Referent Resolution for Zero Pronouns in Thai. Southeast Asian Linguistic Studies in Honour of Vichin Panupong. (Abramson, Arthur S., ed.). pp. 11-24. Chulalongkorn University Press, Bangkok. ISBN 974-636-995-4 (1997)
3. Aroonmanakun, W.: Zero Pronoun Resolution in Thai: A Centering Approach. In Burnham, Denis, et.al. Interdisciplinary Approaches to Language Processing: The International Conference on Human and Machine Processing on Human and Machine Processing of Language and Speech. NECTEC: Bangkok, 127-147 (2000)
4. Chamnirokasant, D.: Clauses in the Thai Language. Unpublished master's thesis, Chulalongkorn University, Thailand (1969)
5. Chaniak, E.: Statistical Techniques for Natural Language Parsing. Department of Computer Science, Brown University. August 7 (1997)
6. Charoensuk, J. and Kawtrakul, A.: Thai Elementary Discourse Unit Segmentation by Discourse Segmentation Cues and Syntactic Information, The Sixth Symposium on Natural Language Processing 2005 (SNLP 2005), Chiang Rai, Thailand, December 13-15 (2005)
7. Charoenporn, T., Sornlertlamvanich, V., Isahara, H.: Building A Large Thai Text Corpus---Part-Of-Speech Tagged Corpus: ORCHID---. Proceedings of the Natural Language Processing Pacific Rim Symposium (1976)
8. Fox, E. A. and Shaw, J. A. Combination of multiple searches. In the second Text Retrieval conference (TREC-2), Gaithersburg, MD, USA,

March 1994. U.S. Government Painting Office, Washington D.C, pages 243-249 (1994)
9. Harman, D. K., editor.: The second Text Retrieval conference (TREC-2), Gaithersburg, MD, USA, March 1994. U.S. Government Painting Office, Washington D.C (1994)
10. Ito, N. Sugimoto, T. Iwasita, S. Kobayashi, I. and Sugeno, M.: A Model of Rhetorical Structure Analysis of Japanese Instruction Texts and its Application to a Smart Help System. In IEEE international Conference on Systems, Man and Cybernetics (2004)
11. Levinson, S., Rabiner, R., and Sondhi, M.: An introduction to the application of the theory of probabilistic function of a Markov proceeds to automatic speech recognition. Bell System Technical Journal, 62:1035-1074 (1983)
12. Mahatdhanasin, D.: A study of sentence groups in Thai essays. Unpublished master's thesis, Chulalongkorn University, Thailand (1980)
13. Mann, W. C. and Thompson, S. A.: Rhetorical structure theory. Toward a functional theory of text organization. Text, 8(3): 243-281 (1988)
14. Marcu, D.: Build Up Rhetorical Structure Threes, American Association for Artificial Intelligence (1996)
15. Marcu, D.: A decision-based approach to rhetorical parsing, The 37th Annual Meeting of the Association for Computational Linguistics, ACL, Maryland, pp. 365-372 (1999)
16. Marcu, D.: The theory and Practice of Discourse Parsing and Summarization. The MIT Press, Cambridge, MA (2000)
17. Panupong, V.: Inter-Sentence Relations in Modern Conversational Thai. The Siam Society, Bangkok (1970)
18. Polanyi, L.: A formal model of the structure of discourse. Journal of Pragmatics, 12, 601-638 (1988)
19. Soricut, R. and Marcu, D.: Sentence Level Discourse Parsing using Syntactic and Lexical Information. In Proceedings of the 2003 Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), May 27-June 1, Edmonton, Canada (2003)
20. Subba, R., Di Eugenio, B., S. N. K.: Learning FOL rules based on rich verb semantic representations to automatically label rhetorical relations. EACL 2006, Workshop on learning Structured Information in Natural Language Applications (2006)
21. Sukvaree, T., Charoensuk, J., Wattanamethanont, M. and Kultrakul, A.: RST based Text Summarization with Ontology Driven in Agriculture Domain. Department of Computer Engineering, Kasetsart University, Bangkok, Thailand (2004)

SOMNUK SINTHUPOUN
DEPARTMENT OF COMPUTER SCIENCE,
MAEJO UNIVERSITY, CHIANGMAI, THAILAND 50290
E-MAIL: <SOMNUK@MJU.AC.TH>

OHM SORNIL2
DEPARTMENT OF COMPUTER SCIENCE,
NATIONAL INSTITUTE OF DEVELOPMENT ADMINISTRATION
BANGKOK, THAILAND 10240
E-MAIL: <OSORNIL@AS.NIDA.AC.TH>

International Journal of Translation

International Journal of Translation – IJT (started in 1989) is a peer-reviewed international journal that has been publishing original research papers devoted to TRANSLATION STUDIES. It is published in March and September every year.

It provides a forum for the discussion of the various theories of translation and interpreting and their practical application; the various processes involved in literary, machine and technical translations; the experience and the points of view of translators while translating & the translated texts from the point of view of cultural and sociolinguistic appropriateness.

Some of the major areas of focus are:

- ◆ Translation theories
- ◆ Translation problems
- ◆ Translation vs. interpreting
- ◆ Computer-assisted translation (CAT)
- ◆ Translation and morphology
- ◆ Natural language processing (NLP)
- ◆ Translation as learning strategy
- ◆ Translation and equivalence
- ◆ Translation and psycholinguistics
- ◆ Translation and communication
- ◆ Machine translation
- ◆ Cultural translation
- ◆ Translation processes
- ◆ Translation and neurology
- ◆ Translation and semantics
- ◆ Translation and cognition

IJT will strive to build up a working group on translators' community across different media; to keep each other informed on what is happening where in translation and to report on translated versions of any major text with a view to analyzing different translated versions from the point of view of contrastive translatology.

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.com>.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 85-101
RECEIVED 28/10/09 ACCEPTED 16/01/10 FINAL 09/03/10

Semantic Analysis using Dependency-based Grammars and Upper-Level Ontologies

AMAL ZOUAQ, MICHEL GAGNON AND BENOÎT OZELL

Ecole Polytechnique de Montréal, Montréal (Québec)

ABSTRACT.

Semantic analysis of texts is a key issue for the natural language processing community. However, this analysis is generally based on a deeply-intertwined syntactic and semantic process, which makes it not easily adaptable and reusable from a practical point of view. This represents an obstacle to the wide development, use and update of semantic analyzers. This paper presents a modular semantic analysis pipeline that aims at extracting logical representations from free text based on dependency grammars and assigning semantic roles to the logical representation elements using an upper-level ontology. An evaluation is conducted, where a comparison of our system with a baseline system shows preliminary results.

Keywords: Semantic Analysis, Logical Analysis, Dependency Grammars, Upper-level Ontologies, Word sense disambiguation.

1 INTRODUCTION

Semantic Analysis is the process of assigning a given sense to the different constituents of a sentence or a text. In the NLP community, most of the approaches, such as HPSG [14] and categorical grammars [10, 15], require the use of a semantic lexicon, i.e. is a dictionary that links words to semantic classes and roles and involves sub-categorization. In fact, this lexicon is the most important component of these grammars, since it is encoded as a set of lexical entries with

syntactic and semantic information (feature structures or type-logical lambda-expressions). In the text mining community, template filling, which also involves knowledge about semantic arguments, is mainly used as a way to assign or extract meaning. Some problems arise from this kind of approach. Firstly, it is generally domain-dependent, especially in the Text Mining Community. This involves repeating the process for each new domain since the identified roles must suit the application domain in order to be accurate. Secondly, the construction of the lexicon implies a huge effort. It is generally language dependent. Thirdly, semantic analysis (involving sub-categorization) can be considered as an intertwined process of syntactic and semantic processing, which make it not easily modularised and updatable.

Based on these issues, we believe that there is a need of a looser coupling between the syntactic and semantic information. This paper presents a reflection on what should be a semantic analysis with the current technologies and formalisms available. It presents a pipeline that separates the process of extracting logical representations (the logical analysis) from the process of assigning semantic roles to the logical representation elements (the semantic annotation). These roles are defined in an upper-level ontology, SUMO [12]. The interest of the pipeline as proposed here is first the modular nature of the syntactic, logical and semantic analyzers, which enables easier updates and focused experimentations that identify the weaknesses of each component of the pipeline, and second, the definition of semantic roles in an ontology, which make the approach more easily interoperable. In fact, one of the major problems of SRL systems is the diversity of semantic roles and their various terminologies and formalisms [8], which hinder their comprehension from one SRL system to another.

The paper is organized as follows. Section 2 presents briefly the state of the art in computational semantics. Section 3 describes the system, including the knowledge model, the steps involved, as well as the required knowledge structures (SUMO and SUMO-WordNet). It also lists the syntactic patterns used in the logical analyzer and presents some of the WSD methods used to assign SUMO senses to the logical elements. Section 4 presents an experiment, shows the results in terms of precision and recall and compares our approach with baseline systems. Finally, section 5 summarizes the paper and outlines implications for NLP research.

2 Related Work

These last years have shown interesting progress in the computational semantics research. While the majority of recent text-based extraction works relies on statistical-based shallow techniques [1], there is still a non negligible amount of research devoted to the implementation of hand-built grammars such as categorical grammars [15], HPSG [14], MRS [4] or TRIPS grammar [1]. These grammars are usually sets of syntactic rules coupled with semantic components, which indicate the role of the rule's arguments in terms of semantics. One drawback of this approach is that the lexicon is not easily obtainable and requires a lot of manual work from computational linguists. This makes the approach not easily scalable and not easily adaptable to new semantic analysis and new models. Moreover, rich grammars such as categorical grammars are not so easily obtainable or reusable.
Other works such as [13] have addressed the extraction of logical forms for semantic analysis as we do but they did not tackle, to our knowledge, the assignment of semantic roles to the logical forms. Finally, very recent works [3] show a growing interest in producing deep semantic representations by taking as input the result of a syntactic parser. This paper is in the same line of research. However our work aims at a looser coupling of the syntactic and semantic features and leaves the deep semantic aspects to a subsequent step of WSD.

3 A Modular Pipeline for Semantic Analysis

The semantic analysis adopted in this paper is a modular pipeline that involves three steps (Fig.1):

1. Syntactic parsing of texts;
2. Logical analysis using a dependency-based grammar;
3. Semantic annotation based on the upper-level ontology SUMO involving word-sense disambiguation.

This modular process is a solution to the above mentioned issues related to current semantic analysis including creating a modular design clearly separating syntactic, logical and semantic annotation or extraction steps, providing a dependency-based grammar that could be comprehensible and reusable by the text mining and NLP community, making this grammar domain-independent and lexicon-independent, and finally using an ontology as a way to formally define semantic roles and make them understandable from a SRL system to another.

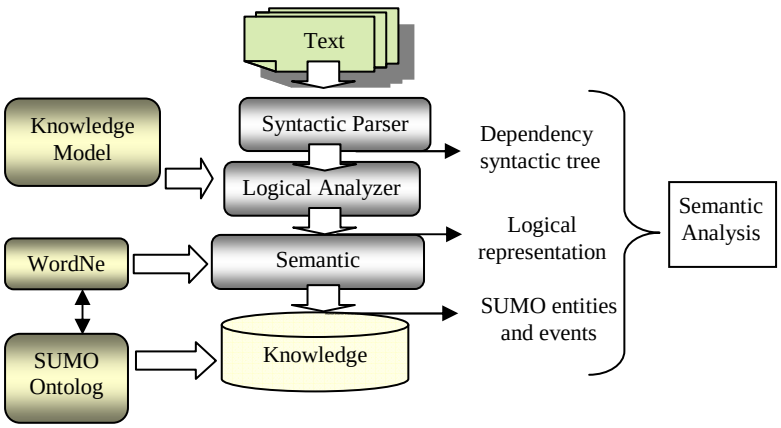


Fig. 1. The Semantic Analysis Pipeline

The following sections explain these steps as well as the linguistic resources needed at each step.

3.1 The Syntactic Analysis

The syntactic analysis is aimed at facilitating the subsequent steps of logical representation and semantic annotation. We believe that this analysis should be based on deep linguistic analysis and should not be limited to simple tagging or surface syntactic parsing. Our goal is to propose a method "easily" reproducible, reusable and able to extract domain-dependent and domain-independent patterns. This should be perfectly handled by dependency parsing.

Dependency parsing outputs grammatical relationships between each pair of words in a sentence. This formalism has proved its efficiency in text mining and we believe that it has the required characteristics of a good grammatical formalism, as it is intuitive, easily understandable, and it enables transparent encoding of predicate-argument structure. Moreover, current state-of-the-art dependency analyzers seem to be sufficiently robust to be considered as reliable tools for knowledge extraction and this is particularly true for the Stanford parser, according to current surveys [16].

Our system uses the basic dependencies format option in the Stanford Natural Language Processing Parser and its dependency component [6] and transforms the output grammatical relations into a

tree, represented as a Prolog term. It also enriches the typed dependencies with the word grammatical categories (such as verb (v), noun (n), …), which are obtained using the Stanford's parts-of-speech output. This operation is important because it enables distinguishing some patterns using their parts-of-speech. For instance, the parser output for the sentence *Banners flap in the wind outside the walls of the city* is presented below. This sentence will be used as an example throughout the whole process of semantic analysis.

nsubj(flap-2, Banners-1);prep(flap-2, in-3);det(wind-5, the-4);pobj(in-3, wind-5);prep(flap-2, outside-6);det(walls-8, the-7);pobj(outside-6, walls-8);prep(walls-8, of-9);det(city-11, the-10);pobj(of-9, city-11)

Banners/NNS flap/VBD in/IN the/DT wind/NN outside/IN the/DT walls/NNS of/IN the/DT city/NN./.

This is transformed into a tree, which is given as input to the logical analyzer:

```
root/tree(token(flap,2)/v,
   [nsubj/tree(token(banners,1)/n,[]),
    prep/tree(token(in,3)/prep,
       [pobj/tree(token(wind,5)/n,
          [det/tree(token(the,4)/d,[])])]),
    prep/tree(token(outside,6)/prep,
       [pobj/tree(token(walls,8)/n,
          [det/tree(token(the,7)/d,[]),
           prep/tree(token(of,9)/prep,
              [pobj/tree(token(city,11)/n,
                 [det/tree(token(the,10)/d,
       [])])])])])]).
```

3.2 The Logical Analysis

The logical analyzer presented in this paper is based on the dependency syntactic tree produced in the syntactic analysis step and is strongly related to the approach presented in [17]. In fact, both systems rely on dependency syntactic patterns to extract logical representations. While these representations were used in [18] to generate domain ontologies using measures from graph theory, they are exploited here as an intermediate step towards an efficient modular semantic analysis. Moreover, one of the new central points of our approach is the use of an upper level ontology as the semantic lexicon for semantic analysis and the use of WSD algorithms in order to assign roles. The key points that should be outlined here is that, first, the logical analysis does not depend on a particular lexicon, a particular vocabulary or a particular

domain. Second, this analysis is based on the compositionality principle, which states that a sentence semantic representation can be obtained by the semantic representation of its parts. Here we consider that a sentence logical representation requires the logical representation of its parts. To our knowledge, there is no previous proposal to create a compositional logical analyzer based on dependency grammars as we propose here.

3.2.1 The Knowledge Model

Although the logical analysis does not require the use of a semantic lexicon, it still needs a conceptual structure made up of a minimal set of categories. In this project, the categories include entities, named entities, events, statements, circumstances, time, number, purpose, measure and attributes. With these categories, chosen to be as general as possible, it is easy to express various information contexts and to remain independent from a particular domain. Although it would be possible to create logical representations using only lexical items, we believe that using these categories can help the semantic analysis.

There is a straightforward map between our knowledge model categories and grammatical relationships. The following table summarizes the mapping involved between the syntactic categories and the knowledge model. Sometimes, the knowledge model element is detected through a part-of-speech (POS) (e.g. verb, noun), but it may also be detected through a number of grammatical relationships (syntactic patterns) necessary to find the relevant element. In the example column, the words in bold indicate the syntactic category related to the knowledge model element. This knowledge model is subject to further enhancements in the future.

Table 1. Mapping knowledge model elements with syntactic categories

Knowledge Model Element	Syntactic Category	Example
Entity	Noun (n)	The **cat** eats.
Event	Verb (v)	The cat **eats**.
Statement	Any pattern involving a clausal complement with external subject (xcomp)	I **like** {to eat in the garden}$_{xcomp}$
Circumstance	Adverbial clause (advcl)	The accident happened {as the night was **falling**}$_{advcl}$
Time	Temporal modifier (tmod)	He swam in the pool {last **night**}$_{tmod}$
Number	Numeric Modifier (num)	**200** people came to the

		party
Attributes	1. Nominal subj. and copula 2. Adjectival complement 3. Adjectival modifier	1. The cat is **big** 2. He looks **tired** 3. He is a **happy** man
Measure	Measure	The director is 55 **years** old

Note that these mappings are not performed in isolation. In fact, relating a knowledge model element to a syntactic category occurs only in the context of detecting specific syntactic patterns. This prevents the system from incorrectly assigning a knowledge model element to a given lexical item. For example, many nominalizations should refer to events instead of entities. Assigning them in the context of a pattern enables us to avoid this confusion. These patterns are explained in the next section.

3.2.2 The dependency-based Grammar: a Pattern Knowledge Base

Besides the link between syntactic categories and knowledge model elements, the dependency-based grammar is composed of a set of patterns coupled with transformational rules. These rules exploit the dependency representation and create logical representations using the general categories introduced in the knowledge model. The grammar is divided into **core** and **modifiers** patterns and is composed, up to now, of 61 rules. Core syntactic patterns, such as the well-known *subject-direct object* pattern, represent main grammatical structures that are necessary for parsing the texts. Modifiers patterns represent modifiers such as prepositions, participial, purpose-clause, temporal modifiers and so on. The patterns are organized into a hierarchy where richer patterns containing the maximum number of relationships are at the top level. In our Prolog implementation, the hierarchy is simply organized as a set of rules where "richer" rules are fired first. It is worth noting that many patterns can be instantiated in a same sentence, including core patterns and modifiers patterns. Also, some patterns extract implicit knowledge. For instance, in the phrase "*the rabbit's head*", the logical analyzer will produce a predicated term *has-attr (rabbit, head*) from the grammatical relationship *poss (possessive)*. At the subsequent step of WSD, the "real" meaning of the relation (*part-of, possess*, etc.) will be assigned.

The following tables show some of these patterns and provide examples, some of them taken from the Stanford dependencies manual [5]. A grammatical relationship between brackets indicates that it is a

child of the preceding relationship. For example, in *nsubj-xcomp[-dobj]*, a *dobj* relationship is a child of the *xcomp* relationship. In the examples column, the words in bold and italics represent the heads (root nodes) of the patterns. Head's syntactic category is indicated in italics in the beginning of each pattern. The reader is referred to [6] to understand the grammatical hierarchy and the corresponding grammatical links.

Table 2. Main syntactic patterns

Patterns	Examples
Verb-nsubj-dobj-iobj	{Mary}$_{\mathbf{nsubj}}$ ***gave*** {Bill}$_{\mathbf{iobj}}$ a {raise}$_{\mathbf{dobj}}$
Verb-nsubj-dobj-xcomp	{The peasant}$_{\mathbf{nsubj}}$ ***carries*** {the rabbit}$_{\mathbf{dobj}}$, {holding}$_{\mathbf{xcomp}}$ it by its ears
Verb-nsubj-dobj	{The cat}$_{\mathbf{nsubj}}$ ***eats*** {a mouse}$_{\mathbf{dobj}}$
Verb-nsubj-xcomp[-dobj]	{Michel}$_{\mathbf{nsubj}}$ ***likes*** to {eat}$_{\mathbf{comp}}$ {fish}$_{\mathbf{dobj}}$
Adjective-nsubj-xcomp	{Benoit}$_{\mathbf{nsubj}}$ is ***ready*** to {leave}$_{\mathbf{xcomp}}$
Verb-csubj-dobj	What Amal {said}$_{\mathbf{csubj}}$ ***makes*** {sense}$_{\mathbf{dobj}}$
Verb-nsubj-expl	{There}$_{\mathbf{expl}}$ ***is*** a small {bush}$_{\mathbf{nsubj}}$
Adjective-nsubj-cop	{Benoit}$_{\mathbf{nsubj}}$ {is}$_{\mathbf{cop}}$ ***happy***
Noun-nsubj-cop	{Michel}$_{\mathbf{nsubj}}$ {is}$_{\mathbf{cop}}$ a ***man***
Verb-nsubj-acomp	{Amal}$_{\mathbf{nsubj}}$ ***looks*** {tired}$_{\mathbf{acomp}}$
Verb-xcomp-ccomp	Michel ***says*** that Benoit {likes}$_{\mathbf{ccomp}}$ to {swim}$_{\mathbf{xcomp}}$
Verb-nsubj	{The cat}$_{\mathbf{nsubj}}$ ***eats***
Verb-dobj	Benoit talked to Michel in order to ***secure*** {the account}$_{\mathbf{dobj}}$
Verb-nsubjpass-prep by	{The man}$_{\mathbf{nsubjpass}}$ has been ***killed*** {by}$_{\mathbf{prep}}$ the police
Verb-csubjpass-prep by	That he {lied}$_{\mathbf{csubjpass}}$ was ***suspected*** {by}$_{\mathbf{prep}}$ everyone
Verb-nsubjpass	{Bills}$_{\mathbf{nsubjpass}}$ were ***submitted***

Table 3. Modifiers patterns

Modifiers Patterns (Modifiers)	Examples
Partmod[prep]	There is garden **surrounded by** houses.
Prep[pcomp]	They heard **about** Mia missing classes.
Prep (after a noun)	Vincent discovered the man **with** a telescope.
Prep (after a verb)	Bills were submitted **to** the man.
Amod	The **white** cat eats
Tmod	Vincent swam in the pool last **night**
Advcl	The accident happened as the night was **falling**.

Ccomp	Michel says that Benoit **likes** to swim.
Purpcl	Benoit talked to Michel in order to **secure** the account.
Infmod	The following points are to **establish**.
Measure	The director is 55 **years** old.
Num	The director is **55** years old.
Poss	The peasant taps the **rabbit**'s head with his fingers.
Quantmod	**About** 200 people came to the party.
Advmod	**Genetically** modified food is dangerous.
Rcmod	Michel loves a cat which Benoit **adores**.

At present, the grammar does not handle anaphora resolution automatically and conjunctions are computed based on a distributive interpretation only, which may not lead to a correct interpretation in some cases. Future work should tackle these issues.

3.2.3 The Transformational Approach

Each pattern is a Prolog rule that builds a logical representation according to the fired pattern. Since we use a compositional approach, each fired rule builds a part of the sentence analysis. The resulting representation is a predicative flat formula composed of predicates (the knowledge model elements) applied to lexical elements, as well as predicates resulting from prepositional relations and predicates indicating if an entity has already been encountered in the discourse or if it is a new entity. Relationships between predicates are represented through their arguments and referential variables are assigned to the instantiated knowledge model elements.

Following our example sentence, the resulting logical representation is: `outside(e1, id3), of(id3, id4), entity(id4, city), resolve_e(id4), entity(id3, walls), resolve_e(id3), in(e1, id2), entity(id2, wind), resolve_e(id2), event(e1, flap, id1), entity(id1, banners), new_e(id1).`

This formula states that there are a number of entities (*city*, *wind*, *etc*.), an event (*flap*) involving the entity *banner* and two relationships "*outside*" and "*of*". "*Outside*" involves the event of *flapping e1* and the entity *walls*. The predicates *new_e* and *resolve_e* are used to indicate if the entity has already been encountered in previous sentences (*resolve_e*) or not (*new_e*). This will help us in anaphora resolution.

An example of a Prolog rule for the *nsubj-dobj* pattern is shown below. The rule involves the discovery of the two relationships of interest (*nsub*j and *dobj*) and calls the *semparse* procedure. This procedure creates a logical representation for the *nsubj* and the *dobj*

sub-trees and finally produces an instance of an event object that combines these two outputs.

```
semparseMainPattern(tree(Node/v,Children),tree(Node
/v,Rest),
Id,SemIn,[event(Id,Node,IdAgent,IdObject)|SemOut]):
-
  select(nsubj/tree(N1/_,C1),Children, R1),
  select(dobj/tree(N2/_,C2),R1, Rest),
  semparse(tree(N1/n,C1),_,IdAgent,SemIn,Sem1),
  semparse(tree(N2/n,C2),_,IdObject,Sem1,SemOut),
  gensym(e,Id).
```

3.3 The Semantic Annotator

The obtained logical representation elements should then be assigned a sense. One of the tasks of a SRL system is to adequately segment predicates and their arguments before their classification into a specific set of roles. Due to the logical analysis, argument and predicate segmentation is already done and the semantic annotator should then focus on assigning an appropriate role to these representations. Here, we mainly focus on entities and events in the logical representations but further work should explore the whole structure.

3.3.1 The SUMO Upper-Level Ontology

One of the difficulties in semantic role labelling is that most of the approaches use very specific subsets of semantic roles and do not agree on the roles to be used. Using an upper-level ontology enables a high-level and formal definition of these roles. Moreover, the interest of using an ontology instead of a flat set of roles is the ability to use its hierarchical and conceptual structure in order to help the WSD process, ascertain the correctness of the identified roles, or reason about the annotated roles. In the context of the Semantic Web, this last point is very important, as the annotated texts will be meaningful to multiple SRL systems which should foster reusability and interoperability.

The Suggested Upper Merged Ontology (SUMO) [11] is an ontology composed of about 1000 terms and 4000 definitional statements. It has been extended with a Mid-Level Ontology (MILO), and a number of domain ontologies, which enable coverage of various application domains. SUMO has also gone through various developments stages and experimentations, which make it stable.

One interesting feature of SUMO is that its various sub-ontologies (base, structural, MILO, and domain ontologies) are independent and

can be used alone or in combination. Here, we only exploit the upper level, meaning that we take into account only the SUMO ontology itself (merge.kif). Another point is its mapping from lexical items (terms) to general high-level concepts. In fact, SUMO [11] is mapped to the widely used WordNet computational lexicon [7]. The SUMO-WordNet mapping links each synset in WordNet to its SUMO sense through three types of relationships: equivalent links, instance links and subsumption links. Despite the fact that this mapping can be error-prone, we believe that it represents an excellent demonstration of how various state-of-the-art resources can be used in a modular pipeline. The other point is that choosing this upper-level ontology is not a limitation and can be extended by a domain ontology if this is required.

3.3.2 Word Sense Disambiguation

Choosing the appropriate role involves the use of WSD algorithms. At this point of our work, we have implemented a number of standard knowledge-based unsupervised WSD methods. The choice of unsupervised methods is guided by the same motivation as for the whole pipeline: avoiding costly and hard-to-build language resources.

The interest of the pipeline at the level of WSD is that the predicates and arguments to be disambiguated are already clearly identified in the logical representations. One step further would be to use the whole logical representation itself as a way to direct the disambiguation process. We are currently working on this.

Among the WSD methods, we used a number of Lesk-derived algorithms namely the Simplified and Original Lesk. We also implemented a version of the [2] algorithm which is based on a semantic network extracted from WordNet to build a context feature vector for the term to be disambiguated. We relied also on a baseline widely used in SRL evaluations: the most frequent sense. Finally, we used an algorithm that relies on co-occurrences frequencies extracted from SEMCOR to determine the number of overlapping terms between these co-occurring terms and the context of the term to disambiguate.

In all these implementations, if the algorithm fails to identify a particular sense, it then backs off to the most frequent sense. Below are the annotated entities and events in our example sentence. Here we only show the SUMO-based annotations but we also keep the WordNet-based annotations in the knowledge base.

This results into the following SUMO-based semantic representation of the example sentence: `outside(e1, id3), of(id3, id4), entity(id4, SUMO:City), resolve_e(id4), entity(id3, SUMO: StationaryArtifact), resolve_e(id3), in(e1,`

```
id2), entity(id2, SUMO: Breathing), resolve_e(id2),
event(e1, SUMO: Motion, id1), entity(id1, SUMO:
Fabric), new_e(id1).
```

Of course, the system can also produce the WordNet-based semantic representation:

```
outside(e1, id3), of(id3, id4),
entity(id4, WN: city%1:15:00::), resolve_e(id4),
entity(id3, WN: wall%1:06:01::), resolve_e(id3),
in(e1, id2), entity(id2, WN: wind%1:04:01::),
resolve_e(id2), event(e1, WN: flap%2:38:00::, id1),
entity(id1, WN: banner%1:06:00::), new_e(id1).
```

4 EVALUATION

Evaluating such a rich pipeline is a challenge in itself. In fact, it involves evaluating the syntactic, logical and semantic annotation. Based on current reviews of the Stanford parser which describe a good performance [16], we decided to focus on the logical and semantic annotation evaluations. Two types of experiments were conducted using the well-known precision and recall metrics:

- A first experiment involving a small corpus of three descriptive texts (185 sentences) manually annotated using SUMO senses in order to build a SUMO gold standard. This corpus helped us in performing the logical form evaluation as well as the semantic annotation;
- A second experiment on the Senseval 3 dataset for the English lexical sample task [9] which enables to test the chosen WSD algorithms on a standard dataset and to compare the results with similar systems. This second experiment does not rely on the previous logical form extraction.

For comparison purposes, we used the most frequent sense baseline in both experiments. Precision and recall are calculated as follows:

Precision = items the system got correct / total number of items the system generated

Recall = items the system got correct / total number of relevant items (which the system should have produced)

4.1 Logical Analyzer Results

This section tests the logical analyzer and the accuracy of the resulting logical formula by measuring the precision and recall of the extracted entities and events in the first corpus using our patterns. These results are summarized in Table 4.

Table 4. The logical analysis results in terms of entities and events

	Precision %	**Recall %**
Entities	94.98	80.45
Events	94.87	85.5

From these experiments, it is clear that our semantic analyzer yields promising results. Most of the time, the incorrect entities and events are due to a wrong syntactic parsing from the Stanford Parser. There are also some patterns that are not yet identified which make the recall lower. These results should be later completed with an evaluation on a bigger corpus, they should be detailed in terms of correctness of predicates, arguments and whole logical formulas [13] and finally, they should include the whole logical representation and not be limited to entities and events.

4.2 Semantic Annotator Results

Semantic annotation (as an isolated module) was tested over the first corpus as well as the Senseval data (English lexical sample task). Various algorithms were tested mainly using knowledge-based methods, including:

- The Simplified and Original Lesk algorithms as well as derivatives such as [2].
- An algorithm computing the most frequent sense based on the WordNet frequencies (extracted from SEMCOR) for the first corpus, and based on term frequencies in the training data for the Senseval dataset.
- An algorithm, dubbed *frequency of co-occurrence*, computing the overlap between the context of the term to be disambiguated and a vector of frequently co-occurring terms for each sense of the term together with their frequency. In the case of the first corpus, these co-occurrences frequencies are extracted from the SEMCOR corpus whereas they are extracted from the Senseval training data in the second corpus.

Many context sizes were tested for all these algorithms including all previous sentences and various words and sentence windows (from 0 to 4) as well as the logical graph structure obtained in the logical analysis.

Due to a lack of space and to the fact that WSD in itself does not represent our contribution in this paper, we simply present the results of

the best performing algorithm together with the most frequent sense baseline without entering into the details of each implemented algorithm (the reader is referred to references and state-of-the-art literature). We did not obtain the best performance using only one algorithm on the three texts (first corpus) but Banerjee and Pedersen algorithm [2] was always among the top-ranking algorithms for entity annotation. Events were best disambiguated using frequency of co-occurrence (text 1), most frequent sense (text 2) and Simplified Lesk (text 3). The following table shows the mean of the precision and recall obtained for the three texts. As can be shown, the algorithm outperforms slightly the most frequent sense baseline. We are seeking better results and future work will explore graph-based WSD disambiguation based on the logical analysis form. Further experiment should also explore the impact of the corpus characteristics on the performance and the choice of WSD algorithms.

Table 5. A comparison of the precision/recall results for the two algorithms (WSD and most frequent sense)

	Best Algorithm (Precision)	Best Algorithm (Recall)	Most Frequent (Precision)	Most Frequent (recall)
SUMO entities	87.08	73.76	84.67	71.65
SUMO events	75.69	68.16	71.54	64.29

Regarding the Senseval corpus, we were able to obtain a precision/recall of 64.1 % (Fine-grained) and 69% (coarse-grained). Based on the overall results of the competition [9], we were able to exceed the most frequent sense performance which was listed as 55.2% (fine-grained) and 64.5% (coarse-grained) using a variant of [2] coupled with frequencies of co-occurrences. We used a two-sentence window around the word to be disambiguated and a cosine similarity. Our results rank us second among the unsupervised algorithms of the competition (although we consider the approach as minimally supervised).

5 CONCLUSION

Current semantic analysis techniques are generally in need of lot of training data, depend on resources such lexicons for their semantic interpretation and lack a uniform way to define roles or labels that should be assigned to sentences constituents. This paper presented a modular pipeline for semantic analysis that relies on state-of-the-art dependency parsing, logical analysis using a dependency-based grammar and finally semantic annotation. This annotation is performed using the upper-level ontology SUMO and the WordNet lexicon, which could be considered as standard resources. Choosing a dependency grammar instead of other formalisms is guided by a practical point of view: it is argued that state-of-the-art analyzers have reached a certain maturity, which makes them a good starting point for a semantic analysis. Moreover, dependency grammars provide an intuitive solution to the identification of logical forms from text as outlined by [13]. The proposed solution does not require costly training or data resources, except some standard resources well-known in the NLP community. The modular nature of the pipeline makes it more easily adaptable and updatable from a software engineering point of view. Finally, the system presented here is intended as a demonstration of what could be a semantic analysis with current methods and tools. Future work includes the enrichment of the dependency-based grammar with new patterns, a better handling of the meaning of conjunctions and other specific constructions (such as idioms), the processing of ambiguous structures and eventive nominalizations as well as the use of a bigger corpus for the evaluation of the logical analysis results. We are currently working on WSD algorithms that could benefit from the logical form not only for argument selection as proposed here, but also for argument annotation.

ACKNOWLEDGEMENTS

The authors would like to thank Prompt Quebec and UnimaSoft Inc. for their financial support.

REFERENCES

1. Allen, J. F., M. Swift, and W. de Beaumont.: Deep Semantic Analysis of Text. In *STEP 2008 Conference Proceedings*, pp. 343-354. College Publications (2008).

2. Banerjee, S. and Pedersen, T.: Extended gloss overlaps as a measure of semantic relatedness. In *Proc. of the 18th Int. Joint Conference on Artificial Intelligence*, Acapulco, Mexico, pp. 805-810 (2003)
3. Bos, J.: Introduction to the Shared Task on Comparing Semantic Representations. In *STEP 2008 Conference Proceedings*, pp 257-261 (2008).
4. Copestake, A., Flickinger, D., Pollard, C. and Sag, I.A.: Minimal Recursion Semantics - an Introduction, *Research on Language and Computation* 3:281–332, Springer (2005).
5. De Marneffe, M-C, and Manning. C.D: Stanford typed dependencies manual, http://nlp.stanford.edu/software/dependencies_manual.pdf (2008)
6. De Marneffe, M-C, MacCartney, B. and Manning. C.D.: Generating Typed Dependency Parses from Phrase Structure Parses. In *Proc. of LREC*, pp. 449-454 (2006)
7. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press (1998)
8. Màrquez, L., Carreras, X., Litkowski, K. and Stevenson, S.: Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2), 145-159 (2008).
9. Mihalcea, R., Chklovski, T. And Kilgarriff, A.: The Senseval-3 English Lexical Sample Task Export, in *Proc. of Senseval-3*, pp. 25--28, Spain, (2004)
10. Moortgat, M.: Categorial grammar and formal semantics. In L. Nagel (ed.) Encyclopedia of Cognitive Science, Vol. 1, pp. 435-447. London, Nature Publishing Group, (2002)
11. Niles, I., and Pease, A.: Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, *Proc. of the IEEE International Conference on Information and Knowledge Engineering*, pp 412--416 (2003)
12. Pease, A., Niles, I., and Li, J.: The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*, Canada (2002)
13. Rus, V.: A First Evaluation of Logic Form Identification Systems, in *Proc. of Senseval-3: Third Int Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pp. 37--40, Spain (2004)
14. Sag, I., Wasow, T. and Bender, E.: Syntactic Theory: A Formal Introduction, 2nd Edition, Center for the Study of Language and Information - Lecture Notes (2003)
15. Steedman M., *The Syntactic Process*, MIT Press (2000)
16. Stevenson, M. and Greenwood, M.A.: Dependency Pattern Models for Information Extraction, Journal of Research on Language & Computation, Springer (2009)

17. Zouaq, A.: Une approche d'ingénierie ontologique pour l'acquisition et l'exploitation des connaissances à partir de documents textuels, *Ph.D. Dissertation*, University of Montreal (2008)
18. Zouaq, A. and Nkambou, R.: Evaluating the Generation of Domain Ontologies in the Knowledge Puzzle Project, *IEEE Transactions on Knowledge and Data Engineering*, 21(11): 1559-1572 (2009).

AMAL ZOUAQ
ECOLE POLYTECHNIQUE DE MONTRÉAL,
C.P. 6079, SUCC. CENTRE-VILLE, MONTRÉAL (QUÉBEC), H3C 3A7
E-MAIL: <AMAL.ZOUAQ@POLYMTL.CA>

MICHEL GAGNON
ECOLE POLYTECHNIQUE DE MONTRÉAL,
C.P. 6079, SUCC. CENTRE-VILLE, MONTRÉAL (QUÉBEC), H3C 3A7
E-MAIL: <MICHEL.GAGNON@POLYMTL.CA>

BENOÎT OZELL
ECOLE POLYTECHNIQUE DE MONTRÉAL,
C.P. 6079, SUCC. CENTRE-VILLE, MONTRÉAL (QUÉBEC), H3C 3A7
E-MAIL: <BENOIT.OZELL@POLYMTL.CA>

International Journal of Communication

International Journal of Communication – IJC is a peer-reviewed international journal that has been publishing original research papers devoted to COMMUNICATION STUDIES, since 1991. It is published in March and September every year.

IJC was started with a special focus on "communication across cultures," "interpersonal communication" and "cognition". It has grown to include later, the various related disciplines and sub-disciplines of communication such as marketing, electronic communication, artificial intelligence, media studies, mass communication, schizophrenia, communication disorder, identity transformation and communicative praxis etc. It has come out in various special numbers such as *Language & Cognition*, *Communication Processes and Disorders*, *Indian Theory of Knowledge and Language*, *Artificial Intelligence and NLP*, *Culture and Cognition*, *Communication and the Recovery of the Real*, *Media & Communication*, *Marketing & Communication*, *Identity Transformation and Communicative Praxis*, *Normative Approaches to Press* & *Normative Foundations of Cultural Dialogue.* It publishes research papers related to the following broad areas:

- Mass communication
- Epistemology
- Advertising & PR
- Cognitive linguistics
- Cybernetics
- Discourse analysis
- Conversational analysis
- Performance studies
- Cultural theory and policy
- Communication and ideology
- Communication theories
- History of consciousness
- Interpersonal communication
- Intra-personal communication
- Schizophrenia & communication disorder
- Media studies
- Hermeneutics
- Literary sciences
- Cognition
- Anthropology
- Propaganda
- Narrative
- Arts, architecture & aesthetics
- Film studies, folklore and folk media
- Communication across cultures
- Marketing & communication
- Intercultural communication
- Organizational communication
- Electronic communication
- Media hegemony
- Journalism
- Sociolinguistics
- Pragmatics
- Semiotics
- Public affairs

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.com>.

Lexical Resources

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 105-119
RECEIVED 26/11/09 ACCEPTED 16/01/10 FINAL 12/03/10

Hypernymy Extraction Using a Semantic Network Representation

TIM VOR DER BRÜCK

IICS, Germany

ABSTRACT

There are several approaches to detect hypernymy relations from texts by text mining. Usually these approaches are based on supervised learning and in a first step are extracting several patterns. These patterns are then applied to previously unseen texts and used to recognize hypernym/hyponym pairs. Normally these approaches are only based on a surface representation or a syntactical tree structure, i.e., constituency or dependency trees derived by a syntactical parser. In this work, however, we present an approach that operates directly on a semantic network (SN), which is generated by a deep syntactico-semantic analysis. Hyponym/hypernym pairs are then extracted by the application of graph matching. This algorithm is combined with a shallow approach enriched with semantic information.

1 INTRODUCTION

Quite a lot of work has been done on hypernymy extraction in natural language texts. The approaches can be divided into three different types of methods:

- Analyzing the syntagmatic relations in a sentence
- Analyzing the paradigmatic relations in a sentence
- Document clustering

The first type of algorithms usually employ a set of patterns. Quite popular patterns were proposed by Hearst, the so–called Hearst patterns [1]. The following Hearst patterns are defined:

– NP_{hyper} such as {{NP_{hypo},}* (and|or)} NP_{hypo}
– such NP_{hyper} as {NP_{hypo},}* {(and|or)} NP_{hypo}
– NP_{hypo} {,NP_{hypo}}*{,} or other NP_{hyper}
– NP_{hypo} {,NP_{hypo}}*{,} and other NP_{hyper}
– NP_{hyper}{,} including {NP_{hypo},}* {and|or} NP_{hypo}
– NP_{hyper}{,} especially {NP_{hypo},}* {and|or} NP_{hypo}

These patterns are applied on arbitrary texts and the instantiated variables NP_{hypo} and NP_{hyper} are then extracted as a concrete hypernymy relation. Several approaches were developed to extract such patterns automatically from a text corpus by either employing a surface representation [2] or a syntactical tree structure [3].

Instead of applying the patterns to an ordinary text corpus, some approaches apply them on the entire Internet by transferring the patterns into Web search engine queries[4, 5]. Pattern learning and application is combined by the system KnowItAll [6] which uses a bootstrapping mechanism to extend patterns and extracted relations iteratively. An alternative approach to pattern matching is the usage of kernel functions where the kernel function defines a similarity measure between two syntactical trees possibly containing a hypernymy or an other semantic relation[7].

Paradigmatic approaches expect that words in the textual context of the hypernym (e.g., neighboring words) can also occur in the context of the hyponym. The textual context can be represented by the set of the words which frequently occur together with the hypernym (or the hyponym). Whether a word is the hypernym of a second word can then be determined by a similarity measure on the two sets [5].

A further often employed method for extracting hypernyms is document clustering. For that, the documents are hierarchically clustered. Each document is assigned a concept or word it describes. The document hierarchy is then transferred to a concept or word hierarchy[8].

In contrast to the formerly mentioned methods, we will follow a purely semantic approach to extract hypernymy relations between concepts (word readings) instead of words which operates on semantic networks (SN) rather than on syntactical trees or surface representations. By using a semantic representation, the patterns are more generally applicable and therefore the number of patterns can be reduced.

In the first step, the entire content of the German Wikipedia corpus is transformed into SNs following the MultiNet[1] formalism[9]. Afterwards, deep patterns are defined which are intended to be matched to that SNs.

[1] MultiNet is the abbreviation of **Multi**layered Extended Semantic **Net**works

Some of them are learned by text mining on the SN representations, some of them are manually defined.

After the patterns are applied on the Wikipedia corpus, the ontological sorts and features of the extracted hyponym and hypernym, as defined by the MultiNet formalism (see Sect. 2), are compared to filter out incorrect concept pairs. Finally, we determine a confidence score for all remaining relations which reflects the likelihood that the hypernymy relation has actually been correctly recognized.

This approach is combined with a shallow method based on Hearst patterns enriched with semantic information if present. The shallow patterns are defined as regular expressions and are applied on the token list which is always present independent of the fact that the SN is successfully constructed.

2 MULTINET

MultiNet is a SN formalism. In contrast to SNs like WordNet [10] or GermaNet [11], which contain lexical relations between synsets, MultiNet is designed to comprehensively represent the semantics of natural language expressions. A SN in the MultiNet formalism is given as a set of nodes and edges where the nodes represents the concepts (word readings) and the edges the relations (or functions) between the concepts. Example SNs are shown in Fig. 1 and Fig. 2. Important MultiNet relations/funtions are [9]:

- SUB: Relation of conceptual subordination (hyponymy)
- AGT: Conceptual role: Agent
- ATTR: Specification of an attribute
- VAL: Relation between a specific attribute and its value
- PROP: Relation between object and property
- *ITMS: Function enumerating a set
- PRED: Predicative concept characterizing a plurality
- OBJ: Neutral object
- SUBS: Relation of conceptual subordination (for situations)

It is differentiated between lexicalized nodes (i.e., associated to entries in the semantic lexicon) and nodes which represents complex situations or individual objects, and are not associated with single lexical entries. The latter nodes are just assigned a unique ID.

MultiNet is supported by a semantic lexicon [12] which defines, in addition to traditional grammatical entries like gender and number, one or more ontological sorts and several semantic features for each lexicon

entry. The ontological sorts (currently more than 40) form a taxonomy. In contrast to other taxonomies ontological sorts are not necessarily lexicalized, i.e., they do not necessarily denote lexical entries. The following list shows a small selection of ontological sorts which are derived from *object*:

- Concrete objects: e.g., *milk*, *honey*
  - Discrete objects: e.g., *chair*
  - Substances: e.g., *milk*, *honey*
- Abstract objects: e.g., *race*, *robbery*

Semantic features denote certain semantic properties for objects. Such a property can either be present, not present or underspecified. A selection of several semantic features is given below:

- ANIMAL
- ANIMATE
- ARTIF (artificial)
- HUMAN
- SPATIAL
- THCONC (theoretical concept)

Example for the concept *house.1.1*[2]: discrete object; ANIMAL -, ANIMATE -, ARTIF +, HUMAN -, SPATIAL +, THCONC -, ...

The SNs following the MultiNet approach are constructed by the deep linguistic parser WOCADI[3][13] for German text analysis. WOCADI employs for parsing a word class functional analysis instead of a grammar.

3 Application of Deep Patterns

The employed patterns are represented as subnets of the SNs where some of the nodes are marked as slots. These slots are filled if the pattern was successfully matched to an SN. In the example depicted in Fig. 1 the hyponym can be extracted by the pattern:

$$SUB(A,B) \leftarrow SUB(C,A) \wedge PRED(E,B) \wedge \\ *ITMS(D,C,E) \wedge PROP(E, other.1.1) \quad (1)$$

where A is instantiated to *secretary*, B to *politician* and C, B and D to non-lexicalized concepts. $*ITMS$ is a MultiNet function which combines several arguments in a conjunction. Disjunctions are combined by

[2] the suffix .1.1 denote the reading numbered .1.1 of the word house

[3] WOCADI is the abbreviation for **wo**rd **cla**ss **di**sambiguation.

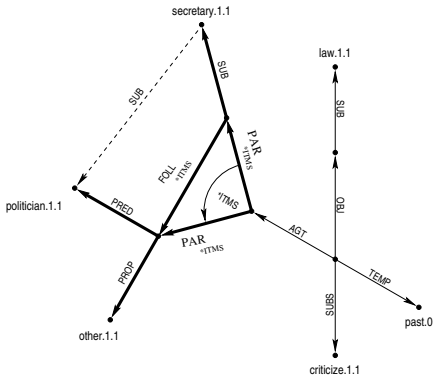


Fig. 1. Hypernymy extraction from the SN representing the sentence: *The secretary and other politicians criticized the law.* The edges matched with the pattern D_1 are printed in bold face. The edge which was inferred by the pattern is printed as a dashed line.

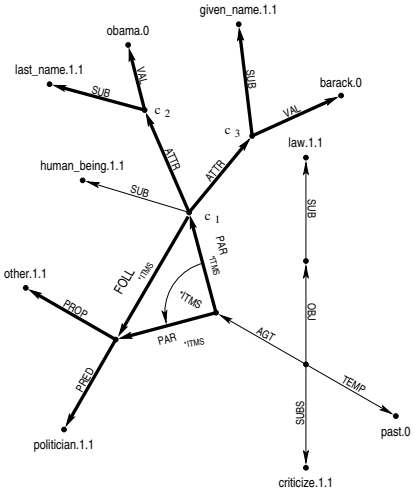


Fig. 2. Hypernymy extraction with an anthroponym in the SN representing the sentence: *Barack Obama and other politicians criticized the law.* The edges matched with the pattern D_2 are printed in bold face.

$*ALTN1/2$. However, this procedure has a serious drawback. The pattern given in Equation 1 is only applicable if the $*ITMS$ function has exactly two arguments (C,E) and one result (D). This means separate patterns are required for three and more arguments. This also implies that the patterns are rather specific, which makes learning them automatically from data difficult. Thus, we convert all functions in an SN with a variable number of arguments like $*ITMS$ and $*ALTN1/2$ to binary relations in the following way:

For each function $x_p = f(x_1, \ldots, x_n)$ with variable arguments as stated above we create n relations $PAR(x_p, x_1)$,...,$PAR(x_p, x_n)$ to represent the parent child relationships between the result and the arguments and $(n\dot{(}n-1))/2$ relations to represent the sequence of the arguments: $FOLL(x_i, x_j) \Leftrightarrow i < j$. Making the above-mentioned modifications the pattern given in Equation 1 changes to:

$$D_1 : SUB(A,B) \leftarrow SUB(C,A) \wedge PRED(E,B) \wedge \underset{*ITMS}{PAR}(D,C) \wedge \underset{*ITMS}{PAR}(D,E) \wedge \underset{*ITMS}{FOLL}(C,E) \wedge PROP(E, other.1.1) \quad (2)$$

Note that different sentences can lead to the same SN. For instance, the semantically equivalent sentences *The secretary and other politicians criticized the law.* and *The secretary as well as other politicians criticized the law.* lead to the same SN, which is displayed in Fig. 1. Thus, the pattern D_1 in Equation 2 can be used to extract the relation

$$SUB(secretary.1.1, politician.1.1)$$

from both sentences. In general, the number of patterns can be considerably reduced by using an SN in comparison to the employment of a shallow representation.

Furthermore, the deep semantic representation allows the simple extraction of hypernymy pairs which involve multi-token anthroponyms, like the fact that *Barack Obama* is a politician, where the extraction of multi-token names is not trivial using shallow patterns. Anthroponyms are already identified by the deep linguistic parser and represented by attribute value pairs (see Fig. 2), which allows to use a similar approach to the extraction of generic hyponyms. In contrast to generic concepts extracted anthroponyms are not stored as binary relations, but as more complex expressions:

Example:

$$\begin{aligned} N :=\ & ATTR(A,F) \wedge SUB(F, last\_name.1.1) \wedge VAL(F,G) \wedge \\ & ATTR(A,H) \wedge SUB(H, given\_name.1.1) \wedge VAL(H,I) \\ D_2 :& N \wedge SUB(A,B) \leftarrow N \wedge PRED(E,B) \wedge \underset{*ITMS}{PAR}(D,A) \wedge \\ & \underset{*ITMS}{PAR}(D,E) \wedge \underset{*ITMS}{FOLL}(A,E) \wedge PROP(E, other.1.1) \end{aligned} \tag{3}$$

If pattern D_2 is applied on the SN shown in Fig. 2 the relations

$$\begin{aligned} ATTR(c_1,c_2) \wedge SUB(c_2, lastname.1.1) \wedge VAL(c_2, obama.0) \wedge \\ ATTR(c_1,c_3) \wedge SUB(c_3, given\_name.1.1) \wedge \\ VAL(c_3, barack.0) \wedge SUB(c_1, politician.1.1) \end{aligned} \tag{4}$$

are extracted, which denote the fact that Barack Obama is a hyponym of the concept *politician.1.1*. Note that we do not differentiate between instances (like person or country names) and hyponyms since instances and hyponyms can be extracted with almost identical patterns (especially for non-anthroponyms and shallow patterns).

4 SEMANTIC-ORIENTED FILTERING TECHNIQUES

The deep patterns described above sometimes extract concept pairs which are not related in a hypernymy relation. A two step mechanism is used to identify such concept pairs. In a first step concept pairs are filtered out if their semantic features and ontological sorts do not meet certain criteria. In the second step, several numerical features are determined for the remaining concepts and combined by the usage of a support vector machine (SVM)[14] to a confidence score. The SVM was trained on a set of annotated hypernymy relation candidates. Concept pairs assigned a high score are likely to express in fact a hypernymy relation. By using this two step approach the number of concept pairs needed to be stored in the database is reduced. In this section we will describe the filtering techniques, the scoring features are introduced in Section 5.

A hyp**o**nym is a specialization of the associated hyp**er**nym. Thus, the hyponym should have all semantic features of the hypernym (with identical values) and the ontological sort of the hyponym has to be subsumed by the sort of the hypernym (equality is allowed too).
Example: *giraffe.1.1* (animal:+) cannot be a hyponym of *house.1.1* (animal:-).

Naturally, this approach only works in all cases if the ontology is monotonic in respect to the employed semantic features. The most prominent example for non-monotonicity is the penguin. It cannot fly although its hypernym *bird.1.1* is associated the property *flying*. To account for such effects and potential misclassifications by the lexicon editor, a small mismatch is allowed.

In the MultiNet formalism, a lexical entry can be marked as a meaning molecule[9, p.292] consisting of several meaning facettes. An example is *school.1.1* which can denote either a building or an institution. If a concept is a meaning molecule, it is associated with more than one semantic feature vector and sort. In this case it is checked if there exists at least one pair of hyponym/hypernym semantic features and sorts which fulfills the above-mentioned subsumption/superset conditions.

Our semantic oriented lexicon contains more than 27 000 deep entries and more than 75 000 shallow entries. Still, in some cases, either the hyponym or hypernym candidate may not be contained which makes a check using semantic features or ontological sorts impossible. If a concept is represented by a compound noun, this problem can be solved by regarding the head instead which can be derived by a morphological analysis.

Different approaches are followed depending on whether the hypernym or the hyponym is not found in the lexicon, but the lexicon does contain its head.

If the hypernym is not contained in the lexicon, it suffices to show that its head concept C is not a hypernym of A to discharge the concept pair (A, B) of being related in a hypernymy relation which is easy to see by contradiction.

The fact that C is the head of B usually implies $SUB(B, C)$. Additionally, let us assume: $\neg\, SUB(A, C)$. Suppose $SUB(A, B)$. Then, due to the transitivity of the hypernym relation, it would follow that $SUB(A, C)$, which is known not to hold.

In the case that the potential hyponym A is not found, a different approach has to be followed. If a tree structure of the ontology is assumed then if A is a hyponym of B, the head C of A can either be a hypernym or a hyponym of B. If both of these cases can be rejected by the comparison of the ontological sorts and semantic features of C and B, the assumption that A is a hyponym of B can be rejected too. Note that theoretically, this approach could fail if the ontology is organized in a directed acyclic graph instead of a tree structure. However, no such problems were observed in practice.

5 FEATURES USED FOR SCORING

We determine a confidence score for each extracted relation, which is computed by combining several numerical features described below.

Correctness Rate: The feature *Correctness Rate* takes into account that the recognized hypernym alone is already a strong indication for the correctness or incorrectness of the investigated relation. The same holds for the assumed hyponym as well. For instance, relations with hypernym *liquid* and *town* are usually recognized correctly. However, this is not the case for abstract concepts. Moreover, movie names are often extracted incompletely since they can consist of several tokens. Thus, this indicator determines how often a concept pair is classified correctly if a certain concept shows up in the first (hyponym) or second (hypernym) position. More formally, we are interested in determining the following probability:

$$p = P(h = t|\, first(rel) = a_1 \wedge sec(rel) = a_2) \tag{5}$$

where

- $first(rel)$ denotes the first concept (the assumed hyponym) in the relation rel
- $sec(ond)(rel)$ denotes the second concept (the assumed hypernym) in the relation rel
- $h(ypernym) = t(rue)$ denotes that a hypernym relation holds

Applying Bayes' theorem to Equation 5 leads to the Equation:

$$p = P(h = t) \cdot \frac{P(first(rel) = a_1 \wedge sec(rel) = a_2 | h = t)}{P(first(rel) = a_1 \wedge sec(rel) = a_2)} \tag{6}$$

For better generalization, we make the assumption that the events $first(rel)$ and $sec(rel)$ as well as $(first(rel)|h = t)$ and $(sec(rel)|h = t)$ are independent. Using these assumptions, Equation 6 can be rewritten:

$$p \approx p' = P(h = t) \cdot \frac{P(first(rel) = a_1 | h = t)}{P(first(rel) = a_1)} \cdot \frac{P(sec(rel) = a_2 | h = t)}{P(sec(rel) = a_2)}$$

$$p' = \frac{P(first(rel) = a_1 \wedge h = t)}{P(first(rel) = a_1)} \cdot \frac{P(sec(rel) = a_2 \wedge h = t)}{P(h = t) \cdot P(sec(rel) = a_2)}$$

$$p' = \frac{1}{P(h = t)} \cdot P(h = t | first(rel) = a_1) \cdot P(h = t | sec(rel) = a_2)$$

If a_1 only rarely occurs in hyponym position in assumed hypernymy relations, we approximate p by $P(h = t|sec(rel) = a_2)$, analogously for

rarely occurring concepts in the hypernym position. As usual, the probabilities are estimated by relative frequencies relying on a human annotation.

First Sentence: The first sentence of a Wikipedia article normally contains a concept definition and thus often expresses a hypernymy relation. Thus, the feature *First Sentence* is set to one, if the associated relation was extracted from a first sentence of a Wikipedia article at least once.

Frequency: The feature *frequency* regards the quotient of the occurrences of the hyponym in other extracted relation in hyponym position and the hypernym in hypernym position. The correlation of this feature with the confidence score is given in Table 1.

This feature is based on two assumption. First, we assume that general terms normally occur more frequently in large text corpora than very specific ones [15]. Second, we assume that usually a hypernym has more hyponyms than vice-versa [9, p.436–437]. Let us consider a simple example. The concept *city* occurs much more often in large text corpora than most cities in the worlds. Furthermore, the number of hyponyms of *city* is very large, since every city in the world is a hyponym of *city*, while the list of hypernyms of a certain city just contains a few concepts like *city*, *location* and *entity*. Therefore, the concept *city* is expected to occur much more often in a hypernym position of an extracted relation than a certain city in the hyponym position. Actually, most cities only occur at most once in an extracted hyponym relation from Wikipedia.

Context: Usually, the hyponym can appear in the same textual context as its hypernym[5]. The textual context can be described as a set of other concepts (or words for shallow approaches) which occur in the neighborhood of the regarded hyponym/hypernym. Analogously to Cimiano, we estimate the semantic similarity between hyponym and hyponym by:

$$hyponym(c_2, c_1) = \frac{|context(c_1) \cap context(c_2)|}{|context(c_1)|} \tag{7}$$

Instead of regarding textual context we investigate the possible properties which can occur at a $PROP$ edge leading from a concept in the SN. This has the advantages that a Word Sense Disambiguation (WSD) was already done and the association between the property and the concept was already established automatically by the SN which may not be trivial if the adjective which is associated to the property is used predicatively.

Pattern Features: For each pattern, an associated pattern feature is defined which is assigned the value one if the relation was extracted by this pattern, otherwise zero. Naturally, the same hypernymy relation can be determined by several patterns. The most strongly correlated pattern features were the feature related to the shallow pattern NP_{hypo} *is a* NP_{hyper} and the deep pattern D_1 shown in Equation 2. Note that in order to get an acceptable recall the pattern NP_{hypo} *is a* NP_{hyper} is only applied on the first sentences of Wikipedia articles.

6 EVALUATION

We applied the patterns on the German Wikipedia corpus from November 2006 which contains 500 000 articles. In total we extracted 391 153 different hypernymy relations employing 22 deep and 19 shallow patterns. The deep patterns were matched to the SN representation, the shallow patterns to the tokens. Concept pairs which were also recognized by the compound analysis were excluded from the results since such pairs can be recognized on the fly and need not be stored in the knowledge base. Thus, these concept pairs are disregarded for the evaluation. Otherwise, recall and precision would increase considerably.

We assigned each extracted concept pair a score calculated by the probability score for relation correctness estimated by a Support Vector Machine[16]. Furthermore, the correlation of all features to relation correctness (1.0 if relation is correct, 0.0 if incorrect) were determined, where a selection of that features is given in Table 1.

The correctness of an extracted relation is given for several confidence score intervals in Table 2 and Fig. 3. There are 89 944 concept pairs with a score of more than 0.7, 3 558 of them were annotated with the information of whether the hypernymy relation actually holds. Note that an extracted relation pair is only annotated as correct if it can be stored in a knowledge base without modification (except from redundancy removal). Thus, a relation is also considered incorrect if

- multi-token expressions are not correctly recognized,
- the singular forms of unknown concepts appearing in plural form are not estimated correctly (this is not trivial for the German language),
- the hypernym is too general, e.g., *word* or *concept*, or
- the wrong reading is chosen by the Word Sense Disambiguation.

We also investigated in which cases deep or shallow patterns were better applicable. Shallow patterns are applied on the tokenizer information of WOCADI. Naturally, shallow patterns are applicable even if a

deep parse was not successful or the sentence was incorrectly parsed. In about 40% of all sentences, a complete SN could not be constructed which is caused either by unknown words, misspellings, grammatical errors or complex grammatical sentence structures.

In contrast, deep patterns are able to extract relations even if additional constituents are located between hyponym and hypernym which is often not possible using shallow pattern. For instance the shallow pattern *X bezeichnet ein Y 'X denotes an Y'* cannot be used to extract the relation $SUB(bajonett.1.1, stoßwaffe.1.1)$ ($SUB(bayonet.1.1, weapon.1.1)$) from the sentence *Bajonett bezeichnet eine auf den Gewehrschaft aufsteckbare Stoßwaffe. 'literally: Bajonet denotes an on the gun stickable weapon.'* while this is possible for the deep pattern

$$\begin{aligned} D_3 : SUB(A,B) \leftarrow & \; SCAR(C,D) \wedge SUB(D,A) \wedge \\ & SUBS(C, bezeichnen.1.2(denote)) \wedge \\ & OBJ(C,E) \wedge SUB(E,B) \end{aligned} \tag{8}$$

Similar considerations hold for the sentence: *Sein Geburtshaus in Marktl ist dasselbe Gebäude, in dem auch Papst Benedikt XVI. zur Welt kam. 'His house of birth in Marktl is the same building in which Pope Benedikt XVI. was born.'*

To handle all such cases with only shallow patterns would require the definition of a tremendous amount of patterns and is therefore not realistically possible in practice.

An example where the normalization from different surface representations and syntactical structures to a single SN proved to be useful:
Auf jeden Fall sind nicht alle Vorfälle aus dem Bermudadreieck oder aus anderen Weltgegenden vollständig geklärt. 'In any case, not all incidents from the Bermuda Triangle or from other world areas are fully explained.'
From the last sentence pair, a hypernymy pair can be extracted by application of rule D_1 (Equation 2) but not by any shallow patterns. The

Table 1. Correlation of features to relation correctness.

Feature	Correlation
Correctness Rate	0.207
Frequency	0.167
Context	0.084
Deep pattern D_1	0.077
Pattern NP_{hypo} *is a* NP_{hyper}	0.074

Table 2. Precision of the extracted hypernymy relations for different confidence score intervals.

Score	≥0.95	≥0.90	≥0.85	≥0.80	≥0.75	≥0.70	≥0.65	≥0.60	≥0.55
Correctness (%)	100.00	87.23	86.49	82.48	82.03	70.49	67.81	57.41	57.03

application of the shallow Hearst pattern NP_{hypo} {,NP_{hypo}}*{,} *und andere/and other* NP_{hyper} fails due to the word *aus 'from'* which cannot be matched. To extract this relation by means of shallow patterns an additional pattern would have to be introduced. This could also be the case if syntactic patterns were used instead since the coordination of *Bermudadreieck 'Bermuda Triangle'* and *Weltgegenden 'word areas'* is not represented in the syntactic constituency tree but only on a semantic level [4].

149 900 of the extracted relations were only determined by the deep but not by the shallow patterns. If relations extracted by one rather unreliable pattern are disregarded, this number is reduced to 100 342. The other way around, 217 548 of the relations were determined by the shallow but not by the deep patterns. 23 705 of the relations were recognized by both deep and shallow patterns. Naturally, only a small fraction of the relations were checked for correctness. In total 6 932 relations originating from the application of shallow patterns were annotated, 4 727 were specified as correct. In contrast, 5 626 relations originating from the application of deep patterns were annotated and 2 705 were specified as correct.

7 Conclusion and Outlook

An approach was introduced for extracting hyponyms by a deep semantic approach. Instead of using the surface representation of sentences, the patterns are defined on a semantic level and are applied on SNs. The SNs are derived by a deep syntactico-semantic analysis. This approach is combined by a shallow method to guarantee an acceptable recall if sentences are not parsable. The evaluation showed that the recall could be considerably improved. In contrast to a shallow representation, the semantic patterns have the advantage of a greater generality which reduces the number of patterns. Furthermore, anthroponyms are already identified and parsed

[4] Note that some dependency parsers employ a semantic-oriented normalization too.

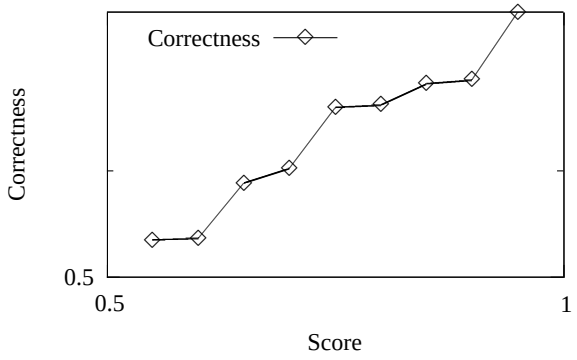


Fig. 3. Precision of the extracted hypernymy relations for different confidence score intervals.

by the SN which simplifies the extraction of instance-of-relations concerning person names.

Further possible improvements are the extraction of other semantic relations using this approach, for instance meronyms or antonyms. Furthermore, validation techniques will be further extended. We plan the usage of the ESPRESSO algorithm [17] as an additional feature and the employment of several deep features.

8 AKNOWLEDGEMENTS

We want to thank our colleagues of our department for their support. Especially we are indebted to I. Glöckner, C. Eichhorn, A. Pils and P. Grimm for proof-reading this work. This work is partly funded by the DFG project *Semantische Duplikatserkennung mithilfe von Textual Entailment* (HE 2847/11-1).

REFERENCES

1. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proc. of COLING, Nantes, France (1992)
2. Morin, E., Jaquemin, C.: Automatic acquisition and expansion of hypernym links. Computers and the Humanities **38**(4) (2004) 363–396
3. Snow, R., et al.: Learning syntactic patterns for automatic hypernym discovery. In: Advances in Neural Information Processing Systems 17. MIT Press, Cambridge, Massachusetts (2005) 1297–1304

4. Sombatsrisomboon, R., et al.: Acquisition of hypernyms and hyponyms from the WWW. In: Proc. of the International Workshop on Active Mining, Maebashi City, Japan (2003)
5. Cimiano, P., et al.: Learning taxonomic relations from heterogeneous sources of evidence. In Buitelaar, P., et al., eds.: Ontology Learning from text: Methods, evaluation and applications. IOS Press, Amsterdam, Netherlands (2005) 59–73
6. Etzioni, O., et al.: Unsupervised named-entity extraction from the web: an experimental study. Artificial Intelligence **165**(1) (2005) 91–134
7. Zhao, S., Grishman, R.: Extracting relations with integrated information using kernel methods. In: Proc. of ACL, Ann Arbor, Michigan (2005) 419–426
8. Quan, T., et al.: Automatic generation of ontology for scholarly semantic web. In: The Semantic Web - ISWC 2004. Volume 4061 of LNCS. Springer, Berlin, Germany (2004) 726–740
9. Helbig, H.: Knowledge Representation and the Semantics of Natural Language. Springer, Berlin, Germany (2006)
10. Fellbaum, C., ed.: WordNet An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts (1998)
11. Hamp, B., Feldweg, H.: Germanet - a lexical-semantic net for german. In: Proc. of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, Madrid, Spain (1997)
12. Hartrumpf, S., Helbig, H., Osswald, R.: The semantically based computer lexicon HaGenLex – Structure and technological environment. Traitement automatique des langues **44**(2) (2003) 81–105
13. Hartrumpf, S.: Hybrid Disambiguation in Natural Language Analysis. PhD thesis, FernUniversität in Hagen, Fachbereich Informatik, Hagen, Germany (2002)
14. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, New York (1998)
15. Joho, H., Sanderson, M.: Document frequency and term specificity. In: Proc. of RIAO, Pittsburgh, Pennsylvania (2007)
16. Chang, C., Lin, C.: LIBSVM: a library for support vector machines. (2001)
17. Pantel, P., Pennachiottti, M.: Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In: Proc. of COLING/ACL, Sydney, Australia (2006)

TIM VOR DER BRÜCK
INTELLIGENT INFORMATION AND
COMMUNICATION SYSTEMS (IICS)
FERNUNIVERSITÄT IN HAGEN, 58084 HAGEN, GERMANY
E-MAIL: <TIM.VORDERBRUECK@FERNUNI-HAGEN.DE>

CREATIVE FORUM

Creative Forum – CF (started in 1988) is a peer-reviewed international journal, which is published in March and September every year. It publishes original research papers pertaining to the late 20th century and current literary practices in India and elsewhere. To outline a few major areas, on which we have centered our previous researches in this journal and intent to invite papers for future issues are:

- Contemporary literature
- Literary theories
- Popular fiction
- Cultural studies
- Feminist studies
- Indian English writing
- Women's writing
- Literary criticism
- Marxist criticism
- Postcolonial studies
- Feminist criticism
- Modernism & postmodernism
- Stylistics
- Structuralism & post-structuralism
- Identity
- Aesthetics

Over a number of years special theme issues devoted to poetry and fiction such as *Recent Indian English Poets*; *Quest for Identity in Indian English Fiction and Poetry*; *Crisis of Identity*; *Commonwealth Literature, Post-Colonial Indian English Literature*; *New Zealand Literature*; *Black Literature*; *Stream of Consciousness in Indian English Fiction*; *Indian Literatures in Translation* & *Partition re/visioned, Perspectives on Women's Writing in India, Popular Culture Studies, Comparative Poetics, Dalit Writings* & *Fiction to Film* etc. have been published.

NOTE FOR THOSE WHO WISH TO CONTRIBUTE

Authors can submit their manuscripts on the above related subjects to the Editors on their e-mail: <bahrius@vsnl.com> in a WORD file according to the camera ready format given on our website: <http://www.bahripublications.com>.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 121-136
RECEIVED 30/11/09 ACCEPTED 16/01/10 FINAL 09/03/10

Linking Named Entities to a Structured Knowledge Base

KRANTHI REDDY. B, KARUNA KUMAR, SAI KRISHNA, PRASAD PINGALI, VASUDEVA VARMA

International Institute of Information Technology, India

ABSTRACT

The task of entity linking aims at associating named entities with their corresponding entries in a knowledge base. The task is challenging because entities, can not only occur in various forms, viz: acronyms, nick names, spelling variations etc but can also occur in various contexts. To extract the various forms of an entity, we used the largest encyclopedia on web, Wikipedia. In this paper, we model entity linking as an Information Retrieval problem. Our experiments using TAC 2009 knowledge base population data set show that an Information Retrieval based approach fares slightly better than Naive Bayes and Maximum Entropy.

1 INTRODUCTION

The rise of web 2.0 technology has provided a platform for user generated content through web blogs, forums etc. This has lead to information overload on the web and it has become an extremely difficult task for users to find the precise information they are looking for. Semi-structured knowledge bases [1] like Wikipedia [2] act as a rich source of information for

[1] Knowledge Base is a structured data base containing entries describing named entities.

[2] Wikipedia is a huge collection of articles. Each article is identified by a unique title. These articles define and describe events and entities.

various user needs and are important for a wide range of applications like search, named entity extraction, text mining etc. But the problem with such structured knowledge bases is that they have to be created manually and updated frequently. For example, " *Information like movies starring Tom cruise have to be updated frequently*". There is also the problem of inconsistency, erroneous values being fed and the information not being up to date.

Automatically updating structured knowledge bases from news articles is a possible solution to this problem since news articles contain the latest information. In view of this strategy, a need arises to address the task of linking named entities from news articles to entries in a knowledge base.

The problem of entity linking shares similarities with cross document co-reference resolution. The task of co-reference resolution is to determine whether two occurrences in a document correspond to the same entity or not. This task becomes more complex when we try to determine whether the instances of two entities across different documents co-refer or not. This is termed as cross document co-reference resolution[1]. This is a challenging problem because the documents can come from different sources and they might also have different conventions and styles[1].

Thus, co-reference resolution of entities across documents plays a critical role towards successfully updating knowledge bases. In 1996 Tipster III program identified cross document co-reference resolution as an advanced area of research since it could be used for multi document summarization and information fusion. It was also identified as one of the potential tasks for the sixth Message Understanding Conference (MUC-6) but was not included as a formal task since it was considered too ambitious at the time[8]. Bagga and Baldwin [1]presented a successful cross document co-reference resolution algorithm to resolve ambiguities between people bearing same names using vector space model.

Following this work, many more contributions have been made to the state of the art. Bhattacharya[2] and Hal Daume[10] construct generative models on how and when entities are shared between documents. Haghighi and Klein[9] propose a fully generative nonparametric Bayesian model which captures co-reference within and across documents. Mann[12] and many of the other previous approaches such as Gooi[7], Malin[11], Chen[3] employ unsupervised learning techniques. Malin[11] considers named-entity disambiguation as a graph problem and constructs a social network graph to learn the similarity matrix. Finin et al.[5] explore the

usage of Wikipedia, DBpedia and free base as a resource for cross document co-reference resolution.

The task of entity linking differs from cross document co-reference resolution in the following aspects. In cross document co-reference resolution, we have a set of documents all of which mention the same entity name. The difficulty lies in clustering these documents into sets which mention the same entity. Whereas, in entity linking, the same entity name could be referred to in different contexts and also using various forms like acronyms, nick names etc. Our problem is to link this named entity to an entry in the knowledge base if present.

Key contributions of our paper are

- We show how variations of an entity, extracted from Wikipedia can be used for linking named entities from news articles to entries in a knowledge base.
- We show that an Information Retrieval based approach is able to perform slightly better than Naive Bayes and Maximum Entropy.

In section 2 we describe the data set and evaluation metrics. We sketch our algorithm in section 3 and describe the experiments conducted in section 4. In section 5 we provide an analysis of our results. We conclude our work in section 6 with a description of our plan for future work.

2 Evaluation Data and Metrics

In this section we describe the data collection and the entity linking tasks and the evaluation metric used. Our experiments were conducted on the Knowledge Base Population(KBP) [3] data set provided as part of the KBP track at Text Analysis Conference [4](TAC) 2009. The KBP data set consists of a reference Knowledge Base (KB) and a document collection. The KB comprises of 818,741 entries where each entry (entity/node) can belong to a Person, Organization, Geo Political Entity or an unknown class. The document collection contains instances of, and information about the target entities for the KBP evaluation queries. A sample KB entry is shown in Fig.1.

KB entries include a name string, an entity type, a unique KB node id, a set of facts and disambiguation text describing the entity.

[3] http://apl.jhu.edu/∼paulmac/kbp.html
[4] http://www.nist.gov/tac/

```
<entity wiki_title="Bud_Abbott" type="PER" id="E0064214" name="Bud Abbott">
<facts class="Infobox actor">
<fact name="name">Bud Abbott</fact>
<fact name="birthname">William Alexander Abbott</fact>...
</facts>
<wiki_text>Bud Abbott William Alexander "Bud" Abbott (October 2, 1895 – April 24, 1974) was an American actor, producer and
comedian born in Asbury Park, New Jersey...
</wiki_text>
</entity>
```

Fig. 1. Knowledge Base Entry

The document collection consists of 1,289,649 data files that contain instances of, and information about the target entities for KBP evaluation queries. The source documents in this collection were taken from various news transcripts and news articles. A sample data file is shown in Fig.2.

```
<DOC>
<DOCID> APW_ENG_20071010.1447.LDC2009T13 </DOCID>
<DOCTYPE SOURCE="newswire"> NEWS STORY </DOCTYPE>
<DATETIME> 2007-10-10 </DATETIME>
<BODY>
<HEADLINE> Peter Fonda auctions memorabilia from `Easy Rider ' film; flag brings about $90,000 </HEADLINE>
<TEXT>
<P> Aside from items offered by the 67-year-old Fonda, the auction included memorabilia related to Peter Frampton, Elvis Presley and
Abbott and Costello.</P>
</TEXT>
</BODY>
</DOC>
```

Fig. 2. Document Collection Data file

The data file consists of a unique document id, the source from where the article has been taken, its headline, and a disambiguation text describing an entity or an event. This text is split into different paragraphs.

The task of entity linking is to determine for each query, if a KB entry exists in the knowledge base. And if it does which KB entry it refers to. A query will consist of a name-string and an associated document-id from the document collection providing context for the name-string. For convenience we refer to name-string as "query string". Query strings can occur as multiple queries using different name variants or in multiple documents. Each query must be processed independently. A sample query is shown in Fig.3.

```
<query id="EL24"><name>Abbott</name><docid>AFP_ENG_19960413.0028.LDC2007T07</docid></query>
<query id="EL25"><name>Abbott</name><docid>APW_ENG_20080725.0086.LDC2009T13</docid></query>
<query id="EL26"><name>Abbott</name><docid>AFP_ENG_20030724.0396.LDC2007T07</docid></query>
```

Fig. 3. Sample Queries

Since the documents can come from different sources, various name variations like acronyms and nick names etc could refer to the same query string. They might also occur in different contexts which makes the problem a challenging one.

Table 1 shows that there are 15 queries with “Abbott/Abbot” as the query string, but they refer to different KB entries which belong to different classes. The query string is associated with 15 different data files showing how varied the context is.

Table 1. Sample Queries

Query string	**KB-id**	**KB Entry title**	**No. of Queries**	**No. of diff. data files**	**Class**
Abbot	E0064214	Bud Abbott	1	1	Person
Abbott	E0064214	Bud Abbott	4	4	Person
Abbott	E0272065	Abbott Lab.	9	9	Unknown
Abbott	E0003813	Abbot, Texas	1	1	Geo-political entity

The following two examples show how varied the context can be.

Context 1: *A spokeswoman for Abbott said it does not expect the guidelines to affect approval of its Xience stent, which is expected in the second quarter.*

Context 2: *Aside from items offered by the 67-year-old Fonda, the auction included memorabilia related to Peter Frampton, Elvis Presley and Abbott and Costello.*

In context 1 “Abbott” refers to “Abbott Laboratories” whereas in context 2 it refers to “Bud Abbott”. The above example shows that the task of entity linking is a challenging one.

To evaluate the effectiveness of the system, we use Micro-Average Score (MAS). MAS is the official metric for evaluating systems performance at TAC 2009. The micro-average score is the precision over all the queries. It is calculated using the following equation.

$$Micro\ Average\ Score = \frac{No.of\ correct\ responses}{No.of\ Queries} \quad (1)$$

Similarly micro-average score can be calculated for nil valued queries and Non-nil valued queries. From Table 2, system output is correct 3 out of 6 times. Hence the Micro Average Score is 3/6 = 0.5 .

Another metric that can be used to evaluate the entity linking task is the Macro-Average Score. In this metric, precision is calculated for each entity (nil and non-nil) and an average is taken across the entities. The main problem with such a metric is that it might be biased towards the system's output. It would be unstable with respect to low-mention-count query entities. The example below explains the calculation of Macro-Average Score.

Table 2. System output for a set of query strings

Query string	**KB-id**	**system output**
Abbott	1	1
Abbott	1	101
Abbott	1	1
Abbott Labs	2	101
Abbott Laboratories	2	nil
Abbott Labs	2	2

From Table 2, the entity corresponding to the KB node with ID=1 was linked correctly 2 of 3 times for a precision of 0.67. The entity with ID=2 was linked correctly 1 of 3 times for a precision of 0.33.The macro-averaged precision is 0.5. $\{(0.67+0.33)/2\}$

3 Our Approach

We break the entity linking task into 3 sub tasks.

1. **Preprocessing:** Since the queries for entity linking can have different name variations, we need to have a knowledge repository of all

the various forms possible for an entity. During this step we build a knowledge repository that contains vast amount of world knowledge for these entities. To create this we can use the web. But parsing and extracting knowledge from the web is a tedious task because of its sparseness and unstructured format. Hence we turned our attention to Wikipedia, which is one of the largest semi-structured knowledge bases on the web[17].

The advantages of using Wikipedia compared to the web or any other resource is that

- It has better coverage of named entities [18]. And since our knowledge base presently has only named entities, Wikipedia acts as a perfect platform for creating our knowledge repository.
- Redirect pages can be used to induce synonyms [18].
- Disambiguation pages can be used to generate a list of candidate targets for homonym resolution[14].
- On analyzing random pages from Wikipedia, we found that the bold text from the first paragraph is a variation of the title. These variations in general are full names, alias names and nick names of the title.
- With over 3 million articles Wikipedia is appropriately sized and big enough to provide us sufficient information to create our knowledge repository.

A lot of previous work on wikipedia mining[6, 15, 16] confirms the fact that valuable information can be mined from Wikipedia .

We use redirect pages[5], disambiguation pages [6] and bold text from the first paragraph to create our knowledge repository, which is simply a collection of different variations of an entity. These heuristics help us in identifying synonyms, homonyms, abbreviations, frequent misspellings and alternative spellings of an entity. Even though we are handling different valid variations, each of the above variations can be misspelled as well. In order to identify these spelling variations we generate metaphonic codes for all the variations using metaphone algorithm[4].

2. **Candidate List Generation:** The entity linking task first determines whether a KB entry exists for a given query string. The query string

[5] A redirect page in Wikipedia is an aid to navigation, it contains no content but a link to another article (target page) and strongly relates to the concept of target page.

[6] Disambiguation pages are specifically created for ambiguous entities, and consist of links to articles defining the different meanings of the entity.

is then searched on the titles of KB nodes and Wikipedia articles in the following two ways.

(a) **Phrase Search:** In this method we see if the exact phrase of the query string or the expanded form of the acronym is present in the article title or not. We add node-ids to candidate list only if we find the exact phrase in article titles. In another variation of phrase search, we allow for a difference of one between the number of tokens in article title and query string. We term this extra token as noise.
For example, If the given query is "UT" and we find the expanded form from our knowledge repository to be "University of Texas"; in exact phrase search we would be retrieving node-ids that have exact phrase "University of Texas" present in the title. Whereas in phrase search with noise we would be retrieving nodes that have the titles "University of Texas at Austin, University of Texas at Dallas" as well.

(b) **Token search:** In this method we do a boolean "AND" search of all the tokens of the query string or the expanded form of the acronym in the article title. If all the tokens are present in an article title we add those node-ids to the candidate list. Another variation is to search with noise.
The difference between phrase search and token search is that in phrase search the token order is constrained where as in token search just the presence of each token is vital and not the order.
For example, If the given query is "CCP" and we find the expanded form from our knowledge repository to be the "Chinese Communist Party"; in token search we would be retrieving nodes with the titles either as "Chinese Communist party" or "Communist Party of China". Note that the entry with the title "Communist Party of China" will not be found as a candidate item in phrase search.

All the KB nodes and Wikipedia articles which satisfy one of the above conditions are added to the candidate list. The addition of Wikipedia articles to the candidate list helps us in the identification of nil valued queries. That is, for a given query if our algorithm maps to the Wikipedia article from the candidate list, we can confirm the non-presence of an entry in KB describing the query string.
The query strings can be categorised as either acronyms or any of the other variations like alias names, nick names etc. If all the characters present in the query are uppercase we consider it to be an acronym

. We follow different approaches for processing the two categories. The algorithm for handling these two cases is as follows.

(a) **Not an Acronym:** If the given query is not an acronym we search for the query string terms directly in the title of the KB entries. If a hit is found we add that entry's node-id to the candidate list. However, if no hits are found, we look for variations of the query string in the knowledge repository and then use them to search the KB and Wikipedia titles. If any hits are found we also add those node-ids to the candidate list.

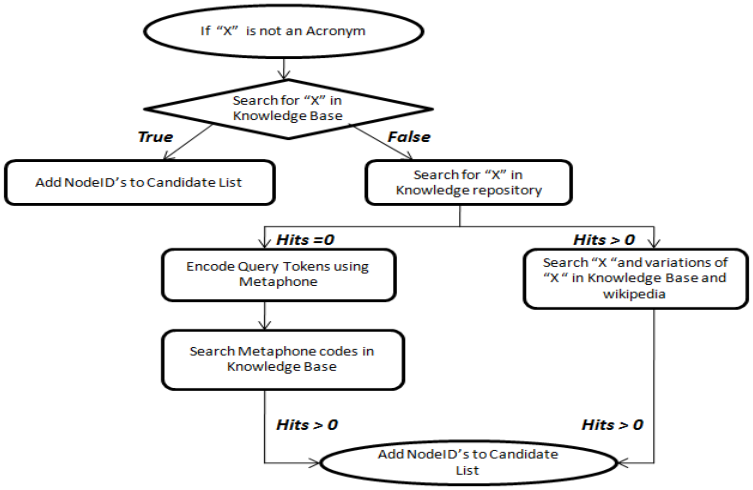


Fig. 4. Flowchart when query is not an Acronym

If no variations are found in the knowledge repository as well, we assume that the query string might have been written using a different spelling. We then generate the metaphone code for each token present in the query. Using these metaphone codes we again search the KB. The flowchart of algorithm is presented in Fig.4.

(b) **Acronym:** If the given query is an acronym we try to get the expanded form from the document content which has been given as disambiguation text for the query string. To find the expanded form, we remove stop words from this disambiguation text and

use an N-Gram based approach. In our N-Gram approach if the length of the acronym is "N" characters, we check if "N" continuous tokens in the disambiguation text have the same initials as the characters of the acronym. If we are successful in finding the expanded form from the disambiguation text, we search the KB using this expanded form. If any hits are found we add the entry's node-id to the candidate list. If we don't find the expanded form from the disambiguation text we search the knowledge repository for the acronym. If the expanded form is found in the knowledge repository, we search the KB and Wikipedia titles using this expanded form and the acronym. If any hits are found we add the entry's node-id to the candidate list. The flowchart of algorithm is presented in Fig.5.

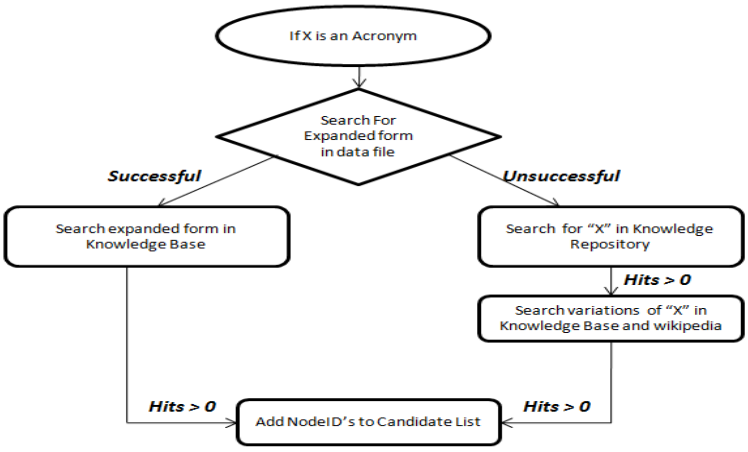


Fig. 5. Flowchart When query is an Acronym

Refining the Candidate list: Articles in KB and Wikipedia are uniquely identified by their titles. With KB being a subset of Wikipedia, for a particular entity the title of articles that describe them would be the same. Thus, if duplicate entries are found, priority is given to KB articles.

3. **Calculating Similarity Score:** We return nil if there are no items in our candidate list or when there are node-ids that belong only to

Wikipedia. If there is only one node-id belonging to KB, it could possibly mean we have only one entry describing the query string. We return this node-id as the possible map.

If there is more than one entry in our candidate list we find the best map using one of the following approaches.

Classification Approach: We have conducted our experiments using Naive Bayes[13] and Maximum Entropy algorithms present in Rainbow Text Classifier[7].

If we consider all the possible candidate items as different classes, we need to find which class is the best map for our query string. Hence, we view this problem as a classification problem with each candidate being a class. Therefore, we built classification models using Naive Bayes and Maximum Entropy algorithms with bag of words as a feature. We use the text describing the candidate item(entity) provided in the KB to train the classification models. We then give the disambiguation text provided along with the query string as test document. This test document is classified into one of the classes and the score obtained is the likelihood of the test document belonging to that class.

Information Retrieval Model: In Information Retrieval, the aim is to retrieve the documents that are closest match for a given natural language query. In our approach, we index each candidate item as a separate document using Lucene[8]. We then form a query from the disambiguation text. Query formulation plays an important role in the success of this approach. While generating the query we try and reduce unwanted tokens. We also try to boost the tokens that seem to be most relevant to our query string. Since the provided disambiguation text has been tagged clearly into different paragraphs, we consider only those paragraphs where the query string is present. The motivation behind this is to capture the context surrounding our query string. We form a boolean "OR" query of all the tokens generated from the disambiguation text neglecting the stop words . We also boost the tokens that are within a window size of 5 terms on either side of the query string. We do this because the tokens closer to the query string are the more prominent tokens describing our entity than the terms that are far off. This is shown in the Fig.6.

[7] http://www.cs.cmu.edu/∼mccallum/bow/rainbow/

[8] Lucene is a high-performance, full-featured text search engine. http://lucene.apache.org/java/docs/.

 text <Query boosting terms> "Query Terms" <Query boosting terms> text

Fig. 6. Token boosting during Query Formulation

If the result node-id belongs to KB, then we return it as the map for the query string. But if the node-id belongs to Wikipedia we return nil, because we don't have an entry in the KB describing our query string.

4 EXPERIMENTS

To generate the candidate list we use one of the following: exact phrase search, phrase search with noise, token search, or token search with noise. Once the candidate list is generated, we use different algorithms to calculate the similarity score between the disambiguation text of the query string and the disambiguation text of the candidate item entries. If more than one item is present in the candidate list belonging to KB and/or Wikipedia, we calculate the similarity score using Naive Bayes, Maximum Entropy from Rainbow Text Classifier or by using an Information Retrieval approach. The algorithm is evaluated using the metrics described in section 2. Table 3 contains the scores for each experiment conducted.

The Micro-Average Score obtained through our algorithm outperforms all the systems submitted at TAC 2009. The average-median score over all the 35 runs submitted at TAC 2009 is 71.08% and the base line score is 57.10% when nil is returned for all the queries. Our best algorithm outperforms median score by as much as 11% and the base line score by 25%.

5 ANALYSIS

Since we have followed a two step process to determine whether for a given query string an entry exists in the knowledge base or not. Therefore for each of these steps we analyze the number of queries that are being incorrectly mapped. For token search with noise, we found that we were unable to find a map for 6.81% (266 of 3904) queries during the candidate list generation, which means that our heuristics failed to capture some query string variations of nicknames, full names, acronyms etc.

Table 3. Results of Various Experiments. IR = Information Retrieval, NB = Naive Bayes, MaxEnt = Maximum Entropy

Alg.	**Noise**	**Phrase/Token search**	**Micro-Average Score**	**nil-valued precision**	**Non-nil valued precision**	**Macro-Average Score**
NB	**1**	**Word Search**	**81.43**	**85.42**	**76.12**	**75.38**
NB	1	Phrase Search	81.25	85.37	75.76	75.10
NB	0	Phrase Search	81.12	85.91	74.75	75.45
NB	0	Word Search	80.87	85.51	74.69	74.96
MaxEnt	**0**	**Word Search**	**78.82**	**87.66**	**67.04**	**75.61**
MaxEnt	1	Word Search	78.46	87.53	66.39	75.58
MaxEnt	0	Phrase Search	78.38	87.93	65.67	76.08
MaxEnt	1	Phrase Search	78.23	87.44	65.97	75.90
IR	**1**	**Word Search**	**82.25**	**86.32**	**76.84**	**75.70**
IR	1	Phrase Search	82.17	86.41	76.54	75.39
IR	0	Phrase Search	81.81	86.90	75.04	75.46
IR	0	Word Search	81.76	86.45	75.52	75.54

While 10.93% (427) were being wrongly mapped during similarity score calculation.

For non-nil valued queries Fig.7 plots the comparison of Precision vs Top "N" search results for the 3 algorithms. It can be seen clearly that as we consider a higher number of hits, the probability of finding the correct map for the query string in the hits list increases. It shows that Information Retrieval and Naive Bayes perform consistently much higher than Maximum Entropy.

Thus we can conclude that the combination of token search with noise for candidate list generation and the Information Retrieval approach for similarity score calculation give the best result. The reason for this approach outperforming all the others is that we are able to generate more candidate items. Though this might also generate more false negatives, but the removal of unwanted paragraphs as noise and the query boosting technique used while calculating similarity score negates this effect.

6 Conclusion and Future Work

In this paper, we explored Information Retrieval and classification based techniques for linking named entities from news articles to entries in a

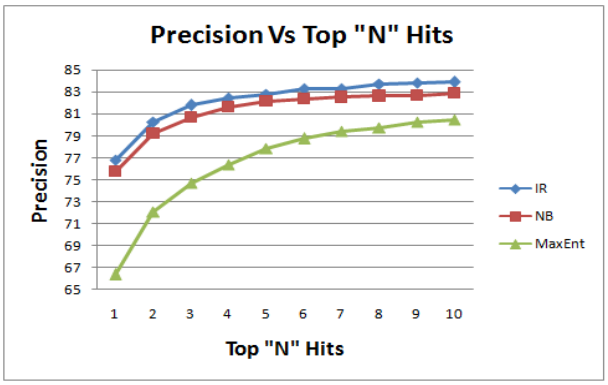


Fig. 7. Precision Vs Top “N” hits.
IR = Information Retrieval, NB = Naive Bayes, MaxEnt = Maximum Entropy

knowledge base. We showed how variations of an entity can be extracted from Wikipedia and used for entity linking. We showed that an Information Retrieval based approach is able to perform slightly better than Naive Bayes and Maximum Entropy approaches. We believe that our approach is promising because, Wikipedia is constantly growing and being updated frequently. With its continuous growth and contribution from users we are guaranteed high quality information. There can be many extensions to the current work. First, using Wikipedia in a better way to create our knowledge repository. We can make use of the infobox tables to extract name variations, nick names etc. Secondly, since news articles always contain the latest information about an entity, we can extract attribute value pairs from them and append them to our KB facts. This will be particularly useful when certain facts keep changing frequently. For example, the number of test matches played by Sachin Tendulkar, number of runs scored by him etc. Thirdly, we can make use of other online resources like DBpedia and Freebase to create our knowledge repository.

7 AKNOWLEDGEMENTS

The authors wish to thank TAC 2009 organisers for providing the data set without which this paper would have never come. The authors also wish to thank anonymous reviewers for their positive feedback. The authors in particular wish to thank Abhilash, Kiran, Praneeth and Arafat for their kind support, help and guidance.

REFERENCES

1. A. Bagga and B. Baldwin. Entity-based cross-document coreferencing using the vector space model. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 79–85. Association for Computational Linguistics Morristown, NJ, USA, 1998.
2. I. Bhattacharya and L. Getoor. A latent dirichlet model for unsupervised entity resolution. In *SIAM International Conference on Data Mining*, pages 47–58, 2006.
3. Y. Chen and J. Martin. Towards robust unsupervised personal name disambiguation. *Proceedings of EMNLP and CoNLL*, pages 190–198, 2007.
4. S. Deorowicz and M.G. Ciura. Correcting spelling errors by modelling their causes. *International journal of applied mathematics and computer science*, 15(2):275, 2005.
5. T. Finin, Z. Syed, J. Mayfield, P. McNamee, and C. Piatko. Book Title: Proceedings of the AAAI Spring Symposium on Learning by Reading and Learning to Read Date: March 23, 2009.
6. E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12, 2007.
7. C.H. Gooi and J. Allan. Cross-document coreference on a large scale corpus. In *Proceedings of HLT/NAACL*, 2004.
8. R. Grishman. Whither written language evaluation. In *Proceedings of the Human Language Technology Workshop*, pages 120–125, 1994.
9. A. Haghighi and D. Klein. Unsupervised coreference resolution in a nonparametric bayesian model. In *ANNUAL MEETING-ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, volume 45, page 848, 2007.
10. Hal Daume III and Daniel Marcu. A bayesian model for supervised clustering with the dirichlet process prior. *J. Mach. Learn. Res.*, 6:1551–1577, 2005.
11. B. Malin. Unsupervised name disambiguation via social network similarity. In *SIAM SDM Workshop on Link Analysis, Counterterrorism and Security*. Citeseer, 2005.
12. G.S. Mann and D. Yarowsky. Unsupervised personal name disambiguation. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 33–40. Association for Computational Linguistics Morristown, NJ, USA, 2003.
13. A.K. McCallum. Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering, 1996.
14. R. Mihalcea. Using wikipedia for automatic word sense disambiguation. In *Proceedings of NAACL HLT*, volume 2007, 2007.
15. D. Milne, O. Medelyan, and IH Witten. Mining domain-specific thesauri from wikipedia: A case study. In *IEEE/WIC/ACM International Conference on Web Intelligence, 2006. WI 2006*, pages 442–448, 2006.

16. K. Nakayama, T. Hara, and S. Nishio. Wikipedia mining for an association web thesaurus construction. *Lecture Notes in Computer Science*, 4831:322, 2007.
17. M. Remy. Wikipedia: The free encyclopedia. *Reference Reviews*, 16.
18. T. Zesch, I. Gurevych, and M. Muhlha user. Analyzing and accessing Wikipedia as a lexical semantic resource. *Data Structures for Linguistic Resources and Applications*, pages 197–205, 2007.

KRANTHI REDDY. B
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY,
HYDERABAD, INDIA
E-MAIL: <BKREDDY@RESEARCH.IIIT.AC.IN>

KARUNA KUMAR
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY,
HYDERABAD, INDIA
E-MAIL: <KARUNA_KY@STUDENTS.IIIT.AC.IN>

SAI KRISHNA
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY,
HYDERABAD, INDIA
E-MAIL: <SAIKRISHNA@RESEARCH.IIIT.AC.IN>

PRASAD PINGALI
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY,
HYDERABAD, INDIA
E-MAIL: <PVVPR@IIIT.AC.IN>

VASUDEVA VARMA
INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY,
HYDERABAD, INDIA
E-MAIL: <VV@IIIT.AC.IN>

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 137-150
RECEIVED 10/11/09 ACCEPTED 16/01/10 FINAL 10/03/10

Brazilian Portuguese WordNet: A Computational Linguistic Exercise of Encoding Bilingual Relational Lexicons

BENTO CARLOS DIAS-DA-SILVA

Universidade Estadual Paulista (UNESP), Brazil

ABSTRACT

This paper describes the methodology of encoding the Brazilian Portuguese WordNet (WN.Br) synsets and the automatic mapping of WN.Br´s conceptual relations of hyponymy, co-hyponymy, meronymy, cause, and entailment relations from Princeton WordNet (WN.Pr). After contextualizing the project and outlining the current lexical database structure and its statistics, it is described the WN.Br editing tool to encode the synsets, its glosses and the equivalence EQ_RELATIONS between WN.Br and WN.Pr synsets, and to select sample sentences from corpora. The conclusion samples the automatic generation of WN.Br´s hyponymy and co-hyponymy conceptual relations from WN.Pr and outlines the ongoing work.

1 INTRODUCTION

Natural language processing (NLP) initiatives to design, build, and compile precise, rich, and robust lexicons for NLP applications are extremely time-consuming and prone to flaws tasks [1] [2], [3] due to the fact that lexicon developers are expected to specify and code huge amounts of specialized and interrelated information as phonetic/graphemic, morphological, syntactic, semantic, and even illocutionary bits of information into computational lexicons [4].

Princeton WordNet (WN.Pr), for example, is a successful sort of a computational lexicon that has set the pattern for compiling bulky relational lexicons systematically. An on-line relational lexical semantic database, WN.Pr combines the designs of a dictionary and of a thesaurus. Similar to a standard dictionary, it covers nouns, verbs, adjectives, and adverbs. After 18 years of research, its 1998 database version (v. 1.6) contained about 94,000 nouns, 10,000 verbs, 20,000 adjectives, and 1,500 adverbs [5].[1] Similar to a thesaurus, words are grouped in terms of lexicalized concepts, which are, in turn, represented in terms of synonym sets (*synsets*), i.e. sets of words of the same syntactic category that share the same concept. Its web structure makes it possible for the user to find a word meaning in terms of both the other words in the same synset and the relations to other words in other synsets as well. Essentially, WN.Pr is a particular semantic network and its sought-after NLP applications have been discussed by the research community [6], [7].

Mirroring WN.Pr´s construction methodology, wordnets of other languages have been under development. EuroWordNet (EWN) [8] is the outstanding multilingual initiative. It is a multilingual wordnet that results from the connection of individual monolingual wordnets by means of encoding the equivalence EQ-RELATIONS (see section 3) between each synset of each individual wordnet and the closest concept represented by the so-called Inter-Lingual-Index (ILI)[2], which enables cross-lingual comparison of words, synsets, concept lexicalizations, and meaning relations from different wordnets [9].

Mirroring both WN.Pr's and EWN's initiatives, and extending the Brazilian Portuguese Thesaurus [10], [11], the Brazilian Portuguese WordNet (WN.Br) project was launched in 2003 and the WN.Br database has been under construction since then. In particular, this paper focuses on the coding of the following bits of information in the database: (a) the co-text sentence for each word-form in a synset; (b) the synset gloss; and (c) the relevant language-independent hierarchical conceptual-semantic relations of hypernymy[3], hyponymy[4], meronymy

[1] The current version (v. 3.0) contains 101,863 nouns, 11,529 verbs, 21,479 adjectives, and 4,481 adverbs. See more details at `http://wordnet.princeton.edu`.

[2] The ILI is an unordered list made up of each synset of the WN.Pr with its gloss (an informal lexicographic definition of the concept evoked by the synset).

[3] The term Y is a hypernym of the term X if the entity denoted by X is a kind of entity denoted by Y.

[4] If the term Y is a hypernym of the term X then the term X is a hyponym of Y.

(part-whole relation), entailment[5] and cause[6] between synsets. Accordingly, section 2 describes the current WN.Br database and its editing tool, an editing GUI (Graphical User Interface), designed to aid the linguist in carrying out the tasks of constructing synsets, selecting co-text sentences from corpora, writing synset glosses, specifying the EQ-RELATIONS, and generating the alignments between the two databases. Section 3, after addressing the issues of cross-linguistic alignment of wordnets by means of the ILI, describes the conceptual-semantic alignment strategy adopted to link WN.Br synsets to WN.Pr synsets by means of the editing tool. Section 4 concludes the paper by exemplifying the automatic mapping of the WN.Pr verb hyponymy and co-hyponymy relations onto the WN.Br verb synsets.

2 THE WORDNET.BR LEXICAL DATABASE

Currently, the WN.Br database contains 44,000 word-forms and 18,500 synsets: 11,000 verbs (4,000 synsets), 17,000 nouns (8,000 synsets), 15,000 adjectives (6,000 synsets), and 1,000 adverbs (500 synsets) [12]. The WN.Br project development strategy assumes a compromise between NLP and Linguistics and, based on the Artificial Intelligence notion of Knowledge Representation [13], [14], applies a three-domain approach methodology to the development of the database. This methodology claims that the linguistic-related information to be computationally modeled, like a rare metal, must be "mined", "molded", and "assembled" into a computer-tractable system [15]. Accordingly, the process of implementing the WN.Br database is developed in three complementary domains: (a) in *the linguistic domain*, the lexical resources (dictionaries and text corpora), the set of lexical and conceptual-semantic relations, and some sort of “natural language ontology of concepts” (e.g. the "Base Concepts" and "Top Ontology" [16]) are mined; (b) in *the computational-linguistic domain*, the overall information that was selected and organized in the preceding domain is molded into a computer-tractable representation (e.g. the "synsets", the "lexical matrix", and the wordnet "lexical database" itself [5]); (c) in *the*

[5] The action A1 denoted by the verb X entails the action A2 denoted by the verb Y if A1 cannot be done unless A2 is done

[6] The action A1 denoted by the verb X is the cause of the action A2 denoted by the verb Y.

computational domain, the computer-tractable representations are assembled by means of the WN.Br editing tool.

2.1 The Linguistic Domain

The WN.Br database architecture conforms to the two key representations of the WN.Pr [5]: the *synset* and the *lexical matrix*: synsets are understood as sets of word-forms built on the basis of the notion of "synonymy in context", i.e. word-form interchangeability in some context [17][7]; the lexical matrix [18] is intended to capture the many-to-many associations between word-form and meaning, i.e. the association of a word-form and the concepts it lexicalize. The lexical matrix is built up by associating each word-form to the synsets to which it is a member. Thus, a polysemous word-form will belong to different synsets, for each synset is intended to represent a single lexicalized concept.

The WN.Br synset developers (a team of three linguists) reused, merged, and tuned synonymy and antonymy information registered in five outstanding standard dictionaries of Brazilian Portuguese (BP) manually ([19], [20], [21], [22], [23, 24])[8], for there are no Brazilian Portuguese machine readable dictionaries (MRDs) and other computer tractable resources available. The NILC Corpus[9] and BP texts available in the web complemented the corpus.

2.2 The Computational-Linguistic Domain

The WN.Br database structures in terms of two lists: the List of Headwords (LH), i.e. the list of word-forms arranged alphabetically, and the List of Synsets (LS) (see Fig.1). Each WN.Br word-form belongs to the LH and is associated to a Sense Description Vector (SDV). Each SDV is co-indexed by three pointers: the "synonymy pointer", which identifies a particular synset in the LS; the "antonymy

---

[7] Antonymy, on the other hand, is checked either against morphological properties of words or their dictionary lexicographical information.

[8] The dictionaries were chosen for their pervasive use of synonymy and antonymy to define word senses, which dictated the strategy to construct the synsets by examining the dictionaries alphabetically, instead of working out synsets by semantic fields.

[9] CETENFolha. Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo. See more details at `http://www.linguateca.pt/`.

pointer", which identifies a particular antonym synset in the LS; and the "sense pointer", which identifies a particular word-form sense number in the SDV. Given such an underlying structure, each synset is linked to its gloss via the “gloss link”, and each word-form is linked to its co-text sentence via the “co-text sentence link”.

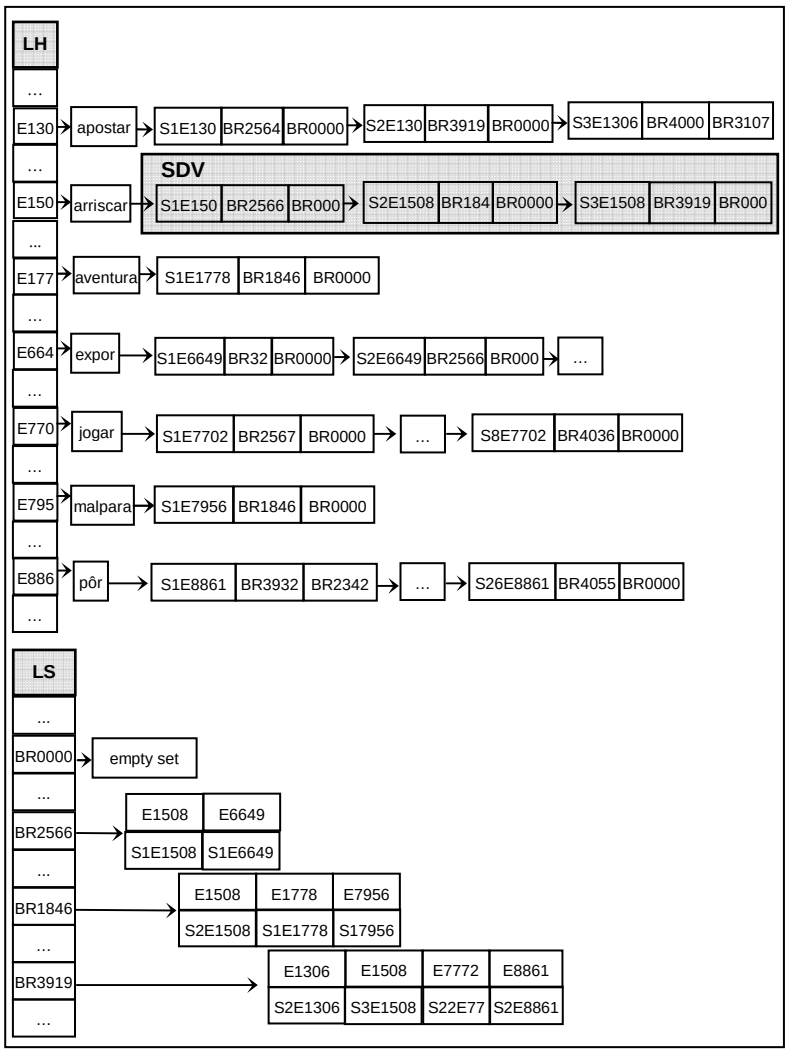


Figure 1: The WN.Br database structure.

2.3 The Computational Domain

The WN.Br editing tool is a Windows®-based GUI that allows the developers (a) to create, consult, modify, and save synsets, (b) to include co-text sentences for each word-form, (c) to write a gloss for each synset (d) to align equivalent synsets equivalence EQ-RELATIONS, (e) to code hyponymy and co-hyponymy relations in the WN.Br automatically, and (f) to generate synset lists (arranged by syntactic category, by number of elements, by the degree of homonymy and polysemy, and by co-text sentence) and WN.Br statistics. Its main functionalities include the storage and bookkeeping of the general information of the database.

The processes of using the editor can be better understood by an example. Fig. 2 shows the basic steps of constructing synsets that contain the BP verb "*lexicalizar*" ("to lexicalize"). In the first dialogue box, the developer selects the appropriate syntactic category and the expected number of synsets to be constructed (i.e. the number of senses); then, s/he clicks on the **Avançar** ("Next") button. In the second dialogue box, the **Todas as Unidades** ("All Unities") field pops up with a list of the word-forms in the WN.Br database. To construct the synset, the developer now selects the appropriate word-forms from the list and clicks on the **Avançar** button. In the third dialogue box, s/he concludes the synset construction by clicking the FIM ("End") button.

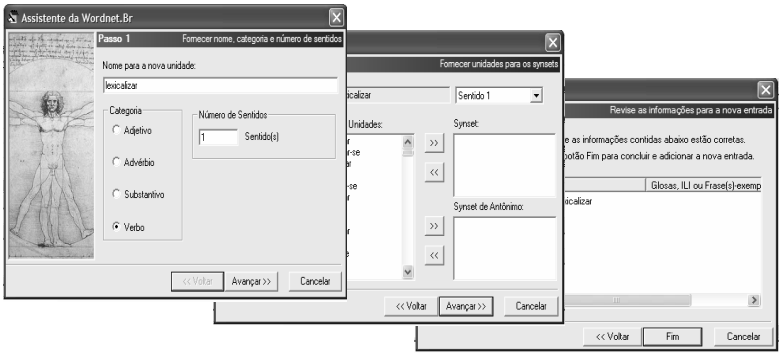


Figure 2: The synset coding wizard.

Co-text sentences, glosses, and ID numbers (see note 10) are pasted/typed in directly in the editor appropriate fields. In Fig.3, the large ellipsis highlights the **Frase(s)-exemplo** ("Sample sentences", i.e

the co-text sentences) field, and the small ellipsis, the **Glosa** ("Gloss") field and the ID number. Currently, the WN.Br database contains 19,747 co-text sentences: Table 1 shows the co-text sentence sources; Table 2 shows the number of co-text sentences per synset.

Table 1: Co-text sentence sources

Source	Nº of Co-text sentences
NILC Corpus	7,659
Aurélio [19]	732
Houaiss [25]	1,761
Michaelis [20]	858
Web	8,052
unknown	685
Total	19,747

Table 2: Co-text sentence statistics

Co-text sentences per synset	Synsets
1	18,604
2	521
3	10

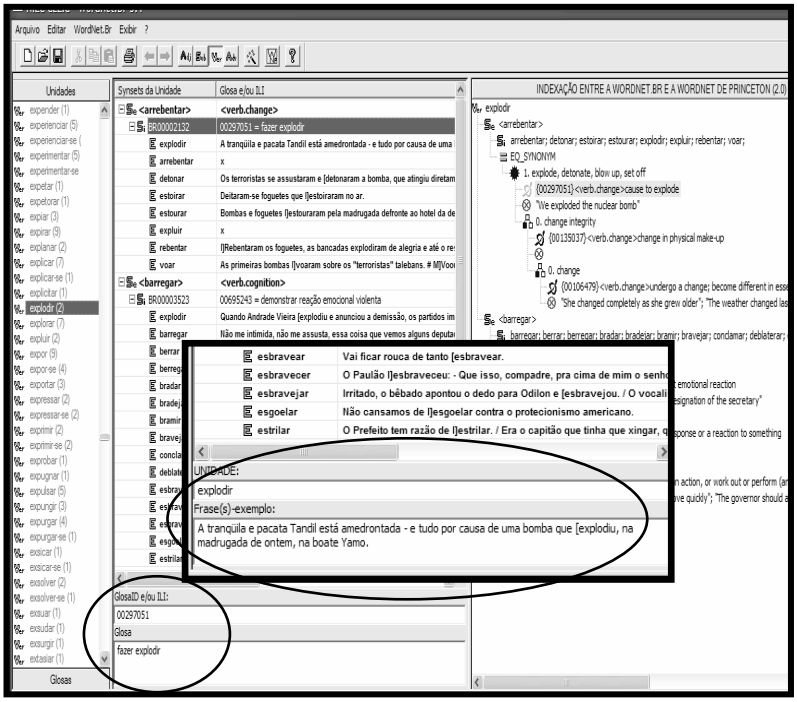


Figure 3: A screenshot with a sample of co-text sentences, glosses, ID alignment) numbers

3 CROSS-LINGUAL ALIGNMENT AND THE WN.BR CONSTRUCTION

The challenge to the WN.Br project has been to specify the equivalence EQ-RELATIONS between WN.Br and WN.Pr (v. 2.0) synsets, for such an alignment is the one that allows researchers to investigate the differences and similarities in the lexicalization processes between BP and English, to develop an English-BP lexical database which can be used in applications such as machine translation systems and cross-language information retrieval involving both languages, and to generate two types of MRDs: a monolingual BP MRD and a bilingual English-Portuguese MRD [12]. Furthermore, and most import for wordnet developers, such an alignment makes it possible the (semi-)automatic specification of the relevant conceptual-semantic relations (e.g. HYPONYMY, TROPONYMY, CO-HYPONYMY, etc.) in the wordnet under construction. In particular, in the WN.Br project, the strategy has been tested successfully to generate such hyponym and co-hyponym relations in the WN.Br verb database (see Fig 6).

The cross-lingual equivalence relations between wordnets are mined in accordance with the types identified in [8], the so-called, self defining EQ-RELATIONS (EQ-SYNONYM, EQ-NEAR-SYNONYM, EQ-HAS-HYPERONYM, and EQ-HAS-HYPONYM). Linguistic mismatches (lexical gaps, due to cultural specificities, pragmatic differences, and morphological mismatches; over/under-differentiation or of senses; and fuzzy-matching between synsets) and technical mismatches (mistakes in the choice of the appropriate EQ-RELATIONS) as have been described in [9] are also accounted for during the alignment procedure. The equivalence EQ-RELATIONS and cross-lingual mismatches are molded into a computer-tractable representation of the ILI-records[10]. The ILI-record is handy for the development, maintenance, future expansion, and reusability of a multilingual wordnet, dispenses with the development and maintenance of huge and complex semantic structures to gather all the senses encoded by each individual wordnet into a multilingual wordnet, and makes the task of adding individual wordnet to a multilingual wordnet less costly [9].

As shown in Fig.4, the structure of the WN.Br database has been extended to encode the cross-lingual equivalence EQ-RELATIONS. Besides the LH and LS lists and the SDV pointers (see 2.2), each synset structure has been augmented with an additional vector to register both the wordnet standard language - independent conceptual - semantic relations (e.g. HYPONYMY,

[10] An ILI-record is a WN.Pr (v. 2.0) synset, its gloss and its ID number [9].

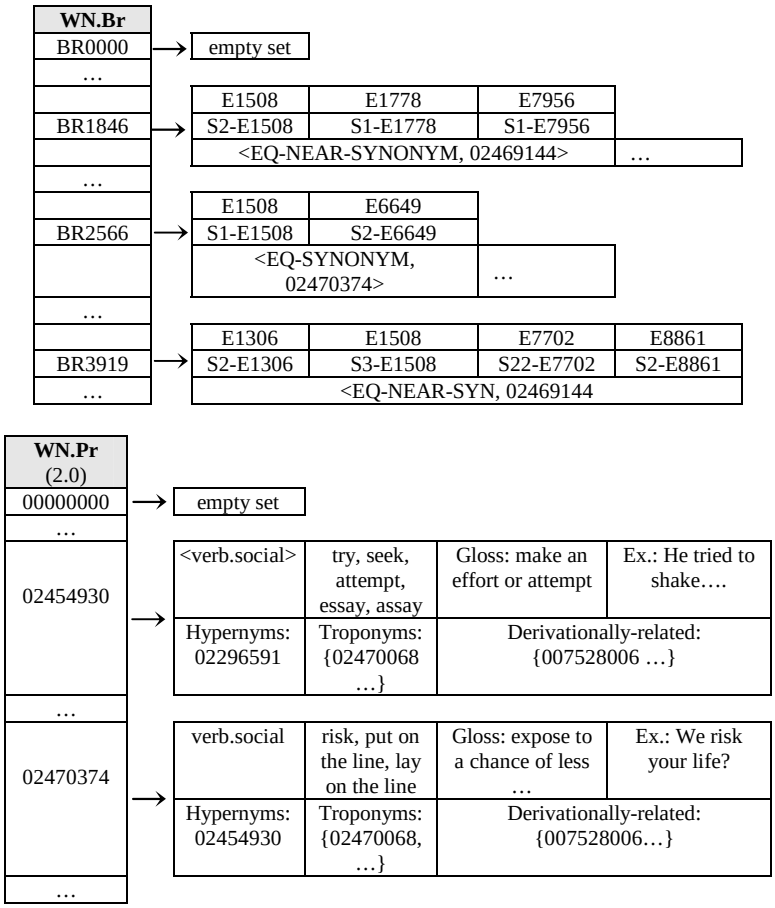


Figure 4: The synset structure augmented with conceptual-semantic EQ-RELATIONS.

TROPONYMY, CO-HYPONYMY, etc.) and the cross-lingual conceptual-semantic EQ_RELATIONS between synsets of the two wordnets. This new vector enriches the WN.Br database structure with the following cross-linguistic information: the "universal" synset semantic type (e.g. <verb.social>), the corresponding English synset (e.g. {*risk, put on the line, lay on the line*}), the English version of the universal gloss (e.g. Expose to a chance of loss or damage), the English co-text sentence (e.g. "Why risk your life?"), and EQ-RELATIONS (e.g. EQ-SYNONYM relation).

The current WN.Br editing tool has three interconnecting modules implemented as a GUI. Each module, in turn, makes it possible for the developer to carry out specific tasks during the procedure for aligning the synsets across the two wordnets: searching the BP-English dictionary, the WN.Br and WN.Pr databases, and the web.

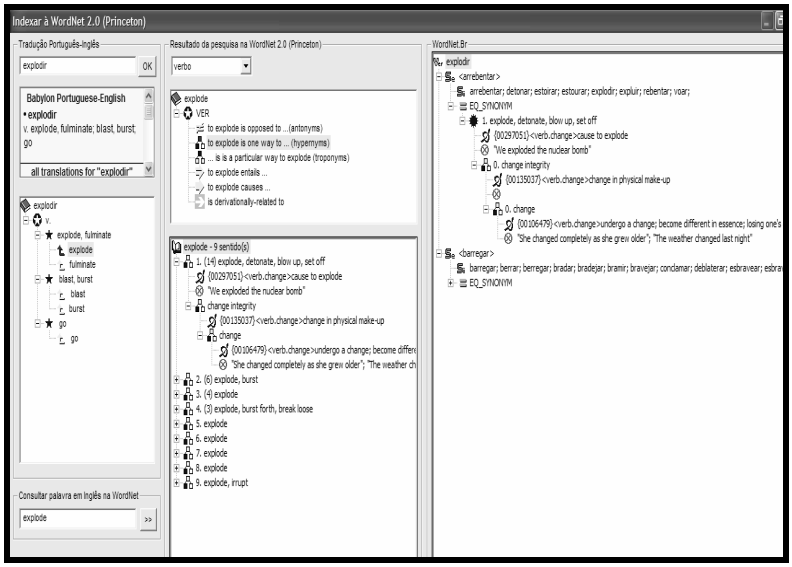


Figure 5: A screenshot of the three-column GUI of the WN.Br tool.

The WN.Br database developer starts off the alignment by right clicking on a target WN.Br word-form. As shown in Fig. 5, the editor in turn displays its three column GUI: on its left, an online bilingual BP-English dictionary and a WN.Pr database search field; in the middle, the selected WN.Pr synset information; on its right, the WN.Br synsets that contain the target word-form. The developer, in the left column, (i) checks all possible English word-forms (e.g. *explode, fulminate, blast, burst, go*) that are equivalent to the target BP word-form (e.g. *explodir*) with recourse to the dictionary and selects the appropriate one (e.g. *explode*); in the middle and right columns, (ii) analyzes the possible types of equivalence EQ-RELATIONS between the two sets of synsets: the ones in the middle column – the sets of synsets of the WN.Pr database (e.g. {*explode, detonate, blow up, set off*}, {*explode, burst*}, etc.) – and the ones in the right column – the sets of synsets of the

WN.Br database that contain the target word-form (e.g. {*arrebentar, detonar, estoirar, estourar, explodir, expluir, rebentar, voar*}, and {*barregar, berrar, berregar, bradar, bradejar, bramir, bravejar, condamar, deblaterar, esbravear, esbravejar,* }). In this particular example, the resulting EQ-SYNONYM alignment is {*explode, detonate, blow up, set off*} and {*arrebentar, detonar, estoirar, estourar, explodir, expluir, rebentar, voar*}

After the specification of alignments such as the one above, Fig. 6 sketches how the WN.Br verb database inherits both hyponym and co-hyponym relations from de WN.Pr verb database automatically. After the manual specification of the following EQ-SYNONYM alignments[11] **tentar=try**, **apostar=gamble**, and **arriscar=risk**, the WN.Br editing tool generates the following relations automatically: **apostar** and **tentar**, **arriscar** and **tentar** are hyponyms; **arriscar** and **apostar** are co-hyponyms.

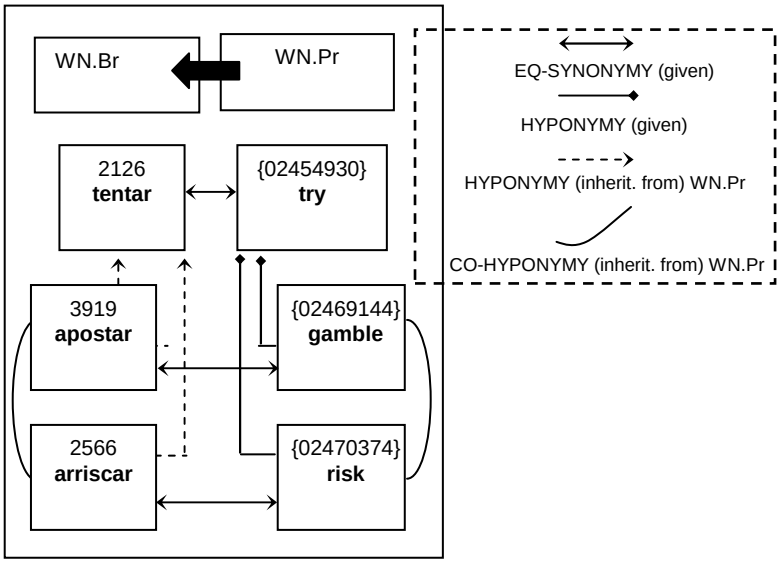


Figure 6: A sample of the automatic encoding of hyponymy and co-hyponymy relations.

[11] For short, Fig. 6 specifies the most representative **word-forms** of each synset: **tentar**: {***tentar***, *ensaiar, experimentar*}; **try**: {***try***, *seek, attempt, essay, assay*}; **apostar**: {*apostar, arriscar, jogar, pôr*}; **gamble**: {***gamble***, *chance, risk, hazard, take chances, adventure, run a risk, take a chance*}; **arriscar**: {*arriscar, aventurar, malparar*}; **risk**: {***risk***, *put on the line, lay on the line*}.

4 FINAL REMARKS

In sum, this paper described the design and content of the current WN.Br database, the procedures and tools for coding synsets, co-text sentences, glosses, language-independent conceptual-semantic relations, and conceptual-semantic equivalence EQ-RELATIONS. The overall procedures for constructing wordnets presented in this paper, though not resorting to reusing existing resources, a current tendency in the field [26], devised a reliable, an efficient, and an automatic way of inheriting WN.Pr´s internal relations in the task of constructing wordnets to other languages.

On the way, besides the specification of the other language-independent conceptual-semantic relations for the verb synsets, it is the encoding of (a) a gloss for each synset of nouns; (b) a co-text sentence for each noun; (c) the mapping of the WN.Br noun synsets to its equivalent ILI-records by means of the following equivalence relations EQ-SYNONYM, EQ-NEAR-SYNONYM, EQ-HAS-HYPERONYM, and EQ-HAS-HYPONYM, and (d) the automatic inheritance from WN.Pr of the relevant conceptual-semantic relations of hyponymy/hypernymy, co-hyponymy, and meronymy/holonymy relations for nouns.

ACKNOWLEDGEMENTS

This project was supported in part by contract 552057/01, with funding provided by The National Council for Scientific and Technological Development (*CNPq*), Brazil; in part by grant 2003/03623-7 from The State of São Paulo Research Foundation (FAPESP), Brazil. My thanks go to all my linguistics students at *CELiC* and the *NILC* developers for their invaluable help in constructing the WN.Br core database. Special thanks to *CAPES*, the *PPG Linguística e Língua Portuguesa, FCL-UNESP/Araraquara*, and the *PROPG-UNESP*.

REFERENCES

1. Palmer, M. (ed.): Miltilingual resources – Chapter 1. In: Eduard Hovy, Nancy Ide, Robert Frederking, Joseph Mariani, and Antonio Zampolli (eds.): Linguistica Computazionale, Vol. XIV-XV (2001)
2. Hanks, P.: Lexicography. In: The Oxford Handbook of Computational Linguistics, R. Mitkov (ed.), Oxford, Oxford University Press (2003)

3. Di Felippo, A., Pardo, T.A.S., Aluísio, S.M. Proposta de uma metodologia para a identificação dos argumentos dos adjetivos de valência 1 da língua portuguesa a partir de córpus. In: Carderno de Resumos do V Encontro de Corpora, São Carlos, São Paulo (2005) 20-21
4. Handke, J.: The structure of the Lexicon: human versus machine. Berlin: Mouton de Gruyter (1995)
5. Fellbaum, C. (ed.): WordNet: An Electronic Lexical Database. The MIT Press, Cambridge (1998)
6. Alonge, A., Calzolari, N., Vossen, P., Bloksma, L., Castellon, I., Marti, M.A., Peters, W.: The Linguistic Design of the EuroWordNet Database. Computers and the Humanities, Vol. 32 (1998) 91-115
7. Gonçalo, J., Verdejo, F., Peters, C., Calzolari, N.: Applying EuroWordNet to Cross-Language Text Retrieval. Computers and the Humanities, Vol. 32 (1998) 185-207
8. Vossen, P.: Introduction to EuroWordNet. Computers and the Humanities, Vol. 32(2,3)(1998) 73-89
9. Peters, W., Vossen, P., Díez-Orzas, P., Adriaens, G.: Cross-linguistic Alignment of Wordnets with an Inter-Lingual-Index. Computers and the Humanities, Vol. 32 (1998) 221-251
10. Dias-da-Silva, B.C., Oliveira, M.F.; Moraes, H.R. Groundwork for the development of the Brazilian Portuguese Wordnet. Advances in natural language processing. Berlin: Springer-Verlag (2002)189-196
11. Dias-da-Silva. B.C.; Moraes, H.R. A construção de thesaurus eletrônico para o português do Brasil. Alfa. São Paulo: Editora Unesp, Vol. 47(2) (2003) 101-115
12. Dias-da-Silva, B.C.: Human language technology research and the development of the Brazilian Portuguese wordnet. In: Proceedings of the 17th International Congress of Linguists – Prague, E. Hajičová, A. Kotěšovcová, J. Mírovský, ed., Matfyzpress, MFF UK (2003) 1-12
13. Hayes-Roth, F.: Expert Systems. In: Encyclopedia of Artificial Intelligence, E. Shapiro (ed.), Wiley, New York (1990) 287-298
14. Durkin, J.: Expert Systems: Design and Development. Prentice Hall International, London (1994)
15. Dias-da-Silva, B. C.: Bridging the Gap Between Linguistic Theory and Natural Language Processing. In: 16th International Congress of Linguists – Paris, B. Caron, ed., Pergamon-Elsevier Science, Oxford (1998) 1-10
16. Rodríguez, H, Climent, S., Vossen, P., Bloksma, L., Peters, W. Alonge, A., Bertagna, F., Roventini, A.: The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top-Ontology. Computers and the Humanities, Vol. 32 (1998) 117-152
17. Miller,G.A.:Dictionaries in the Mind. Language and Cognitive Processes,Vol.1(1986)171-185
18. Miller, G.A., Fellbaum, C.: Semantic Networks of English. Cognition 41 (1991) 197-229

19. Ferreira, A. B. H.: Dicionário Aurélio Eletrônico Século XXI. Lexicon, São Paulo, CD-ROM (1999)
20. Weiszflog, W. (ed.): Michaelis Português – Moderno Dicionário da Língua Portuguesa. DTS Software Brasil Ltda, São Paulo, CD-ROM (1998)
21. Barbosa, O.: Grande Dicionário de Sinônimos e Antônimos. Ediouro, Rio de Janeiro, 550 p. (1999)
22. Nascentes, A.: Dicionário de Sinônimos. Nova Fronteira, Rio de Janeiro (1981)
23. Borba, F.S. (coord.): Dicionário Gramatical de Verbos do Português Contemporâneo do Brasil. Editora da Unesp, São Paulo, 600 p. (1990)
24. Borba, F.S.: Dicionário de usos do português do Brasil. São Paulo: Ed. da UNESP (2002)
25. Houaiss, A.: Dicionário Eletrônico Houaiss da Língua Portuguesa. FL Gama Design Ltda., Rio de Janeiro CD-ROM (2001)
26. Rigau, G., Eneko, A.: Semi-automatic methods for WordNet construction. In: 1st Intenational WordNet Conference Tutorial, Mysore, India (2002)

BENTO CARLOS DIAS-DA-SILVA
CENTRO DE ESTUDOS LINGÜÍSTICOS
E COMPUTACIONAIS DA LINGUAGEM - CELIC1,
FACULDADE DE CIÊNCIAS E LETRAS,
UNIVERSIDADE ESTADUAL PAULISTA (UNESP)
CAIXA POSTAL 174 – 14.800-901, ARARAQUARA, SP, BRAZIL
E-MAIL: <BENTO@FCLAR.UNESP.BR>

Parsing and Disambiguation

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 153-168
RECEIVED 24/11/09 ACCEPTED 16/01/10 FINAL 11/03/10

Identifying Different Meanings of a Chinese Morpheme through Latent Semantic Analysis and Minimum Spanning Tree Analysis

BRUNO GALMAR, JENN-YEU CHEN

National Cheng Kung University, Taiwan

ABSTRACT

A character corresponds roughly to a morpheme in Chinese, and it usually takes on multiple meanings. In this paper, we aimed at capturing the multiple meanings of a Chinese morpheme across polymorphemic words in a growing semantic micro-space. Using Latent Semantic Analysis (LSA), we created several nested LSA semantic micro-spaces of increasing size. The term-document matrix of the smallest semantic space was obtained through filtering a whole corpus with a list of 192 Chinese polymorphemic words sharing a common morpheme (港 gong1). For each of our created Chinese LSA space, we computed the whole cosine matrix of all the terms of the semantic space to measure semantic similarity between words. From the cosine matrix, we derived a dissimilarity matrix. This dissimilarity matrix was viewed as the adjacency matrix of a complete weighted undirected graph. We built from this graph a minimum spanning tree (MST). So, each of our LSA semantic space had its associated MST. It is shown that in our biggest MST, paths can be used to infer and capture the correct meaning of a morpheme embedded in a polymorphemic word. Clusters of the different meanings of a polysemous morpheme can be created from the minimum spanning tree. Finally, it is concluded that our approach could model partly human knowledge representation and acquisition of the different meanings of Chinese polysemous morphemes. Our work is thought to bring some insights to the Plato's problem and additional evidence towards the plausibility of words serving as ungrounded symbols. Future directions are sketched.

1 INTRODUCTION

Polymorphemic Chinese words are composed of the binding of two Chinese characters (e.g. 王公) or more (e.g. 公路赛) . We proposed a computational approach to extract the different meanings of 公 in a list[1] of 192 polymorphemic 公 words which occur in a corpus.

A Chinese character like 公 corresponds roughly to a morpheme in Chinese, and it usually takes on multiple meanings. For example, an etymology dictionary offers the following 16 senses[2] -16 etymological dimensions of meaning- for the character 公 (gong1) :

unselfish / unbiased / fair / to make public / open to all / public / the first of old China's five-grades of the nobility / an old Chinese official rank / the father of one's husband (one's husband's father) / one's father-in-law / one's grandfather / a respectful salutation / the male (of animals) / office / official duties / a Chinese family name

公 can take one of these meanings in the words in which it occurs. In the word 公平 (fair) the meaning of 公 is fair. In this case, the meaning of the morpheme is identical with the one's of the bimorphemic word. This *"fair"* meaning of 公 is different from the meaning of 公 in 公园 (public park,park) which is *"public, open"*.

Our computational approach to infer the meaning of 公 in polymorphemic words can be unfolded in five steps:

1. Through filtering a Chinese corpus by three nested list of words, we created three nested term-document matrices, weighted them and computed reduced Singular Value Decomposition (SVD) on them to obtain three nested Latent Semantic Analysis (LSA) semantic spaces.
2. For each LSA semantic space we computed the cosine matrix and the dissimilarity matrix for all terms.
3. We used each dissimilarity matrix as the adjacency matrix of a complete weighted undirected graph.
4. We built the Minimum Spanning Tree of each graph.
5. We browsed and analyzed paths in the Minimum Spanning Tree for extraction of the meaning of 公 in the polymorphemic words.

We reviewed Chinese computational morphology and Chinese word sense disambiguation literature and found no prior work proposing such a com-

[1] Actually, this list includes some idioms like 天公不作美 which could not be satisfactorily labeled as polymorphemous words.

[2] source: www.chineseetymology.org/ The list is still not exhaustive!

putational approach for meaning identification of a polysemous morpheme in Chinese words.

We know of no Chinese dictionary or database which lists for each meaning of a polysemous morpheme all the Chinese words embedding the morpheme with this meaning. For example, the Chinese Wordnet of the Academica Sinica[3] proposes a list of some of the different meanings of 公 but provides no listing of all the 公 words with a same given meaning of 公 e.g. *"fair"*.

Our primary research goal is to design tools for Chinese cognitive scientists and linguists who study the semantic interaction between Chinese morphemes and polymorphemic words. Our tools will serve to prepare experimental materials for lexical decision tasks and relatedness judgment tasks involving the repetition of a same Chinese polysemous morpheme embedded with a fixed identified meaning in different Chinese words. [1,2].

2 THE NESTED SEMANTIC LATENT SEMANTIC ANALYSIS SPACES

We used the Academica Sinica Balanced Corpus (ASBC), a five million words corpus based on Chinese materials from Taiwan. The corpus is made of 9183 documents which are considered as semantically meaningful units. Most of the functional words were removed from the corpus.[4]

In the ASBC corpus, 公 as a monomorphemic word occurs with 5 different POS tags: "公(Vh)", "公(Nb)", "公(Nc)", "公(Na)" and "公(A)". These 5 公 words and 187 additional polymorphemic 公 words constitute the list of 192 公 words under study.

2.1 *The First Term-Document Matrix (192 Words 3716 Documents)*

The first and smallest of our term-document matrices was obtained through filtering a whole corpus with a list of 192 公 words. The resulting term-document matrix is made of 192 rows - representing the 192 公 words - and 3716 columns - representing all the documents in which at least one of the 公 words occurs -. The term frequency of each 公 word in each document is stored in that term-document matrix. At that level, we know

[3] http://cwn.ling.sinica.edu.tw/

[4] Words with the following POS tags were removed: Dk Di Caa Cbb Nep Nh P Cab Cba DE I T SHI Neu. For more information about the meaning of the tags, please refer to CKIP Technical Report 95-02/98-04

how the 192 公 words co-occur in the ASBC corpus and we voluntary ignore both the huge number of remaining terms in the corpus and the set of documents in which the 公 words do not occur. This minimalist term-document matrix will serve after Latent Semantic Analysis to create our smallest LSA semantic space. This space is thought to be the worst or poorest representation of the semantic relationships between the 192 公 words.

2.2 *The Second Term-Document Matrix (202 Words 4327 Documents)*

We wanted our second LSA semantic space to contain at least ten words that represent *10 etymological dimensions* of 公. These 10 dimensions words were thought to be able to serve as attractors of semantically similar 公 words and eventually as centroids of 公 clusters. These words could serve later to infer the meaning of 公 in 公 words. We first devised a list of twelve words: (公正, 公平, 公開, 公共, 無私, 貴族, 爵位, 父, 岳父, 雄性, 機關, 機構). These twelve words capture 10 relatively different dimensions of meaning of 公. Both the pairs (貴族, 爵位) and (父, 岳父) are semantically redundant. For example the words 貴族 (noble, nobility) and 爵位 (order of feudal nobility) capture the same meaning of nobility. The word 岳父 (father's in law) is an hyponym of 父 (father), they both capture the fatherhood's relationships meanings of 公. Later we could observe which word in each pairwise behaves as the strongest attractor. In the twelve words list, the first four words are 公 words already present in the first semantic space. Thus to create the second semantic space, we added to the initial list of 192 公 words, the words (貴族, 爵位, 父, 岳父, 雄性, 機關, 機構). We also included the words (國際, 國際性) to attract 公 words referring to international metric units (e.g. 公克 (gram), 公分 (centimeter), 公升 (liter)). After filtering the whole corpus with the new list of 202 words, we obtained a term-document matrix of 202 terms and 4327 documents.

2.3 *The Third Term-Document Matrix (283 Words 6798 Documents)*

To create the third LSA semantic space we added to the precedent list of 202 words:

1. words which are key-words occurring in a Chinese dictionary's definitions of some of the 187 polymorphemic 公 words. For example the definition for åĚňéĞŇ (kilometer) is "量詞。計算長度的單位". The words (量詞, 計算, 長度, 單位) were all added for building the third list of terms.

2. words which share some common morphemes to the multimorphemic words of the initial list. (e.g. 女王 shares the 王 morpheme with 王公)
3. a few words (國家(country), 事物 (thing)) which occur in category labels created by two Taiwanese participants in a pilot study of the subjective sorting of the 187 polymorphemic 公 words.
4. and words[5] which were thought to be potential attractors of certain 公 words (e.g. 動物 (animal) for 公鹿 (male deer), 豬公 (male pig), 公里 (cock) or 七矮人 (seven dwarfs) for 白雪公主 (White-Snow)).

After filtering the whole corpus with a new list of 283 words, we obtained a term-document matrix of 283 terms and 6798 documents. This matrix will serve to compute our biggest micro-semantic space. This third semantic space was thought to be semantically complete and rich enough to embed meaningful semantic relationships between the words its contains. Such a micro-size space could be a better start than a whole corpus semantic space to investigate the different meanings of 公 in 公 words.

2.4 *The Three Weighting Schemes*

To each of our three term-document matrices we applied a total of three weighting schemes:

1. The term-document matrix containing the term frequencies m_{ij} was logarithmised by computing:

$$\log(m_{ij}+1) \tag{1}$$

   as a local weighting scheme. The benefit is to reduce the frequency effect between terms in a same document.
2. As a global weighting scheme, we used the Inverse Document Frequency scheme[3,4]. Every row i - representing the term frequencies of $term_i$ - of the term-document matrix is multiplied by:

$$\log_2\left(\frac{Number\ of\ documents\ in\ the\ corpus}{Numbers\ of\ documents\ in\ which\ the\ term_i\ appears}+1\right) . \tag{2}$$

   Such a weighting scheme gives more weight to words with a global low frequency.

[5] Automatic selection of these words is still to be done. These words were added for testing purposes. They can be removed.

3. At the document level - the columns of the term-document matrix - we also applied a weighting scheme. To reduce the effect of the size difference between documents, we multiplied each column of the term-document matrix by:

$$\log_2\left(\frac{Max\,document\,size}{Document\,size}+1\right) \quad . \tag{3}$$

More weight is given to small documents. This document level weighting scheme is preferred to resizing the corpus's meaning unit from the original entire document to paragraph of a given size. Resizing could result in splitting meaningful units in different documents.

2.5 *Singular Value Decomposition And Reduced SVD*

After applying the three weighting schemes to the term-document matrices, we computed their reduced Singular Value Decomposition (SVD).

Given $U=[u_1,...,u_m]\,\varepsilon\,R^{m*n}$ and $V=[v_1,...,v_n]\,\varepsilon\,R^{n*n}$ two orthogonal matrices, the SVD of a term-document matrix A can be written:

$$A=U\Sigma V^T=\sum_{i=1}^{p}\sigma_i u_i v_i^T \text{ with } \Sigma=diag(\sigma_1,...,\sigma_p)\,\varepsilon\,R^{m*n}\ ,\ p=min\{m,n\} \quad . \tag{4}$$

where $\sigma_1\geq\sigma_2\geq\sigma_p\geq 0$ are the singular values.

For example, for the third term-document matrix, we have $m=283$ and n= 6798.

Thus, the full SVD represents terms and documents in a 283 dimensions space.

After several trials[6], we decided to reduce the dimensionality of the LSA spaces by taking into account only the first one hundred singular values. So for our three term-document matrices we operated a reduced SVD to obtain three 100 dimensions spaces. This can be written:

$$A\simeq A_{100}=U_{100}\Sigma_{100}V_{100}^T \quad . \tag{5}$$

where $\Sigma_{100}=diag(\sigma_1,...,\sigma_{100})$ and $\sigma_1\geq\sigma_2\geq\sigma_{100}>0$ are the 100 first non-zeros singular values.

We termed $A_{192,100}$, $A_{202,100}$ and $A_{283,100}$ the three reduced LSA semantic spaces containing respectivelly 192, 202 and 283 words.

[6] We tried different values, including a dimension equal to the lowest dimension of the term-document matrix -the number of terms- but these choices were discarded while comparing the quality of results described in part 4.

2.6 *Cosine Matrix*

To compare semantic similarity between two words in a LSA space, the cosine measurement of the two vectors v_i, v_n representing the two terms is computed as:

$$\cos(v_i, v_n) = \frac{v_i \cdot v_j}{\|v_i\|\|v_j\|} \quad . \tag{6}$$

For each of our created Chinese LSA space, we computed the whole cosine matrix C of all the terms to measure semantic similarity between words.

$$C = \begin{pmatrix} 1 & \cdots & \cos(v_1, v_j) & \cdots & \cos(v_1, v_m) \\ \vdots & \ddots & & & \\ \cos(v_j, v_1) & & \ddots & & \\ \vdots & & & \ddots & \\ \cos(v_m, v_1) & & & & 1 \end{pmatrix} \quad . \tag{7}$$

C is symmetric due to $\cos(v_i, v_j) = \cos(v_j, v_i)$. We computed the three cosine matrices C_{192}, C_{202} and C_{283} whose dimensions are respectively 192*192, 202*202 and 283*283.

2.7 *Dissimilarity Matrix*

From the cosine matrix C, the dissimilarity matrix D is derived.

$$D = \begin{pmatrix} 1 & \cdots & \cdots & \cdots & 1 \\ \vdots & \ddots & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & & \ddots & \vdots \\ 1 & \cdots & \cdots & \cdots & 1 \end{pmatrix} - C \quad . \tag{8}$$

$$\text{with } d_{ij} = 1 - \cos(v_i, v_j) \geqq 0$$

We computed the three dissimilarity matrices D_{192}, D_{202} and D_{283} whose dimensions are respectively 192*192, 202*202 and 283*283.

3 GRAPH-THEORY BASED APPROACH

3.1 *A Few Definitions*

A *graph G* = (V,E) is an ordered pair, where V is a set whose elements are called *vertices*, and where E is a set of pairs of distinct vertices. Given *p* and *q* two vertices of V, the element {p,q} ε E is called an *edge* and link the vertices *p* and *q*.

When edges are given a weight - a real number here -, the graph is said to be *weighted*. If no orientation is assigned to edges, the graph is said to be *undirected*. When for every pair of vertices V_i, V_j, there is a sequence of edges allowing to join V_i and V_j, then the graph *G* is said to be connected. If every pair of vertices in *G* is directly connected through an edge, the graph is said to be *complete*. Two vertices V_i and V_j linked by an edge are said to be *adjacent*.

The *adjacency matrix A* of a complete weighted graph *G* is the matrix whose entry A_{ij} is 0 if $i = j$ and otherwise is w_{ij} the weight assigned to the edge V_i, V_j[5,6].

A *tree* of a graph *G* is a connected subgraph of *G* with no cycle. A *spanning tree* (ST) of a graph *G* is a tree of *G* which contains all the vertices of *G*.

A *minimum spanning tree (MST)* of a graph *G* is a spanning tree (ST) of *G* whose the sum of edges is minimum[5,6]. This can be written:

$$\sum_{e \varepsilon MST} w(e) = \min_{ST \varepsilon G} \left(\sum_{e \varepsilon ST} w(e) \right) \quad . \tag{9}$$

3.2 *Applying Graph Theory to the Dissimilarity Matrix*

The dissimilarity matrix *D* introduced in Âğ2.7 can be viewed as the adjacency matrix of a complete weighted undirected graph *G*. The rows and the columns of the adjacency matrices represent the words under study. Each word is a vertex of *G* and each edge of *G* linking two vertices v_i and v_j is weighted by d_{ij}. Thus we have:

$$\forall i \, , \, d_{ii} = 0 \text{ and } \forall (i, j) \text{ with } i \neq j \, , \, d_{ij} = 1 - \cos(v_i, v_j) \quad . \tag{10}$$

From each of the three dissimilarity matrices D_{192}, D_{202} and D_{283}, we used Prim's algorithm to build three minimum spanning trees MST_{192}, MST_{202} and MST_{283} [7]. Hence, each of our LSA semantic space $A_{192,100}$,

$A_{202,100}$ and $A_{283,100}$ has an associated minimum spanning tree. Uniqueness of the MST of a graph G is ensured if each edge of G has a different weight. By removing edges of comparatively high weights in the MST, clusters can be formed [8].

Lemma 1. *[9]*
Any two vertices in a tree are connected through a unique path

Therefore in a MST, the path connecting two vertices is unique. The length of the path between two vertices could be measured by:

1. summing the weights of all the edges composing the path.
2. combining the precedent sum with the total number of intermediary nodes.
3. qualitatively summing the number of concepts composing the path.

Length can serve as an indicator of similarity between two words. This similarity can be interpreted as semantic, situational or of other nature. The shorter the length of the path between two words, the closer is their similarity relationship.

We studied the paths from any of the åĚň words to the twelve words representing the etymological dimensions of åĚň. We also looked at the paths from the twelve dimensions words to the five åĚň morphemes with different POS tags.

4 RESULTS

4.1 *Uniqueness of the Three MST*

For each of the three adjacency matrices D_{192}, D_{202} and D_{283}, some edges have a same weight. Therefore, we concluded than none of our three minimum spanning trees MST_{192}, MST_{202} and MST_{283} is unique.

4.2 *A 192 Vertices MST MST_{192}*

The first MST contained only all the 亼 words.

For Chinese native readers, few of the 191 edges of the MST_{192} between polymorphemic 亼 words bear relevant semantic similarity information. We listed some examples of such edges in Table 1.

Table 1. Edges capturing genuine semantic similarity

Edge	Weight
公秉(nf) – 公升(nf)	4.675626e-03
有限公司(nc) – 公司法(na)	1.304338e-02
公民權(na) – 公民(na)	3.671904e-01
公費生(na) – 公費(na)	9.534522e-02
豬公(na) – 大豬公(na)	8.965163e-06
蘇花公路(nc) – 橫貫公路(nc)	5.542596e-03

MST_{192} captures some hyponomic relationships: 公秉 (kiloliter) and 公升 (liter), or situational relationships: 聖誕老公公 (Father Christmas) and 百貨公司 (department store) as Father Christmas can be found in department store around Christmas.

4.3 *A 202 Vertices MST* MST_{202}

MST_{202} embeds all the 公 words and the selected words representative of dimensions of meanings of 公. In MST_{202}, on average, words belongs to two edges. The twelve dimensions words, on average, also share two edges with other words. Of all the dimensions words, only 父 and 爵位 serve as hypothesized as strong 公 attractors by attracting respectively 4 and 5 words. For a Chinese reader, there are no genuine semantic relationships between 爵位 and the words forming edges with it. 貴族 - representing the same meaning dimension as 爵位 - behaves as a weak attractor by sharing only one edge with a 公 word. 國際 failed to attract international metric units.

The hyponomic relationship in Table 2 between 岳父 and 父 is captured by one edge between the two words.

Table 2. Edge capturing the "father's in law"–"father" hyponomic relationship

Edge	Weight
父(na) – 岳父(na)	8.157458e-02

Except that the five 公 monomorphemic words are outliers, clustering does no provide additional insightful information than simple browsing of the MST.

4.4 *A 283 Vertices Tree LSA 100 dimensions MST_{283}*

In MST_{283}, on average, words belongs to 2 edges as for MST_{202}. The average for the dimensions words is in increase, slightly over 2 (e.g. 2.16). Compared to the two precedent MST, MST_{283} can be used to extract genuinely the meaning of a 公 in a 公 word.

INFERRING THE MEANING OF 公 IN 公鹿 (MALE DEER). Table 3 lists the three edges forming the path from 公鹿 (male deer) to 雄性 (male or maleness) and one edge joining 雄性 and one of the monomorphemic 公 word 公(A).

Table 3. Sequence of Edge capturing the "maleness" meaning of 公 in 公鹿 (male deer)

Edge	Weight
雌(a) – 公鹿(na)	2.363839e-03
雌性(na) – 雌(a)	6.576066e-03
雌性(na) – 雄性(na)	6.728410e-02
雄性(na) – 公(a)	1.832874e-01

Figure 1 represents graphically these four edges. The morpheme 公(A) also shares two additional edges with two polymorphemic 公 words - 公Word in Fig. 1 -.

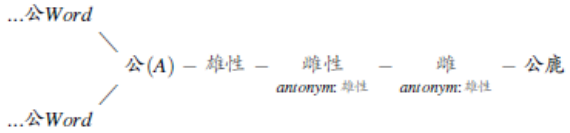


Fig. 1. Paths from 公鹿 (male deer) to 雄性 (maleness) and from 公(A) to 雄性

The two intermediary words 雌性 (female or femaleness) 雌 (female) between 公鹿 and 雄性 are non-公 words and are both antonyms to 雄性. We can say that the path from 公鹿 to 雄性 is conceptually of length 1: only one concept (femaleness) separates the concept of 公鹿 and 雄性.

Besides, 雄性 which is one of our dimension word has attracted a monomorphemic 公 word 公(A). This can mean that one of the meaning of 公(A) is related to 雄性. 公(A) shares two other edges with the words 外公 and 公使. All the three edges are listed in table 4.

Table 4. The three edges of 公(A)

Edge	Weight
公使(na) – 公(a)	6.327506e-01
外公(na) – 公(a)	9.058475e-01
雄性(na) – 公(a)	1.832874e-01

The edge {雄性, 公} has the smallest weight. Thus we can attach 雄性 as a primary meaning to 公.

From the three propositions:

1. In all MST_{283}, 公鹿 shares only one edge with a word: 雌
2. Only one concept (femaleness) separates the concept of 公鹿 and 雄性.
3. 雄性 is a primary meaning of 公(A).

We can infer that in MST_{283}, the closest meaning of 公 in 公鹿 (male deer) is 雄性 (male, maleness). Every Chinese speaker will agree on the meaningfulness and correctness of such a conclusion.

VISUAL REPRESENTATION OF MST_{283} AND CLUSTERING.

MST_{283} is plotted on Fig.2. åĚň dimension words and monomorphemic words are represented with bigger circles to ease their localisation in the MST. The MST_{283} contained the paths between any pairs of words. By removing some of the edges of MST_{283}, clusters[7] can be formed. For example, the five word {公鹿, 雄性, 雌性, 雌, 公(A)} of the example detailed in Âğ 4.4.1 constitute one of the clusters. The mean size of the 30 clusters is 3 and 190 out of 283 words were classified as outliers.

Actually clusters to be efficiently used for meaning extraction, should be represented as subgraphs and not just as sets of words. In the latter case, clustering results are an impoverished representation of the whole knowledge embedded in the minimum spanning tree. The main reason is that the path structure - sequence of vertices to go from one word to another - is not present in clusters. However, considering the cluster {公鹿, 雄性, 雌性, 雌, 公(A)}, it is still possible to infer that the meaning of 公 in 公鹿 is represented by a common conceptual meaning of the three words (雄, 雌性, 雌).

[7] [10,8] showed that clustering from the minimum spanning tree is equivalent to single-linkage clustering.

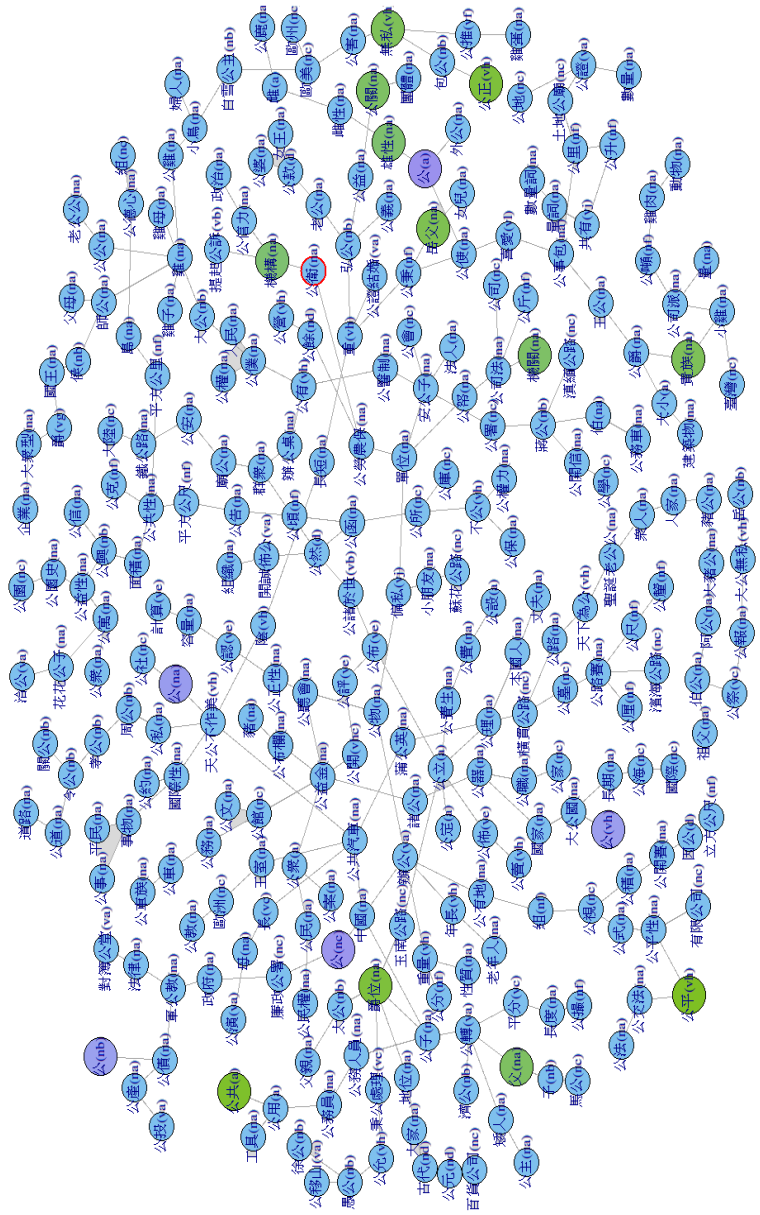


Fig. 1. MST_{283}

5 General Conclusion

Of the three minimum spanning trees, only the biggest - the one which embeds words from the dictionary's definition of the 人 words - can capture the meaning of 人 in 人 polymorphemic words in a way that is satisfactory for a native Chinese reader. In addition to capturing what appears for the observer to be semantic relationships, the edges of the minimum spanning trees can also embed situational relationships.

Finally, it is concluded that our approach is a first step in modeling partly representation and acquisition of the different meanings of Chinese polysemous morphemes. This work is thought to bring some insights to the Plato's problem and additional evidence towards the plausibility of words serving as ungrounded symbols[11,3]. More practically, this work could serve to add a new feature to current Chinese Wordnets: the listing of all the Chinese words embedding a same polysemous morpheme with a fixed identified meaning. Such a listing will help cognitive scientists studying the effects of repetitive exposure to Chinese polysemous morphemes embedded in compound words.

6 Future Directions

Firstly, we aimed at replicating that work using the Chinese Wikipedia instead of the Academica Sinica Balanced Corpus. The Chinese Wikipedia could reflect more adequately the representation of human knowledge as it has a semantic organization and its content and files structure follow categorization meaningful to human.

Secondly, we are presently investigating how to build minimum spanning trees satisfying constraints. For example, we aim at selectively build a MST which would warranty that a maximum of attractors words share edges with a 人 monomorphemic word and with a maximum of 人 words. Such a MST will serve to extract the meaning of a maximum of 人 words.

Finally, instead of using Latent Semantic Analysis to create the nested semantic spaces, we could use the following alternatives:

1. Fiedlar retrieval: [12] proposed that by considering the term-document matrix as a bipartite graph between the set of words and the set of documents, computing a set of the smallest eigenvalues of the Laplacian matrix of the bipartite graph, one can perform an enhanced kind of LSA analysis where unlikely to traditional LSA, documents and terms are considered equivalent and cohabiting in a same space.

2. Probabilistic models of semantic analysis: Latent Dirichlet Allocation (LDA) or Probabilistic LSA. They are probabilistic successors of LSA which have been found to outperform LSA[13,14,15] .

7 Acknowledgments

We thanked Iris Huang for her suggestions about the functional words to remove from the corpus and fruitful discussions and Train Min Chen for sharing subjects' data of her pilot study of a categorization task of 大 words. We also thanked Ingo Feinerer and Fridolin Wild from the R project, for their help in fixing some problems while we were creating and experimenting with our Chinese LSA spaces.

All of the data presented in this paper is freely available from the first author.

References

1. Chen, J.Y., Galmar, B., Su, H.J.: Semantic satiation of chinese characters in a continuous lexical decision task. In: The 21st Annual Convention of the Association For Psychological Science. (2009)
2. Galmar, B., Chen, J.Y.: Can neural adaptation occur at the semantic level? a study of semantic satiation. In: The 12th annual meeting of the Association for the Scientific Study of Consciousnesss. (2007)
3. Landauer, T., Dumais, S.: A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychological review **104**(2) (1997) 211–240
4. Landauer, T., McNamara, D., Dennis, S., Kintsch, W.: Handbook of latent semantic analysis. Lawrence Erlbaum (2007)
5. Foulds, L.: Graph theory applications. Springer (1995)
6. Gross, J., Yellen, J.: Graph theory and its applications. CRC press (2006)
7. Graham, R., Hell, P.: On the history of the minimum spanning tree problem. Annals of the History of Computing **7**(1) (1985) 43–57
8. Zahn, C.: Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on computers **20**(1) (1971) 68–86
9. Wu, B., Chao, K.: Spanning trees and optimization problems. Chapman & Hall (2004)
10. Gower, J., Ross, G.: Minimum spanning trees and single linkage cluster analysis. Journal of the Royal Statistical Society. Series C (Applied Statistics) **18**(1) (1969) 54–64
11. Glenberg, A., Robertson, D.: Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. Journal of Memory and Language **43**(3) (2000) 379–401

12. Hendrickson, B.: Latent semantic analysis and Fiedler retrieval. Linear Algebra and its Applications **421**(2-3) (2007) 345–355
13. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research **3** (2003) 993–1022
14. Blei, D., Griffiths, T., Jordan, M., Tenenbaum, J.: Hierarchical topic models and the nested Chinese restaurant process. Advances in neural information processing systems **16** (2004) 106
15. Blei, D., Lafferty, J.: A correlated topic model of science. Annals of Applied Statistics **1**(1) (2007) 17–35

BRUNO GALMAR
NATIONAL CHENG KUNG UNIVERSITY,
TAIWAN
E-MAIL: <HSUYUESHAN@GMAIL.COM>

JENN-YEU CHEN
NATIONAL CHENG KUNG UNIVERSITY,
TAIWAN
E-MAIL: <PSYJYC@MAIL.NCKU.EDU.TW>

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 169-182
RECEIVED 23/11/09 ACCEPTED 16/01/10 FINAL 09/03/10

Phrase-level Polarity Identification for Bangla

AMITAVA DAS AND SIVAJI BANDYOPADHYAY

Jadavpur University, India

ABSTRACT

In this paper, opinion polarity classification on news texts has been carried out for a less privileged language Bengali using Support Vector Machine (SVM)1. The present system identifies semantic orientation of an opinionated phrase as either positive or negative. The classification of text as either subjective or objective is clearly a precursor to determining the opinion orientation of evaluative text since objective text is not evaluative by definition. A rule based subjectivity classifier has been used. The present system is a hybrid approach to the problem, works with lexicon entities and linguistic syntactic feature. Evaluation results have demonstrated a precision of 70.04% and a recall of 63.02%.

Keywords: Opinion Mining, Polarity Identification, Bengali and Phrase Level Polarity Identification.

1 INTRODUCTION

Emotion recognition from text is a new subarea of Natural Language Processing (NLP) and has drawn considerable attention of the NLP researchers in recent times. Several subtasks can be identified within opinion mining; all of them involve tagging at document/sentence/phrase/word level according to expressed opinion. One such subtask is based on a given opinionated piece of text on one single issue or item, to classify the opinion as falling under one of two

[1] http://chasen.org/~taku/software/TinySVM/

opposing sentiment polarities, or locate its position in the continuum between these two polarities. A large portion of work in sentiment-related classification/regression/ranking falls within this category. The binary classification task of labeling an opinionated document as expressing either an overall positive or an overall negative opinion is called sentiment polarity classification or polarity classification. Much work on sentiment polarity classification has been conducted in the context of reviews (e.g., "thumbs up" or "thumbs down" for movie reviews) [2]. While in this context "positive" and "negative" opinions are often evaluative (e.g., "like" vs. "dislike"), there are other problems where the interpretation of "positive" and "negative" is subtly different. But development of a complete opinion mining system needs an automatic subjectivity detection module (it is a classification module that can differentiate among subjective or objective texts) followed by polarity classifier. Assuming that all texts are opinionated may cause the system development easier but the resultant system will be unable to meet real life goal. Very little attempt could be found in literature to develop a complete opinion mining system. Rather people concentrate on specific sub problems. The present system has been developed on news corpus which is more generic than review corpus. The system evaluation has shown the precision and recall values are 70.04% and 63.02% for Bengali respectively.

In this paper, a complete opinion mining system is described that can identify subjective sentences within a document and an efficient feature based automatic opinion polarity detection algorithm to identify polarity of phrases. Related works are described in Section 2. Resource acquisition has been discussed in Section 3. The feature extraction technique has been described in Section 4. Conclusion has been drawn in Section 6.

2 RELATED WORKS

"What other people think" has always been an important piece of information for most of us during any decision-making process. An opinion could be defined as a private state that is not open to objective observation or verification [3]. Opinion extraction, opinion summarization and opinion tracking are three important techniques for understanding opinions. Opinion-mining of product reviews, travel advice, consumer complaints, stock market predictions, real estate

market predictions, e-mail etc. are areas of interest for researchers since last few decades.

Most research on opinion analysis has focused on sentiment analysis [4], subjectivity detection ([5], [6], [7],[8]), review mining [9], customer feedback [10] and strength of document orientation [11]. Methods on the extraction of opinionated sentences in a structured form can be found in [12]. Some machine learning text labeling algorithms like Conditional Random Field (CRF) ([13],[14]), Support Vector Machine (SVM) [15] have been used to cluster same type of opinions. Application of machine-learning techniques to any NLP task needs a large amount of data. It is time-consuming and expensive to hand-label the large amounts of training data necessary for good performance. Hence, use of machine learning techniques to extract opinions in any new language may not be an acceptable solution.

Opinion analysis of news document is an interesting area to explore. Newspapers generally attempt to present the news objectively, but textual affect analysis in news documents shows that many words carry positive or negative emotional charge [16]. Some important works on opinion analysis in the newspaper domain are [17], [18] and [19], but no such efforts have been taken up in Indian languages especially in Bengali.

Various opinion mining methods have been reported that use lexical resources like WordNet [20], SentiWordNet [21] and ConceptNet [22] etc.

3 RESOURCE ACQUISITION

To start opinion mining task for a new language demands sentiment lexicon and gold standard annotated data for machine learning and evaluation. The detail of resource acquisition process for annotated data, subjectivity classifier, sentiment lexicon and the dependency parser are mentioned below.

3.1 Data

Bengali is the fifth popular language in the World, second in India and the national language in Bangladesh. Automatic opinion mining or sentiment analysis task mainly concentrated on English language till date. Bengali is a less computational privileged language. Hence Bengali corpus acquisition is an essential task for any NLP system

development. For the present task Bengali news corpus has been identified. News text can be divided into two main types: (1) news reports that aim to objectively present factual information, and (2) opinionated articles that clearly present authors' and readers' views, evaluation or judgment about some specific events or persons. Type (1) is supposed to be the common practice in newspapers, and Type (2) appears in sections such as 'Editorial', 'Forum' and 'Letters to the editor'. 'Reader's opinion' section or 'Letters to the Editor Section' from the web archive of a popular Bengali newspaper are identified as the relevant corpus in Bengali. A brief statistics about the corpus are reported in the Table 1. The corpus is then manually annotated and used for training and testing respectively. Detailed reports about this news corpus development in Bengali can be found in [23]. The annotation scheme that has been used to annotate the corpus is reported in Table 2. The positive algebraic sign in the feature structure ("<fs af=,+,") depict the phrase polarity as positive.

Table 1. Bengali News Corpus Statistics

Total number of documents in the corpus	20
Total number of sentences in the corpus	447
Avgerage number of sentences in a document	22
Total number of wordforms in the corpus	5761
Avgerage number of wordforms in a document	288
Total number of distinct wordforms in the corpus	3435

Table 2. Bengali News Corpus Annotation Scheme

2	((	CCP	
2.1	যেমন	CC	
	))		
3	((	NP	<fs af=',+,,,,,,' name='?'>
3.1	মঙ্গলজনক	NN	
3.2	,	SYM	
	s))		

3.2 Subjectivity Classifier

The subjectivity classifier as described in [1] has been used. The resources used by the classifier are sentiment lexicon, Theme clusters and POS tag labels.

The classifier first marks sentences bearing opinionated words. In the next stage the classifier marks theme cluster specific phrases in each sentence. If any sentence includes opinionated words and theme phrases then the sentence is definitely considered as subjective. In the absence of theme words, sentences are searched for the presence of at least one strong subjective word or more than one weak subjective word for its consideration as a subjective sentence. The recall measure of the present classifier is greater than its precision value. The evaluation results of the classifier are 72.16% (Precision) on the NEWS Corpus.

The corpus is then validated by a human annotator and is effectively used during training and testing of the polarity classifier.

3.3 Sentiment Lexicon

A typical approach to sentiment analysis is to start with a lexicon of positive and negative words and phrases. In these lexicons, entries are tagged with their prior polarity: out of context, does the word seem to evoke something positive or something negative. For example, happy has a positive prior polarity, and sorrow has a negative prior polarity. However, the contextual polarity of a phrase in which a word appears may be different from the word's prior polarity. There are two main lexical resources widely used in English: SentiWordNet [21] and Subjectivity Word List [24] for Subjectivity Detection. SentiWordNet is an automatically constructed lexical resource for English which assigns a positivity score and a negativity score to each WordNet synset. Positivity and negativity orientation scores range within 0 to 1. Release 1.1 of SentiWordNet for English was obtained from the authors of the same. The subjectivity lexicon was compiled from manually developed resources augmented with entries learned from corpora. The entries in the subjectivity lexicon have been labeled for part of speech as well as either strong subjective or weak subjective depending on reliability of the subjective nature of the entry.

A word level translation process followed by error reduction technique has been used for generating the Bengali Subjectivity lexicon from English.

A subset of 8,427 opinionated words has been extracted from SentiWordNet, by selecting those whose orientation strength is above the heuristically identified threshold of 0.4. The words whose orientation strength is below 0.4 are ambiguous and may lose their subjectivity in the target language after translation. A total of 2652

words are discarded [24] from the Subjectivity word list as they are labeled as weakly subjective.

For the present task, a English-Bengali dictionary (approximately 102119 entries) developed using the Samsad Bengali-English dictionary[2] has been chosen. A word level lexical-transfer technique is applied to each entry of SentiWordNet and Subjectivity word list. Each dictionary search produces a set of Bengali words for a particular English word. The set of Bengali words for an English word has been separated into multiple entries to keep the subsequent search process faster. The positive and negative opinion scores for the Bengali words are copied from their English equivalents. This process has resulted in 35,805 Bengali entries.

3.4 Dependency Parser

Dependency feature in opinion mining task has been first introduced by [25]. This feature is very useful to identify intra-chunk polarity relationship. It is very often a language phenomenon that modifiers or negation words are generally placed at a distance with evaluative polarity phrases. But unfortunately dependency parser for Bengali is not freely available. In this section we describe the development of a basic dependency parser for Bengali language.

The probabilistic sequence models, which allow integrating uncertainty over multiple, interdependent classifications and collectively determine the most likely global assignment, may be used in a parser. A standard model, Conditional Random Field (CRF)[3], has been used. The tag set that has been used here is same as NLP Tool Contest in ICON 2009[4]. The input file in the Shakti Standard Format (SSF)[5] includes the POS tags, Chunk labels and morphology information. The chunk information in the input files are converted to B-I-E format so that the begin (B) / inside (I) / End (E) information for a chunk are associated as a feature with the appropriate words. The chunk tags in the B-I-E format of the chunk with which a particular chunk is related through a dependency relation are identified from the training file and noted as an input feature in the CRF based system. The corresponding relation name is also another input feature associated

[2] http://dsal.uchicago.edu/dictionaries/biswas_bengali/
[3] http://crfpp.sourceforge.net
[4] http://ltrc.iiit.ac.in/icon2009/nlptools.php
[5] http://www.docstoc.com/docs/7232788/SSF-Shakti-Standard-Format-Guide

with the particular chunk. Each sentence is represented as a feature vector for the CRF based machine learning task. After a series of experiments the following feature set is found to be performing well as a dependency clue. The input features associated with each word in the training set are the root word, pos tag, chunk tag and vibhakti.

Root Word: Some dependency relations are difficult to identify without the word itself. It is better to come with some example.

AjakAla NN NP X k7t

In the previous example, there is no clue except the word itself. The word itself is noun, chunk level denotes a noun phrase and there is no vibhakti attached to the word. For these cases, word lists of temporal words, locations names and person names have been used for disambiguation [26]. Specifically identification of k7t relation is very tough because the word itself will be a common noun or a proper noun but the information of whether the word denotes a time or a location helps in the disambiguation.

Part of Speech: Part of speech of a word always plays a crucial role to identify dependency relation. For example dependency relations like k1 and k2 in most of the cases involve a noun. It has been observed through experiments that not only POS tag of present word but POS tags of the context words (previous and next) are useful in identifying the dependency relation in which a word takes part.

Chunk label: Chunk label is the smallest accountable unit for detection of dependency relations and it is an important feature. But during the training sentences are parsed into word level, hence chunk label are associated to the appropriate words with the labels as B-X (beginning), I-X (Intermediate) and E-X (End) (where X is the chunk label).

Vibhakti: Indian languages are mostly non-configurational and highly inflectional. Grammatical functions (GFs) are predicted by case inflections (markers) on the head nouns of noun phrases (NPs) and postpositional particles in postpositional phrases (PPs). In the following example the '0_janya' vibhakti inflection of the word "pAoyZAra" leads to rh (Hetu - causal) case inflections. However, in many cases the mapping from case marker to GF is not one-to-one.

4 FEATURES EXTRACTION

SVM treats opinion polarity identification as a sequence tagging task. SVM views the problem as a pattern-matching task, acquiring symbolic patterns that rely on both the syntax and lexical semantics of a phrase. We hypothesize that a combination of the two techniques would perform better than either one alone. With these properties in mind, we define the following features for each word in an input sentence. For pedagogical reasons, we may describe some of the features as being multi-valued (e.g. stemming cluster) or categorical (e.g. POS category) features. In practice, however, all features are binary for the SVM model. In order to identify features we started with Part Of Speech (POS) categories and continued the exploration with the other features like chunk, functional word, SentiWordNet in Bengali[1], stemming cluster, Negative word list and Dependency tree feature. The feature extraction pattern for any Machine Learning task is crucial since proper identification of the entire features directly affect the performance of the system. Functional word, SentiWordNet (Bengali) and Negative word list is fully dictionary based. On the other hand, POS, chunk, stemming cluster and dependency tree features are extractive. Classifying polarity of opinionated texts either at the document/sentence or phrase level is difficult in many ways. A positive opinionated document on a particular object does not mean that the author has positive opinions on all aspects. Likewise, a negative opinionated document does not mean that the author dislikes everything. In a typical opinionated text, the author writes both positive and negative aspects of the object, although the general sentiment on the object may be positive or negative. Document-level and sentence-level classification does not provide such information. To obtain such details, there is a need to go to the object feature level.

4.1 Part Of Speech (POS)

Number of research activities like [6], [27] etc. have proved that opinion bearing words in sentences are mainly adjective, adverb, noun and verbs. Many opinion mining tasks, like the one presented in [28], are mostly based on adjective words.

4.2 Chunk

Chunk level information is effectively used as a feature in supervised classifier. Chunk labels are defined as B-X (Beginning), I-X (Intermediate) and E-X (End), where X is the chunk label. It has been noted that it is not unusual for two annotators to identify the same expression as a polar element in the text, but they could differ in how they mark the boundaries, such as the difference between 'such a disadvantageous situation' and 'such…disadvantageous' (Wilson and Wiebe, 2003). Similar fuzziness appeared in our marking of polar elements, such as 'কেন্দ্রীয় দলের দূর্নীতিতে' (corruption of central team) and 'দূর্নীতিতে' (corruption). Hence the hypothesis is to stick to chunk labels to avoid any further disambiguation. A detailed empirical study reveals that polarity clue may be defined in terms of chunk tags.

4.3 Functional word

Function words in a language are high frequency words and these words generally do not carry any opinionated information. But function words help many times to understand syntactic pattern of an opinionated sentence. A list of 253 entries is collected from the Bengali corpus. First a unique high frequency word list is generated where the assumed threshold frequency is considered as 20. The list is manually corrected keeping in mind that a word should not carry any opinionated or sentiment feature.

4.4 SentiWordNet

Words that are present in the SentiWordNet carry opinion information. The developed Sentiment Lexicon is used as an important feature during the learning process. These features are individual sentiment words or word n-grams (multiword entities) with polarity values either positive or negative. Positive and negative polarity measures are treated as a binary feature in the supervised classifier. Words which are collected directly from SentiWordNet are tagged with positivity or negativity score.

4.5 Stemming cluster

Several words in a sentence that carry opinion information may be present in inflected forms. Stemming is necessary for such inflected words before they can be searched in appropriate lists. Due to non availability of good stemmers in Indian languages especially in Bengali, a stemmer based on stemming cluster technique has been evolved. This stemmer analyzes prefixes and suffixes of all the word forms present in a particular document. Words that are identified to have same root form are grouped in a finite number of clusters with the identified root word as cluster center. Details could be found in [30].

4.6 Negative words

Negative words like no (না), not (নয়) etc does not carry any opinion information but those relationally affect the resultant polarity of any polar phrase. A manually generated list has been prepared and used as a binary feature in the SVM classifier.

4.7 Dependency tree feature

Dependency feature has been successfully used here to identify modifier relationship of any polar phrase within a sentence. The analysis of Bengali corpora reveals that people generally use negation words/modifiers with any positive polar phrases. As an example

সে আদৌ ভালো নয় (He is not good enough)

The feature extractor module searches the dependency tree using breadth-first search to identify syntactically related nodes. The purpose of the feature is to encode dependency structure between related polar phrases.

5 EVALUATION

The evaluation result of the SVM-based polarity classification task for Bengali is presented in Table 3. The evaluation result of the system for each polarity class i.e., positive and negative are mentioned separately in the table 4.

Table 3. Results of Polarity classification.

Language	Domain	Precision	Recall	
Bengali	NEWS	70.04%	63.02%	

Table 4. Polarity wise System Evaluation.

Polarity	Precision	Recall	
Positive	56.59%	52.89%	
Negative	75.57%	65.87%	

6 CONCLUSION

One limitation of log-linear function models like SVM is that they cannot form a decision boundary from conjunctions of existing features, unless conjunctions are explicitly given as part of the feature vector. To maintain the granularity, features are explicitly mentioned as a classical word lattice model. A post-processor finally assigns the polarity value to the chunk head depending upon the chunk head's resultant polarity. We are now working on improving the performance of the present system. Future task will be in the direction of development techniques for creation of opinion summaries according to their polarity classes.

REFERENCES

1. Amitava Das and Sivaji Bandyopadhyay. Theme Detection an Exploration of Opinion Subjectivity. In Proceeding of Affective Computing & Intelligent Interaction (ACII 2009).
2. Peter Turney, Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In Proceeding of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics.
3. Randolph Quirk, Sidney Greenbaum, Geoffry Leech and Jan Svartvik. A comprehensive Grammar of the English Language. Longman, New York. (1985)
4. Tomohiro Fukuhara, Hiroshi Nakagawa and Toyoaki Nishida. Understanding sentiment of people from news articles: Temporal sentiment analysis of social events. Proceedings of the International Conference on Weblogs and Social Media (ICWSM), 2007.

5. Baroni M and Vegnaduzzo S. Identifying subjective adjectives through web-based mutual information. Proceedings of Konvens, pages 17-24, 2004.
6. Vasileios Hatzivassiloglou and Janyce Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In Proceedings of the International Conference on Computational Linguistics (COLING), 2000.
7. Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the Association for Computational Linguistics (ACL), pages 271-278, 2004.
8. Soo-Min Kim and Eduard Hovy. Automatic detection of opinion bearing words and sentences. In Companion Volume to the Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP), 2005.
9. Hu and Liu. Mining and summarizing product reviews. Proceedings of 10th ACM SigKDD, 2004.
10. Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. Proceedings of the International Conference on Computational Linguistics (COLING), 2004.
11. Esuli A and Sebastini F. Determining the semantic orientation of terms through gloss analysis. Proceedings of CIKM, 2005.
12. Nozomi Kobayashi, Kentaro Inui and Yuji Matsumoto. Opinion Mining from Web documents: Extraction and Structurization. Journal of Japanese society for artificial intelligence, Vol.22 No.2, special issue on data mining and statistical science, pages 227-238, 2007.
13. Yejin Choi, Clarie Cardie, Ellen Riloff and Siddharth Patwardhan. Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), pages 355-362, 2005.
14. Andrew Smith, Trevor Cohn and Miles Osborne. Logarithmic Opinion Pools for Conditional Random Fields. In Proceeding of the 43rd Annual Meeting of the ACL, pages 18-25, 2005.
15. Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 412-418, 2004.
16. Gregory Grefenstette, Yan Qu, James G. Shanahan and David A. Evans. Recherché d'Information Assistée par Ordinateur. In Proceedings of RIAO, 7th International Conference on 2004.
17. S. Argamon-Engelson, M. Koppel, and G. Avneri. Style-based text categorization: What newspaper am I reading?. In Proceedings of the AAAI Workshop on Text Categorization, pages 1-4, 1998.
18. L.-W. Ku, Y.-T. Liang, and H.-H. Chen. Opinion extraction, summarization and tracking in news and blog corpora. In Proceeding of

AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 100-107, 2006.
19. A. Stepinski and V. Mittal. A fact/opinion classifier for news articles. In Proceedings of the ACM Special Interest Group on Information Retrieval (SIGIR), pages 807-808, 2007.
20. A. Esuli and F. Sebastiani. Page Ranking WordNet synsets: An application to opinion mining. In Proceedings of the Association for Computational Linguistics (ACL), 2007.
21. Esuli and Sebastini. Sentiwordnet: a publicly available resource for opinion Genova, Italy. 2006.
22. Nathan Eagle, Push Singh and Alex (Sandy) Pentland .Common sense conversations: understanding casual conversation using a common sense database. In Proceedings of the Artificial Intelligence, Information Access, and Mobile Computing Workshop (IJCAI 2003).
23. Ekbal, A., Bandyopadhyay. S. A Web-based Bengali News Corpus for Named Entity Recognition. Language Resources and Evaluation Journal. pages 173-182, 2008.
24. Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In Proceedings of CICLing 2005 (invited paper).
25. Choi, Yejin and Cardie, Claire and Riloff, Ellen and Patwardhan, Siddharth, Identifying Sources of Opinions with Conditional Random Fields and Extraction Patterns. In Proceedings of HLT-EMNLP-05, the Human Language Technology Conference/Conference on Empirical Methods in Natural Language Processing. Pages 355-362, 2005.
26. Asif Ekbal, Rejwanul Haque, Amitava Das, Venkateswarlu Poka and Sivaji Bandyopadhyay. 2008. Language Independent Named Entity Recognition in Indian Languages. In Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages. pages 33-40.
27. Paula Chesley, Bruce Vincent, Li Xu, and Rohini Srihari. Using verbs and adjectives to automatically classify blog sentiment. In AAAI Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW), pages 27-29, 2006.
28. Tetsuya Nasukawa and Jeonghee Yi. Sentiment analysis: Capturing favorability using natural language processing. In Proceedings of the Conference on Knowledge Capture (K-CAP), pages 70-77, 2003.
29. Amitava Das and Sivaji Bandyopadhyay, Subjectivity Detection in English and Bengali: A CRF-based Approach. In ICON 2009.

AMITAVA DAS
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JADAVPUR UNIVERSITY, KOLKATA 700032, INDIA
E-MAIL: <AMITAVA.SANTU@GMAIL.COM>

SIVAJI BANDYOPADHYAY
DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
JADAVPUR UNIVERSITY, KOLKATA 700032, INDIA
E-MAIL: <SIVAJI_CSE_JU@YAHOO.COM>

Machine Translation and Multilinguism

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 185-199
RECEIVED 23/11/09 ACCEPTED 16/01/10 FINAL 10/03/10

Exploiting Charts in the MT Between Related Languages

PETR HOMOLA, VLADISLAV KUBOŇ

Institute of Formal and Applied Linguistics, Czech Rep.

ABSTRACT

The paper describes in detail the exploitation of chart-based methods and data structures in a simple system for the machine translation between related languages. The multigraphs used for the representation of ambiguous partial results in various stages of the processing and a shallow syntactic chart parser enable a modification of a simplistic and straightforward architecture developed ten years ago for MT experiments between Slavic languages in the system Česílko. The number of translation variants provided by the system inspired an addition of a stochastic ranker whose aim is to select the best translations according to a target language model.

1 INTRODUCTION

Using graphs has a long tradition in the field of machine translation (MT). It is very difficult to trace back the first attempts to represent some linguistic phenomena by means of charts, but it is not difficult to find a clear historical example of usefulness of such representation. This example is probably the most famous MT system of all times, the first really successful and commercially exploited system, METEO [1,2]. There were many reasons why METEO worked so well that it served for decades as a positive example for the whole MT community demonstrating that machine translation is possible after all. The formalism used in the system, Colmerauer's Q-systems [3], is definitely among those reasons.

Q-systems are in fact a mechanism for transformation of trees which label the edges of an oriented chart. The transformations are controlled by a grammar which contains declarative rewriting rules. Each rule may be applied to a continuous set of edges and the result of its application is a new edge or a continuous set of new edges starting and ending in the same nodes as the original sequence. The grammar may be divided into more parts which constitute a sequence in which the output of a previous phase (in the form of a chart) serves as an input of the subsequent one. At the end of each phase the system deletes all edges which were used on a left hand side of some rule and the edges which do not constitute a part of a path leading from the starting to the final node. This mechanism thus very naturally cleans all partial results and at the same time it allows to maintain ambiguity whenever it is necessary in between two particular phases.

2 CHARTS IN THE MT BETWEEN RELATED LANGUAGES

Apart from METEO, Q-systems were used as well in one of the first systems of MT between related languages, in the Czech-to-Russian MT system RUSLAN [4]. The ability of the chart-based analyzers to deal with ambiguities at various levels (morphology, syntax, semantics) and to preserve them across the levels (certain morphological ambiguities cannot be resolved without syntactic clues) was fully exploited in this MT system.The system used a traditional transfer-based architecture with full-fledged syntactic analysis involving even some semantics. The relatedness of both languages was not reflected in the architecture of the system.

The last decade witnessed a growing interest in MT between related languages for different language groups—Slavic [5,6], Scandinavian [7], Turkic [8], and languages of Spain [9]. The main advantage of translating between related languages is the possibility to use much simpler means, in most cases some kind of “shallow” methods, most prominently in parsing or in transfer. This is actually the case of experiments for Slavic languages and the languages of Spain, where both systems follow a very simple architecture originally designed for the Czech-to-Slovak system Česílko. A morphological tagger disambiguates the input, individual lemmas and tags are translated and transfered into a target language and a morphological synthesis creates a target language sentence. This rather simplistic approach chosen both in the system Česílko and Apertium has a substantial drawback in the fact that the morphological and lexical ambiguity is solved early in the translation process with all

the consequences—the taggers used are still not sufficiently precise (the best taggers for highly inflected languages quite naturally still have precision inferior to their counterparts for English) and thus they introduce translation errors which cannot be removed in the subsequent stages of the translation process. The architecture also does not allow to cope with lexical ambiguity, another source of frequent translation errors.

In the following sections we would like to describe in detail how the exploitation of chart-based methods may improve the MT between closely related languages. The description will concern all important processing stages of the system: morphological analysis, shallow syntactic analysis and transfer. The experiments are conducted on a group of Slavic languages with Czech as a source language and Slovak as a primary target language.

3 CHARTS IN MORPHOLOGY

As mentioned above, the simplistic architecture of Česílko exploits a morphological tagger for a (complete) disambiguation of ambiguous word-forms. In our experiments we have decided to replace the tagger by a shallow syntactic chart parser which helps to (partially) disambiguate the ambiguous input on the basis of the local context and, at the same time, it preserves those ambiguous variants which cannot be resolved in such a way. In order to keep the ambiguities wherever necessary, our system uses a multigraph (i.e., a graph allowing parallel edges between a pair of nodes).

In morphology, the advantage of the multigraph is obvious especially for highly inflected languages. Individual word-forms are very often ambiguous with regard to the gender, number and case and the possibility to keep all the variants as long as necessary (until the ambiguity is resolved in later stages of the processing), is really an important advantage.

The use of a multigraph in a chart parser also has certain hidden drawbacks which have to be handled by workarounds or tricks. Let us discuss the most crucial issue.

Let us consider the Czech sentence *Starý hrad se tyčí nad řekou* "The old castle towers over the river". The phrase *starý hrad* is morphologically ambiguous (both forms can be used in both nominative and accusative case). After this phrase has been recognized as the subject of the main verb, we know that the case is nominative in this context. And since there is no other reading where it would be accusative, the parser can remove this wrong reading.

But what would have happened if we had the isolated phrase *staré hrady* "old castles"? There would be again two possible readings (nominative and accusative) which cannot be resolved due to the lack of context. Nevertheless there are still other meanings for each of the words independently (disregarding the dependence between them). In this case, these edges will not be removed during the final cleaning of edges although the parser has analyzed the whole phrase. We can use a simple workaround in this case: we can insert a new edge (*shackle*) between edges which represent two word forms of the input sentence. These artificial edges will link both clusters of edges representing different morphological readings. If there is at least one analysis which connects both words, the parser will remove the shackle during the cleaning phrase and thus only the complete parse will be preserved for further processing because the 'false' edges will not lie on a valid path any more and will be deleted as well (the adjective would have more morphological meanings; for the sake of simplicity, the multigraph contains only one edge with different gender).

It is obvious that if we modify the multigraph by adding 'shackles' between all edges labelled with morphological information about individual input words we also have to modify all grammar rules accordingly.

4 CHARTS IN SYNTAX

In this section we would like to discuss typical issues of exploiting the chart parser in a syntactic analysis. One of the most important issues which may substantially reduce the parsing efficiency of chart parsers is their natural tendency to create redundant identical results.

4.1 *Elimination of identical results*

The application of grammar rules to the multigraph is non-deterministic, the rules are being applied in an order which may look very close to random. As a result, the application of several different sequences of rules may lead to identical results, as illustrated in Figure 1:

There are two possible parses:

1. The rule identifying direct objects is applied first, the rule identifying subjects is applied afterwards.
2. The rule identifying subjects is applied first, the rule identifying direct objects is applied afterwards.

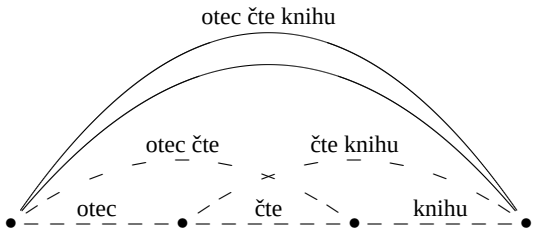


Fig. 1. Example of duplicate parses of a sentence

Theoretically, we would get two edges spanning the whole sentence and labelled with identical dependency trees (of course, if we adhere to constituent trees, both structures will reflect the order of application of grammar rules and they will be different, but let us not forget that we are primarily talking about the MT system between Slavic languages where the use of dependency notation has a long tradition). In our implementation of the parser, this kind of duplicity is recognized automatically to avoid exponential explosion.

4.2 *Multigraph clean-up and further optimization*

As long as a rule can be applied to the multigraph, edges are added to it but no existing edge is removed. The new edges represent (are labelled with) intermediary feature structures that may be used in further parsing or they may be candidates for the final result. Once the multigraph cannot be extended by any rule (according to a particular grammar), the intermediary edges need to be discarded from the multigraph since we want only the most complex feature structures to be processed in the transfer phase. This clean-up is somewhat similar to garbage collection in programming languages with automatic memory management.

As an example, let us consider the following Czech verb phrase as the input of the parser:

(1) *auta* *jezdila*
cars-NEUT,NOM,PL move-PAST,NEUT,PL
"The cars moved/were moving."

The input of the parser is the morphologically preprocessed multigraph (the multisets of edges between the same pair of nodes reflect the morphological ambiguity of a word form), which can be found in the upper part of Figure 2.

After the application of one particular rule, namely the one that attaches a noun in nominative (the subject) to its predicate (a resultative participle in this case), we will get the multigraph from the lower part of Figure 2 as the result of the syntactic analysis (dotted lines denote used edges, circles denote used nodes[1]).

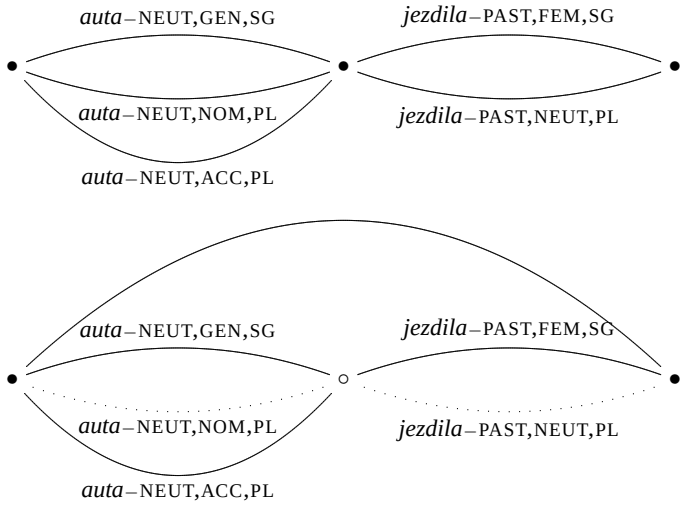


Fig. 2. The input and the result of the syntactic analysis

Now we need to get rid of all obsolete edges:

1. First of all, we remove all used edges (denoted by dotted lines).

[1] We define used node as a node that has at least one used edge to the left and at least one used edge to the right.

2. We remove all edges which start or end in a used node (i.e., the edges that reflect morphological variants of a used edge which are morphologically misanalyzed in the given context according to the used grammar).
3. For each path p from the initial node to the end node, we calculate the number $u(p)$ of used edges it contains. Then we assign each edge e the score $s(e) = \min_{p \in P} u(p)$. The score for the whole graph is defined as $s = \min_{e \in E} s(e)$. Finally, we remove all edges where $s(e) > s$.[2]

The last step ensures that every edge which remains in the multigraph lies on a path from the initial node to the end node. The resulting graph represents the output of the module of shallow syntactic analysis and as such it is passed to the subsequent module which is the transfer. At the same time, all complex feature structures (that represent syntactic trees) that label the edges of the multigraph are being syntactically synthesized.

Processing of long sentences may result in very large multigraphs with the number of edges growing exponentially. If we had to translate the Russian phrase *старый замок* "old castle" into Czech, the transfer would give the two features structures from Figure 3.

$$\begin{bmatrix} \text{"замок"} & \\ \text{ADJ} & [\ \text{"старый"}\ ] \end{bmatrix} \rightarrow \left\{ \begin{bmatrix} \text{"hrad"} & \\ \text{ADJ} & [\ \text{"starý"}\ ] \end{bmatrix}, \begin{bmatrix} \text{"zámek"} & \\ \text{ADJ} & [\ \text{"starý"}\ ] \end{bmatrix} \right\}$$

Fig. 3. Lexical transfer of feature structures

The syntactically synthesized multigraph is shown in Figure 4.

As the two edges with the feature structure for the adjective *starý* are identical, we can optimize the spatial complexity of the multigraph by contracting identical edges that have at least one common node. We call this process *compacting* the multigraph. It is obvious that in complex multigraphs, the number of edges can be lowered significantly. Immediately before morphological synthesis, the optimization can be even more efficient if we do not contract only edges with identical feature structures

[2] If there is at least one path from the initial node to the end node consisting only from unused edges then the algorithm is equal to the one described in [3], i.e., all used edges are deleted as well as edges that do not belong to a path from the initial node to the end node.

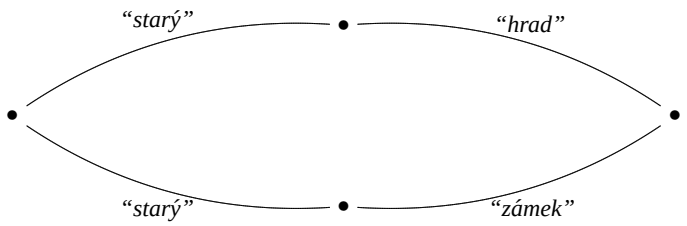


Fig. 4. The result of a transfer and corresponding feature structures

but also those with identical surface form in the target language (there is an extensive syncretism in Slavic languages).

5 TRANSFER AND SYNTACTIC SYNTHESIS

Transfer and syntactic synthesis are performed jointly in one module. The task of the transfer module is to adapt complex structures created by the parser which cover the whole source sentence continuously to the target language lexically, morphologically and syntactically. In the following sections we describe the phase of the lexical transfer and the structural transfer, the latter being split further in structural preprocessor and syntactic decomposer.

5.1 *Lexical transfer*

The aim of the lexical transfer is to ‘translate a feature structure lexically’, i.e., the lemmas associated with feature structures are translated. Morphological features may be adapted as well wherever appropriate.

In order to demonstrate the nature of the data contained in the dictionary, let us present a fragment of the dictionary used in lexical transfer between Czech and Slovenian:

Example 1.
```
hvězda|zvezda
dodat|dodati
kůň|konj
strom|drevo|gender=neut;
```

Let us have a brief look at the last line of the example. The Czech noun *strom* "tree" is in masculine gender while the gender of its Slovenian counterpart *drevo* is neuter, that is why there is the additional information *gender=neut* which instructs the transfer module to adapt the feature *gender* of the corresponding feature structure so that it can be correctly synthesized morphologically.

5.2 *Structural transfer*

The task of the structural transfer is to adapt the feature structures of the source language (their properties and mutual relationship) so that the synthesis generates a grammatically well-formed sentence with the meaning of the source sentence. It is necessary to admit that the well-formedness can generally be guaranteed only locally for the part of the sentence the feature structure covers (this is caused by the decision to exploit shallow parsing instead of a full-fledged one).

When changing the structure, the transfer may do one of the following actions:

- to change values of atomic features in the feature structure, to add atomic features with a specific value or to delete some atomic features;
- to add a node to the syntactic tree;
- to remove a node from the syntactic tree.

5.3 *Translation of multiword expressions*

It is a well known fact that some words of a source language are translated as multiword expressions in the target language and vice versa, for example:

Example 2. *babička* "grandmother" (Cze) → *stará mama* (Slv)
zahradní jahoda "garden strawberry" (Cze) → *truskawka* (Pol)

Since these cases require removing or adding of a subordinated feature structure (for the adjective) which is equivalent to removing or adding a node from/to the syntactic tree, such cases are handled by special rules in the structural transfer.

6 RANKER

As shown in the previous sections, the multigraph is a very usefull data structure allowing to keep multiple variants of a translated sentence through all stages of automatic translation. In order to fully exploit this property, it is necessary to add a module which would be able to select the best translation from all variants contained in the multigraph. For this purpose we have decided to apply a stochastic ranker. The advantage of using a ranker instead of a tagger is obvious—apart from inserting errors caused by the imperfection of the tagger into the input sentence the tagger also disambiguates the input too early and makes the translation process too straightforward. If we replace it with the ranker we are able to propagate more translation candidates through the system.

The reason why we are using a stochastic module in a system which relies a lot on hand-written rules is pretty obvious—it would be very complicated (if possible at all) to resolve the degree of ambiguity preserved in the multigraph by hand-written rules. The stochastic post-processor is able to select one particular sentence that suits best the given context.

6.1 *Ranking*

We use a simple language model based on trigrams (trained on word forms without any morphological annotation) which is intended to sort out "wrong" target sentences (these include grammatically ill-formed sentences as well as inappropriate lexical mapping). For example, the language model for Slovak has been trained on a corpus of 18.8 million words which have been randomly chosen from the Slovak Wikipedia[3].

Let us present an example of how this component of the system works. Let us suppose thet there is the following Czech segment (matrix sentence) in the source text:

Example 3. *Společnost* *ve* *zprávě*
company-FEM,SG,NOM in report-FEM,SG,LOC
uvedla
inform-LPART,FEM,SG
"The company informed in the report…"

The rule-based part of the system is supposed to generate (the shallow grammar contains no rules for VPs) four Slovak (target) segments that collapse to the following two after morphological synthesis:

[3] http://sk.wikipedia.org

1. *Spoločnosť vo správe uviedli*,
2. *Spoločnosť vo správe uviedla*.

The Czech word *uvedla* is ambiguous (fem.sg and neu.pl). According to the language model, the ranker will choose the second sentence as the most probable result.

There are also many homonymic word forms that result in different lemmas in the target languages. For example, the Czech word *pak* means both "then" and "fool-pl.gen", the word *tři* means "three" and the imperative of "to scrub", *ženu* means "wife-sg.acc" and "(I'm) hurrying out" etc. The ranker is supposed to sort out the contextually wrong meaning in all these cases if it has not been resolved by the parser.

6.2 *Evaluation*

We have evaluated the system of the Czech-to-Slovak MT on hundreds of sentences mainly from newspapers. The metrics we are using is the Levenshtein edit distance between the automatic translation and a reference translation. The reason why we do not use some more standard evaluation metric such as BLEU [10] is simple—there is no sufficiently large set of good quality testing data which would contain multiple translations of each particular source sentence into Slovak. As it has already been shown in several articles (e.g. [11]), the correlation of BLEU with the human judgment is not as high as it was generally believed. On top of that, the reliability of BLEU decreases significantly if only a single reference translation is used. The edit distance has one more advantage—while the BLEU score does not provide any clue how complicated is the post-editing of the result, the Levenshtein metric is pretty straightforward in this respect and thus it is more suitable for really practical evaluation of the MT output.

There are three basic possibilities of the outcome of translation of a segment.

1. The rule-based part of the system has generated a 'perfect'[4] translation (among other hypotheses) and the ranker has chosen it.
2. The rule-based part of the system has generated a 'perfect' translation but the ranker has chosen another one.
3. All translations generated by the rule-based part of the system need post-processing.

[4] By 'perfect' we mean that the result does not need any human post-processing.

In the first case, the edit distance is zero, resulting in accuracy equal to 1. In the second case, the accuracy is $1 - d$ with d meaning the edit distance between the segment chosen by the ranker and the correct translation divided by the length of the segment. In the third case, the accuracy is calculated as for (2) except that we use the reference translation to obtain the edit distance.

Given the accuracies for all sentences we use the arithmetic average as the translation accuracy of the whole text. The accuracy is negatively influenced by several aspects. If a word is not known to the morphological analyzer, it does not get any morphological information which means that it is practically unusable in the parser. Another possible problem is that a lemma is not found in the dictionary. In such a case, the original source form appears in the translation, which naturally decreases the score. Finally, sometimes the morphological synthesis component is not able to generate the proper word form in the target language (due to partial incompatibility of tagsets for both languages). In such a case, the target language (Slovak) lemma appers in the translation.

The results are summarized in Table 1. The results obtained by our system are compared with the results of an original system for Czech-to-Slovak MT. The numbers clearly support the claim that the change of the architecture enabled by an exploitation of a multigraph in all phases of the translation mentioned in our paper improves the system performance. The improvement can be attributed both to the shallow parser as well as the ranker, one without the other provides worse results.

Table 1. Czech-to-Slovak evaluation

accuracy	original	ranker & chunker	ranker & parser
character based	93.9%	96.3%	96.4%
word based	81.1%	87.8%	88.3%

7 SEGMENTATION

Due to morphological, syntactic and lexical ambiguity, the number of edges in a chart may grow exponentially during processing a sentence. Especially for languages with rich inflection, such as Czech and other Slavic languages, this fact may seriously influence the effectivity of the

translation process, thus it would be helpful to optimize the processing of sentences that are too long. Since the method described in this paper is based on shallow NLP and parts of source sentences are processed independently, using smaller translation units rather than whole sentences would speed up the translation process without necessarily lowering translation accuracy. In our experiment, we have exploited the corpus of Czech sentences with manually annotated clause structure [12] to see how the segmentation of compound sentences could help.

The Prague Dependency Treebank[5] [13] is a large and elaborated corpus with rich syntactic annotation of Czech newspaper texts. A part of this corpus was manually annotated with respect to structure of sentences—the concept of segments, easily automatically detectable and linguistically motivated units was adopted [14]. Segments are understood as maximal non-empty sequences of tokens that do not contain any punctuation mark or coordinating conjunction. The sentence annotation captures the level of embedding for individual segments. This concept of linear segments serves as a good basis for the identification of clauses—single clause consists of one or more segments with the same level of embedding; one or more clauses then create(s) a complex sentence.

The definition of segments adopted in the project is based on very strict rules for punctuation in Czech. Generally, the beginning and end of each clause must be indicated by a boundary. This holds for embedded clauses as well. In particular, there are only very few exceptions to a general rule saying that there must be some kind of a boundary between two finite verb forms of meaningful verbs.

In the pilot phase of the project, 3,443 sentences from PDT were annotated with respect to their sentence structure which gives 7,975 segments and 5,003 clauses. While most sentences contain only one or two clauses, the maximal observed number of clauses in a sentence is 11.

An experiment that used segments of the corpus instead of whole sentences as translation units has shown that the translation process was 3–4 times faster (depending on the set of syntactic rules) while the accuracy of the translation did not change. Thus the only remaining problem is to refine the algorithm that automatically segments compound sentences of the source language.

[5] `http://ufal.mff.cuni.cz/pdt2.0/`

8 CONCLUSIONS

The results achieved in the experiments with machine translation between two very closely related languages (Czech and Slovak) described in this paper seem to support the hypothesis that the change of the rather simplistic architecture of the original system Česílko enabled by an exploitation of a multigraph and a shallow chart parser combined with a stochastic ranker of the target language sentences generated by the system resulted in improved translation quality. The use of a chart-based technique in several phases of the translation process is a crucial factor for the improvement.

ACKNOWLEDGMENTS

The presented research has been supported by the grant No. 1ET100300517 of the GAAV ČR and partially supported by the grant No. 405/08/0681 of the GAČR.

REFERENCES

1. Chandioux, J.: METEO, an operational system for the translation of public weather forecasts. In: American Journal of Computational Linguistics, FBIS Seminar on Machine Translation, Rosslyn, Virginia (1976) 27–36
2. Thouin, B.: The METEO system. In: Proceedings of a conference Practical experience of machine translation. Ed.Veronica Lawson (Amsterdam, New York, Oxford: North-Holland Publishing Company, 1982), London, England (1981) 39–44
3. Colmerauer, A.: Les systèmes Q ou un formalisme pour analyser et synthétiser des phrases sur ordinateur. Technical report, Mimeo, Montréal (1969)
4. Oliva, K.: A parser for Czech implemented in Systems Q. Technical report, MFF UK, Prague (1989)
5. Marinov, S.: Structural Similarities in MT: A Bulgarian-Polish case (2003)
6. Homola, P., Kuboň, V.: A translation model for languages of acceding countries. In: EAMT Workshop, Malta (2004)
7. Dyvik, H.: Exploiting Structural Similarities in Machine Translation. Computers and Humanities **28** (1995) 225–245
8. Altintas, K., Cicekli, I.: A Machine Translation System between a Pair of Closely Related Languages. In: Proceedings of the 17th International Symposium on Computer and Information Sciences (ISCIS 2002), Orlando, Florida (2002) 192–196

9. Corbi-Bellot, A., Forcada, M., Prtiz-Rojas, S., Perez/Ortiz, J.A., Remirez-Sanchez, G., Martinez, F.S., Alegria, I., Mayor, A., Sarasola, K.: An Open-Source Shallow-Transfer Machine Translation Engine for the Romance Languages of Spain. In: Proceedings of the 10th Conference of the European Association for Machine Translation, Budapest (2005)
10. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, Pennsylvania (2001) 311–318
11. Callison-Burch, C., Osborne, M., Koehn, P.: Re-evaluating the Role of BLEU in Machine Translation Research. In: Proceedings of the EACL'06, Trento, Italy (2006)
12. Lopatková, M., Klyueva, N., Homola, P.: Annotation of sentence structure; capturing the relationship among clauses in czech sentences. In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), Suntec, Singapore, Association for Computational Linguistics (2009) 74–81
13. Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.: Prague Dependency Treebank 2.0. LDC (2006)
14. Kuboň, V., Lopatková, M., Plátek, M., Pognan, P.: A Linguistically-Based Segmentation of Complex Sentences. In Wilson, D., Sutcliffe, G., eds.: Proceedings of FLAIRS Conference, AAAI Press (2007) 368–374

PETR HOMOLA
INSTITUTE OF FORMAL AND APPLIED LINGUISTICS,
CZECH REP.
E-MAIL: <HOMOLA@UFAL.MFF.CUNI.CZ>

VLADISLAV KUBOŇ
INSTITUTE OF FORMAL AND APPLIED LINGUISTICS,
CZECH REP.
E-MAIL: <VK@UFAL.MFF.CUNI.CZ>

International Journal of Mind, Brain & Cognition

The *International Journal of Mind, Brain, and Cognitive Science* (IJMBC) is a biyearly journal for the multidisciplinary study of minds and other intelligent systems. It seeks to present state of the art research in the broad areas of Cognitive Science, Neuroscience, Linguistics, Computer Science, Artificial Intelligence as well as from the humanities where facts blended with theory explore the workings of the mind-brain and throw new light on our cognitive behaviour. The study of the human mind, from the neurophysiological as well as computational perspectives, has given us exciting data about our own inner mental life and also how we -- with our bodies and brains -- interact with the environment and cognize. With development of theoretical frameworks in several disciplines along with sophisticated brain mapping techniques, it is now possible to test hypotheses concerning the workings of the mind and eventually see answers to more difficult questions like consciousness and several specific cognitive mechanisms. This journal will provide a platform for research that integrates disciplinary frameworks and proposes theoretical as well as experimental work that is, wherever possible, truly multidisciplinary.

As an international journal of Cognitive Science, this journal will showcase research of the best quality that stands solidly on well-developed theoretical foundations and faces some of the challenging research tasks. Full length research articles that seek rigorous answers to questions of mind-brain interaction and hence cognitionand thought-provoking theoretical articles that challenge received views and explore new options will receive priority. Research that is merely confirmatory will be given less importance. The journal will also publish careful book reviews in these domains. All submissions will be rigorously peer reviewed and will maintain highest standards of quality.

The following kinds of articles are appropriate for the journal: (a) theories or theoretical analyses of cognitive processes and brain theory; (b) experimental studies relevant to theoretical issues in cognitive science; (c) computational models of neural and cognitive processes and (d) discussions of new problem areas or methodological issues in cognitive science.

The journal will publish four categories of articles. Regular articles have a word limit of 10000 words. Brief reports have a target length of about 4,000 words. Letters to the editor are expected not to exceed 1,000 words, and may include commentaries on articles, responses to commentaries, and discussion items of general relevance to the cognitive science community. Book reviews should not exceed 1,000 words.

NOTE FOR CONTRIBUTORS

Authors can submit their manuscripts to the Editor-in-Chief, Prof. Probal Dasgupta on his e-mail <probal53@yahoo.com> in a WORD file according to the APA 5th edition format.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 201-216
RECEIVED 21/11/09 ACCEPTED 16/01/10 FINAL 09/03/10

Manipuri-English Example Based Machine Translation System

THOUDAM DOREN SINGH[†*], SIVAJI BANDYOPADHYAY[*]

[†]*Center for Development of Advanced Computing (CDAC), India*
[*]*Jadavpur University, India*

ABSTRACT

The development of a Manipuri to English example based machine translation system is reported. The sentence level parallel corpus is built from comparable news corpora. POS tagging, morphological analysis, NER and chunking are applied on the parallel corpus for phrase level alignment. The translation process initially looks for an exact match in the parallel example base and returns the retrieved target output. Otherwise, the maximal match source sentence is identified. For word level mismatch, the unmatched words in the input are translated from the lexicon or transliterated. Unmatched phrases are looked into the phrase level parallel example base; the target phrase translations are identified and then recombined with the retrieved output. The system is currently not handling multiple maximal matches or no match (full or partial) situation. The EBMT system has been evaluated with BLEU and NIST scores of 0.137 and 3.361 respectively, better than a baseline SMT system with the same training and test data.

Keywords: Example based machine translation, Manipuri – English, Sentence Level Parallel Corpus, Phrase Level Alignment, Evaluation

1 INTRODUCTION

Machine Translation (MT) is the process of translating text or speech units from a source language (SL) into a target language (TL) by using computers while preserving the meaning and interpretation. Various MT paradigms have so far evolved depending upon how the translation knowledge is acquired and used. The main drawback of Rule Based MT systems is that sentences in any natural language may assume a large variety of structures and hence translation requires enormous knowledge about the syntax and semantics of both the SL and TL. On the other hand, SMT techniques depend on how accurately various probabilities are measured. Realistic measurements of these probabilities can be made only if a large volume of sentence aligned parallel corpora is available. The requirement of SMT system for big parallel corpus and inability to get back the original translation used during training prompted the use of the EBMT paradigm for Manipuri-English MT system. An EBMT system stores in its example base the translation examples between the SL and TL. These examples are subsequently used as guidance for future translation tasks. In order to translate a new input sentence in SL, all matching SL sentences that match any fragment of the input SL sentence are retrieved from the example base, along with their translation in TL. These translation examples are then recombined suitably to generate the translation of the given input sentence.

Manipuri is a less privileged Tibeto-Burman language spoken by approximately three million people mainly in the state of Manipur in India as well as its neighboring states and in the countries of Myanmar and Bangladesh and is in the VIII Schedule of Indian Constitution with little resource for NLP related research and development. Some of the unique features of this language are tone, the agglutinative verb morphology and predominance of aspect than tense, lack of grammatical gender, number and person. Other features are verb final in word order, lack of numeral classifier and extensive suffix with more limited prefixation. Different word classes are formed by affixation of the respective markers. This is the first attempt to develop Manipuri-English machine translation using example based approach.

There is no parallel corpus available to develop Manipuri-English MT system at the first place. In our present work Manipuri-English news parallel corpora is being developed from web as an initial step using a semi-automatic approach. The translation methodology incorporated in our system is to search and identify for (a) complete

sentence match (b) phrase level and finally (c) word levels and using entries from the lexicon after applying suffix removal/addition operations using a suffix adaptation module. The EBMT system developed so is compared with a baseline SMT system using Moses decoder. The rest of the paper is organized in such as way that related works are discussed in section 2, parallel corpus development at section 3, EMBT system methodology in section 4, evaluation in section 5 and conclusion in section 6.

2 Related Works

Aligning sentences in bilingual corpora based on a simple statistical model of character lengths using the fact that longer sentences in one language tend to be translated into longer sentences in the other language, and that shorter sentences tend to be translated into shorter sentences is reported in [6]. Reliable measures for extracting valid news articles and sentence alignments of Japanese and English are reported in [12]. Statistical alignment tool such as GIZA++ [26] are used for words and phrase alignment of statistical machine translation systems. The EBMT system as reported by Makoto Nagao at a 1981 conference identified the three main components: matching fragments against a database of real examples, identifying the corresponding translation fragments and then recombining these translation fragments to give the target text. Researchers [25], [10] have considered EBMT to be one major and effective approach among different MT paradigms, primarily because it exploits the linguistic knowledge stored in an aligned text in a more efficient way. Example-based Machine Translation [13] makes use of past translation examples to generate the translation of a given input. [4] learn translation templates from English-Turkish translation examples. They define a template as an example translation pair where some components (e.g. word stems and morphemes) are generalized by replacing them with variables in both sentences. The use of morphemes as units allows them to represent relevant templates for Turkish. There is currently no template implementation in our EBMT system. EBMT systems are often felt to be best suited to a sublanguage approach, and an existing corpus of translations can often serve to define implicitly the sublanguage which the system can handle [25]. EBMT for highly inflected language with free order sentence constituents like Basque to English [18] are reported using morphemes for basic analysis. Hybrid Rule-Based – Example-Based MT using sub-sentential translation units

are reported in [17]. There are reports on translating from poor to rich morphology languages [2], namely English to Czech and English to Hindi in Indian context [1]. Phrasal EBMT System for Translating English to Bengali is found at [27].

3 PREPARATION OF EXAMPLE BASES

Manipuri is a less computerized language and the parallel corpora, annotated corpora, dictionary and other lexical resources are generally not available. The following three example bases have been developed as part of the present work:

1. Manipuri-English Parallel corpora of 16919 sentences
2. Manipuri-English dictionary of 12229 entries which includes 2611 transliterated words
3. Manipuri-English – 57629 aligned phrases

3.1 Sentence alignment

The Manipuri-English parallel corpus is collected from news available in both Manipuri and English in a noisy form from http://www.thesangaiexpress.com/ . The corpora is comparable in nature as identical news events are described in both Manipuri and English news stories. There are 23375 English and 22743 Manipuri sentences respectively in the noisy corpus. A semiautomatic parallel corpus extraction approach is applied to align the corpora in order to make it usable for the Machine Translation system. As part of the process, the articles are aligned and dynamic programming approach [6] is applied to achieve the sentence pairs after making sure that there are equal numbers of articles on both sides. Based on the similarity measures [12], we allow 1-to-*n* or *n*-to-1 (1<=n<=6) alignments when aligning the sentences. Let M*i* and E*i* be the words of Manipuri and English sentences for *i*-th alignment. The similarity between *Mi* and *Ei* is calculated as:

$$SIM(M_i,E_i) = \frac{co(M_i \times E_i) + 1}{l(M_i) + l(E_i) - 2co(M_i \times E_i) + 2} \quad \textbf{(1)}$$

where,

$l(X) = \sum_{x \epsilon X} f(x)$, $f(x)$ is the frequency of x in the sentences.

$co(Mi \times Ei) = \sum(m,e) \in Mi \times Ei \ min(f(m), f(e))$

$Mi \times Ei = \{(m, e)|m \in Mi, e \in Ei\}$ *and* $Mi \times Ei$ is a one-to-one correspondence between Manipuri and English words.

A Manipuri stemmer is used in order to make use of a medium size dictionary since there is no Manipuri Wordnet available. After the parallel alignment and cleaning, there are 16919 parallel news sentences. The Manipuri-English dictionary [7] is being digitized and currently contains 9618 Manipuri words. Use of transliterated English words in Manipuri is very prominent and there are 2611 transliterated words.

3.2 Morphological Processing

In Manipuri, words are formed by three processes called affixation, derivation and compounding. The majority of the roots found in the language are bound and the affixes are the determining factor of the word class in the language. In this agglutinative language the numbers of verbal suffixes are more than that of the nominal suffixes. Works on morphological processing in Manipuri are found in [3] and [19].

Verb morphology does not indicate number, person, gender or pronominal agreement between the verb and its arguments. There are two derivational prefixes: an attributive prefix which derives adjectives from verbs and a nominalizing prefix which derives nouns from verbs.

A noun may be optionally affixed by derivational morphemes indicating gender, number and quantity. A noun may have one of the 5 semantic roles: agent, actor, patient, reciprocal/goal and theme. Actor and theme roles are not indicated morphologically, while all other semantic roles are indicated by an enclitic. Word class and sentence identification using morphological information is reported in [20].

3.3 POS Tagging and Chunking

Works on the POS tagging for Manipuri have been reported in [21] that describes Morphology Driven POS tagger of Manipuri as well as in [22] that uses Support Vector Machines (SVM) and Conditional Random Fields (CRF). The Manipuri tagset is the same as the 26 tagset defined for the Indian languages. The POS tagger with 26[1] tags using SVM methodology is identified as more viable for the present system

[1] http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.pdf

because of its detailed 26 tags. The English sentences are POS tagged and chunked using fnTBL [14].

There is no evidence for a verb phrase constituent in Manipuri. The Manipuri verb clause consists of a verb (V) and its argument (i.e., noun phrase) this verb subcategorizes for. Given below are the phrase structure rules which derive sentences in Manipuri.

(1) S→ NP* V
NP* → NP NP NP …

Example of a Manipuri sentence is given here.

অপিকপা অমোৎপা অশোনবা অঙাংদু কপ্পী|
apikpa amotpa asonba angangdu kappi
|----------NP----------| V
Small dirty weak that child is crying
'The small, dirty, weak boy is crying'

A noun phrase may consist of a noun followed by derivational and inflectional morphology or a noun and adjectives, numerals and/or quantifiers. The order of these constituents within the noun phrase is relatively free.

(2) NP→ N (Adj*) (Num/Quant)
NP→ (Adj*) N (Num/Quant)

For example,

উচেক	অচৌবদু	ফজে |
uchek	*achoubadu*	*phajei*
That bird	big	is

beautiful.

|---------NP----------|

Grammatically, a sentence must consist of an inflected verb, which is a verb root and an inflectional suffix. An adverbial clause can be derived through the suffixation of clausal subordinators to a nominalized clause. The phrase structure rule which is used to generate adverbial clause is

(3) AdvP→ S' CS

S' is the sentence and CS is the clausal subordinator. It can be a locative marker দা (da) . e.g.,

ঐখোয়দা	লাকপদা
eikhoida	*lakpada*
To our place	upon coming home

'when coming to our place'

The SVM based chunker [11] is used. The training process has been carried out by YamCha[2] toolkit, an SVM based tool for detecting classes in documents and formulating the chunking task as a sequential labeling problem. For classification, we have used TinySVM-0.07[3] classifier that seems to be the best optimized among publicly available SVM toolkits. We train the system with 1,600 sentences of 35,120 words and used the model.

3.4 NER module

The NER system for Manipuri [23] is developed using Support vector machine considering the four major named entities tags, namely Person name, Location name, Organization name and Miscellaneous name. The training process has been carried out by YamCha toolkit, an SVM based tool for detecting classes in documents and formulating the NER tagging task as a sequential labeling problem trained with 28,629 sentences. For classification, we have used TinySVM-0.07 classifier that seems to be the best optimized among publicly available SVM toolkits. Experimental results show the effectiveness of the proposed approach with the overall average Recall, Precision and F-Score values of 93.91%, 95.32% and 94.59% respectively. The named entities are transliterated into the target language using modified joint source channel model for transliteration [28].

3.5 Word and Chunk alignment

Each Manipuri word has no one-to-one correspondence with the words of English sentences and also there is no direct equivalence of Manipuri case markers to English. Words and phrases are aligned using GIZA++, a statistical word alignment toolkit [26]. The high quality aligned phrases are extracted in order to feed into the generation module of the system. A word in Manipuri can correspond to several English words and vice versa. Some of the examples are:

ৱাখোক লান্থোক (*wathok lanthok*)←→crisis
লৌখবা (*louthaba*)←→take something down
লৌথোক লৌশিন (*louthok lousin*)←→ give and take
চাইখায়বা (*chaikhayba*)←→ scatter

[2] http://chasen.org/~taku/software/yamcha/
[3] http://chasen.org/~taku/software/TinySVM/

Also some Manipuri to English translation variations with additional suffixes but maintaining the same meaning is observed as given below:

ছেৎহনবা (*chet-han-ba)* / ছেৎশনহনবা (*chet-sin-han-ba)* ←→ tighten
ঙহাক (*Ngahak)* / ঙহাক্তং (*Ngahak-tang)* ←→ a while

The variations of the verb part are caused by the inclusion/exclusion of derivational suffixes. The verbal suffixes are used to indicate the mood, aspect and not only indicating the type of sentences.

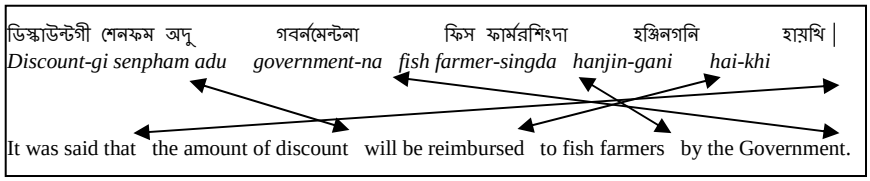


Figure 1: Equivalence between Manipuri and English components

Chunks are aligned using a dynamic programming "edit-distance style" alignment algorithm. In the following, *a* denotes an alignment between a target sequence *e* and a source sequence *f*, with $I = |e|$ and $J = |f|$. Given two sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, P(a|e, f) = \underset{a}{\operatorname{argmax}}\, P(a, e|f).$$

Considering alignments such as those obtained by an edit-distance algorithm, i.e.

$a = (t_1, s_1)(t_2, s_2) \ldots (t_n, s_n)$,

with $\forall\ k \in [1, n]$, $t_k \in [0, I]$ and $s_k \in [0, J]$, and $\forall\ k < k'$:

$t_k \leq t_{k'}$ or $t_{k'} = 0$,

$s_k \leq s_{k'}$ or $s_{k'} = 0$,

$I \subseteq \bigcup_{k=1}^{n}\{t_k\}$, $J \subseteq \bigcup_{k=1}^{n}\{s_k\}$,

where $t_k = 0$ and $s_k = 0$ denote a non-aligned target and source chunks. We then assume the following model:

$$P(a, e|f) = \prod_k P(t_k, s_k, e|f) = \prod_k P(e_{t_k}|f_{s_k}),$$

where $P(e_0|f_j)$ and $P(e_i|f_0)$) denote an "insertion" and "deletion" probabilities respectively.

Assuming that the parameters $P(et_k|fs_k)$ are known, the most likely alignment is computed by a simple dynamic-programming algorithm which is a classical edit-distance algorithm in which distances are replaced by inverse-log-conditional probabilities. Moreover, this algorithm can be simply adapted to allow for block movements, in the context of MT evaluation [8]. This adaptation is necessary to take into account the potential differences between the order of constituents in Manipuri and English. We compute these parameters by relying on the information contained within the chunks considering word to word probabilities and chunk labels. Relationships between chunks are then computed using the model:

$$P(e_i|f_j)=\sum_{a_c} P(a_c, e_i|f_j) \simeq \max_{a_c} P(a_c, ei|fj)= \prod_k \max_{l} P(e_{il}|f_{jk}).$$

In the case of chunk labels, a simple matching algorithm is used. It is possible to combine several sources of knowledge in a log-linear framework, in the following manner:

$$log P(e_i|f_j) = \sum \lambda_k log P_k(e_i|f_j) - logZ,$$

where $Pk(.)$ represents a given source of knowledge, λ_k the associated weight parameter and Z a normalization parameter. To produce a higher quality, the aligned phrases generated using GIZA++ are also added to the aligned chunks extracted by the chunk alignment module.

4 MT SYSTEM DEVELOPMENT METHODOLOGY

This is the first attempt to build MT system for Manipuri to English. While the EBMT employ pattern matching technique to translate subparts of the given input sentence , two fundamental problems of developing Manipuri to English EBMT system are (a) wide syntactic divergence between the source and target languages (b) higher degree of agglutination and richer morphology of Manipuri compared to English. Considering the first problem, we resolve it by adapting the following approach of reordering the input Manipuri sentence. Manipuri follows verb final in word order and there is lack of grammatical relation between subject and object. For example, the

following sentence pair follows the same meaning (Tomba drives the car), though with different emphasis.

তোম্ব-না	কার-দু	খৌই
Tomba-na	*Car-du*	*thou-i*
Tomba-nom	Car-distal	drive

কার-দু	তোম্ব-না	খৌই
Car-du	*Tomba-na*	*thou-i*
Car-distal	Tomba-nom	drive

The identification of subject and object in both the sentences are done by the suffixes না (*na*) and দু (*du*). The case markers are the critical part of conveying right meaning during translation though the most acceptable order is SOV. The basic difference of phrase order compared to English is handled by reordering the input sentence following the rule [16]:

$C'_mS'_mS'O'_mO'V'_mV' \rightarrow SS_mV\ V_mOO_mC_m$

where, S: Subject
O: Object
V : Verb
C_m: Clause modifier
X': Corresponding constituent in Manipuri, where X is S, O, or V
X_m: modifier of X

The phrase reordering program is written using the perl module Parse::RecDescent.

There is no direct equivalence of the Manipuri case markers in English. So, establishing a word level similarity between Manipuri and English is more tedious if not impossible. Essentially, all morphological forms of a word and its translations have to exist in the parallel example bases, and every word has to appear with every possible case marker, which will require an impossibly huge amount of example base. Dealing at sub-sentence level replicates more complexity even at the level of chunking, before the actual process kicks off. One major advantage of EBMT is that it requires neither a huge parallel corpus as required by SMT, nor it requires framing a large rule base required by RBMT. Study of EBMT is therefore feasible for us as we do not have access to such linguistics resources. The translation steps incorporated in our system is to search and identify for (a) complete sentence match (b) phrase level and finally (c) word levels and using entries from the lexicon after applying suffix removal/addition operations using a suffix

adaptation module of the source language input sentence, translate the corresponding units individually to the target language and finally arranging the translated phrases to form the target language equivalent of the source language sentence. While relating the Manipuri and English noun phrases (NPs), NPs usually end with a case-morpheme that contains information about case and number. For example, স্কুলদুগী নুপামচাশিংদুনা (*school-du-gi nupamacha-sing-du-na,* "by the boys of the school") is related as follows:

স্কুল *(school,* school*)* + দু*(du- distal marker,* the*)* + গী*(gi- case marker,* of*)* নুপামচা*(nupamacha,* boy*)*+শিং*(sing - plural marker,* s*)*+ দু*(du- distal marker,* the*)*+না*(na –case marker,* by*)*

The input sentence is passed through a stemmer in order to separate the significant suffixes along with the corresponding information for the phrase level and word adaptations. Basic sentence types in Manipuri are determined through illocutionary mood markers, all of which are verbal inflectional suffixes, with the exception of the interrogative which is an enclitic.

The simple aspect markers are -ই -y, -মি -mi, -নি -ni, -পি -pi, -ঙি -ngi, -লি -li. The progressive aspect makers are -রি –ri, -লি -li . The perfect aspect markers are -রে –re, -লে -le. To handle the various surface words of the input text, a stemmer is plugged in to maximize the matches. In the matching module, there is establishment of correspondence between units in a bilingual text at sentence, phrase or word level. Sentences can, however, be quite long. And the longer they are, the less possible it is that they will have an exact match in the translation archive, and the less flexible the EBMT system will be. In practice, EBMT systems that operate at sub-sentence level involve the dynamic derivation of the optimum length of segments of the input sentence by analyzing the available parallel corpora [5]. This requires a procedure for determining the best "cover" of an input text by segments of sentences contained in the database. It is assumed that the translation of the segments of the database that cover the input sentence is known. What is needed, therefore, is a procedure for aligning parallel texts at sub-sentence level. If sub-sentence alignment is available, the approach is fully automated but is quite vulnerable to the problem of low quality as mentioned above, as well as to ambiguity problems when the produced segments are rather small. The problem of multiple phrase matches will be handled later using the language model of the target language by picking up proper target phrase. The other alternative that will be experimented in future will be to look for most probable maximal

match using frequency information for each parallel pair. If there is no match, either partial or full, in the example base, the future plan is to go for phrasal EBMT system. The algorithmic steps followed are depicted below:

a. If there is Sentence level match
   Produce exact output translation
b. Else process the input – POS, Morph, NER and Chunks
   For maximal match (find the sentence in the Example Base that matches most with the input)
   i. one maximal match
      - phrase level mismatch - look for phrase level match and return output, replace this translated phrase in the retrieved target for the maximal match sentence as the parallel sentence level Example Base is phrase aligned
      - word level mismatch - look into the bilingual lexicon or transliteration
      - above is applicable for more than one word or phrase mismatch
   ii. more than maximal match
      - carry out the above process for all the maximal match pairs. The best target among multiple outputs is selected using the language model.
      - take the pair that occurs most in the Example Base – keep frequency information for each pair, then do as in one maximal match.
   iii. no match in the sentence level and maximal
      - go for phrasal EBMT

Finally, the translated fragments obtained so are stitched together to form the target sentence following the reordering rules as per the target language.

5 EVALUATION

The EBMT system is developed with parallel 15319 sentences, 57629 phrases and 12229 words and evaluated with 900 gold standard test

sentences. We use BLEU and NIST scores for the evaluation of our system. A higher BLEU score indicates better translation. We develop a Manipuri-English baseline SMT system with the same example base used for EBMT and compare the result with EBMT system developed as shown in table 3. There is no previous report available of Manipuri-English SMT system either. The Moses [9] decoder is used. The English (trigram) language model is trained on the English portion of the training data, using the SRI Language Modeling Toolkit [24] with modified Kneser-Ney smoothing.

The two experiments of EBMT and SMT are done using 15319 sentences plus 12229 words. The testing is done three fold taking 300 sentences each.

Table 1 : Statistics of corpus used

	#sentences	#words
Parallel corpus	15319	366728
Test corpus	(300+300+300)=900	20190

Table 2: Evaluation result

Technique	Test#1		Test #2		Test#3		Average	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
Baseline SMT	0.134	3.405	0.125	3.12	0.126	3.06	**0.128**	**3.195**
EBMT system	0.150	3.513	0.131	3.25	0.132	3.32	**0.137**	**3.361**

6 CONCLUSION

The result of initial experiment of Manipuri-English EBMT system is quite encouraging with NIST score of 3.361 and BLEU score of 0.137 which is better than a baseline SMT system. Since, the source side of the translation is highly agglutinative and morphologically rich, incorporating the morphological information could help improving the system. However, the performance of the overall system can be improved further with the addition of other modules such as word sense disambiguation, multiword expression etc. By proper handling of divergence and adaptation of Manipuri-English EBMT performance can be further improved.

REFERENCES

1. Ananthakrishnan, R., Choudhary, H., Ghosh, A., Bhattacharyya, P.: Case markers and Morphology: Addressing the crux of the fluency problem in English-Hindi SMT, Proceedings of IJCNLP, Singapore, (2009)
2. Avramidis, E., and Koehn, P.: Enriching Morphologically Poor Languages for Statistical Machine Translation, Proceedings of ACL-08, HLT, (2008)
3. Choudhury, S. I., Singh, L. S., Borgohain, S., Das, P.K.: Morphological Analyzer for Manipuri: Design and Implementation, In proceedings of AACC, Kathmandu, Nepal, pp 123-129. (2004)
4. Cicekli, I., and G¨uvenir, H. A.: Learning translation templates from bilingual translation examples. In M. Carl and A. Way, editors, Recent Advances in Example-Based Machine Translation, pages 255–286. Kluwer Academic Publishers, Dordrecht, The Netherlands. (2003)
5. Cranias, L., Papageorgiou, H. and Piperidis, S.: 'A Matching Technique in Example-Based Machine Translation', In proceedings of Coling (1994), pages. 100–104. (2004)
6. Gale, W. A., Church, K. W.: A program for aligning sentences in bilingual corpora. Computational Linguistics, 19(1):75–102. (1993)
7. Imoba, S.: Manipuri to English Dictionary. Published by:- S. Ibetombi Devi, Imphal (2004)
8. Leusch, G., Ueffing, N., and Ney, H.: CDER: Efficient MT evaluation using block movements. In proceedings of EACL-06, pages 241-248 (2006)
9. Koehn, P., Hieu, H., Alexandra, B., Chris, C., Marcello, F., Nicola, B., Brooke, C., Wade, S., Christine, M., Richard, Z., Chris, D., Ondrej, B., Alexandra, C., Evan, H.: Moses: Open Source Toolkit for Statistical Machine Translation, Proceedings of the ACL 2007 Demo and Poster Sessions, pages 177–180, Prague. (2007)
10. Kit, C., Pan, H. and Webster, J.: Example-Based Machine Translation: A New Paradigm, Translation and Information Technology, Chinese U of HK Press, pages. 57-78. (2002)
11. Kudo, T., and Matsumoto, Y.: Use of Support Vector Learning for Chunk Identification, In Proceedings of CoNLL-2000. (2000)
12. Utiyama, M., and Isahara, H.: Reliable Measures for Aligning Japanese-English News Articles and Sentences, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1, Pages: 72 – 79, Sapporo, Japan. (2003)
13. Nagao, M.: A framework of a mechanical translation between Japanese and English by analogy principle. In Proceedings of the International NATO Symposium on Artificial and Human Intelligence, pages 173–180, Lyon, France. (1984)
14. Ngai, G. and Florian, R.: Transformation-based Learning in the Fast Lane, Proceedings of NAACL. (2001)
15. Papineni, K., Roukos, S., Ward, T., and Zhu, W.: BLEU: a Method for Automatic Evaluation of Machine Translation, IBM Research Report, Thomas J. Watson Research Center (2001)

16. Rao, D., Mohanraj K., Hegde, J., Mehta, V. and Mahadane, P.: A Practical Framework for Syntactic Transfer of Compound-Complex Sentences for English-Hindi Machine Translation, Proceedings of KBCS 2000. (2000)
17. Sánchez-Martínez, F., Forcada, M. L., and Way, A.: Hybrid Rule-Based - Example-Based MT: Feeding Apertium with Sub-sentential Translation Units, In Proceedings of the 3rd International Workshop on Example-Based Machine Translation, pages 11-18, Dublin, Ireland. (2009)
18. Stroppa, N., Groves, D., Way, A., Sarasola, K.: Example-Based Machine Translation of the Basque Language, In Proceedings of the 7th Conference of the Association for Machine Translation in the Americas, pages 232-241, Cambridge, Massachusetts, USA (2006)
19. Singh, T. D., Bandyopadhyay, S.: Manipuri Morphological Analyzer, In the Proceedings of the Platinum Jubilee International Conference of LSI; December 6-8, 2005, University of Hyderabad, India. (2005)
20. Singh, T. D., Bandyopadhyay, S.: Word Class and Sentence Type Identification in Manipuri Morphological Analyzer. In Proceedings of MSPIL, IIT Bombay, pages 11-17. (2006)
21. Singh, T. D., Bandyopadhyay, S.: Morphology Driven Manipuri POS Tagger, IJCNLP-08 Workshop on NLP for Less Privileged Languages. Proceedings of the Workshop. AFNLP, pages 91-98, 11. January, 2008, IIIT, Hyderabad, India. (2008)
22. Singh, T. D., Ekbal, A., Bandyopadhyay, S.: Manipuri POS Tagging using CRF and SVM: A Language Independent Approach, In the proceedings of ICON-2008, Pune, India, pages 240-245 (2008)
23. Singh, T. D., Kishorjit, N., Ekbal, A,. Bandyopadhyay, S.: Named Entity Recognition for Manipuri using Support Vector Machine, In proceedings of PACLIC 23, Hong Kong (2009)
24. Stolcke, A.: SRILM – An extensible language modeling toolkit. In Proceedings of the International Conference on Spoken Language Processing, pages 901–904, Denver, Colorado. (2002)
25. Somers, H.: Review article: Example-Based Machine Translation, Machine Translation 14, pages 113-158. (1999)
26. Och, F., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1):19–51. (2003)
27. Naskar, S. K., Bandyopadhyay, S.: A Phrasal EBMT System for Translating English to Bengali, In the Proceedings of MT SUMMIT X, Phuket, Thailand. (2005)
28. Ekbal, A., Naskar, S.K., Bandyopadhyay, S.: A Modified Joint Source-Channel Model for Transliteration, Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 191–198, Sydney (2006)

THOUDAM DOREN SINGH
CENTER FOR DEVELOPMENT OF ADVANCED COMPUTING (CDAC)
GULMOHAR CROSS ROAD NO 9, JUHU, MUMBAI-400049, INDIA
E-MAIL: <THOUDAMDS@CDACMUMBAI.IN>

SIVAJI BANDYOPADHYAY
COMPUTER SCIENCE AND ENGINEERING DEPARTMENT
JADAVPUR UNIVERSITY, KOLKATA-700032, INDIA
E-MAIL: <SIVAJI_CSE_JU@YAHOO.COM>

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 217-230
RECEIVED 24/11/09 ACCEPTED 16/01/10 FINAL 08/03/10

Bilingual Document Clustering using Translation-Independent Features

CLAUDIA DENICIA-CARRAL[1], MANUEL MONTES-Y-GÓMEZ[1], LUIS VILLASEÑOR-PINEDA[1] AND RITA M. ACEVES-PÉREZ[2]

[1]National Institute of Astrophysics, Optics and Electronics, Mexico.
[2] Polytechnic University of Altamira (UPALT), Mexico.

ABSTRACT

This paper focuses on the task of bilingual clustering, which involves dividing a set of documents from two different languages into a set of thematically homogeneous groups. It mainly proposes a translation independent approach specially suited to deal with linguistically related languages. In particular, it proposes representing the documents by pairs of words orthographically or thematically related. The experimental evaluation in three bilingual collections and using two clustering algorithms demonstrated the appropriateness of the proposed representation, which results are comparable to those from other approaches based on complex linguistic resources such as translation machines, part-of-speech taggers, and named entity recognizers.

1 INTRODUCTION

In recent years, due to the globalization phenomenon, there is an increasing interest for organizing and classifying documents from different languages. In this scenario, document clustering aims to identify subsets of documents thematically related in spite of their source language.

The traditional approach for document clustering is based on the assumption that it is possible to establish the topic of documents solely

from the frequency of their terms. This basic approach is appropriate for monolingual clustering since all documents may be represented using the same set of words; nevertheless, in a multilingual situation, where documents belong to different languages, it is useless. An immediate solution to this problem is the application of a translation process which allows to construct a common representation for all documents, and, therefore, to apply any existing clustering method.

Even though the translation-based approach is the common strategy for multilingual document clustering (MDC), there are certain linguistically related languages in which it would be possible to apply a translation-independent approach. Particularly, we refer to languages that belong to the same linguistic family (like romance languages), or that by historical reasons or geographic closeness have borrowed a number of words (as the case of Spanish and English). For this kind of languages, it is possible to construct a joint representation of their documents based on words such as common named entities, cognates and foreign words[1].

Taking advantage of the above circumstance, in this paper we explore a translation-independent bilingual clustering approach that represents documents by a set of pairs of related words. We mainly consider two kinds of pairs of related words: on the one hand, orthographically related words such as “presidente-president” or “presidente-presidential”, and, on the other hand, thematically related words such as “candidato-voters” or “presidente-elections”, which may be extracted from the contexts of the firsts. Therefore, the main contribution of this paper is a method for the extraction of these kinds of pairs of words (herein referred as translation-independent features) and the evaluation of their usefulness as document features in bilingual clustering tasks.

The rest of the paper is organized as follows. Section 2 presents some works on multilingual document clustering. Section 3 details the method for the extraction of translation-independent features. Sections 4 and 5 describe the experimental configuration and results respectively. Finally, Section 6 presents our conclusions and some ideas for future work.

[1] Common (or cognate) named entities such as “Barack Obama” which are equally written in Spanish and English; cognates such as “presidente” and “president”; and foreign words such as “software” that is an English word normally used Spanish.

2 RELATED WORK

As we previously mentioned, the translation-based approach is the traditional strategy for MDC. Methods from this approach differentiate one from another by the kind of resources they use for translation as well as by the parts of the texts they translate. There are methods that achieve the translation by means of automatic translation machines [3, 6, 7, 13], and methods that use a bilingual thesaurus or dictionary [12, 14]. Similarly, some of these methods translate the whole documents [6], whereas some others only translate some specific keywords or parts of speech [3, 7, 9, 13].

Motivated by the fact that the performance of this kind of methods is affected by the quality of the automatic translation, Montalvo et al. [8, 9] proposed a translation-independent clustering method that takes advantage from the lexical similarities existing in linguistically related languages. In particular, they proposed using cognate named entities as document features. Their results in a bilingual corpus consisting of documents describing a common set of news events indicate that this kind of features leads to good results in bilingual document clustering.

A possible criticism to the above conclusion might be that it was drawn from a restrictive experimental scenario, where named entities hold a very important role. However, it is expected that for other kind of collections about more general topics, the presence of cognate named entities will be lower, causing the generation of sparse document representations and, therefore, a degradation of the clustering quality. In order to tackle this problem, in this paper we propose to represent documents by a broader set of orthographically similar pairs of words, allowing features such as “presidente-presidential”, which are not a translation of each other, but show a clear semantic relation. In addition, we propose enriching the representation by including some thematically related pairs of words such as “presidente-elections”, which do not present any orthographic similarity, but may be extracted from the contexts of orthographically similar pairs of words.

In order to confirm our claims about the robustness of the proposed features, we present an evaluation that considers three bilingual collections of news reports from the same thematic category but that describe very different events. Somehow, by this experiment, our aim is to investigate the limits of translation-independent features in the task of bilingual document clustering.

3 EXTRACTION OF TRANSLATION-INDEPENDENT FEATURES

As we previously mentioned, our proposal is mainly supported on the idea that, for two linguistically related languages, a pair of words having a high orthographic similarity tend to maintain a semantic relation, and, in addition, that the contexts of these words tend to be similar and thematically consistent.

Based on the above assumptions we designed a method for extracting a set of translation-independent features from a given bilingual document collection. This method considers two main steps. The first step focuses on the identification of all orthographically similar pairs of words, whereas, the second uses these pairs of words in order to discover others that tend to co-occur in their contexts, and, therefore, that maintain a "possible" thematic relation.

At the end, we represent the documents from the given bilingual collection using all extracted features, being each feature defined as a pair of related words (w_1, w_2), where w_1 is a word from language L_1 and w_2 is a word from language L_2.

The following two sections describe in detail the extraction of both kinds of features, orthographically and thematically related. Then, Section 3.3 formalizes the representation of documents by the proposed set of features.

3.1 Features based on Orthographic Similarity

Given a document collection (D) containing documents from two different languages (L_1 and L_2), the extraction of this kind of features is carried out as follows:

1. Divide the collection in two sets (D_1 and D_2); each one containing the documents from one single language.
2. Determine the vocabulary (i.e., set of different words) from each language, eliminating the stop words. We mention these sets V_1 and V_2 respectively.
3. Evaluate the orthographic similarity for each pair of words from the two languages; $sim_{ort}(w_i \in V_1, w_j \in V_2)$. In our experiments we measured this similarity by the quotient of the length of their longest common subsequence (LCS) and the length of the largest word. For instance, the LCS of the words "*australiano*" (in Spanish) and "*australien*" (in English) is "*a·u·s·t·r·a·l·i·n*", and, therefore, their similarity is 9/11.

4. Select as candidate features all pair of words ($w_i \in V_1$, $w_j \in V_2$) having an orthographic similarity greater than a given specified threshold. That is, we consider that the pair of words (w_i, w_j) is a candidate translation-independent feature if $sim_{ort}(w_i, w_j) \geq \alpha$.
5. Eliminate candidate features (w_i, w_j) that satisfy one of the following conditions: $sim_{ort}(w_i, w_j) < sim_{ort}(w_k \in V_1, w_j)$ or $sim_{ort}(w_i, w_j) < sim_{ort}(w_i, w_k \in V_2)$. The purpose of this final step is to select only the strongest relation for each word, avoiding the generation of many irrelevant features.

At this point it is important to comment that this initial step of our method is similar to other existing approaches for automatic extraction of cognates [2, 5, 10]. It also determines the relation of two words by their orthographic similarity, however, it extracts these pairs of words from the own target document collection avoiding the use of a parallel corpus or bilingual dictionary. Because of this characteristic, the proposed method can extract a great number of related words, some of them incorrect but the vast majority useful for the MDC task. In particular, it may extract pairs of words that are not cognates in a strict sense but that maintain some semantic relation such as "presidencia" (presidency in Spanish) and "president" (in English).

In addition to the extraction of a great number of related pairs of words, this method does not require applying processes for POS tagging or named entity recognition, and, therefore, it may be easily adapted to several pair of languages.

3.2 Features based on Thematic Closeness

As stated in the beginning of Section 3, this second step of the extraction method is based on the idea that the semantic relatedness of two words may be calculated according to their lexical neighbors. Therefore, it considers that a pair of words from different languages ($w_i \in L_1$, $w_j \in L_2$) may be thematically related if they tend to co-occur with the same set of orthographically similar words. In order to illustrate the idea behind the method consider the following example.

Given a bilingual collection formed by documents in Spanish and English, and once extracted a set of orthographically similar features {(presidente, president), (Obama, Obama), …, (congreso, congress)}, it may be possible to assume that the word "elecciones" (elections in Spanish) and "voters" (in English) are thematically related given that "elecciones" tend to co-ocurr with words such "presidente, Obama and

congreso", whereas "voters" co-occur with "president, Obama and congress".

The following lines describe the general process for the extraction of this kind of features.

Given a collection of documents D with documents written in two different languages, called L_1 and L_2, the extraction of thematically related pairs of words is carried out as follows:

1. Divide the collection in two sets (D_1 and D_2); each one containing the documents from one single language.
2. Determine the vocabulary (i.e., set of different words) from each language, eliminating the stop words. We mention these sets V_1 and V_2 respectively.
3. Select the subset of orthographically "equal" features (E) extracted in the previous step; $E = \{(w_i, w_j) | sim_{ort}(w_i, w_j) = 1\}$.
4. Represent each word from D by a vector $w_i = <p_{i1}, p_{i2}, \ldots, p_{i|E|}>$, where p_{ij} indicates the number of documents in which word w_i co-occurs with one of the words from feature j.
5. Compute the similarity for each pair of words from the two languages; $sim_{ocr}(w_i \in V_1, w_j \in V_2)$. In our experiments we measured this similarity based on the vector representations defined in (4) and using the cosine formula.
6. Select as features all pair of words ($w_i \in V_1$, $w_j \in V_2$) having a co-occurrence similarity greater than a given specified threshold. That is, we consider that the pair of words (w_i, w_j) is a translation-independent feature if $sim_{ocr}(w_i, w_j) \geq \beta$.

3.3 Representation of Documents using the Proposed Features

We describe the documents from the bilingual collection D using all extracted features. In particular, we represent each document by a vector $d_i = <p_{i1}, p_{i2}, \ldots, p_{i|D|}>$, where p_{ik} indicates the relevance of feature f_k in document d_i. We compute this relevance based on the TF-IDF weighting scheme as indicated below.

Considering that feature f_k is represented by the pair of words (w_{1k}, w_{2k}), where w_{1k} belong to language L_1 and w_{2k} belong to language L_2, p_{ik} is calculated as follows:

$$p_{ik} = TF_{ik} \times IDF_k = \frac{\#(w_{xk}, d_i \in D_x)}{|d_i|} \times \log\left(\frac{|D|}{\#(w_{1k}, D_1) + \#(w_{2k}, D_2)}\right)$$

where #(w_{xk}, d_i) indicates the number of occurrences of the word w_{xk} in document d_i, #(w_{xk}, D_x) the number of documents from language Lx in which w_{xk} occurs, $|d_i|$ the length of document d_i and $|D|$ the number of documents in the whole collection.

4 EXPERIMENTAL SETUP

4.1 Evaluation Corpora

The document collection used in the experiments is a selection of news reports from the Reuters Multilingual Corpus Vol. 1 and Vol. 2[2]. This selection includes documents from three languages, namely, Spanish, English and French, and from 16 different categories. Table 2 shows some numbers about this collection.

It is important to remember that all experiments were done using a pair of languages; therefore, we carried out three bilingual experiments: one for Spanish-English considering 922 documents, other for Spanish-French considering 955 documents and another for English-French with 895 documents.

Table 1. Corpora Statistics

Language	**Documents**	**Vocabulary without stop words**	**Words per document (*average*)**	**Phrases per document (*average*)**
Spanish	491	13437	49.19	3.87
English	431	11169	41.06	3.03
French	464	13076	47.34	3.67

4.2 Clustering Algorithms

Given that our aim was to evaluate the usefulness of the proposed features as an individual factor in the task of BDC, we considered a common platform for all experiments, which uses the same weighting scheme for all types of features (TF-IDF), the same similar measure for comparing the documents (cosine measure), as well as two different clustering algorithms.

[2] http://trec.nist.gov/data/reuters/reuters.html

From the vast diversity of clustering algorithms (for a survey refer to [15]), we decided using the Direct algorithm [4] (a prototype-based approach) and the Star algorithm [1] (a graph-based approach) because:

On the one hand, these algorithms impose different input restrictions; while the first requires knowing the number of desire clusters, the second only needs to consider a minimum threshold (σ) for document similarity.

On the other hand, the Direct algorithm has been previously used in BDC works [8, 9], and the Star algorithm has been recently used in monolingual document clustering tasks [11].

4.3 Evaluation Measure

The used evaluation measure was the *F* measure. This measure allows comparing the automatic clustering solution against a manual clustering (reference solution). It is traditionally computed as described below, where a value of $F = 1$ indicates that the automatic clustering is identical to the manual solution, and a value of $F = 0$ indicates that both solutions do not have any coincidence.

$$F = \sum_{\forall i} \frac{n_i}{n} \max\{F(i,i)\}$$

$$F(i,j) = \frac{2 \times recall(i,j) \times precision(i,j)}{recall(i,j) + precision(i,j)}$$

In this formula, $recall(i,j) = n_{ij}/n_i$ and $precision(i,j) = n_{ij}/n_j$; where n_{ij} is the number of elements of the manual cluster i in the automatic cluster j, n_j is the number of elements of the automatic cluster j and n_i is the number of elements of the manual cluster i.

5 EXPERIMENTAL RESULTS

In order to evaluate the appropriateness of the proposed representation we performed three bilingual experiments and considered two different clustering algorithms. Tables 2 and 3 shows the results corresponding to the best experimental configuration indicated by a particular combination of values of α (orthographic similarity threshold), β (co-occurrence similarity threshold) and σ (document similarity threshold

for the Star algorithm)[3]. In addition, these tables also include two baseline results: on the one hand, the results achieved by a translation-based method, and, on the other hand, the results from a translation-independent approach using cognate named entities as document features [8, 9]. For the first case we used the translation machine available from Google[4] and applied a document frequency (*DF*) threshold for dimensionality reduction[5] [16], whereas, for the second we performed the recognition of named entities using FreeLing for Spanish, Lingpipe for English and Lia_NE for French[6].

The obtained results show that the proposed method clearly outperforms the approach considering the use of cognate named entities as document features; in average, the MAP scores are 11.6% and 8.6% greater when using the Direct and Star algorithms respectively. From these tables, it is also possible to notice that results from the proposed method are very similar to those from the translation-based method, indicating that our proposal is a competitive alternative when dealing with bilingual collections from linguistically related languages, but having the advantage of not requiring any language processing resource or tool.

Table 2. Results obtained with the Direct clustering algorithm

Languages	**Experiment**	***F* measure**	**Best combination**
English-Spanish	Using translation	0.21	-
	Using translation (with DF)	0.24	*DF*=5
	Using cognate named entities	0.27	(α = 0.7)
	Using the proposed representation	**0.37**	(α = 0.6; β = 0.9)
French-Spanish	Using translation	0.33	-
	Using translation (with DF)	0.34	*DF*=5
	Using cognate named entities	0.21	(α = 0.7)
	Using the proposed	**0.36**	(α = 0.8; β =

[3] We considered the following values for these thresholds: α = {1, 0.9,0.8,0.7,0.6}, β = {1,0.9,0.8}, and σ ={0.1,0.2,0.3,0.4,0.5,0.6}.

[4] www.google.com.mx/language_tools

[5] For the experiments we used $DF \geq 1$, $DF \geq 5$ and $DF \geq 10$; the best results were reached using $DF \geq 5$.

[6] These tools are available from the following web sites: http://garraf.epsevg.upc.es/freeling/, http://alias-i.com/lingpipe/, http://lia.univ-avignon.fr/.

	representation		0.8)
French-English	Using translation	0.39	-
	Using translation (with DF)	**0.40**	*DF*=5
	Using cognate named entities	0.25	(α = 0.6)
	Using the proposed representation	0.35	(α = 0.7; β = 0.9)

Table 3. Results obtained with the Star algorithm

Languages	**Experiment**	***F* measure**	**Best combination**
Spanish-English	Using translation	0.29	(σ = 0.1)
	Using translation (with DF)	**0.30**	(*DF* = 5, σ = 0.1)
	Using cognate named entities	0.25	(α = 0.7; σ = 0.1)
	Using the proposed representation	**0.30**	(α = 0.7; β = 0.9; σ = 0.1)
French-Spanish	Using translation	0.25	(σ = 0.1)
	Using translation (with DF)	0.29	(*DF* = 5, σ = 0.1)
	Using cognate named entities	0.21	(α = 0.8; σ = 0.2)
	Using the proposed representation	**0.30**	(α = 0.9; β = 0.9; σ = 0.1)
French-English	Using translation	0.27	(σ = 0.1)
	Using translation (with DF)	**0.31**	(*DF* = 5, σ = 0.1)
	Using cognate named entities	0.17	(α = 0.7; σ = 0.5)
	Using the proposed representation	0.29	(α = 0.8; β = 0.9; σ = 0.2)

From Tables 2 and 3 it may be argued that the proposed method is sensitive to the selection of the two/three threshold values. In order to clarify the extent of the influence of this selection in the achieved results, Table 5 shows the average and the standard deviation of the *F* measure for all the experiments using the proposed representation and the translation-based approach. These results indicate that the proposed method obtained better average values as well as less standard deviation, allowing to conclude that our method is slightly more robust than the translation-based approach, or, in other words, that all approaches tend to be similarly sensitive to the selection of their parameters.

Table 4. Variability of the results using the Star algorithm (considering all values of α, β and σ for our proposal and $DF = 5$ all values of σ for the translation-based approach)

Language	Experiment	*F* measure	
		Average	***Standard Deviation***
Spanish-English	Translating all to Spanish	0.16	0.08
	Translating all to English	0.17	0.07
	Using the proposed representation	0.19	0.05
French-Spanish	Translating all to Spanish	0.12	0.07
	Translating all to English	0.12	0.07
	Using the proposed representation	0.16	0.06
French-English	Translating all to Spanish	0.15	0.07
	Translating all to English	0.15	0.07
	Using the proposed representation	0.17	0.05

6 CONCLUSIONS AND FUTURE WORK

In this paper we presented a translation-independent bilingual clustering approach that represents documents by a set of pairs of related words. Particularly, we considered two kinds of pairs of related words: orthographically related and thematically related words.

In spite of the complexity of the task –as demonstrated by the achieved results– the representation based on translation independent features shown to be an alternative to the translation-based approach. The results demonstrated that proposed representation is suitable for the clustering task, having the advantage of not depending on any linguistic resource. However, it is important to remember that the application of our proposal is limited to linguistically related languages that belong to the same linguistic family or that by historical reasons or geographic closeness have borrowed a number of words.

Even though the proposed method may be applied to general domain collections, we consider it is more adequate for specific domain document sets, where specialized terms are abundant and tend to be orthographically similar. Regarding this hypothesis, as future work we plan to apply our method to this kind of collections. In addition, we plan to extend the proposed representation to deal with multilingual collections that include documents in more than two languages.

ACKNOWLEDGMENTS

This work was done under partial support of CONACYT (Project Grant CB-2007-1-83459 and scholarship 165323).

REFERENCES

1. Aslam J., Pelekhov, K., and Rus, D. A practical clustering algorithm for static and dynamic information organization. In: *Proceedings of the Symposium on Discrete Algorithms*, 208-217. Washington, D.C., US. 2001.
2. Bergsman, S., Kondrak, G. Multilingual Cognate Identification using Integer Linear Programming. In: *Proceedings of the International Workshop on Acquisition and Management of Multilingual Lexicons*, 11-18. Borovets, Bulgaria. 2007.
3. Chen, H.-H., & Lin, C.-J. A Multilingual News Summarizer. In: *Proceedings of 18th International Conference on Computational Lingüstics*, 159-165. 2000.
4. Karypis, G. CLUTO: A Clustering Toolkit. *Technical Report: 02-017*. University of Minnesota, Department of Computer Science. 2002.
5. Kondrak G. Combining Evidence in Cognate Identification. In: *Proceedings of 17th Conference of the Canadian Society for Computational Studies of Intelligence*, 44-59, London, UK. 2004.
6. Leftin, L. J. Newblaster Russian-English Clustering Performance Analysis. *Technical Report*, Computer Science, Columbia University. 2003.
7. Mathieu, B., Besancon, R., and Fluhr, C. Multilingual Document Clustering Discovery. In *Proceedings of RIAO-04*, Avignon, France, 1-10. 2004.
8. Montalvo, S., Martínez, R., Arantza, C., & Fresno, V. Multilingual News Document Clustering: Two Algorithms Based on Cognate Named Entities. Text, Speech and Dialogue. *Lecture Notes in Artificial Intelligence*, Vol. 4188, 165-172. 2006.
9. Montalvo, S., Martínez, R., Casillas, A., & Fresno, V. Multilingual news clustering: Feature translation vs. identification of cognate named entities. *Pattern Recognition Letters* 28 , 2305-2311. 2007.
10. Mulloni A. and Pekan V. Automatic Detection of Orthografics Cues for Cognate Recognition. In: *Proceedings of LREC*, Genoa, Italy. 2387-2390. 2006.

11. Pérez-Suarez, A., Martinez-Trinidad, F., Carrasco-Ochoa, A., & Medina-Pagola. A New Graph-based Algorithm for Clustering Documents. In: *Proceedings of the Foundations of Data Mining workshop* (FDM'08), at ICDM'08 Workshops. Italia. 2008.
12. Pouliquen, B., Steinberger, R., Ignat, C., Käsper, E., and Temnikova, I. Multilingual and Cross-lingual News Topic Tracking. In *Proceedings of the 20th International Conference on Computational Linguistics*, 959-965, Geneva, Switzerland. 2004.
13. Rauber, A., Dittenbanch, M., Merkl, D. Towars Automatic Content-based Organization of Multilingual Digital Libraries: an English, French and German View of the Russian Information Agency Novosti News. In: *Third All-Russian Conference Digital Libraries: Advanced Methods and Technologies*, Digital Collections Petrozavodsk. 2001.
14. Steinberger, R., Pouliquen, B., & Hagman, J. Cross-lingual Document Similarity Calculation Using the Multilingual Thesaurus EUROVOC. *Lecture Notes in Computer Science*, Vol. 2276, 101-121. 2002.
15. Tan, P. N., Steinbach, M., and Kumar, V. Cluster Analysis: Basic Concepts and Algorithms (chapter 8). In: *Introduction to Data Mining*. Addison-Wesley. 2006.
16. Yang, Y. and Pedersen J. A comparative study on feature selection in text categorization. In: *International Conference on Machine Learning*, 412–420. 1997.

CLAUDIA DENICIA-CARRAL
LABORATORY OF LANGUAGE TECHNOLOGIES,
NATIONAL INSTITUTE OF ASTROPHYSICS,
OPTICS AND ELECTRONICS (INAOE), MEXICO.
E-MAIL: <CDENICIA@INAOEP.MX>

MANUEL MONTES-Y-GÓMEZ
LABORATORY OF LANGUAGE TECHNOLOGIES,
NATIONAL INSTITUTE OF ASTROPHYSICS,
OPTICS AND ELECTRONICS (INAOE), MEXICO.
E-MAIL: <MMONTESG@INAOEP.MX>

LUIS VILLASEÑOR-PINEDA
LABORATORY OF LANGUAGE TECHNOLOGIES,
NATIONAL INSTITUTE OF ASTROPHYSICS,
OPTICS AND ELECTRONICS (INAOE), MEXICO.
E-MAIL: <VILLASEN@INAOEP.MX>

RITA M. ACEVES-PÉREZ
DEPARTMENT OF ELECTRONIC ENGINEERING,
POLYTECHNIC UNIVERSITY OF ALTAMIRA (UPALT), MEXICO.
E-MAIL: <RITA.ACEVES@UPALT.EDU.MX>

Applications

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 233-247
RECEIVED 03/12/09 ACCEPTED 16/01/10 FINAL 10/03/10

Interactive QA using the QALL-ME Framework

IUSTIN DORNESCU AND CONSTANTIN ORĂSAN

University of Wolverhampton, UK

ABSTRACT

One of the main concerns when deploying a real-world QA system is user satisfaction. Despite the relevance of criteria such as usability and utility, mainstream research usually overlooks them due to their inherent subjective, user-centric nature and the difficulty of the evaluation involved. This problem is particularly important in the case of real-world QA systems where a "0 results found" answer is not very useful. This paper presents how interaction can be embedded into the QALL-ME framework, an open-source framework for implementing closed-domain QA systems. The changes necessary to improve the framework are described, and an evaluation of the feedback returned to the user for questions that have no answer is performed.

1 INTRODUCTION

The need to access information and find answers to questions in vast collections of documents led to the emergence of the field of Question Answering (QA). Despite extensive research in this field, the accuracy of open domain question answering system, i.e., systems that can answer any question from any collection of documents, is still rather modest. For this reason, real-world question answering systems are usually closed domain, which means that they are built for very specific domains and exploit domain knowledge to answer questions [1].

One of the main concerns when deploying a real-world QA system is user satisfaction. Despite the relevance of criteria such as usability and utility, mainstream research usually overlooks them due to their

inherent subjective, user-centric nature and the difficulty of the evaluation involved. QA benchmarks and evaluation fora (such as TREC[1] or CLEF[2]) usually focus on achieving highly accurate and robust systems, and quantify their performance in terms of precision and recall. We argue that a successful deployment of QA systems should not solely rely on the correctness of the answers, but also on how they interact with users and satisfy their needs. This was acknowledged by the increasing interest in interactive question answering [2–4].

The QALL-ME project [5] has developed a framework for implementing question answering systems for restricted domains. The first implementation of this framework was for the domain of tourism, but it is not bound in any particular way to this domain as demonstrated by its adaptation to the domain of bibliographical information [6]. Despite its flexibility, the framework lacks built-in support for user interaction. This paper demonstrates how it is possible to embed interactivity in the existing framework. The remainder of the paper is structured as follows: Section 2 briefly presents the overall QALL-ME project and framework. The technique used to embed interaction in the framework is presented in Section 3. An evaluation of user satisfaction of the interactivity is presented in Section 4, and the paper finishes with a discussion and conclusions.

2 THE QALL-ME PROJECT AND FRAMEWORK

QALL-ME (Question Answering Learning technologies in a multiLingual and Multimodal Environment) is an EU-funded project with the objective of developing a shared infrastructure for multilingual and multimodal open domain Question Answering.[3] It allows users to express their information needs in the form of multilingual natural language questions using mobile phones and returns a list of ranked specific answers rather than whole web pages.

Language variability, one of the main difficulties of dealing with natural language, is addressed in QALL-ME by reformulating it as a textual entailment recognition problem. In textual entailment, a text (T) is said to entail a hypothesis (H), if the meaning of H can be derived from

[1] http://trec.nist.gov/

[2] http://www.clef-campaign.org/

[3] More information about the QALL-ME project can be found at http://qallme.fbk.eu

the meaning of T. To this end, each question is treated as the text and the hypothesis is a procedure for answering the question [7].

The QALL-ME framework is an architecture for multilingual question answering (QA) systems that can answer questions from structured data sources for freely specifiable domains.[4] In a closed-domain QA system, a question can be viewed as a composition of constraints regarding instances, types and the relations between them. The QALL-ME Framework does the following:

- reliably identifies constraints with respect to a domain modelled by an ontology
- creates the SPARQL query corresponding to the question interpretation
- retrieves the results from a data repository

The first implementation of the framework was for the domain of tourism which will be used for the examples in this paper. The QALL-ME framework is based on a Service Oriented Architecture (SOA) which, for this domain, is realised using the following web services:

1. **Context providers**: used to anchor questions in space and time in this way enabling answers to temporally and spatially restricted questions
2. **Annotators**: identify different types of entities in the input question. Currently three types of annotators are available:
   - named entity annotators which identify names of cinemas, movies, persons, etc.
   - term annotators which identify hotel facilities, movie genres and other domain-specific terminology
   - temporal annotators that are used to recognise and normalise temporal expressions in user questions
3. **Entailment engine**: used to overcome the problem of user question variability and determine whether a user question entails a retrieval procedure associated with predefined question patterns.
4. **Query generator**: relies on an entailment engine to generate a query that can be used to extract the answer to a question from a database. For the tourism demonstrator the output of this web service is a SPARQL query.[5]

[4] http://qallme.sourceforge.net/

[5] SPARQL is a query language defined by the W3C RDF Data Access Working Group which can be used for accessing RDF graphs. It is defined

5. **Answer pool**: retrieves the answers from a database using the query produced by the query generator. In the case of the tourism demonstrator, the answers are extracted from RDF encoded data using SPARQL queries.

The answers, encoded as an RDF graph, are passed to a presentation module which is domain dependent and is not defined in any way by the framework. The interaction between services and the cross-lingual capabilities of the system are realised with the help of a domain ontology, which in the case of the first prototype is described in [8]. The ontology is also used to determine the format in which data is stored and to construct SPARQL queries that are used to access the RDF graph.

The services described above are called by a QPlanner that decides which one should be called depending on the setting: monolingual or cross-lingual (for more details see [5]). One of the drawbacks of the existing QPlanner is that it is only feed-forward, meaning that if a question does not have an answer there is no way to inform the user and allow any form of interaction. In the context of QALL-ME, [9] proposed a way to interact with the user, but the approach does not use the existing framework and requires a completely new implementation. The next section discusses how it is possible to integrate interaction with minimum changes to the existing services.

3 PROVIDING SUPPORT FOR INTERACTION WITH THE USER

Most QA systems have a very basic level of interactivity consisting of independent (question, response) pairs. This type of interaction can quickly become frustrating for the user unless the accuracy of the system is very high. Unlike their open-domain counterparts, closed-domain systems embed enough knowledge to successfully address most correct questions relevant to the domain. However, misunderstandings can occur and the system should provide feedback regarding its 'understanding' of the question, thereby helping the user quickly identify the source of misunderstanding. If the interaction medium is extended to accommodate the user's feedback using either natural language templates (e.g. *No, I did not ask about ... I wanted to know ...*) or via an interactive user interface (Web, mobile clients), then a feedback loop is created which promotes

in terms of the W3C's RDF data model and will work for any data source that can be mapped into RDF. More information can be found at: http://www.w3.org/2001/sw/DataAccess/

a more natural interaction with the user and continuously improves responses from the system.

The treatment of questions that yield no answer should be an important part of any real-world QA system. The first step in clarifying why there are *0 results found* consists of providing feedback regarding the interpretation of the question, as mentioned above. When the cause is not a misunderstanding, but an overly-specific question, the system should explain why there is no answer and suggest possible changes to the original question, encouraging the user to pose additional questions, e.g. by suggesting more general questions with relaxed constraints, or suggesting alternative constraints which do yield relevant answers. For example, if the user asks *Where can I see Matrix in Wolverhampton tonight?* and there is no such screening, it is not useful just to display *No results*. Users may find an answer such as *There is no screening of the movie Matrix in Wolverhampton tonight. Do you want to find out "What movies can I see in Wolverhampton tonight?"* more appropriate. This gives them the opportunity to either accept the suggestion and be presented with the information, or pose a different question initiating a new cycle.

In QALL-ME, processing a question ends once the data is retrieved from the database. Results are presented to the user by a presentation module which is specialised for a particular interaction medium. To enable the behaviour suggested above, the presentation module needs more than just a SPARQL query and the actual results. This is mainly because the SPARQL query only encodes the semantics of the question implicitly, while the presentation module needs explicit meta-data about the system's understanding/interpretation to suggest viable alternatives to the user. In the current QALL-ME architecture, the actual interpretation occurs during the query generation and the entailment engine stages. For this reason, the semantic information is not directly accessible to the QPlanner. The solution is therefore to augment the output returned by the Query Generator (i.e. the SPARQL query) with meta-data which makes the question interpretation explicit, e.g. in terms of EAT and the constraints identified in the question.

In this section, we show how interaction with the user can be achieved without changing the architecture of the QALL-ME Framework. The proposed mechanism consists of two main parts: 1) injecting meta-data explicitly encoding the system's understanding of the question with respect to the underlying domain ontology, and 2) formulating

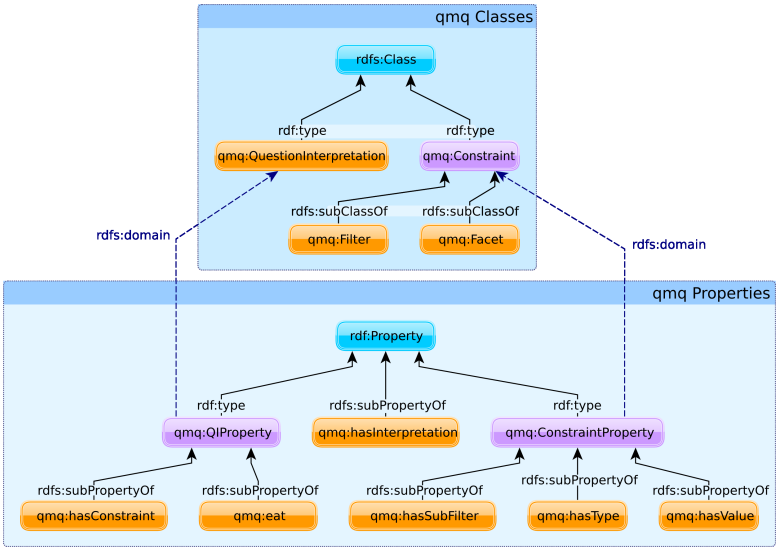


Fig. 1. The qmq terminology

informative answers based on the results given and the question interpretation meta-data. Each of these is discussed below.

3.1 *The qmq mechanism (injecting meta-data)*

In order to provide maximum flexibility, we chose to encode the necessary meta-data using an RDFS terminology, and leave implementation details and extensions to be tailored for each actual application. The terminology contains only the basic concepts: expected answer type, question constraint and question interpretation. The mechanism does not require any changes to the current Web Services specification of the QALL-ME framework, being compatible with the current prototype implementations. An added bonus is that the RDFS terminology used for representing the semantic interpretation is extensible, in line with the generic character of the QALL-ME framework, allowing other types of semantic interpretations to be added in future, without breaking existing applications. Figure 1 presents the `qmq` terminology.

As mentioned above, the output of the Question Generation service is a SPARQL query that can extract the answer to a question. Schematically this query is represented by the following template:

```
[[prefix declarations]]
CONSTRUCT{
  [[triples containing the results: qmo]]
  [[additional information: qmo]]
  [[answer meta-data: qma]]
}WHERE{
  [[triples for identifying solutions]]
  [[filters for grounding the constraints]]
}
```

In order to accommodate the new features, the base SPARQL template is changed by injecting additional information:

```
[[prefix declarations]]
CONSTRUCT{
  [[triples containing the results: qmo]]
  [[additional information: qmo]]
  [[answer meta-data: qma]]
  [[interpretation meta-data: qmq]]
}WHERE{OPTIONAL{
  [[triples for identifying solutions]]
  [[filters for grounding the constraints]]
}}
```

Adding the encompassing OPTIONAL keyword ensures that the meta-data triples from the CONSTRUCT part are generated when querying the data-store, even if no actual solution is found. This means that the presentation module can use this extra information to generate informative answers.

3.2 *Generating Feedback: question interpretation*

Providing feedback regarding the system's understanding helps the user to easily identify misinterpretations, allowing them to rephrase the question in a way that would eliminate the cause of ambiguity or error. The presence of the question interpretation meta-data in the retrieved RDF graph (even in the absence of actual results), enables

the presentation module to describe the systems understanding of the question to the user. Instead of saying: *0 results found*, the presentation module can use the meta-data to say: *No action movies are shown between 12 and 18 October in Wolverhampton, West Midlands*.

For the above example, the triples added to the CONSTRUCT section to enable feedback generation are:

```
prefuri:qi rdf:type qmq:QuestionInterpretation;
       qmq:hasInterpretation "[GENRE] movies
               that will be showed
               during TIME] in [DESTINATION]"@en;
       qmq:eat qmo:Movie.
```

The `prefuri` prefix can be a standard prefix or a dereferenceable URI associated with the user session, enabling real dialogue interaction. The qmq:hasInterpretation property contains a natural language explanation, which is a form of textual feedback to be shown to the user, but it can also be the URI of a resource from a custom repository encoding more complex information (e.g. HTML generation templates).

The actual implementation is application dependent. The content generation templates can be part of a resource which uses/extends the qmq terminology to accommodate different presentation media, multi-linguality, dialogue management, etc. We chose the RDFS semantics for specifying the terminology in order to maintain flexibility. An actual implementation could use a richer representation schema if necessary.

3.3 *Informativeness: Filters*

To find information quickly, users need a certain level of familiarity with the system such as how to best pose questions, the kind of requests the system can address and the data that the system can access. In order to facilitate and create a more natural interaction, the system should go beyond displaying lists of results by generating informative answers. A straight-forward way to do this is by extending the meta-data describing the constraints.

In cases where a question does not yield any results, the system should be able to suggest ways in which the constraints can be successfully relaxed, to find some results. The qmq:Filter instances should therefore mark the value that enforces the constraint. In the following listing we give an example of such meta-data:

```
prefuri:c1 rdf:type qmq:Filter;
       qmq:hasInterpretation "Movies
                    in [DESTINATION]";
       qmq:hasType qmo:Destination;
       qmq:hasProperty qmo:name;
       qmq:hasValue '''FILTER (?destName =
                    "[DESTINATION]").'''.

prefuri:c2 rdf:type qmq:Filter;
       qmq:hasInterpretation "Movies during
                    [TIMEX]";
       qmq:hasType qmo:DatePeriod;
       qmq:hasProperty qmo:startDate;
       qmq:hasValue '''[TIMEX2]'''.
```

Each of the filters indicates the SPARQL filter clause enforcing it via the property qmq:hasValue. Using this meta-data, the presentation module can remove the filter from the SPARQL and pre-emptively check if ignoring this constraint yields any/additional results. This is particularly useful in cases where questions yield no answers: the system can suggest alternatives by removing the filtering clause from the SPARQL, querying the data-store again and identifying alternatives. For example, by relaxing the spatial constraint `c1`, the system can find which Destinations (qmq:hasType) actually yield results, and extract their names (qmq:hasProperty).

The way such information is displayed depends on the application and the medium used. A textual answer could be: *No movies found during 'this week' within 'Wolverhampton'. Try another DatePeriod(5 movies) or another Destination(12 movies)*. On a mobile device, a map can be displayed with the number of movies available for every Destination, and on the Web the interface can be much richer: a full list of answers with reviews, ratings, times. The system can have several strategies for generating answers and only display the highest ranked ones based on their informativeness.

4 EVALUATION OF THE ANSWER FEEDBACK COMPONENT

As explained in the previous section, one way to deal with questions with no answers is to relax or remove their constraints. However, there are various ways in which these constraints can be changed. In this section

we present an evaluation where a set of 6 questions, 3 from the domain of movies/cinema and 3 from the domain accommodation, were shown to users telling them they have no answers and showing them alternative questions that can be generated by the system as part of the response. Users were asked to rate each alternative question on a scale from 1 to 4, 1 indicating a very bad alternative, and 4 corresponding to an excellent alternative. 31 participants were involved in the experiment.

The questions considered for this experiment contained constraints about time (*this weekend*, *tonight*), location (*Wolverhampton*), movie name (*Matrix*), facilities (e.g. *disabled access*), movie genre (*horror movie*), hotel rating and room price. We selected these constraints as they are important in user questions and cover a wide range of concepts from the ontology. The following list of questions was used:

1. Where can I see Matrix in Wolverhampton tonight?
2. Where can I see Matrix in Wolverhampton this weekend?
3. What is the name of a hotel in Wolverhampton with disabled access?
4. What horror movie can I see in Birmingham on Friday night?
5. Where can I find a four star hotel in Wolverhampton?
6. What is the name of a hotel in Wolverhampton where single rooms cost less than £57?

For each of these questions between 4 and 6 alternative questions were produced, in addition to a reply indicating that there are no answers. For example, the following alternatives were proposed for the first question:

1. There is no screening of the movie Matrix in Wolverhampton tonight.
2. Do you want to find out **What movies** can I see in Wolverhampton tonight?
3. Do you want to find out **Where** can I see Matrix tonight?
4. Do you want to find out **When** can I see Matrix in Wolverhampton?
5. Do you want to know Where can I see Matrix in Wolverhampton **tomorrow**?
6. Do you want to know Where can I see Matrix in Wolverhampton **tomorrow evening**?
7. Do you want to find out Where can I see Matrix in **Birmingham** tonight?

A score was calculated for each alternative option for a question as the average rating assigned to it by the users. Tables 1 and 2 present these

Table 1. Results for the question in the movie/cinema domain

		1	2	3	4	5	6	7
Q1	avg.score	2.61	3.42	3.35	3.23	2.45	*2.16*	2.32
	rank	2	1	1	1	2	*3*	2,3
Q2	avg.score	2.65	3.29	3.13	3.06	2.55	2.35	
	rank	2	1	1	1	2	2	
Q4	avg.score	2.61	2.90	2.45	3.13	*1.65*	2.39	
	rank	2	1,2	2	1	*3*	2	

scores. Paired T-test was used to calculate whether there is a statistically significant difference between the answers.

In our domain, spatially and temporally restricted questions are very common, therefore it is important to suggest follow-up questions that are likely to be useful to users. In the first two questions we investigate if the granularity of the temporal constraint from the question (*tonight* vs. *this weekend*) has an impact on the usefulness of alternative time spans. Table 1 presents the average score for each option. The difference between options in the same rank is not statistically relevant. Suggesting a particular alternative value for the temporal constraint (e.g. *tomorrow* instead of *tonight*) was considered less useful than showing all the available alternatives (options 5 and 6 vs. 4 in both questions). This is also true for the spatial constraint (option 7 vs. 3 in Q1). In both questions, the three options that inform the user of all the available alternatives (options 2, 3 and 4) were consistently rated the most useful. This suggests that the users want to know what their options are before committing to a decision.

In question 4 there are three constrains: temporal, spatial and film genre. The results show that option 4 (which movies are available, regardless of the genre) is very useful, while option 5 (which romantic comedies are available) is the least useful and the other options are not statistically distinguishable from the reference option. The results are consistent with our findings so far: ignoring the genre constraint and listing the available movies (option 4) is more useful than pre-emptively modifying the initial question with a viable alternative (options 2, 3, 5 and 6). Users found *romantic comedies* (option 5) a much less desirable alternative to *horror* movies than *action thrillers* (option 6), suggesting

Table 2. Results for the question in the accommodation domain

		1	2	3	4	5	6
Q3	avg.score	2.45	*1.26*	2.23	3.13	*1.55*	3.58
	rank	3	*5*	3	2	*4*	1
Q5	avg.score	2.61	2.84	3.19	3.32	2.29	
	rank	2,3	2	1	1	3	
Q6	avg.score	2.68	*2.03*	3.03	2.94	3.35	
	rank	2	*3*	2	2	1	

that it is not only the constraint type that is important, but also the value specified by the user.

In the accommodation domain we would expect the users to be more rigid regarding temporal and spatial constraints, with factors such as price and star rating also being important.

In question 3 we investigated: alternative facilities - swimming pool (option 2), ignoring the facilities constraint (option 3), alternative destination - Birmingham (option 4), alternative site - cinema (option 5) and alternative type - bed and breakfast (option 6). Only options 1 and 3 do not have statistically significant differences, while options 2 and 5, as expected, have a very low score. The results show that the *disabled access* facility is the most important constraint in the question: users would accept a bed and breakfast or a different city, before considering giving it up. However not all facilities are this important: at the other end of the scale we can imagine room facilities such as *ironing board* or *complimentary newspaper* which usually reflect preference rather than necessity.

In question 5 we investigated: alternative city - Birmingham (option 2), ignoring rating (option 3), alternative rating (option 4), and alternative type - bed and breakfast (option 5). The results suggest that it is useful to inform the users about what hotels are available regardless of star rating, and that a bed and breakfast is less desirable, in this case contrary to the previous question. Therefore, option 4 suggests that when the alternative values are well known (e.g. star ratings for hotels), suggesting an alternative is useful. The last two questions show that the usefulness is influenced not only by the type of the constraints present in the questions, but also by the constrained values.

In question 6, also option 5 which gave the user the full list of hotels and the price charged by each was the most useful (statistically significant). Option 2 was statistically less useful because it just offered a list of hotels, without factoring the price. Picking particular alternatives for constraints (options 3 and 4) scored better than the reference response, but did not prove statistically different.

The analysis presented above shows that suggesting follow-up questions is more useful than just informing the users there were no answers. In most cases, users prefer more general questions which give them information about the available options. Suggesting questions in which one of the initial constraints is changed was not very useful as it may not match the user's preferences. The analysis also revealed that for most constraint types, the usefulness of alternative values depends on the actual values specified in the question, confirming that usefulness depends on the context. This suggests that in order to factor all possible contexts, a system has to automatically learn from users' choices in order to improve its performance.

As a result of the analysis, the prototype was updated to generate responses which help users acknowledge their options in a shortened version allowing preference data to be collected: *The movie 'Matrix' is not on 'tonight' in 'Wolverhampton'. You can either see other **movies**, other **times** (when it is available), or other **destinations** (where it is available).*

To provide such an answer, the system must pre-emptively check that the alternatives proposed actually yield answers. If the user's input matches one of the three follow-up words (e.g. *The movies!*), the system does not have to run the entire pipeline again, but instead, simply returns the answers which have been determined already. In these cases, the interaction has more turns than the initial (question, response) pair.

When two of the constraints cannot be satisfied, the system can say: *The movie 'Matrix' is not on in 'Wolverhampton' (regardless of time). You can see other destinations, or other movies available tonight in Wolverhampton.* However, the generation templates become increasingly complex and the system needs a ranking of constraints to create natural language formulations which are informative and make sense. As result, a rich user interface is perhaps easier to generate.

5 CONCLUSIONS

This paper presented an RDF-based approach for implementing interaction in the QALL-ME framework. An analysis of domain questions revealed that they can be represented as a composition of constraints. This usually takes the form of a conjunction of predicates, as in the following question, *What action movies with Bruce Willis are on in Wolverhampton?* which can be represented as: `hasType(x,qmo:Movie) && hasGenre(x,qmo:action) && hasActor(x,qmo:BruceWillis) && inDestination(x,qmo:Wolverhampton)`. These Boolean predicates correspond to constraints identified in the question by an Entailment Engine which is used to address the problem of language variability. Their truth value can be tested against a data repository by means of inference rules determined by the domain ontology. The Query Generation Web service combines the premises of these rules when generating the WHERE block of a SPARQL query which is used to retrieve the answers to the question, at the same time preserving the semantics of the question interpretation.

The proposed mechanism for allowing user feedback consists of injecting an RDF representation of the question interpretation into the triples of the SPARQL query. Having direct access to this interpretation means that the presentation module can provide feedback, suggest alternative ways in which the question can be asked, and even answer those variations and pre-emptively include the findings in informative answers. The interaction is also more natural from the users' point of view, increasing user satisfaction.

An evaluation of the feedback revealed that simply suggesting follow-up questions is useful, but that usually users want to know all their options in cases where a precise answer cannot be provided. Context is also important, and in order to make competent suggestions the system needs to learn from the choices made by its users. A study is under way to determine whether generating informative answers based on the satisfiability of constraints is useful for the user. A simple feedback loop will be created to allow more interaction based on a single question, via the addition, modification and removal of constraints. This means that the user does not need to pose long or repetitive questions every time they want more information.

REFERENCES

1. Harabagiu, S., Moldovan, D.: Question answering. In Mitkov, R., ed.: Oxford Handbook of Computational Linguistics. Oxford University Press (2003) 560 – 582
2. Hersh, W.: Evaluating interactive question answering. In Strzalkowski, T., Harabagiu, S., eds.: Advances in Open Domain Question Answering. Springer (2006) 431 – 455
3. Rieser, V., Lemon, O.: Does this list contain what you were searching for? Learning adaptive dialogue strategies for interactive question answering. Natural Language Engineering **15**(1) (January 2009) 55–72
4. Quarteroni, S., Manandhar, S.: Designing an interactive open-domain question answering system. Natural Language Engineering **15** (2009) 73–95
5. Sacaleanu, B., Orasan, C., Spurk, C., Ou, S., Ferrandez, O., Kouylekov, M., Negri, M.: Entailment-based question answering for structured data. In: Coling 2008: Companion volume: Posters and Demonstrations, Manchester, UK (2008) 29 – 32
6. Orăsan, C., Dornescu, I., Ponomareva, N.: QALL-ME needs AIR: a portability study. In: Proceedings of Adaptation of Language Resources and Technology to New Domains (AdaptLRTtoND) Workshop, Borovets, Bulgaria (2009) 50 – 57
7. Negri, M., Magnini, B., Kouylekov, M.O.: Detecting expected answer relations through textual entailment. In: Proceedings of 9th International Conference on Intelligent Text Processing and Computational Linguistics, Heidelberg, Germany, Springer (2008) 532–543
8. Ou, S., Pekar, V., Orăsan, C., Spurk, C., Negri, M.: Development and Alignment of a Domain-Specific Ontology for Question Answering. In European Language Resources Association (ELRA), ed.: Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), Marrakech, Morocco (2008)
9. Magnini, B., Speranza, M., Kumar, V.: Towards interactive question answering: An ontology-based approach. In: Proceedings of the Workshop on Semantic Computing and Multimedia Systems (SCMS 2009), Berkeley, California (September 2009)

IUSTIN DORNESCU
RESEARCH GROUP IN COMPUTATIONAL LINGUISTICS
UNIVERSITY OF WOLVERHAMPTON, UK
E-MAIL: <I.DORNESCU2@WLV.AC.UK>

CONSTANTIN ORĂSAN
RESEARCH GROUP IN COMPUTATIONAL LINGUISTICS
UNIVERSITY OF WOLVERHAMPTON, UK
E-MAIL: <C.ORASAN@WLV.AC.UK>

International Conference on Intelligent Text Processing & Computational Linguistics CICLing 2011

CALL FOR PAPERS

12th International Conference on Intelligent Text Processing and Computational Linguistics, will be held in Spring 2011, see www.CICLing.org. Its proceedings are anticipated to be published by Springer in *Lecture Notes in Computer Science*, and the poster session in a special issue of a journal. The following awards are anticipated: best paper, best student paper, best presentation, and best poster.

Its areas of interest include all topics related with computational, linguistics, natural language processing, human language technologies, information retrieval, etc.

Excellent keynote speakers. The most prominent scientists of the field are invited for keynote talks that are published *in extenso* in the proceedings. Each keynote speaker also organizes an additional tutorial or discussion. They usually participate in the tours where you can interact with them in an informal environment.

General interest. The conference covers nearly all topics related to computational linguistics and text processing. This makes it attractive for people from different areas and leads to vivid and interesting discussions and exchange of opinions.

Informal interaction. The conference is intended for a rather small group of professionals. This allows for informal and friendly atmosphere, more resembling a friendly party than an official event. At CICLing, you can pass hours speaking with your favorite famous scientists who you scarcely could even greet in the crowd at large conferences.

Excellent cultural program. The conference is intended for people feeling themselves young in their souls, adventurous explorers of both science and life. Its cultural program brings the participants to unique marvels of history and nature hidden from ordinary tourists.

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 249-267
RECEIVED 23/11/09 ACCEPTED 16/01/10 FINAL 08/03/10

Using Linguistic Knowledge for Fine-tuning Ontologies in the Context of Requirements Engineering

JÜRGEN VÖHRINGER[1], DORIS GÄLLE[1], GÜNTHER FLIEDL[1], CHRISTIAN KOP[1], MYKOLA BAZHENOV[2]

[1]*Alpen-Adria-Universität Klagenfurt*

[2] *National Technical University "Kharkov Polytechnic Institute"*

ABSTRACT

Nowadays ontology creation is on the one hand very often hand-knitted and thus arbitrary. On the other hand it is supported by statistically enhanced information extraction and concept filtering methods. Automatized generation in this sense very often evokes "shallow ontologies" including gaps and missing links. In the requirements engineering domain fine-granulated domain ontologies are needed; therefore the suitability of both hand-knitted and automatically generated gap-afflicted ontologies for developing applications can not always be taken for granted. In this paper we focus on fine-tuning ontologies through linguistically guided key concept optimization. In our approach we suggest an incremental process including rudimentary linguistic analysis as well as various mapping and disambiguation steps including concept optimization through word sense identification. We argue that the final step of word sense identification is essential, since a main feature of ontologies is that their contents must be shareable and therefore also understandable and traceable for non-experts.

Keywords: requirements engineering, ontology engineering, rule mapping, incremental linguistic analysis, WordNet querying

1 INTRODUCTION/MOTIVATION

In the last ten years the job of creating ontologies moved from an Artificial-Intelligence question to a central topic of the exploding semantic web community [17]. Usually ontology creation is very often hand-knitted and thus arbitrary or supported by statistically enhanced information extraction and concept filtering methods. This shift resulted in an uncontrolled growth of ontologies on the one hand and a heightened degree of ontology generality on the other hand. Both developments entail that many existing ontologies are not usable in real-world-applications like requirements engineering.
For supporting systematic and application oriented ontology engineering we previously researched and proposed linguistic guidelines for structuring concept and property notions in OWL represented ontologies [1]. But these guidelines only support ontology generation if domain expertise is sufficiently available in a clearly decoded manner. Obviously specific domain information quite often exists in a not explicit and ambiguous textual format. In this case the elicitation of domain specific concepts still poses many difficulties to the ontology designer.

Hence we developed a linguistic system for supporting the ontology designer to make implicit information easier to trace. Our methodology includes algorithms for tagset mapping, multi-level chunking and wordesense identification. The tagging task is carried forward to QTAG, a probabilistic tagger written in Java [2]. The mapping engine we developed for splitting up standard tags into ontologically relevant tags and specific attributes generates unique input for our rule based chunker.

Some chunking heuristics needed for grouping words to morphological units and syntactical chunks are then used for decoding linguistic candidates for conceptualization nodes in the ontology layer. The main contribution of our research work presented in this paper is an efficient method for incrementally including contextual information in the ontology representation. By combining standard natural language processing methods with certain expansion strategies we definitely improve the usability of standard ontologies. Our approach preserves the basic and partially generic knowledge format for storing domain knowledge and its guided updating.
The approach consists of the following three main steps:

1) linguistic preprocessing: extracting words and phrases from natual language text

2) linguistically guided incremental ontology engineering
3) filling up ontology concept description slots through WordNet based word sense identification

The paper is structured as follows. In section 2 we give an overview of related work. In section 3 we roughly present our theoretical approach including the description of the linguistic pre-processing layers, in particular the mapping, and multi-level chunking steps. In section 4 the output of our ontology refinement tool is shown and an ontology example is described. We also propose a list of rules for ontology element creation. In section 5 we describe our Wordnet based tool for incremental optimization of standard ontologies through wordsense disambiguation. Section 6 gives a summary of the proposal presented in this paper.

2 Related Work

[29] argue that the accuracy and robustness of automatically or semi-automatically engineered ontologies needs to be improved for real-world applications and they propose fuzzy algorithms for real-world-ontology engineering. [28] proposes the use of glosses in ontology engineering for improving the accuracy. We agree that for real-world applications like ontology engineering in requirements engineering projects, automatically generated ontologies might not be suitable and we therefore propose linguistic heuristics for supporting ontology creation and fine-tune ontologies through step-by-step integration of domain knowledge.

Concerning linguistic preprocessing the most relevant linguistic methods used in our approach are tagging and chunking. For tagging English free texts many open source systems like the decision based "Treetagger" [3], the rule- and transformation-based "Brill tagger" [4], the maximum-entropy "Stanford POS Tagger" [5], the trigram based probabilistic "QTAG" [2] etc. are available. For chunking some NLP toolkits exist, e.g. "MontyLingua" [8], "MontyKlu" (an online-version of "MontyLingua" developed by members of our research group in Klagenfurt [9]), the OpenNLP chunker [10] and the "NLTK Toolkit" [11]. These systems mainly provide standardized and acceptable output, but as we know according to practical requirements engineering needs they have not been tested yet.

3 LINGUISTIC PREPROCESSING

3.1 Extended Tagging format

We have chosen “QTAG” as the basis for our extended tagging format which we have adopted for these special purpose. Since QTAG is a java-based, extendable, trainable, language independent tagger, it was easy to integrate in our engineering toolset [6,7]. We extract relevant information from the QTAG output and transform it into the extended tagset format described below. Therefore, we have to use some additional methods and heuristics to elicit semantic information needed during the further processing steps of the engineering workflow. Our enriched tagset consists of standard POS-categories with lists of additional specialized attributes (e.g. v0 with subclass attribute “tvag2”[3]). These attributes are necessary for identifying ontological key relations. Table 1 shows how typical standard part-of-speech tags are extracted from the QTAG output and reassigned using the NIBA tagset notation[4]. Additional information about concrete part-of-speech instances is presented by using fine-granulated attributes[5]. As an example, the verb “is” in QTAG gets the tag <BEZ>. This tag decodes, that “is” is an auxiliary verb with the inherent morphosyntactic values present tense, singular, third person and having “be” as the base form.

Table 1. Mapping Rules for mapping standard tags to attributed tags

BEZ	<=>	v0	verbclass="aux" temp="pres" form="ind" num="sg" ps="3" baseform="be"
NPS	<=>	n0	type="proper" num="pl"

[3] We use “tvag2” for annotating a mono-transitive verb with agentive subject

[4] Central NIBA tags are e.g. v0 (= main verbal element), n0 (= noun), a0 (= adjective) etc.

[5] Typical tag internal attributes are “base form = go” or “type = common” etc.

3.2 Chunking rules

Based on some variants of the X-bar Theory [24] and on some core definitions in the existing NIBA Tag system [25] we composed a set of chunking rules for English for the production of syntactically and morphosynctactically motivated chunks (Table 2).

Table 2. Extended Chunking Rules

Rule (Summands → Result)	**Rule level**	**Rule descriptions**	**Examples**
n0+n0 → n0	1	Compound Noun	blood pressure
[pt0]+a0 → a2	1	Adjective Phrase	very nice
[a0]+a0 → a2	1	Adjective Phrase	bright green
[pt0]+q0 → q2	1	Quantor Phrase	very many
[q0]+q0 → q2	1	Quantor Phrase	one million
[pt0]+adv0 → adv2	1	Adverb Phrase	very often
[adv0]+adv0 → adv2	1	Adverb Phrase	yesterday noon
pron0(type=pers) → n3	1	Noun Phrase	she
v0(verbclass=aux)+[adv0]+v0 → v0(type=complex)	1	Complex Verb	will certainly go
v0(verbclass=aux)+pt0(type= neg)+v0 →v0(type=complex)	1	Complex Verb	would not write
v0+pt0(type=verbal) → v0(type=complex)	1	Complex Verb	wake up
q2+q2 → q2	2	Quantor Phrase	two hundred million
pron0(type=poss)+n0 → n3	2	Noun Phrase	his mother
[det0]+[a2]+[q2]+n0 → n3	3	Noun Phrase	the nice two girls
[det0]+[q2]+[a2]+n0 → n3	3	Noun Phrase	the three busy scientists
p0+n3 → p2	4	Prepositional Phrase	of blood pressure measurement

There are several types of chunking rules, which are arranged in a certain order that should be followed during the chunking process. Summands are the array of input nodes which are needed for building the next resulting upper node of the chunking tree. Some of summands are strictly required for rule producing, they are written without square brackets, but some are not obligatory, they are placed inside brackets.

3.3 Identification of Semantic Roles

Due to the fixed and transparent subject-verb-object (SVO) structure of English, the identification of semantic roles in chunked sentences is by default a quite simple and straightforward task. According to [31] we propose automatic role labeling using partially mainly propbank and verbnet information. The Verbclass Tag in column a in Table 3 of a concrete verb triggers the assignment of a role in column c to a N3(P2) in d via indexation from left to right.

Table 3. Verb classes and their (morpho)syntactic and semantic features[23]

Nr.	Tag (a)	Verbclass (b)	PAS[6] (c) (Argument Structure)	Syntactic context[7] (d)
1	aux	Auxiliary verb	V-fin	_V0
2	eV	Ergative verb	$[TH_i]$[8]	$N3_i$_
3	iV	Intransitive verb	$AG_i/TH_i[]$	$N3_i$_

[6] PAS = "Predicate Argument Structure"; it includes the verb class specific semantic roles and brackets, which decode the argument status of these roles. They can have an external status (subject function) or an internal status (object function).

[7] P2 stands for prepositional phrases; N3 decodes nominal phrases in our framework; N2 is a reduced nominal phrase in predicative function; N3 A2 decodes an adjective phrase in our framework. For further explanation see [23].

[8] The acronyms for the default semantic roles are TH = Thema (neutral object), AG = Agens(the Actor of an Action), GO = Goal (the final point of a process), SO = Source (the starting point of a process), LOC = Location and EXP = Experiencer (a person, who undergoes the process of experiencing something).

4	lokV	Locations verb	TH_i [LOC_j]	$N3_i$_$P2_j$
5	possV	Possessive verb	GO_i [TH_j]	$N3_i$_$N3_j$
6	psychV	Mental verb	TH_i [GO_j]	$N3_i$ _$N3_j$
7	tvag2	Monotransitive verb with agent subject	AG_i [TH_j]	$N3_i$ _$N3_j$
8	tv3	Ditransitive verb	AG_i[TH_j,GO_k /SO_k]	$N3_i$_$N3_j$ P2 $(N3)_k$
9	sentV	Perception verb	EXP_i [TH_j]	$N3_i$_$N3_j$
10	copV	Copula verb	TH_i [$Pred_j$*/$Pred_k$*]	$N3_j$_$A2_j$/$N2_j$
11	tv2	Monotransitive verb without agent subject	TH_i_TH_j	$N3_i$ _$N3_j$

Nevertheless we have to take into account that the phrasal structure sometimes inhibits simple solutions like for example left to right counting of nouns. Thus we used the following algorithm to cope with the problem of phrasal complexity:

- create a set of rules which can operate on simple singular term subjects and objects (e.g. proper nouns and personal pronouns);
- consult the exception database with already assigned verbal subclass tags using training sentences which include higher level argument patterns referring to more complex phrases;
- reconstruct the structure of the primarily assigned phrases if relevant morphosyntactic features don't fit;
- leave open the possibility to manually change wrong/exceptional assignments or to add new information about verb classes, noun phrases and other patterns;

4 Our approach: Linguistically guided incremental Ontology Creation

To avoid using non-fitting ontologies for specific domain relevant demands, particularily in requirements engineering, we take textual descriptions as a starting point for our processing. These texts are generated by filtering those text segments from extensive, domain-relevant documents, in which key words or key phrases occur, which

can be accepted as candidates for concept- or relation-notions in the ontology. Utilizing various filtering strategies, in a first step kewords in a text are identified, which are deemed important for a specific domain. Afterwards sentences from the original requirements that contain those keywords are filtered. A precondition is that these sentences form a cohesive text block. For further information about this process see [30]. It was produced.

In the following an examplary text segment from the medical domain is given, that was automatically selected from the original domain-related requirements text using the previously mentioned keyword filtering strategy:

```
With regard to the monitoring of blood
pressure measurements, it is important to
clearly define time and date at which the blood
pressure of a hemodialysis patient is measured
in each hemodialysis session.
```

We perform the steps of linguistic preprocessing as proposed in section 3 as a first step of transforming the textual input into a domain ontology:

- QTAG output (standard tags)
- Standard Tags transformed to enriched tags
- Chunking output

The XML output of the linguistic preprocessing can be seen in Fig. 1.

This output contains linguistic tags for words, e.g. n0: “regard”, some attributes (e.g. base-form=”regard”, type=”common”, corelex=”coa” etc.) and chunk-tags (e.g. n3: “the monitoring of blood pressure measurements”). This extended linguistic representation of the input text allows mapping and interpretation in the sense of ontology conceptualization. The Table 4 lists rules used for identifying and creating ontology elements from the preprocessed texts.

In the Table 4 some rules for the most relevant linguistic categories like N3, N0, P0 are listed. The interpretation example in the right column shows that an explicit mapping from text to class names is possible. To sum up: all relevant ontology element types are identified in an unambiguous way. The above listed rules transform linguistic annotators to ontological tags. The tags specify words in a unique manner. The output text below shows strings class candidates, relation designators, attribute identifiers and stop word material, which is filtered out during transformation:

With regard to the **monitoring** of **blood pressure measurements**, it is important to clearly define time and date at which the **blood pressure** of a **dialysis patient** is measured in each **hemodialysis session.**[9]

```
<p2>
<p0 idiom="pof1" idiomphrase="with regard to ">With</p0>
<n3>
<n2>
<n0 num="sg" idiom="pof2" derivedPOS="v0" idiomphrase="with
regard to " base-
form="regard" type="common" corelex="coa">regard</n0>
<p2>
<p0 idiom="pof3" derivedPOS="ip0" idiomphrase="with regard to
">to</p0>
<n3>
<det0 form="general" type="def">the</det0>
<n2>
<n0 num="sg" derivedPOS="v0" base-
form="monitor" type="common">monitoring</n0>
<p2>
<p0>of</p0>
<n3>
<n0 desc="compound" type="common">
<n0 desc="compound" type="common">
<n0 num="sg" derivedPOS="n0" base-
form="blood" type="common">blood</n0>
<n0 num="sg" base-
form="pressure" type="common">pressure</n0>
</n0>
<n0 num="pl" base-
form="measurement" type="common" corelex="ate">measuremen
ts</n0>
</n0></n3></p2></n2></n3></p2></n2></n3></p2>
```

Fig. 1. XML output of linguistic preprocessing for a fragment of the example text

[9] Underlined words decode relations , dotted underlines words function as attributes, **Bold words** are interpreted as classes; All other elements are categorized as stop words and filtered out.

Table 4. Rules for mapping of linguistic categories to ontology elements

Rule Nr	***Rule***	***Description***	***OWL Type***	***Example***
1	N0 → Class	Default-Rule for N0 if no Exception applies (see Rule 2)	Class	*"monitoring" → monitoring (class)*
2	N0 (Exception) → Functional Property	Exception for N0 applies, ifN0 is found (according to rules described in [26])	Functional Property	*"time" → time (slot in class blood-pressure)*
3	N3 → Class	Rules 3 to 5 always apply for N3	Class	*"hemodialysis session" → hemodialysis session (class)*
4	N3 → is_a + Class[10]	Rules 3 to 5 always apply for N3	subClassOf Class	*"hemodialysis session" → isa (subClassOf) Seassion (class)*
5	N3 → belongs_to + Class[11]	Rules 3 to 5 always apply for N3	Functional property Class	*"hemodialysis session" → belongs_to (Functional Property) hemodialysis (class)*
6	(P0 |		Functional	*"blood*

[10] The head (right-most part of the compound) becomes a new class

[11] After Rule 4 the head is removed and the remaining of the original N3 becomes a new class

	AUX0_V0) + Singular → Functional Property		Property	*pressure of a hemodialysis patient" →"bp_of"*
7	(P0 | AUX0_V0) + Plural → Object Property	Cardinality of connection is n	Object Property	*"monitoring of blood pressure measurements" →"m_of*"
8	(N0 | N3) +"corelex=hum" → is_a + Class "human"		Functional property Class	*"hemodialysis patient" → isa (subClassOf) human (class)*
9	tvag2 → Functional property + class "agens"		Functional property Class	*"blood pressure is measured" → is_measured (Functional Property) agens (class)*

Fig. 2 shows an ontology fragment which is generated by applying the transformation rules in Table 4. For representation of ontology relevant knowledge OWL [15] and RDF [16] are commonly used [12]. We chose Protégé for representing our ontology example.

[12] Exemplary modern toolkits for ontology engineering are Protégé [12], NeOn [13] and Chimaera [14].

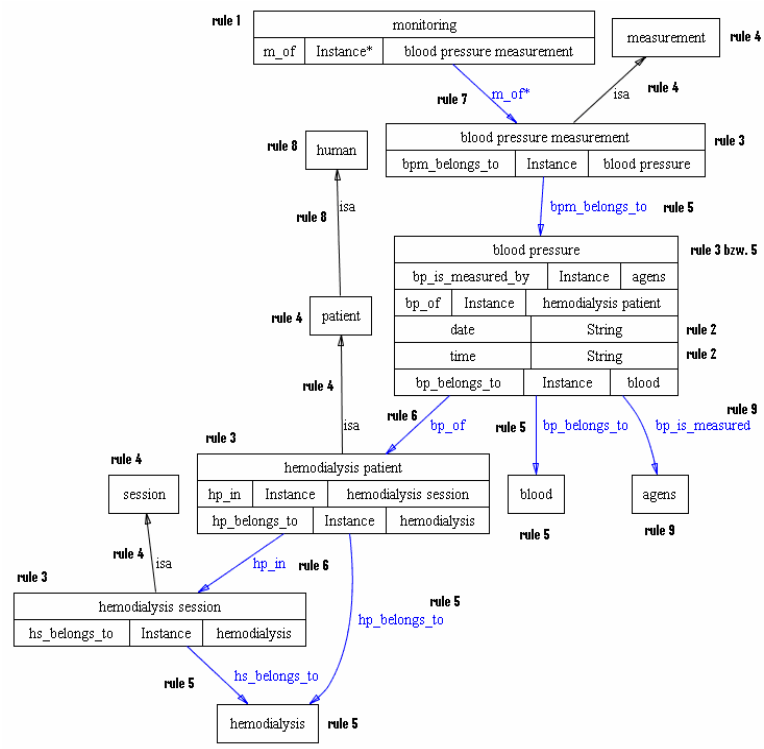


Fig. 2. OWL-Visualization of the above text via the Protégé [12]-Plugin Ontoviz

5 ADDING CONCEPT DESCRIPTIONS THROUGH LEXICALLY DRIVEN WORD SENSE IDENTIFICATION

According to Gruber [27] an ontology is "[…] a formal explicit specification of a shared conceptualization". From this follows that ontology contents has to be sharable and reusable among and across projects within the domain. The ontology fragments that were created stepwise from natural language text based on the heuristics described in section 3 still have empty description slots and are therefore not easily shareable and understandable for non-domain experts. Furthermore missing empty description slots make similarity calculation based on WordNet difficult, since they presuppose a known word sense. Such

similarity calculations are important when new concepts are matched with existing ones in further ontology integration steps. For this reason we propose additional measures for fine-tuning the ontology by refining concept notions that are crucial for the specific domain. The description slots of the ontology (e.g. the example ontology in figure 2) are filled by providing a WordNet related engineering mechanism, which will be described in the following. For ontology concepts that have an empty description slot, WordNet is queried regarding the available word senses and their definitions. The following cases are distinguished:

Case 1: The concept is identified in WordNet and has exactly one meaning. This meaning is automatically assigned to the concept. Example: the concept *blood pressure* was identified from the natural language text, is new to the domain ontology and therefore has an empty description slot. Querying WordNet returns one possible meaning:
blood pressure *-- the pressure of the circulating blood against the walls of the blood vessels; results from the systole of the left ventricle of the heart; sometimes measured for a quick evaluation of a person's health; "adult blood pressure is considered normal at 120/80 where the first number is the systolic pressure and the second is the diastolic pressure"*
This definition is chosen and assigned to the concept, but can still be manually adapted.

Case 2: The concept is identified in WordNet but has more than one possible meaning. In this case the correct meaning is chosen from the list of available word senses. Example: the concept *blood* has an empty description slot and querying WordNet returns the following possible Word Senses, ordered by probability of appearance:
1. ***blood*** *-- the fluid (red in vertebrates) that is pumped by the heart; "blood carries oxygen and nutrients to the tissues and carries waste products away"; "the ancients believed that blood was the seat of the emotions"*
2. lineage, line, line of descent, descent, bloodline, blood line, ***blood****, pedigree, ancestry, origin, parentage, stemma, stock -- the descendants of one individual; "his entire lineage has been warriors"*
3. ***blood*** *-- temperament or disposition; "a person of hot blood"*
4. rake, rakehell, profligate, rip, ***blood****, roue -- a dissolute man in fashionable society*
5. ***blood*** *-- people viewed as members of a group; "we need more young blood in this organization"*

In the medical domain the first, literal, sense of the word is chosen (*fluid that is pumped by the heart*) and assigned to the concept.

Case 3: The concept is not found in WordNet. This usually means that the concept is too specialized and the probability is high that we are dealing with a compound. By applying percolative rules on endocentric compounds we determine the head of the compound, for which again the definitions are determined. Example: the concept *hemodialysis session* is too specialized and hence has no match in WordNet. However it is an endocentric compound with the head *session* and the modifier *hemodialysis.* A search for *session* returns a word sense list as described in case 2:
1. ***session*** *-- a meeting for execution of a group's functions; "it was the opening session of the legislature"*
2. school term, academic term, academic session, ***session*** *-- the time during which a school holds classes; "they had to shorten the school term"*
3. ***session*** *-- a meeting devoted to a particular activity; "a filming session"; "a gossip session"*
4. seance, sitting, ***session*** *-- a meeting of spiritualists; "the seance was held in the medium's parlor"*
The sense 3 (*meeting devoted to a particular activity*) is selected. Regarding the modifier, one word sense is returned as described in case 1:
hemodialysis*, haemodialysis -- dialysis of the blood to remove toxic substances or metabolic wastes from the bloodstream; used in the case of kidney failure*
The definition of hemodialysis session is thus constructed from the definition of its parts:
hemodialysis session *-- a meeting devoted to dialysis of the blood to remove toxic substances or metabolic wastes from the bloodstream; used in the case of kidney failure;*

Case 4: Although the concept is found in WordNet, the description is considered too specialized by a domain expert and therefore inadequate. In this case the chosen description is either manually adapted or the hypernym definition is automatically determined through WordNet querying of its hypernym's concept definition. Example: the concept *hemodialysis* returns the definition
hemodialysis*, haemodialysis -- dialysis of the blood to remove toxic substances or metabolic wastes from the bloodstream; used in the case*

of kidney failure
Since it is considered too specialized the following hypernym definition is established:
dialysis *-- separation of substances in solution by means of their unequal diffusion through semipermeable membranes*

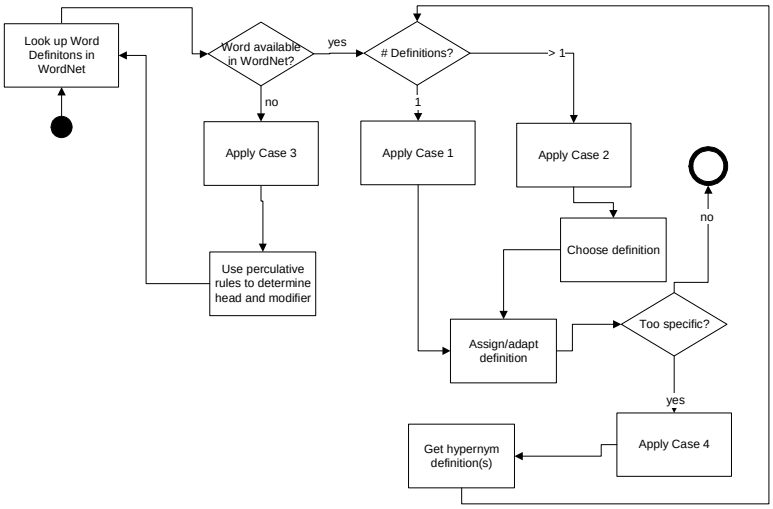


Fig. 3. Word Sense identification based on WordNet

The process (summarized in figure 3) is a guided way of fine-tuning ontologies not according to the quantity but the quality of concepts by adding semantics in order to make them easier understandable for non-experts and facilitate reuse. Using a general lexicon like WordNet allows the standardization of definitions. A prototype implementation is available that utilizes Perl for WordNet querying and allows among other things the listing of available word senses in WordNet, the determination of hypernym definitions and the adaption of definitions where required. Furthermore word sense identification is a bidirectional process as gaps in WordNet (see case 3 above) can be identified and filled. The process above is not limited to WordNet: every lexicon providing definitions can be utilized. More comprehensive lexicons are preferable, for this reason WordNet is a good default choice.

6 SUMMARY

In the requirements engineering domain fine granulated ontologies are necessary for efficient generation of models that can be further used in the application engineering steps. In this paper we proposed a step by step strategy of ontology engineering emanating from manually produced or already statistically filtered text.

Our approach focuses on the diversification of standard tags for optimizing the automatic elicitation of classes, relations and attributes in domain ontologies. Doing this with free text input can only be successful, if certain NLP standard techniques like probabilistic tagging get combined with special procedures like filtering, tag-enriching and chunking. The involved procedures are heuristically founded and follow a multilevel chunking strategy. We described a framework for mapping automatically generated linguistic categories to ontology concepts. Beyond that we showed in detail how these concepts can be refined and therefore optimized based on WordNet, in order to ensure their shareability. Our arguments are supported by a tool set that was developed in our research group for linguistically enhanced requirements engineering The output graph of our example (see chapter 4) proves that creating ontology fragments with linguistic fine-tuning is suitable in the context of requirements engineering.

REFERENCES

1. Fliedl, G., Kop, C., Vöhringer, J.: From OWL Class and Property Labels to Human Understandable Natural Language. In: Lecture Notes in Computer Science, 12th International Conference on Applications of Natural Language to Information Systems, NLDB 2007, Paris, France (2007)
2. The QTAG POS-Tagger http://morphix-nlp.berlios.de/manual/node17.html
3. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods – In Language Processing (1994)
4. Brill, E.: Unsupervised learning of disambiguation rules for part of speech tagging. In Proceedings of the 3rd Workshop on Very Large Corpora, pages 1–13. (1995)
5. Goldwater, S., Griffiths, T.: A fully Bayesian approach to unsupervised part-of-speech tagging. In Proceedings of ACL (2007)

6. Fliedl G., Kop Ch., Vöhringer J., Winkler Ch.: NIBA Project: Overview. In: ER 2005 24th International Conference on Conceptual Modeling. Alpen-Adria-Universität Klagenfurt (2005)
7. Kop, C., Mayr, H.C. and others.: Tool Supported Extraction of Behavior Models. In: Information Systems Technology and ist Applications, ISTA' 2005 4th. International Conference, 23. - 25. May 2005, Palmerston North, New Zealand. GI 2005, 114-123, (2005)
8. Montylingua :a free, commonsense-enriched natural language understander, http://web.media.mit.edu/~hugo/montylingua
9. Monty Klu Web v0.1, http://montyklu.knospi.com
10. OpenNLP project website: http://opennlp.sourceforge.net/
11. Bird. S., Klein, E., Loper, E.: Natural Language Processing in Python, 2008, http://nltk.org/doc/en/book.pdf
12. Holger Knublauch, Ray W. Fergerson, Natalya F. Noy and Mark A. Musen: The Protege OWL Plugin: An Open Development Environment for Semantic Web Applications
13. Haase P., Lewen H., Studer R., Tran T. Erdmann M. d'Aquin M., Motta E.: The NeOn Ontology Engineering Toolkit. In WWW 2008 Developers Track. April 2008.
14. McGuinness, D. L.. "Conceptual Modeling for Distributed Ontology Environments." In Proceedings of the Eighth International Conference on Conceptual Structures Logical, Linguistic, and Computational Issues (ICCS 2000). Darmstadt, Germany. August 14-18, 2000.
15. OWL Web Ontology Language Overview , Deborah L. McGuinness and Frank van Harmelen, Editors. W3C Recommendation, 10 February 2004, http://www.w3.org/TR/2004/REC-owl-features-20040210/. Latest version available at http://www.w3.org/TR/owl-features/.
16. The Resource Description Framework (RDF): http://www.w3.org/TR/rdf-concepts/
17. Mustafa Jarrar and Robert Meersman (2008). "Ontology Engineering -The DOGMA Approach". Book Chapter (Chapter 3). In Advances in Web Semantics I. Volume LNCS 4891, Springer.
18. SNOMED Clinical Terms: Overview of the Development Process and Project Status: Michael Q. Stearns, MD', Colin Price, MPhil, FRCS, Kent A. Spackman, MD, PhD" , Amy Y. Wang, MD'
19. The Unified Medical Language System: an informatics research collaboration.: Humphreys BL, Lindberg DA, Schoolman HM, Barnett GO.
20. Tom Gruber (1993). "A Translation Approach to Portable Ontology Specifications". In: *Knowledge Acquisitions* 5, (May): 199-220.
21. Natalya F. Noy and Deborah L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Stanford University,
22. Asunción Gómez-Pérez, Mariano Fernández-López, Oscar Corcho (2004). *Ontological Engineering: With Examples from the Areas of Knowledge Management, E-commerce and the Semantic Web*. Springer, 2004.

23. Fliedl G., Kop Ch., Mayr H. C., Hölbling M., Horn Th., Weber G., Winkler Ch.:
Extended Tagging and Interpretation Tools for Mapping
24. Black, C.A.: A step-by-step introduction to the Government and Binding theory of a syntax. In: Notes on Linguistics 2 (1996), Mai, Nr. 73
Weber, G.: NIBA<tag> Aspekte der Implementierung eines erweiterten Taggers für die automatische Textannotation in NIBA. Master Thesis. Alpen-Adria-Universität Klagenfurt (2007)
25. Mayr H. C., Kop C.: A User Centered Approach to Requirements Modeling. In Proceedings of Modellierung 2002, Köllen Verlag, Bonn 2002, pp. 75 – 86.
26. Gruber T.: A translation approach to portable ontologies. Knowledge Acquisition, 5(2), pp. 199-220, 1993
27. Jarrar M.: Towards the Notion of Gloss, and the Adoption of Linguistic Resources in Formal Ontology Engineering Proceedings of the 15th International World Wide Web Conference, Edinburgh, Scotland, pp. 497-503, ACM Press, May 2006
28. Lau R. Y. K., Li Y., Xu Y.: Mining Fuzzy Domain Ontology from Textual Databases, In: WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence, Publisher: IEEE Computer Society, November 2007
29. Perkonigg M.: Linguistische Aspekte des Attempto Controlled English (ACE). Masterarbeit, Alpen-Adria-Universität Klagenfurt, 2009.
30. Giuglea A.-M., Moschitti A.: Semantic role labeling via FrameNet, VerbNet and PropBank In: ACL-44: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics; Publisher: Association for Computational Linguistics; July 2006

JÜRGEN VÖHRINGER
INSTITUTE OF APPLIED INFORMATICS,
RESEARCH GROUP APPLICATION ENGINEERING,
ALPEN-ADRIA-UNIVERSITÄT KLAGENFURT

DORIS GÄLLE
INSTITUTE OF APPLIED INFORMATICS,
RESEARCH GROUP APPLICATION ENGINEERING,
ALPEN-ADRIA-UNIVERSITÄT KLAGENFURT

GÜNTHER FLIEDL
INSTITUTE OF APPLIED INFORMATICS,
RESEARCH GROUP APPLICATION ENGINEERING,
ALPEN-ADRIA-UNIVERSITÄT KLAGENFURT

CHRISTIAN KOP
INSTITUTE OF APPLIED INFORMATICS,
RESEARCH GROUP APPLICATION ENGINEERING,
ALPEN-ADRIA-UNIVERSITÄT KLAGENFURT

MYKOLA BAZHENOV
COMPUTER AIDED MANAGEMENT SYSTEMS DEPARTMENT,
NATIONAL TECHNICAL UNIVERSITY
"KHARKOV POLYTECHNIC INSTITUTE"

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 269-283
RECEIVED 03/12/09 ACCEPTED 16/01/10 FINAL 10/03/10

Incorporating TimeML into a GIS

MARTA GUERRERO NIETO, MARÍA JOSÉ GARCÍA RODRÍGUEZ, ADOLFO URRUTIA ZAMBRANA, WILLINGTON SIABATO, MIGUEL-ÁNGEL BERNABÉ POVEDA

Universidad Politécnica de Madrid, Spain

ABSTRACT

This study approaches a methodology for the integration of temporal information belonging to a historical corpus in a Geographic Information System (GIS), with the purpose of analyzing and visualizing the textual information. The selected corpus is composed of business letters of the Castilian merchant Simón Ruiz (1553-1597), in the context of the DynCoopNet Project (Dynamic Complexity of Cooperation-Based Self-Organizing Commercial Networks in the First Global Age), that aims to analyze the dynamic cooperation procedures of social networks.

The integration of historical corpus into a GIS has involved the following phases: (1) recognition and normalization of temporal expressions and events in 16th century Castilian following the TimeML annotation guidelines and (2) storage of tagged expressions into a Geodatabase. The implementation of this process in a GIS would allow to later carrying out temporal queries, dynamic visualization of historical events and thus, it addresses the recognition of human activity patterns and behaviours over time.

Keywords: TimeML, historical corpus, GIS, semantic annotation.

1 INTRODUCTION

Events are placed in time and space; both are needed to get a full representation of historical events. Traditionally these two components have been studied separately by History and Geography, although both disciplines require for their understanding and reasoning the joint consideration of space and time of any given phenomenon. The Geographic Information Systems (GIS) have greatly facilitated management, editing analysis and visualization of geographic data related to the territory. However, the use of GIS as a tool of spatio-temporal analysis and dynamic representation of historical facts with the purpose of reviewing and strengthening many aspects of geographic history [1] is an issue which has been contemplated since de 70's [2]. At the present time one of the subjects which is currently an open research line is the incorporation of reasoning and quantification of time and the recognition of temporal patterns.

One of the objectives of the DynCoopNet Project, in which this research study is framed, is to inquire into the dynamics of cooperation commercial networks that were established in the First Global Age (1400-1800). Our contribution to that project is to encourage the use of GIS in social science and humanities and approach the studies of confrontation and review of historical events, incorporating for that end tools capable of carrying out analysis in a temporal GIS.

In this paper we propose a methodology for incorporation of the time variable into a GIS. First, we aim at identifying temporal expressions using TimeML that allows describing both definite and indefinite temporal expressions; it also allows defining the events and establishing temporal relationships inspired by Allen's temporal algebra [3]. Second, we propose the incorporation of temporal concepts included in written texts into a GIS. For that purpose we intend to extract the temporal information from a historical corpus made up of letters written in 16th century Castilian by the merchant Simón Ruiz by using Natural Language Processing (NLP) tools.

In the next section the temporal component will be studied in depth taking into account the two research areas as described. Subsequently the annotation guide used to extract temporal information will be described, and in the fourth section the methodology used in the identification and normalization of temporal expressions of the Spanish language will be shown, also paying attention to the integration of TimeML into the Geodatabase. Finally the conclusions of the study and future work will be outlined.

2 TIME

The temporal information has been researched from different disciplines. From the computational viewpoint, temporal information processing has aroused great interest in the scientific community, as attested by the large number of workshops which have taken place in the area of creation of extraction and temporal analysis tools (TERQAS [4], TANGO [5], DAGSTUHL [6], MUC [7]); in the area of temporal semantic annotation languages (TIDES [8], TimeML [9]); in the area of annotation systems (TERSEO [10], TARSQI [29]) and in different evaluation workshops (TERN [11], TempEval [12]). Likewise, in the field of geographic information a large number of studies have approached this subject. The incorporation of the temporal variable into GIS has been investigated along the 1990 decade. The first studies focussed on the management of time in the data bases [13] [14]; recent research is focussed on spatio-temporal modelling. This data modelling is being carried out from a conceptual framework and a technical viewpoint [15] [16]. Many models are based on the addition of the temporal variable within the spatial databases, restricted to individual layers, such as the 'Spatio-temporal Cube' Model [17], 'Snapshot' Model [18] or the Composite Spatio-temporal Model [19]. The most recent spatio-temporal models are associated to objects: Moving Object Data Models [20], Spatio-Temporal Object-Oriented Data Model [21], and Object-Relationship Model [22]). Other studies are currently developing advances in spatio-temporal databases, such as the Intentionally-Linked Entities Model (ILE) [23], which allows representing complex entities and establishing a relational context. However, although great conceptual efforts have been made for the building of databases and prototypes based on spatio-temporal databases and their implementation for a particular application, there is no global model as yet that might be used for any application.

As discussed above, this study will be used markup languages to integrate temporal information of historical phenomena. The specific markup language for geographic information is GML (Geographic Markup Language), delineated by the Open Geospatial Consortium (OGC) in 2000. This language has been defined for modelling, transportation and storage of geographic information [24]; however, even though it has a temporal reference system, it lacks a detailed description. Actually there have been initiatives to extend the geographic markup language over the temporal domain so as to being able to represent this type of information [25]. The choice of the

temporal markup language has been made subservient to the geographic tool since an already known language was needed and used for this tool, at the same time enabling description of the temporal variable and information interchange.

Temporal information stored as metadata of the geographic data of a document may be appropriate for queries related to the date of that document but they are insufficient if event duration is queried or other dates other than the publishing date wants to be obtained [26]. For the incorporation of temporal expressions coming from a natural language document into a database, those expressions must be presented with a certain structure and they must be subjected to normalization. To this end the TimeML markup language has been used.

3 TIMEML TEMPORAL MARKUP LANGUAGE

The TimeML temporal semantic annotation is a linguistic specification to annotate events and temporal expressions, that it provides a systematisation for the extraction and representation of temporal information and for information interchange. It came into being with the aim of annotating newspaper articles, although, as we will see, it may be extended to another type of text information. The most characteristic properties of this language are: interpretation of temporal expressions, temporal annotation of events, and arrangement of events to others through a temporal anchorage. TimeML, developed in 2002 [4] [5], is being consolidated as an ISO standard (ISO WD 24617-1:2007), and it is compatible with ISO 8601 which specifies the standard notation to store dates. It should be noted that it has been approved as an annotation language for TempEval, whose objective is to evaluate the automatic systems in text semantic analysis [12].

TimeML combines and extends characteristics of other temporal annotation standards such as STAG [27] (guide to annotate events and time in newspaper texts, whose tag for temporal information is TIMEX) and TIDES [8], developed to mark temporal expressions of a document and identify the value of the temporal expression (TIMEX2). In TimeML the temporal expressions are marked with the TIMEX3 tag, which intends to indicate an improvement in relation to previous tags.

For treatment of the different *timexes* there are different annotation languages and an annotated corpus for the English language, TimeBank, made up of 183 articles of the US press [28]. There are also automatic tools of temporal annotation, TARSQI [29] and TERSEO

[10] and temporal ontologies, among them Time Ontology and its forerunner DAML Time 2006 [20] standing out since it is related to TimeML. Yet the majority of those resources cannot be used for the Spanish language or they have ended up being obsolete, so the creation of Spanish corpora annotated in TimeML and the development of specific tools would be necessary.

3.1 TimeML: description and characteristics

This markup language has three basic tags: TIMEX3, EVENT, SIGNAL and three link subtypes: TLINK, ALINK and SLINK. Next a brief explanation of each tag is presented:

- TIMEX3 is used to mark temporal expressions: *21st March 2001, yesterday, at 6 PM, next year.*
- EVENT is used to mark events mentioned in a text: *to occur, to believe, to study, to begin.*
- SIGNAL is used to annotate temporal signals: *before, after, during*.
- TLINK is used to mark temporal relationships: *Louise went to Romania from the 21st to the 27th of March* (the temporal information is related to the event *to go*).
- ALINK is used to annotate aspectual relationships: *Mary will begin presentation of her paper at 12 noon* (the verb *will begin* is showing a phase of the event).
- SLINK is used to annotate relationships of modality or evidentiality: *John said he would go to Romania in March. (conjecture is made before the realisation of the event).*

TimeML offers the possibility of expressing different granularities. It owns four types for time expression (TIMEX3):

- DATES is used for expressions referring to a calendar: *on the 22nd of March 2010, last Sunday, yesterday morning.*
- DAY TIMES is used for a temporal expression less than a day: *this afternoon, at twenty minutes to three*. Attention should be paid to the distinction between these two types of times because of the different granularity of the expressions.
- DURATION is used to describe a duration in time: *for four days, two years ago.*

- SET is used for expressions of repetition in time: *twice a week, every eight days.*

The natural language does not have a single way of expressing a specific granularity but there may be different temporal expressions referring to the same granule. Granularity is the level of detail with which the time is measured; it may be stated that natural language does not have a canon for time expression [26]; however it is known that the granularity of linguistic temporal expressions differs adapting to the Gregorian calendar. Equivalencies may be found between the natural language and this calendar, at least in western languages. Hence the differences proposed to model time following the calendar [30] [31].

In order to ensure the consistency of data structure in all the documents, a DTD has been used. It allows defining the data format and the document structure, its elements and tag nesting. Thus, those documents may be validated since the element structure of the elements and their description may be known.

4 METHODOLOGY FOR INCORPORATION OF TEMPORAL EXPRESSIONS INTO A GIS

The purpose of this methodology is the incorporation of the temporal variable, described by means of a temporal markup language coming from a text information, into a GIS. The corpus used comes from a selection of 20 letters of the Spanish merchant Simón Ruiz, dated in the 16th century.

The procedure has been divided into three steps:

- Automatic identification of the temporal expressions of the corpus using the *GeoParser* server.
- Manual normalization of the temporal expressions with TimeML.
- Incorporation of TimeML into a Geodatabase.

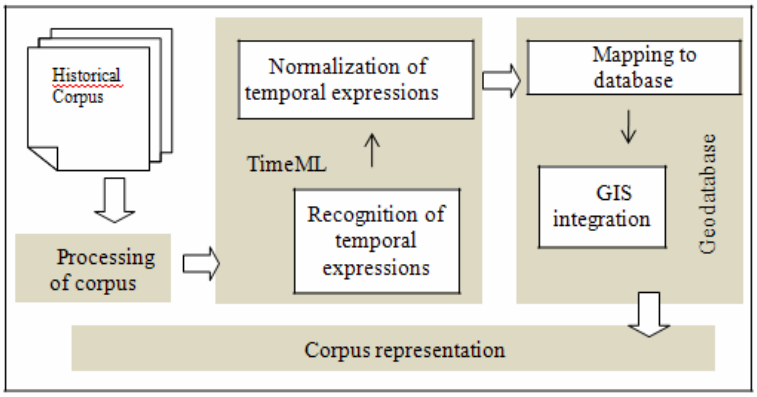


Fig. 1. Diagram of the processes included in the proposed methodology. *Source : author of the study.*

4.1 Identification of the temporal expressions

No automatic tool associated to TimeML has been developed yet for recognition of temporal expressions in the Spanish language. In view of this lack, we have opted to use *GeoParser*, a text processing tool that identifies these expressions and recognizes the geographic scope of the document, as well as the whole geographic entities mentioned throughout the document. This tool, a spinoff of the DIGMAP Project [32], has been used with the aim of assisting in the detection of temporal expressions, disambiguation and assignment of a geographic environment to those expressions. In this case the document to deal with is written in old Castilian, so that it is necessary to carry out some modifications in the lexicon of the application, so that these expressions are properly recognized, therefore normalized.

This recognition process implements the Named Entity Recognition Method (NER), based on seeds and supplemented by searches in gazetteers and register queries containing names of historical periods. This approach has been extended with rules of expression for recognition of dates, durations and frequencies extracted from the TimeML annotation schema. This process has been carried out manually, marking the temporal expressions recognized by the *GeoParser*, as well as those that were not identified.

The procedure for recognition of these temporal expressions is facilitated by the `DateTime` tag, with which *GeoParser* marks and describes the already identified expressions.

An additional advantage of the utilization of *GeoParser* is that it facilitates the incorporation of the expression identified in the GIS since not only identifies but it also infers the geographic context to which it is related. For this purpose it relies on a gazetteer having a register of placenames, historical periods and description of their properties, e.g. it contains place types, coordinates, temporal intervals, hierarchies, alternative names and semantic associations. The gazetteer used [33] integrates data from multiple sources, among them GeoNames and the directory of historical periods ECAI [34].

4.2 Normalization of temporal expressions in TimeML

The normalization of temporal expressions has been carried out semi-automatically due to the fact that at the present time *GeoParser* does not normalize recognized expressions; besides, since we are dealing with a historical corpus of old Castilian, the adaptation of TimeML has been necessary in accordance with that linguistic variety and the incorporation of its rules. This adaptation has already been carried out for the English, Chinese and Italian languages, the adaptation for the Spanish language being underway [35] [36]. It is relevant to note that for the time being no studies have been undertaken about the old correspondence in TimeML. So far the ground on which work about temporal extraction has been done is newspaper or legal texts [37].

After identification of temporal expressions in Renaissance Castilian, the next step has been the TimeML annotation of those expressions that had not been previously normalized. From the beginning XML has been chosen as language, without forgetting that TimeML is not dependent on this format.

Next an example is shown of the corpus of normalization of these temporal expressions in TimeML where the guide values appear: `TIMEX3, EVENT` and `TLINK.`

"A primero de agosto recibi la de v.m. de 12 del pasado" (*On the first of August I received yours of the 12th of the past month from your honour)*

```
<TIMEX3 tid="tid12" type="DATE" value="1570-08-
01" anchorTimeID="tid11">primero de
agosto</TIMEX3>

<SIGNAL sid="sid1"> a </SIGNAL>

<EVENT eid="eid28" aspect="PERFECTIVE"
mood="NONE" pos="VERB" vform="NONE"
class="OCCURRENCE" tense="PAST"
stem="RECIBIR">recibi</EVENT>

<TIMEX3 tid="tid13" type="DATE" value="1570-07-
12" anchorTimeID="tid11">12 del pasado </TIMEX3>

<SIGNAL sid="sid2"> de </SIGNAL>

<TIMEX3 tid="tid11" type="DATE" value="1570-08-
08"> ocho de agosto de 1570</TIMEX3>

<TLINK relType="INCLUDEs" lid="lid31"
timeID="tid12" relatedToEventInstance="eid28"
signalId="sid1"/>

<TLINK relType="BEFORE" lid="lid32"
timeID="tid13" relatedToEventInstance="eid28"
signalId="sid2"/>
```

The type of temporal information we are likely to find in the letters is varied and rich due to the rhetoric of that time and the type of document, with temporal expressions of the type: "*a tantos dias*" (*within that many days*), "*de pocos días a esta parte*" (*from a few days hither*) or "*diez del que viene*" (*tenth of next*). As can be observed by the example, the linguistic expressions used in the letters may be deictic, i.e. knowledge of the narrative moment in which the expressions are framed is needed to be able to pin down the time interval comprised by the expression. The corpus used allows using the temporal metadata in order to determine at which moment the events occur, so as to be able to locate them on a timeline. This is achieved with the `AnchorTime` attribute, as in the example, allowing establishment of a temporal axis.

In order to arrange the events of the corpus, there are two ways of proceeding: extrinsic and intrinsic. The former means arranging the letters of Simón Ruiz only taking into account the document publishing dates, i. e. the metadata. The latter consists of arranging all the temporal expressions appearing in the document. Since this one is a more sophisticated process, it is necessary to deal with the entire information of the corpus.

The tag marking the temporal relationships is the `TLINK`, those relationships being based on the thirteen binary relationships of Allen's temporal algebra. The `TLINKs` represent the temporal relationships existing between two events, two times, or between an event and a time. In the example, the event would be "*recibi*" (*I received)*, which is accompanied by two temporal expressions, "*primero de agosto" (the first of August)* and " *12 del pasado" (the 12th of the past month)*. The temporal relationships between these three elements are marked with the `TLINK` tag, as can be seen in the example.

It is relevant to remember that the XML is not inherent to the TimeML since the latter may turn into other formats; as a matter of fact, a web annotation tool is being developed that generates data in database tables from text annotation [39].

4.3 Incorporation of TimeML as part of a geodatabase

After having annotated the corpus in TimeML we go on with the integration of the text in the geographic information system. The GIS have different data formats, all of which assume a database-oriented structuring of information: geodatabases, tables in MS Access, tables in MS Excel (with certain restrictions), etc. The porting of TimeML annotation into any of these formats is guaranteed by the fact that TimeML does not allow for recursive entities, and hence it provides a stable, predictable structure, so that a relational database could be designed to store the information contained by TimeML annotation.

In order to automate the transfer of information, creation of a mapping algorithm between both structures (database and corpus) is required with the purpose of saving and extracting the information freely. Such tool could be implemented as an internal module of the database manager or as an independent software component [38]. Consequently both entities (geodatabase and the DTD defining TimeML) would be practically identical. This would facilitate introduction of the information, and in addition, the annotated expressions of the corpus would not undergo any change. The XML and the geodatabase turn into the two faces of the storage of the temporal expressions.

Finally, having the information within the geodatabase, the representation of the annotated corpus will depend on peculiarities the historical events described; for example if we were dealing with the binnacle of a ship's captain, the representation of shipping routes, goods transported, oceanic currents, winds and storms could be

emphasized, but if the corpus was made up of texts describing land voyages, the representation details would be substantially different, highlighting other aspects. In this regard, as far as representation of the linguistic annotation of the corpus is concerned, we may add that this is a future line of research which involves the dynamic display of events.

5 CONCLUSION AND FUTURE WORK

A methodology has been designed for the recognition and normalization of temporal expressions following the TimeML specifications; the procedure followed has been presented and the union of the two scopes for the development of the temporality in the GIS has been pointed out. Likewise the methodology for link linguistic corpora and the Geographic Information Systems has been presented.

The functional advantages of integrating document texts in natural language and the representation of their temporality in a GIS have been exposed.

The advantages of utilization of the TimeML have been described: its standard character, its format as a database, its applicability to any language by providing a defined grammar and above all, by allowing the arrangement of events on a timeline. TimeML, as other markup languages, allows massive treatment of text information.

Some of the limitations to carry out the proposal have been described: (a) to achieve the representation of the temporal information tagged in the corpus, the Geographic Information System should have a spatio-temporal database allowing storage and querying of the information coming from the corpus, i.e. a temporal GIS reflecting the TimeML; (b) corpora tagged in TimeML are scarce for languages other than English which prevents the use of automatic learning techniques and gives rise to the use of semi-automatic and manual annotation; (c) adaptation of TimeML guide to old Castilian is needed to facilitate identification of temporal expressions in this type of texts.

As future work we intend to get the recognition, normalization and quantification of temporal expressions in wider Spanish historical corpora as well as the integration of temporal and spatial linguistic annotation. In addition we have anticipated the creation of an analysis tool allowing the utilization of temporal expressions at the time of specifying the query within a spatio-temporal GIS, as well as the extension of the SQL queries with diffuse temporal expressions and

temporal proper names, i.e. we seek facilitate the implementation of natural languages queries containing temporal expressions in a GIS.

ACKNOWLEDGMENTS

This research has been developed within the framework of the Dyncoopnet Project, financed by a Complementary Action of the Spanish Ministry of Science and Innovation (HUM2007-31128-E). The authors wish to express their sincere thanks to Dr. Roser Saurí from Barcelona Media for her support and guidance in this study.

REFERENCES

1. Gregory, I.N., Ell, P.S.: Historical GIS: Technologies, Methodologies and Scholarships. Cambridge Univesity Press (2007)
2. Sack R.D.: Chronology and Spatial Analysis. Annals of the Association of American Geographers, vol.64, pp.439--452 (1974)
3. Allen, J. F.: Maintaining knowledge about temporal interval. Communications of ACM, 26, 11, pp. 832--843 (1983)
4. Pustejovsky, J.: TERQAS: Time and Event Recognition for Question Answering Systems. ARDA Workshop, MITRE, Boston (2002).Available at http://www.timeml.org/site/terqas/index.html
5. TANGO (TimeML Annotation Grafhical Organizer), http://www.timeml.org/site/tango/index.html
6. Dagstuhl Seminar Proceedings. Annotating, Extracting and Reasoning about Time and Events, http://drops.dagstuhl.de/opus/volltexte/2005/313/
7. Advanced Research Projects Agency. Proceedings of the Sixth Message Understanding Conference (MUC-6) (1995). Software and Intelligent Systems Technology Office.
8. Ferro, L., Gerber, L., Mani, I., Sundheim, B., & Wilson, G.: TIDES 2005 Standard for the Annotation of Temporal Expressions. The MITRE Corporation (2005)
9. Pustejovsky, J., Castaño, J., Ingria, R., Saurí, R., Gaizauskas, R., Setzer, A., Katz, G., Radev, D.: TimeML: Robust specification of event and temporal expressions in text. In: AAAI Spring Symposium on New Directions in Question-Answering (Working Papers), Stanford, CA, pp. 28--34 (2003)
10. Saquete, E., Martínez-Barco, P., Muñoz, R., Negri, M., Speranza, M., Sprugnoli, R.: Automatic resolution rule assignment to multilingual Temporal Expressions using annotated corpora. In: Proceedings of the Thirteenth International Symposium on Temporal Representations and Reasoning, pp. 218--224 (2006)

11. DARPA TIDES (Translingual Information Detection, Extraction and Summarization). The TERN evaluation plan: Time Expression Recognition and Normalization. Working papers, TERN Evaluation Workshop (2004). Available at http://timex2.mitre.org/tern.html
12. Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Katz, G., Pustejovsky, J.: SemEval-2007. Task 15: TempEval Temporal Relation Identification. In: Proceedings of SemEval 2007, 4th International Workshop on Semantic Evaluation, ACL, Prague, pp.75--80 (2007) Available at http://nlp.cs.swarthmore.edu/semeval/tasks/index.php
13. Roddick, J.F. y Patrick, J.D.: Temporal semantics in information systems-A survey. Information systems, vol. 17, pp. 249--267 (1992)
14. Tansel, A.U.: Temporal databases: theory, design, and implementation, Benjamin/ Cummings series on database systems and applications. Benjamin/ Cummings Pub. Co., Redwood City, Calif. (1993)
15. Langran G.: Time in Geographic Information System. Taylor & Francis (1992)
16. Peuquet, D.J.: Representations of Space and Time. Guilford Publications, New York (2002)
17. Hägerstrand, T.: What about people in regional science? Papers of the Regional Science Association, vol. 24, pp. 7--21 (1970)
18. Armstrong, M. P.: Temporality in spatial databases. In: Proceedings GIS/LIS'88, San Antonino, Texas, USA, Vol. 2, pp. 880 --889 (1988)
19. Langran, G. y Chrisman, N.: A framework for temporal geographical information systems. Cartographica, vol. 25, no. 3, pp. 1--14 (1988)
20. Erwig, M., Guting, R.H., Schneider, M., Vazirgiannis, M.: Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. GeoInformatica, vol. 3(3), pp-- 265-291 (1999)
21. Montgomery, L. D.: Temporal Geographic Information Systems Technology and Requirements: Where We are Today. Thesis, Department of Geography, Ohio State University, USA (1995)
22. Claramunt, C., Parent, C., Spaccapietra, S., Theriault, M.: Database Modeling for environmental and Land Use Changes. Geographical Information and Planning, Chapter 20, Springer-Verlag (1998)
23. Kantabutra, V.: A new type of database system: Intentionally-Linked Entities: a detailed suggestion for a direct way to implement the entity relationship data model. CSREA: EEE 2007, pp. 258--263 (2007)
24. OpenGIS, Geography Markup Language (GML). Encoding Standard, OGC 07-036, v. 3.2.1. Available at http://portal.opengeospatial.org
25. Zipf, A., Krüger, S.: TGML: Extending GML by Temporal Constructs: A Proposal for a Spatiotemporal Framework in XML. In: ACM-GIS 2001, the Ninth ACM International Symp. on Advances in Geographic Information Systems, Atlanta, USA (2001)
26. Llidó Escrivá, D. M.: Extracción y recuperación de la información temporal, Thesis Universidad Jaume I, Castellón, Spain (2002)
27. Setzer, A., Gaizauskas, R.: Annotating Events and Temporal Information. In: Newswire Text, In LREC 2000, pp. 1287--1294, Athens (2000)

28. Pustejovsky, J., Hanks, P., Saurí, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., Lazo, M.: The TIMEBANK Corpus. In: Proceedings of Corpus Linguistics 2003, pp. 647--656 (2003). Available at http://www.timeml.org/site/timebank/documentation-1.2.html
29. Verhagen, M., Mani, I., Saurí, R., Littman, J., Knippen, R., Jang, S. B., Rumshiský, A., Phillips, J., Pustejovsky, J.: Automating temporal annotation with TARSQI. In: 43rd Annual Meeting of the Association for Computational Linguistics, Ann Arbor, Michigan (2005)
30. Hobbs, J. R., Pustejovsky, J.: Annotating and Reasoning about Time and Events. In: Proceedings, AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning, Stanford, California (2003)
31. Martins, B., Manguinhas, H., Borbinha, J., Siabato, W.: A geo-temporal information extraction service for processing descriptive metadata in digital libraries, e-Perimetron. In: International web journal on sciences and technologies affined to history of cartography and maps, vol.4 no.1 pp. 25--37 (2009)
32. Han, B., Gates, D., Levin, L.: Rom Language to Time: A temporal Expression Anchorer. In Proceedings of the 13th International Symposium on Temporal Representation and Reasoning (TIME 2006), Budapest, Hungary (2006)
33. Manguinhas, H., Martins, B., Borbinha, J., Siabato, W.: The DIGMAP Geo-Temporal Web Gazetteer Service, e-Perimetron. In: International web journal on sciences and technologies affined to history of cartography and maps, vol. 4 no.1, pp. 9--24 (2009)
34. Petras V., Larson R. Buckland M.: Time period directories: a metadata infrastructure for placing events in temporal and geographic context. In: Proceedings of the 6th ACM/IEEE-CS, Joint Conference on Digital Libraries (2006)
35. Saurí, R., Saquete, E., Pustejovsky, J.: Annotating Time Expressions in Spanish, TimeML Annotation Guidelines (in Press).
36. Saurí, R., Batiukova, O., Pustejovsky, J.: Annotating Events in Spanish, TimeML Annotation Guidelines (in Press).
37. Schilder, F., and Mcculloh, A.: Temporal information extraction from legal documents. In: Katz et al. http://drops.dagstuhl.de/opus/volltexte/2005/318/
38. Elmasri, R.: Fundamentals of database systems. Pearson Education, 4th ed., pp. 842--856 (2004)
39. Verhagen, M.: BAT: Brandeis Annotation Tool, v. 4.2, (2009) http://www.timeml.org/site/bat/

MARTA GUERRERO NIETO
GRUPO MERCATOR - UNIVERSIDAD POLITÉCNICA DE MADRID
Cª DE VALENCIA, KM.7, 28031 MADRID, SPAIN
E-MAIL: <MGUERRERO@TOPOGRAFIA.UPM.ES>

MARÍA JOSÉ GARCÍA RODRÍGUEZ
GRUPO MERCATOR - UNIVERSIDAD POLITÉCNICA DE MADRID
Cª DE VALENCIA, KM.7, 28031 MADRID, SPAIN
E-MAIL: <MJOSEGR@TOPOGRAFIA.UPM.ES>

ADOLFO URRUTIA ZAMBRANA
GRUPO MERCATOR - UNIVERSIDAD POLITÉCNICA DE MADRID
Cª DE VALENCIA, KM.7, 28031 MADRID, SPAIN
E-MAIL: <ADOLFO.URRUTIA@TOPOGRAFIA.UPM.ES>

WILLINGTON SIABATO
GRUPO MERCATOR - UNIVERSIDAD POLITÉCNICA DE MADRID
Cª DE VALENCIA, KM.7, 28031 MADRID, SPAIN
E-MAIL: <W.SIABATO@UPM.ES>

MIGUEL-ÁNGEL BERNABÉ POVEDA
GRUPO MERCATOR - UNIVERSIDAD POLITÉCNICA DE MADRID
Cª DE VALENCIA, KM.7, 28031 MADRID, SPAIN
E-MAIL: <MA.BERNABE@UPM.ES>

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 285-304
RECEIVED 26/11/09 ACCEPTED 16/01/10 FINAL 09/03/10

A Dialogue System for Indoor Wayfinding Using Text-Based Natural Language

HERIBERTO CUAYÁHUITL, NINA DETHLEFS, KAI-FLORIAN RICHTER, THORA TENBRINK, AND JOHN BATEMAN

University of Bremen, Germany

ABSTRACT

We present a dialogue system that automatically generates indoor route instructions in German when asked about locations, using text-based natural language input and output. The challenging task in this system is to provide the user with a compact set of accurate and comprehensible instructions. We describe our approach based on high-level instructions. The system is described with four main modules: natural language understanding, dialogue management, route instruction generation and natural language generation. We report an evaluation with users unfamiliar with the system — using the PARADISE evaluation framework — in a real environment and naturalistic setting. We present results with high user satisfaction, and discuss future directions for enhancing this kind of system with more sophisticated and intuitive interaction.

1 INTRODUCTION

Wayfinding in (partially) known environments poses a considerable challenge for humans. This fact is not only confirmed by a substantial body of research [1, 2] but also by the ubiquity and high demand for incremental navigation assistance systems, as well as web-based services providing in-advance information about routes. However, most information provided by such systems is tailored for large-scale navigation using cars or public transport [3]. Indoor wayfinding assistance is not a trivial issue

and has not been addressed widely so far. Related work includes the following. Kray et al. [4] present an interactive display system mounted on walls providing visual navigation support to building users. Callaway [5] describes indoor navigation help while navigating rather than in-advance directions as explored here. A modelling software proposed by Münzer and Stahl [6] generates dynamic visual route information. Hochmair [7] reports a desktop usability study comparing various modes of indoor navigation aids. Becker et al. [8] and Ohlbach and Stoffel [9] present models for representing the complex spatial configurations adequately for navigation and route assistance. Kruijff et al. [10] present and discuss a human-robot interaction scenario set within an office environment. Automatic systems generating natural language-based route descriptions in-advance have therefore received little attention to date.

In the following we present a first substantial step in this direction: a dialogue system that automatically generates indoor route instructions in German when asked about locations, using text-based natural language input and output. The challenging task in this system is to provide the user with a compact set of accurate and comprehensible instructions suitable for navigating in a complex indoor setting. Our test environment is a campus building which, due to a range of asymmetries and unconventional architectural features, poses a range of navigational challenges.

2 SYSTEM ARCHITECTURE

This dialogue system aims to provide users with route descriptions in German for navigating in a particular building of our university that is generally recognised as presenting significant navigational challenges to both new and infrequent visitors. A pipeline architecture of this system is shown in the high-level diagram of Figure 1. First, the user interacts with a Graphical User Interface (GUI) by asking questions about route directions using text-based natural language. Second, the language understanding module applies OpenCCG parsing [11] and keyword spotting — the latter is used in case of unparsed inputs — to the user utterance in order to extract a user dialogue act. Third, the dialogue management module specifies the system's behaviour by mapping knowledge-compact dialogue states (extracted from the knowledge base) to machine dialogue acts such as 'request' , 'clarify' or 'present_info'. Fourth, the language generation module provides high-level route instruction through the use of pCRU that generates logical forms that are then given to the KPML language generator [12], which in turn outputs text to be shown in the

GUI (see Figure 2). Finally, the knowledge base maintains the history of the interaction. These modules were integrated under the DAISIE framework, which provides support for building situated dialogue systems [13]. These modules are described in more detail in the rest of this section.

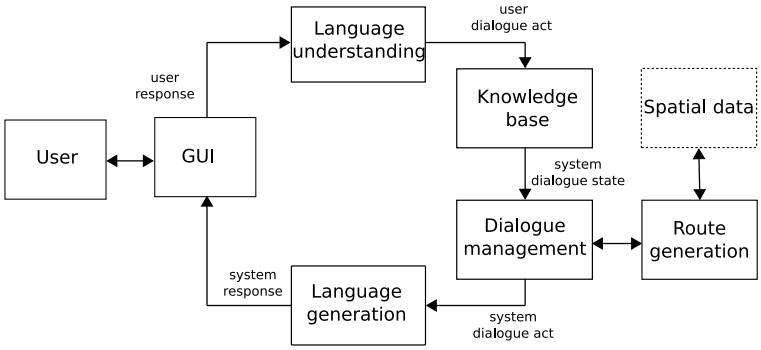


Fig. 1. *A pipeline architecture of our dialogue system for indoor wayfinding.*

2.1 *Natural language understanding*

For parsing the textual user inputs we used the OpenCCG parser [11]. We use a grammar for German and represent the user input in a structure called 'Frame Object Structure' [14], using semantic types derived from the Generalized Upper Model [15]. A sample structure for the sentence 'Wie komme ich zu Raum a3440?' (How do I get to room a3440?) is represented as

```
@g1:gs-NonAffectingDirectedMotion(gum-Coming ^
  <uio-hasSurfaceFunction>uio-Wh-Question ^
  <gum-actor>(p1:gum-ConsciousBeing ^ Pron) ^
  <gum-processInConfiguration>(g2:gum-Process ^
                                gum-Coming) ^
  <gs-motionDirection>(x1:gs-GeneralizedLocation ^
    <gs-hasSpatialModality>
          (z1:gs-GeneralDirectionalNearing ^ zu) ^
    <gs-relatum>(r1:slm-Room ^ Raum ^
      <num>sg ^
      <Modifier>(a1:slm-Office ^ a3440))))
```

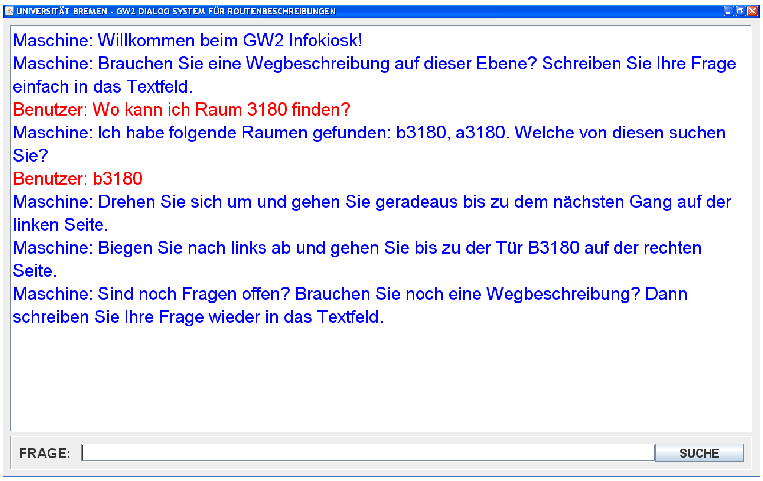


Fig. 2. *A screenshot of our text-based dialogue system for indoor wayfinding. A translation to English of this dialogue is provided in Table 2.*

In addition, we used a keyword spotter to identify locations in case of sentences without parse in the CCG grammars. The task of the keyword spotter is to identify names of locations or names of people and to treat the remaining words as fillers. The output of this module is a user dialogue act represented by a used dialogue act type ('ask', 'provide', 'confirm' , 'silence') and slot-value pairs. The dialogue act for the sample above can be described as 'ask(destination=room a3440)'. We used the same format for describing system dialogue acts.

2.2 *Dialogue management*

Our dialogue manager is based on the Markov Decision Process (MDP) model, but we use a deterministic mechanism for action-selection. The MDP model is used to optimize stochastic sequential decision making problems and is defined as a 4-tuple $< S, A, T, R >$, where S is a finite set of states, A is a finite set of actions, T is a state transition function, and R is a reward function. The solution to an MDP is to find a policy $\pi(s)$ that maps states s to actions a. Because we use deterministic action-selection, we can omit the reward function. This form of control is typically used as baseline for learnt dialogue strategies [16, 17].

We applied this model to our system as follows: (1) the space of dialogue states is represented with a vector of state variables as shown in Table 1, (2) the action space is represented with dialogue act types shown in Equation 1, (3) the state transitions are modelled by observing dialogue states from the knowledge base, and (4) the deterministic dialogue policy is defined in Equation 1. A sample human-machine dialogue illustrating this form of interaction is shown in Table 2. This dialogue is described with wordings in German and English and corresponds to the dialogue shown in Figure 2.

Table 1. *Representation of the dialogue state. Each state variable has either three or four possible values, resulting in* $3^5 \times 4^2 = 3888$ *states.*

Variable	ID	Domain Value
Salutation	X_1	0=null, 1=greeting, 2=closing
Origin	X_2	0=unknown, 1=requested, 2=known
Destination	X_3	0=unknown, 1=requested, 2=known
NumTuples	X_4	0=null, 1=one, 2=more-than-one
Instructions	X_5	0= unknown, 1=known, 2=provided
UserUtterance	X_6	0= unknown, 1=parsed, 2=unparsed, 3=spotted
MoreInstructions	X_7	0=null, 1=empty, 2=yes, 3=no

$$\pi(s) = \begin{cases} \text{opening} & \text{if } s \in \{0000000\} \\ \text{request} & \text{if } s \in \{1000000, 1000012\} \\ \text{other\_request} & \text{if } s \in \{1220210, 1220220\} \\ \text{query\_route} & \text{if } s \in \{1220210, 1220220, 1110030, 1210030, \\ & \qquad 1220030, 1211030, 1221030\} \\ \text{present\_info} & \text{if } s \in \{1221110, 1221130, \} \\ \text{clarify} & \text{if } s \in \{1112100, 1112030, 1212030, 1222211, \\ & \qquad 1222231\} \\ \text{apologize} & \text{if } s \in \{1110020, 1210020, 1220220, 1210210\} \\ \text{confirm} & \text{if } s \in \{1112010, 1112030\} \\ \text{closing} & \text{if } s \in \{1 * * * * * 3\} \\ \text{wait} & \text{otherwise} \end{cases} \tag{1}$$

Table 2. *Sample dialogue in the wayfinding domain, where the dialogue state is formed by the state variables shown in Table 1, and a set of actions (shown in Equation 1) is available per state. The action-selection policy is defined by Equation 1.*

Agent	Dialogue State	Action	Sample Utterance
SYS	0000000	opening	Willkommen beim <Hidden> Infokiosk! [Welcome to the <Hidden> info kiosk!]
	1000000	request	Brauchen Sie eine Wegbeschreibung auf dieser Ebene? Schreiben Sie Ihre Frage einfach in das Textfeld. [Do you need a route description on this level? Just write your question into the text field.]
USR			*Wo kann ich Raum 3180 finden?* [Where can I find room number 3180?]
SYS	1212030	clarify	Ich habe folgende Raeume gefunden: b3180, a3180. Welche von diesen suchen Sie? [I have found the following rooms: b3180, a3180. Which of these are you looking for?]
USR			*b3180*
SYS	1220010	query_route	<query the route generation module>
SYS	1221110	present_info	Drehen Sie sich um und gehen Sie geradeaus bis zu dem naechsten Gang auf der linken Seite. [Please turn round and go straight on until the next hallway on the left hand side.] Drehen Sie sich nach links und gehen Sie bis zu der Tuer B3180 auf der rechten Seite. [Turn left and go until the door B3180 on the right hand side.]
SYS	1220210	other_request	Sind noch Fragen offen? Brauchen Sie noch eine Wegbeschreibung? Dann schreiben Sie Ihre Frage wieder in das Textfeld. [Do you have further questions? Do you need another route description? In that case please write your question again into the text field.]

2.3 *Route instruction generation*

Our system employs a computational process called GUARD (Generation of Unambiguous, Adapted Route Directions) [18] for generating context-specific route instructions. Context-specific route directions account for

environmental characteristics and a route's properties. They adapt communication to the action to be taken in the current surrounding environment. Such instructions are termed "context-specific" because of the explicit adaptation to the structure and function in wayfinding [19]. GUARD unambiguously describes a specific route to a destination, with instructions adapted to environmental characteristics. Selection of the route is not part of GUARD itself. GUARD originally has been developed for providing route instructions in outdoor environments. Figure 3 provides an overview of the generation process.

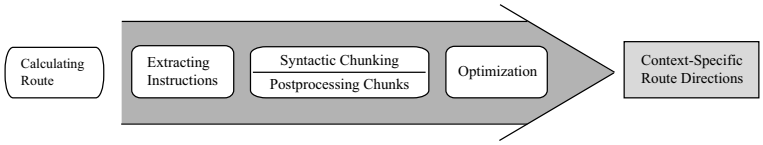


Fig. 3. Overview of GUARD, the generation process for context-specific route directions.

GUARD works on a network representation of paths in an environment. This graph is annotated with information on landmarks, for example, their location and shape. GUARD accounts for different types of landmarks in generating instructions whose role in the route instructions depends on their location relative to the route [20, 21]. Landmarks are associated with decision points based on a heuristic that accounts for distance and potential obstruction. When generating instructions, each associated landmark is tested for whether it can be used as a reference object in the instruction, which depends on its functional role in the given spatial configuration [21].

The generation of context-specific route instructions is a three-step process. First, for every decision point of the route all instructions that unambiguously describe the route segment to be taken are determined. This results in a set of possible instructions for each decision point. Next, GUARD performs spatial chunking. Spatial chunking refers to combining instructions for several consecutive decision points into a single instruction, for example, "turn left at the third intersection" instead of "straight, straight, left." GUARD is flexible with respect to the principles used in chunking (e.g., [22, 3]). Finally, in the third step, the actual context-specific route directions are generated. Here, from all possible instruc-

tions, those that best describe the route are selected. As this is realized as an optimization process, "best" depends on the chosen optimization criterion. Just as with the chunking principles, GUARD is flexible with respect to the criterion used. As a default, it aims for instructions that contain the least number of chunks, i.e., that require the least number of individual instructions[18]. Optimization results in a sequence of chunks that cover the complete route from origin to destination. Due to the aggregation of instructions performed in chunking, instructions for some decision points will be represented implicitly, thus, reducing the amount of communicated information.

In summary, the approach to context-specific route directions finds the best instruction sequence according to the optimization criterion, but for a previously given route. Recently, there has been work on using GUARD's principles in a path search algorithm finding the routes that are also the easiest to describe [23].

2.4 *Natural language generation*

GENERATION OF HIGH-LEVEL INSTRUCTIONS. Our approach for generating high-level route instructions is described in Algorithm 1. Briefly, it operates with the following steps: (a) it receives the output of the route instruction generator; (b) segments the received low-level instructions based on major changes of direction such as left or right; (c) obtains a landmark and direction for the current segment; (d) generates a turning instruction (cf. line 10); (e) generates a go instruction until the current landmark (cf. line 11); (f) unifies the previous two instructions; and (g) generates the language for the unified instruction (cf. line 13). Whilst steps d and e are processed with the pCRUs described in the next subsection, step g is processed with the KPML language generation system [12]. An example of this process using 'corridors' as non-terminal landmarks is illustrated in Figure 4.

GENERATION OF ROUTES WITH PCRU. For the generation of route descriptions, we distinguish different route-associated actions that need to be performed in different segments of a route, for example, turning actions or following actions. While these could be verbalised by a template-based approach, we instead use full NLG and aim to make our descriptions more natural by allowing appropriate variation in the realisation of route segments, so as to reflect the same tendencies found in human de-

Algorithm 1 Generator of high-level textual route instructions

```
1: function GENERATOROFHIGHLEVELINSTRUCTIONS(lowLevelInstructions)
2:     segments ← segment low-level instructions based on major changes of
                  directions such as left and right.
3:     for each segment do
4:         if non-terminal segment then
5:             landmark ← destination landmark for the current segment
6:         else
7:             landmark ← target destination
8:         end if
9:         direction ← direction of the current landmark (e.g. left, right, in
               front)
10:        spl1 ← obtain a turning direction (e.g. turn around, turn left, turn
               right)
11:        spl2 ← obtain a go direction to the landmark with corresponding
               direction
12:        instruction ← aggregation of spl1 and spl2
13:        Generate the textual description corresponding to the current instruc-
               tion
14:    end for
15: end function
```

scriptions. We achieve this by using the pCRU framework described in the rest of this section.

Probabilistic context-free representational underspecification (pCRU) [24] is an approach to resolving the nondeterminacy that typically arises in generation between a semantic representation and its possible linguistic surface forms. This relationship is almost always one-to-many as can be illustrated by the following example. Consider the following SPL [25], which serves as an input to the KPML generation system [12].

```
(v0 / |space#NonAffectingOrientationChange|
   :|actor| ( hearer / |person| )
      :|space#direction| (sd /
                |space#GeneralizedLocation|
         :|space#hasSpatialModality| (lp /
                |space#LeftProjection| ) ) )
```

This semantic representation expresses a simple turning action to the left. A small subset of possible realisations are (1)-(5) below, which differ

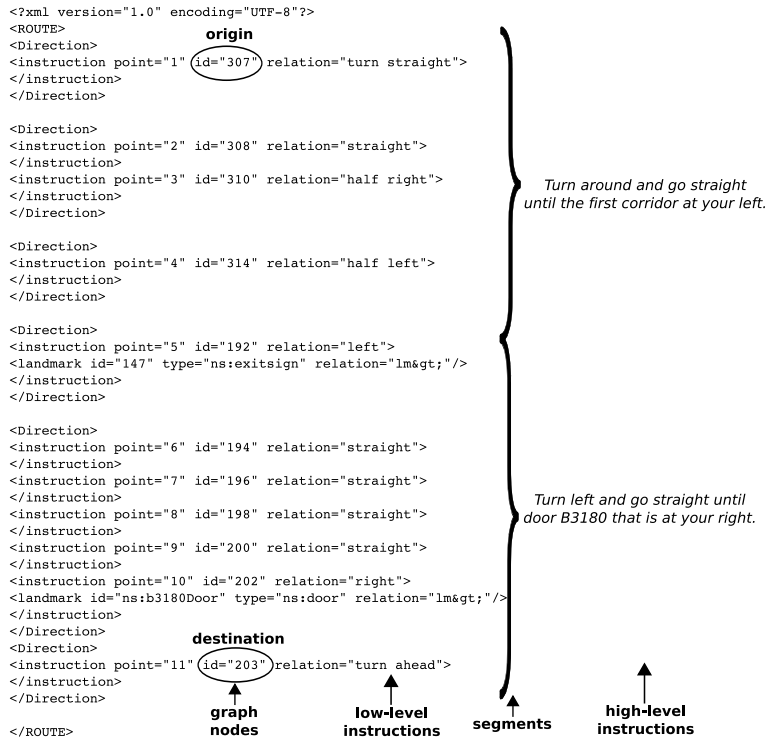


Fig. 4. *Sample route with high-level instructions derived from applying Algorithm 1.*

along several dimensions, such as the choice of speech function (imperative versus declarative), tense (present versus present continuous), the phoricity of the direction attribute (PP versus AP), or whether or not to use ellipsis or the exact choice of the verb.

(1) “Turn left.”
(2) “Turn to the left.”
(3) “You are turning left.”
(4) “Left.”
(5) “Go left.”

Under the pCRU framework, we formalise the above variation in a context-free grammar (CFG) consisting of a set of terminal symbols *W*,

a set of nonterminal symbols N, a start symbol S with $S \in N$ and a set of production rules R of the form $n \rightarrow \alpha$, with $n \in N$, $\alpha \in (W \cup N)$* and W and N being disjoint. This leads to the following CFG for a `TurningSimple` action.

```
TurningSimple = CONFIGTYPE PROCESS ACTOR SPEECHFUN
                  TENSE DIR (0.7)
TurningSimple = CONFIGTYPE PROCESS ACTOR SPEECHFUN
                  TENSE ":ellipsis full" DIR (0.3)
CONFIGTYPE = "|space#NonAffectingOrientationChange|"
                  (1.0)
PROCESS = ":lex turn" (0.8)
PROCESS = ":lex go" (0.2)
ACTOR = "( hearer / |person| )" (1.0)
TENSE = ":tense present" (0.9)
TENSE = ":tense present-continuous" (0.1)
SPEECHFUN = ":speechact command" (0.9)
SPEECHFUN = ":speechact assertion" (0.1)
DIR = :|space#route| (gr / |space#GeneralizedRoute|
     :|space#direction| (sd /
        |space#GeneralizedLocation| :phoric-q phoric
          :|space#hasSpatialModality| (sm /
                  LOCATION-DIRECTION ) ) (0.7)
DIR = :|space#route| (gr / |space#GeneralizedRoute|
     :|space#direction| (sd /
      |space#GeneralizedLocation| :phoric-q notphoric
            :|space#hasSpatialModality| (sm /
                  LOCATION-DIRECTION ) ) (0.3)
```

This representation allows us to capture all arising variation within a single formalism as well as control the application of the respective expansion rules by attaching probabilities to them which indicate each rule's probability of application.

3 DIALOGUE SYSTEM EVALUATION

This evaluation aimed to investigate the performance of our text-based approach for indoor wayfinding. For such a purpose, the dialogue system described above was implemented and tested with a set of users in a real building. This building is complex to navigate; although it has several floors, only one floor was tested.

3.1 *Evaluation methodology*

We evaluated our dialogue system using objective and subjective metrics mostly derived from the PARADISE framework [26]. This framework is commonly used for assessing the performance of spoken dialogue systems, and can be used for evaluating dialogue systems with different modalities in the wayfinding domain.

The groups of quantitative metrics are described as follows. First, the group of *dialogue efficiency* metrics includes 'system turns', 'user turns', and 'elapsed time' (in seconds). The latter includes the time used by both conversants, from the first user utterance until the last system utterance. Second, the group of *dialogue quality* metrics consists of percentages of parsed sentences, sentences with spotted keywords, and unparsed sentences. Third, the group of *task success* metrics includes the typical binary task success expressed as

$$\text{BinaryTaskSuccess} = \begin{cases} 1 \text{ for finding the target location} \\ 0 \quad \text{otherwise.} \end{cases} \tag{2}$$

In this group we proposed two additional metrics in order to penalize the degree of difficulty in wayfinding. The first is referred to as '3-valued Task Success (TS)' defined as

$$\text{3-ValuedTS} = \begin{cases} 1 \text{ for finding the target location} \\ 1/2 \text{ for finding the target location with small-medium problems} \\ 0 \quad \text{otherwise,} \end{cases} \tag{3}$$

and the second is referred to as '4-valued task success' defined as

$$\text{4-ValuedTS} = \begin{cases} 1 \text{ for finding the target location} \\ 2/3 \text{ for finding the target location with small-medium problems} \\ 1/3 \text{ for finding the target location with severe problems} \\ 0 \quad \text{otherwise.} \end{cases} \tag{4}$$

The value of 1 is given when the user finds the target location without hesitation, the value with small-medium problems is given when the user finds the location with slight confusion(s), and the value with severe problems is given when the user gets lost but eventually finds the target location. Finally, the group of quantitative metrics are described in Table 3. The sum of scores from these metrics represents the overall user satisfaction score.

Table 3. Subjective measures for evaluating indoor wayfinding, adapted from [26].

Measure	Question
Easy to Understand	Was the system easy to understand?
System Understood	Did the system understand what you asked?
Task Easy	Was it easy to find the location you wanted?
Interaction Pace	Was the pace of interaction with the system appropriate?
What to Say	Did you know what you could write at each point?
System Response	Was the system fast and quick to reply to you?
Expected Behaviour	Did the system work the way you expected it to?
Future Use	Do you think you would use the system in the future?

3.2 *Experimental setup*

Our experiments evaluated the dialogue system described above with a user population of 26 native speakers of German. They were university students (16 female, 10 male) aged 22.5 on average. Each user was presented with six wayfinding tasks, resulting in a total of 156 route dialogues. They were asked in each case to find a particular location based on the route instruction generated by the dialogue system on request by the user. The locations were spatially distributed. Two tasks used 2 High-Level Instructions (HLIs), two tasks used 3 HLIs, and two tasks used 4 HLIs. The dialogue tasks were executed pseudorandomly (from a uniform distribution). At the beginning of each session, participants were asked about their familiarity with the building using a 5-point Likert scale, where 1 represents the lowest familiarity and 5 the highest. This resulted in a familiarity score of 2.4. Then, our participants received the following set of instructions: (a) you can ask the system using natural language, (b) you can take notes from the received instructions, (c) follow the instructions as precisely as possible, (d) you are not allowed to ask anyone how to get to the target location, and (e) you can give up anytime after trying without success by telling that to the assistant that will follow you. At the end of each wayfinding task, participants were asked to fill a questionnaire (Table 3) for obtaining qualitative results using a 5-point Likert scale, where 5 represents the highest score.

3.3 *Experimental results*

According to dialogue efficiency metrics, it can be observed from Table 4 that the user-machine interactions involved short dialogues in terms of

system turns, user turns and time. These results suggest that once users receive instructions to find a given location, they tend not to ask further questions. We can also observe a large number of words per system turn mostly due to the textual instructions, where the longer the number of high-level instructions the longer the textual output. In addition, although some users used only keywords in the textual input, overall they asked questions.

According to dialogue quality, it can be noted that our grammars did not have wide coverage. There are many different ways to ask for a given location, including sentences with ungrammatical structures and sentences with words absent in the lexicon. However, the keyword spotter then was crucial for identifying the users' target location.

According to task success, our dialogue system obtained a very high binary task success, but this measure does not take into account how hard it was for the user to find the given locations. In contrast, whilst our 3-valued task success measure penalizes more strongly, our 4-valued task success measure is between the other two metrics. From these metrics, we found that the latter generated more faithful scores because it predicts more closely user satisfaction. This argument can be validated with statistical analysis, but this is left as future work.

Our qualitative results report very high scores for user satisfaction, mainly for the dialogues with 2 High-Level Instructions (HLIs) and 3 HLIs. However, users found it harder to follow the dialogues with 4 HLIs. One can think that the reason was due to the length of the instructions, but we observed that it was more due to ambiguity in which corridors to follow. The lower scores in the following qualitative metrics support this argument: easy to understand, task easy, expected behaviour and future use. Nevertheless, we found that a dialogue system for indoor wayfinding using language processing capabilities — with only text input and output — can obtain very high overall scores in user satisfaction.

Finally, we included an additional question in the survey filled after each dialogue: 'Did you find the location only based on the given instructions by the system or did you use additional help such as signs?' This question also used a 5-point Likert scale, where 5 represents the highest score for strictly following only the system instructions. This resulted in an average value of 4.3, which suggests that the results described above were derived from following almost entirely the system's instructions.

Table 4. Average results of our wayfinding system for dialogues with different amounts of High-Level Instructions (HLIs), organized according to the following groups of metrics: dialogue efficiency, dialogue quality, task success and user satisfaction.

Measure	2 HLIs (52 dialogues)	3 HLIs (52 dialogues)	4 HLIs (52 dialogues)	All (156 dialogues)
Avg. System Turns	2.25	2.38	2.28	2.30
Avg. User Turns	1.30	1.61	1.64	1.52
Avg. System Words per Turn	34.05	40.04	49.59	41.30
Avg. User Words per Turn	4.06	5.34	4.84	4.79
Avg. Time (in seconds)	20.69	19.77	25.87	22.14
Parsed Sentences (%)	23.8	4.3	22.5	16.7
Spotted Keywords (%)	74.6	91.4	73.2	79.9
Unparsed Sentences (%)	1.6	4.3	4.2	3.4
Binary Task Success (%)	96.2	100.0	88.5	94.9
3-Valued Task Success (%)	92.3	88.5	63.5	81.4
4-Valued Task Success (%)	94.9	92.3	75.6	87.6
Easy to Understand	4.65	4.6	4.08	4.46
System Understood	4.71	4.62	4.62	4.65
Task Easy	4.60	4.54	3.73	4.29
Interaction Pace	4.71	4.65	4.52	4.63
What to Say	4.71	4.63	4.65	4.66
System Response	4.60	4.62	4.58	4.56
Expected Behaviour	4.64	4.50	4.21	4.45
Future Use	4.46	4.37	4.12	4.31
User Satisfaction (sum)	37.1	36.5	34.5	36.0
User Satisfaction (%)	92.7	91.2	86.3	90.0

4 CONCLUSIONS AND FUTURE WORK

In this paper we have presented a dialogue system for indoor wayfinding in a complex building using text-based natural language input and output. The system was described with four main components: natural language understanding, dialogue management, route instruction generation and natural language generation. In the latter we described our approach based on high-level instructions. A key advantage of our dialogue system is its support for language-independence, only parsing and generation grammars have to be added in order to support a new language, the rest is reused. Experimental results — using the PARADISE evaluation framework — in a real environment with 26 participants (156 dialogues) pro-

vide evidence to support the following claims: (a) text-based dialogues resulted in very short interactions, they mostly consist of question and answer, though eventually clarifications or apologies occurred; (b) keyword spotting was an essential component to assist the parser with unparsable utterances; (c) our proposed 4-valued task success metric predicts better user satisfaction than binary task success or 3-valued task success; and (d) a text-based dialogue system for indoor wayfinding can obtain very high overall scores in user satisfaction. To the best of our knowledge, this is the first evaluation of its kind in the indoor wayfinding domain.

We suggest the following avenues for future research:

First, text-based language processing, spoken language processing and graphical interfaces (such as maps) can be combined into principled frameworks for building effective wayfinding systems. Such systems should be evaluated as in this paper in order to assess the performance across different system versions. In this way, systematic evaluations can be made by varying different conditions under a given framework. This is an important and useful step to take that has not so far been achieved in indoor navigation.

Second, the dialogue manager is responsible for controlling the system's dialogue behaviour. When the system's behaviour becomes complex, it is less recommendable to use hand-crafted behaviour because it is non-adaptive and labour intensive. Machine learning methods such as reinforcement learning can be used to induce the system's behaviour automatically [27, 17, 28]. This is relevant for learning adaptive and complex behaviour such as learning to ground, learning to clarify, learning to present information, learning multimodal strategies and learning to negotiate route directions.

Third, in the case of indoor route directions, future work can entail covering paths that cross multiple floors. This will require both handling a graph with dedicated transition nodes between floors and a clear communication of floor changes in the route directions. In the present work, we used corridors as main landmarks; however, a principled mechanism to rank indoor landmarks can be investigated. In addition, providing route instructions for new spatial environments is possible by providing spatial representations of additional environments in the form of a route graph.

Finally, future work in language generation can aim to enhance the adaptiveness of route descriptions along three dimensions: (a) to make descriptions more tailored towards a particular user by taking their familiarity of the environment into closer consideration [29]; (b) to present information for users with different cognitive styles for users familiar or

unfamiliar with a given environment [30, 31, 32]; and (c) to investigate how to incorporate interactive alignment [33].

REFERENCES

[1] Golledge, R.: Wayfinding behavior: Cognitive mapping and other spatial processes. Johns Hopkins Press, Baltimore, MD (1999)

[2] Passini, R.: Wayfinding: A conceptual framework. Urban Ecology **5**(1) (1981) 17–31

[3] Dale, R., Geldof, S., Prost, J.P.: Using natural language generation in automatic route description. Journal of Research and Practice in Information Technology **37**(1) (2005) 89–105

[4] Kray, C., Kortuem, G., Krüger, A.: Adaptive navigation support with public displays. In Amant, R.S., Riedl, J., Jameson, A., eds.: Proceedings of IUI 2005. ACM Press, New York. (2005) 326–328

[5] Callaway, C.: Non-localized, interactive multimodal direction giving. In van der Sluis, I., Theune, M., Reiter, E., Krahmer, E., eds.: Proceedings of the Workshop on Multimodal Output Generation MOG 2007, Centre for Telematics and Information Technology (CTIT), University of Twente (2007) 41–50

[6] Münzer, S., Stahl, C.: Providing individual route instructions for indoor wayfinding in complex, multi-level buildings. In Probst, F., Keßler, C., eds.: Proceedings of the 5th Geographic Information Days, Münster, IfGIprints (2007) 241–246

[7] Hochmair, H.H.: Pda-assisted indoor-navigation with imprecise positioning: Results of a desktop usability study. In Meng, L., Zipf, A., Winter, S., eds.: Map-based Mobile Services - Interactivity, Usability and Case Studies. Springer, Berlin, Heidelberg (2008) 228–247

[8] Becker, T., Nagel, C., Kolbe, T.H.: A multilayered space-event model for navigation in indoor spaces. In Lee, J., Zlatanova, S., eds.: Proceedings of the 3rd International Workshop on 3D Geo-Information, Seoul, Korea, Berlin, Heidelberg, Springer (2008)

[9] Ohlbach, H.J., Stoffel, E.P.: Versatile route descriptions for pedestrian guidance in buildings: Conceptual model and systematic method. In: 11th AGILE International Conference on Geographic Information Science, University of Girona, Spain (2008)

[10] Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Situated Dialogue and Spatial Organization: What, Where... and Why? International Journal of Advanced Robotic Systems **4**(1) (March 2007) 125–138 Special Issue on Human-Robot Interaction.

[11] Clark, S., Hockenmaier, J., Steedman, M.: Building Deep Dependency Structures using a Wide-Coverage CCG Parser. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). (2002) 327–334

[12] Bateman, J.A.: Enabling technology for multilingual natural language generation: the KPML development environment. Journal of Natural Language Engineering **3**(1) (1997) 15–55
[13] Ross, R.J., Bateman, J.A.: Daisie: Information state dialogues for situated systems. In: TSD. Volume 5729 of LNCS., Springer (2009) 379–386
[14] Collier, R.W.: Agent Factory: A Framework for the Engineering of Agent Oriented Applications. PhD thesis, University College Dublin (2001)
[15] Bateman, J., Hois, J., Ross, R., Tenbrink, T., Farrar, S.: The Generalized Upper Model 3.0: Documentation. SFB/TR8 internal report, Collaborative Research Center for Spatial Cognition, University of Bremen, Germany (2008)
[16] Lemon, O., Georgila, K., Henderson, J.: Evaluating efectiveness and portability of reinforcement learned dialogue strategies with real users: The TALK TownInfo evaluation. In: SLT, Palm Beach, Aruba (Dec 2006) 178–181
[17] Cuayáhuitl, H.: Hierarchical Reinforcement Learning for Spoken Dialogue Systems. PhD thesis, School of Informatics, University of Edinburgh (January 2009)
[18] Richter, K.F.: Context-Specific Route Directions - Generation of Cognitively Motivated Wayfinding Instructions. Volume DisKi 314 / SFB/TR 8 Monographs Volume 3. IOS Press, Amsterdam, The Netherlands (2008)
[19] Klippel, A.: Wayfinding choremes. In Kuhn, W., Worboys, M., Timpf, S., eds.: Spatial Information Theory. Foundations of Geographic Information Science, Berlin, International Conference COSIT, Springer (2003) 320–334 LNCS 2825.
[20] Hansen, S., Richter, K.F., Klippel, A.: Landmarks in OpenLS - a data structure for cognitive ergonomic route directions. In Raubal, M., Miller, H., Frank, A.U., Goodchild, M.F., eds.: Geographic Information Science - Fourth International Conference, GIScience 2006, Springer; Berlin (2006) 128–144 LNCS 4197.
[21] Richter, K.F.: A uniform handling of different landmark types in route directions. In Winter, S., Duckham, M., Kulik, L., Kuipers, B., eds.: Spatial Information Theory. LNCS 4736, Springer; Berlin (2007) 373–389 COSIT.
[22] Klippel, A., Tappe, H., Kulik, L., Lee, P.U.: Wayfinding choremes — a language for modeling conceptual route knowledge. Journal of Visual Languages and Computing **16**(4) (2005) 311–329
[23] Richter, K.F., Duckham, M.: Simplest instructions: Finding easy-to-describe routes for navigation. In Cova, T.J., Miller, H.J., Beard, K., Frank, A.U., Goodchild, M.F., eds.: Geographic Information Science - 5th International Conference, GIScience 2008. LNCS 5266, Springer; Berlin (2008) 274–289
[24] Belz, A.: Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. Natural Language Engineering **1** (2008) 1–26

[25] Kasper, R.: SPL: A Sentence Plan Language for text generation. Technical report, USC/ISI (1989)
[26] Walker, M., Kamm, C., Litman, D.: Towards developing general models of usability with PARADISE. Natural Language Engineering **6**(3) (2000) 363–377
[27] Singh, S., Litman, D., Kearns, M., Walker, M.: Optimizing dialogue management with reinforcement learning: Experiments with the NJFun system. JAIR **16** (2002) 105–133
[28] Cuayáhuitl, H., Renals, S., Lemon, O., Shimodaira, H.: Evaluation of a hierarchical reinforcement learning spoken dialogue system. Computer Speech and Language **24**(2) (2010) 395–429
[29] Tenbrink, T., Winter, S.: Variable Granularity in Route Directions. Spatial Cognition and Computation **9** (2009) 64–93
[30] Lovelace, K.L., Hegarty, M., Montello, D.R.: Elements of good route directions in familiar and unfamiliar environments. In: COSIT '99: Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science, London, UK, Springer-Verlag (1999) 65–82
[31] Burnett, G., Smith, D., May, A.: Supporting the navigation task: characteristics of good landmarks. In: Proceedings of the Annual Conference of the Ergonomics of Society. Turnin, Italy. (2001) 441–446
[32] May, A.J., Ross, T., Bayer, S.H.: Driver's Information Requirements when Navigating in an Urban Environment. Journal of Navigation **56** (2003) 89–100
[33] Pickering, M.J., Garrod, S.: Toward a mechanistc psychology of dialog. Behavioral and Brain Sciences **27** (2004) 169–226

HERIBERTO CUAYÁHUITL
TRANSREGIONAL COLLABORATIVE RESEARCH CENTER,
SFB/TR 8, UNIVERSITY OF BREMEN
ENRIQUE-SMITDH-STR. 5, 28215, BREMEN, GERMANY
E-MAIL: <HERIBERTO@UNI-BREMEN.DE>

NINA DETHLEFS
FB10 FACULTY OF LINGUISTICS AND LITERARY SCIENCES,
UNIVERSITY OF BREMEN
BIBLIOTHEKSTRASSE 1, 28359, BREMEN, GERMANY

KAI-FLORIAN RICHTER
TRANSREGIONAL COLLABORATIVE RESEARCH CENTER,
SFB/TR 8, UNIVERSITY OF BREMEN
ENRIQUE-SMITDH-STR. 5, 28215, BREMEN, GERMANY

THORA TENBRINK
TRANSREGIONAL COLLABORATIVE RESEARCH CENTER,
SFB/TR 8, UNIVERSITY OF BREMEN
ENRIQUE-SMITDH-STR. 5, 28215, BREMEN, GERMANY
AND
FB10 FACULTY OF LINGUISTICS AND LITERARY SCIENCES,
UNIVERSITY OF BREMEN
BIBLIOTHEKSTRASSE 1, 28359, BREMEN, GERMANY

JOHN BATEMAN
TRANSREGIONAL COLLABORATIVE RESEARCH CENTER,
SFB/TR 8, UNIVERSITY OF BREMEN
ENRIQUE-SMITDH-STR. 5, 28215, BREMEN, GERMANY
AND
FB10 FACULTY OF LINGUISTICS AND LITERARY SCIENCES,
UNIVERSITY OF BREMEN
BIBLIOTHEKSTRASSE 1, 28359, BREMEN, GERMANY

IJCLA VOL. 1, NO. 1-2, JAN-DEC 2010, PP. 305-317
RECEIVED 30/10/09 ACCEPTED 16/01/10 FINAL 08/03/10

A Case Study of Rule Based and Probabilistic Word Error Correction of Portuguese OCR Text in a "Real World" Environment for Inclusion in a Digital Library

BRETT DRURY AND J. J. ALMEIDA

LIAAD-INESC and University of Minho, Portugal

ABSTRACT

The transfer of textual information from large collections of paper documents to electronic storage has become an increasingly popular activity for private companies and public organizations. Optical Character Recognition (OCR) software is a popular method to effect the transfer of this information. The latest commercially available OCR software can be very accurate with reported accuracy of 97% to 99.95%[6]. These high accuracy rates lower dramatically when the documents are in less than pristine condition or if the typeface is non-standard or antiquated. In general, OCR recovered text requires some further processing before it can be used in a digital library. This paper documents an attempt by a private company to apply automatic word error correction techniques on a "real world" 12 million document collection which contained texts from the late 19th Century until the late 20th Century.

The paper also describes attempts to increase the effectiveness of word correction algorithms through the use of the following techniques: 1. Reducing the text correction problem to a restricted language domain, 2. Segmenting the collection by document quality and 3. Learning domain specific rules and text characteristics from the document collection and operator log files. This case study also considers the commercial pressures of the project and the effectiveness of both rule based and probabilistic word error

correction techniques on less than pristine documents. It also provides some conclusions for researchers / companies considering multi-million document transfers to electronic storage.

1 INTRODUCTION

Real world document collections are not always in pristine condition. The document may have surface contamination which can be due to the age of the document, the quality of the media, the type of media and other material affixed to the document such as official stamps. The typeface may be antiquated which may further degrade the accuracy of OCR software. The recovered text may contain too many errors to be used in a digital library. Frequently, some further correction of the text is required. This paper will describe an attempt by a commercial company to correct Portuguese text which had been recovered by OCR software for inclusion in a digital library.

1.1 *Document Collection Characteristics*

The document collection contained over 12 million documents which was created over a hundred year period. The quality of the documents ranged from the very good (clear type face and no surface contamination) to the very poor (illegible and heavy surface contamination). The collection contained some homogeneous text, for example correspondence. The correspondence was mainly letters, which on occasion had images as an attachment. This correspondence also included bill and product information which in some circumstances was in a language other than Portuguese. The document collection also contained some non-standard items such as reports.

1.2 *Processing Documents*

The paper documents were scanned using large commercial scanners which were capable of processing a large number of documents per hour. The scanners produced images of documents in Tagged Information File Format (TIFF) and were in monochrome. The images were then sorted by a simple algorithm and organized into folders which contained related

images. Each image was given a unique number within the folder. The images were then pre-processed (deskew and despeckle) in preparation for the OCR process. The images were processed by the OCR software which ran on two powerful computers which functioned 24 hours a day. The OCR software was set on the slowest and most accurate setting. The OCR software required nearly 2 years to process the 12 million documents. The text from the OCR process was inserted into a Database Management System (DBMS). The text was then subject to a post-processing correction process. The text was to be used in a full text index which would be used for searching, consequently stop words such as "por" could be excluded from the correction process.

1.3 *Initial Correction Attempts*

A popular approach is to use human operators to correct text. This can be slow. It was reported that an efficient company in Romania with 25 staff could process 600,000 documents a year.[10]. This mirrored our initial experience with a completely manual approach. A software application was built which used the Microsoft Word API to identify word errors and their possible replacements. The operator corrected the text one word error at a time. The mean time for each operator to correct one document was approximately 180 seconds. This was too slow as it would have taken a team of 5 operators approximately 72 years to complete the task. This was not only unacceptably slow, but would have represented a potential enormous cost to the company.

The operators mean time to process each document was reduced with a modified manual approach, which was to correct popular spelling and characters errors automatically. The performance of the application was increased by a multi-threaded approach. The errors and potential word candidates were cached by one thread, whilst another thread updated the user display whilst another thread was responsible for updating and fetching text from the DBMS. The error caching thread was significantly faster than the human operator, consequently there was no delay when moving from one error to the next.[1] Fetching the error and word candidates di-

[1] Although this improved the operators' mean time, the operators found it difficult to work with the application as the operators had to concentrate 100% of the time. If I were to write the application again I would add random delays to give the operators a small break in concentration.

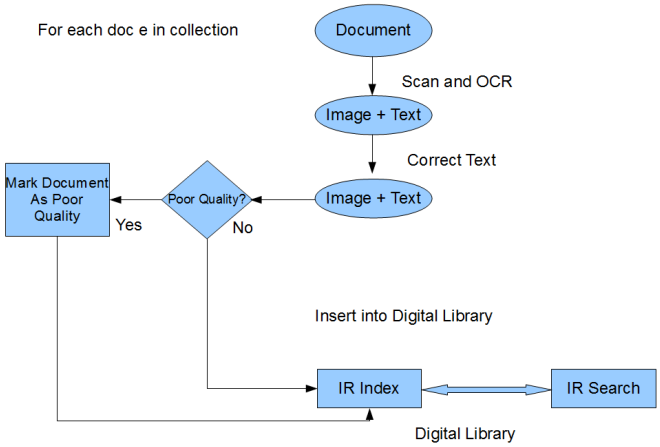


Fig. 1. Process For Transferring Documents

rectly from the Microsoft Word API without caching introduced a significant delay. The mean time was reduced to 30 seconds a document. A team of 5 operators working full-time could process a million documents a year, which was significantly faster than the Romanian case study[10]. This efficiency improvement was still not fast enough as it would have taken 12 years to complete the task and would have represented a cost of approximately 500,000 Euros in labour. An automated process was required to process a significant number of texts, not only to reduce the time required to complete the project, but to ensure the company realised a profit from the project.

It should be noted that the operators required significant supervision. There was pressure for each operator to reduce their times to process each text. The less able operators simply cheated by marking documents as complete when the document had not been processed or marking a good document as unprocessable. This would lower the mean time of the operator. It was necessary to review at regular intervals a statistically significant sample of each operators output to identify which operators were "honest" and which were "cheating".

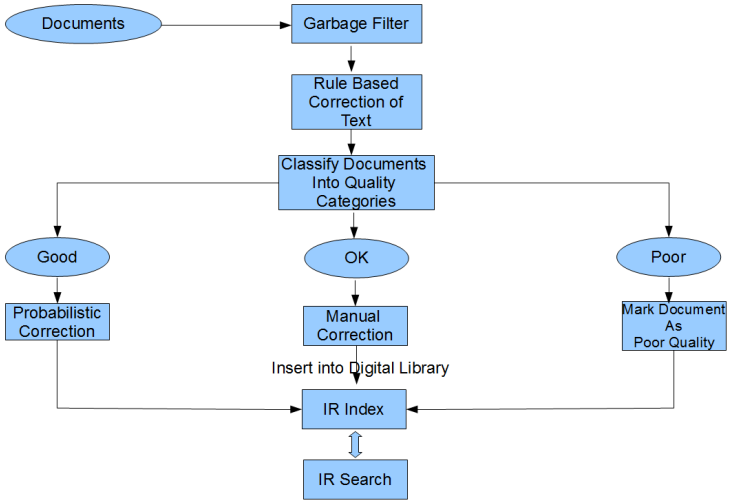


Fig. 2. Modified Process For Transferring Documents

1.4 *Summarization of Problem*

The initial process is described in figure: 1. This process was too slow and costly. There was a demand to move to a partial automatic system, as described in figure: 2. The rest of the paper will discuss the transformation to the automatic word correction process as described in figure: 2. This will include:

1. Reducing the text correct problem to a restricted language domain.
2. Segmenting the collection by document quality.
3. Learning domain specific rules and text characteristics from the document collection and operator log files.

2 Document Collection Preparation

2.1 *Assessing The Document Collection*

A "quality measure" was assigned to each document, so that it was possible to measure the performance of the error correction techniques. A

simple measure was used, which was the number of correct words divided by the total number of words. A statistically significant sample of the documents was manually verified against the quality score. This simple measure provided an accurate reflection of the document's quality. Low scoring documents had heavy surface contamination or were handwritten. High scoring documents were free from contamination with a clear typeface. The distribution of quality effectively followed a normal distribution, with the bulk of the documents having a quality score between 0.5 and 0.7.

$$Quality = NumberofCorrectWords \div TotalNumberofWords$$

2.2 *Rule Induction from Operator Generated Log Files*

The five employees had processed the document collection with the modified manual system for three years. Three million documents were processed. The spelling corrections were logged for each operator. It was possible to categorize the error types from the log files into the following categories: 1. Substitution of Characters, 2. Elimination of Characters, 3. Insertion of Characters, 4. Split Word Errors, 5. Joined Word Errors. A number of frequently occurring errors were unique to a Latin based language, for example, the inaccurate splitting or joining of reflexive words, for example "da-me" would be joined as "dame".

2.3 *Pre-processing of Text (rule based correction)*

GARBAGE REMOVAL A large number of documents were printed on ruled paper which was interpreted by the OCR software as miscellaneous symbols. A filter was constructed which attempted to remove text which was generated by physical markings other than text.

JOIN/SPLIT WORDS A number of rules to detect and correct split and join errors were inferred from the log files. Join errors were detected by identifying "word boundaries" in continuous text, for example capital letters or punctuation. The text was split on the word boundary and the resulting words checked against a dictionary. If they were both correct than the words were accepted. Split errors were detected by joining two continuous errors and evaluating the resultant text with a dictionary. If the text was a correct word then it was accepted.

CORRECTING COMMON WORD AND CHARACTERS ERRORS The join and split word rules were incorporated into a pre-processor application with the hard coded rules from the modified manual system for popular word and character substitutions. Two runs were made, the first was "strict" where the resultant words had to be correct. The second was "permissive", where there was a tolerance of one edit distance. The preprocessor was relatively successful and moved the "bulge" of the normal distribution for the quality to the right with a mean average improvement of quality of 0.2, i.e on average a document which scored 0.5 would score 0.7 after the pre-processor runs.

3 PROBABILISTIC ERROR CORRECTION

In recent years there has been a number of advances in probabilistic error correction for text produced by OCR systems. These techniques assume that text recovered by OCR is semi-determinate[9]. The assumption is that OCR systems will consistently identify identical/similar markings on a document as the same character. This semi-determinate nature allows a certain degree of predictability of the errors produced by the software and that some types of errors are more frequent than others.

The following three techniques were utilized in this case study.

CHARACTER CONFUSION MATRIX A character confusion table provides a list of transformation probabilities from one character to another, for example c $\rightarrow$ ç would be high where as c $\rightarrow$ w would be low. A probability of a word candidate substituting an error was achieved by a simple summing of the individual character probabilities and calculating the mean value [3].

The character confusion matrix in this project was built from the operator log files which documented all word error changes over a three year period. The substitution errors were calculated by comparing error and correction words of the same length. Insertion and deletion errors were calculated by comparing error and correction words which had a difference in length of 1 character.

DICTIONARY THINNING Dictionary thinning allows the reduction of possible word candidates. A custom dictionary was developed which con-

tained only the correct words which were in the document collection and their frequency. The frequency was important because word frequency in a document collection obeys Zipf distribution [7] and may provide an indication of likelihood of the word candidate being correct [8].

The dictionary was constructed by parsing the whole document collection and comparing the words to the J-Spell dictionary. The words which were not in the J-Spell[2] dictionary were initially rejected and written to a file with their frequency. The remaining words were written to another text file which was our initial dictionary. The top 1,000 most frequent errors were analysed by a human operator. The operator identified words which were incorrectly rejected, for example surnames and names of companies. These words were reintroduced into the dictionary.

WORD N GRAMS Word n-grams provided an indication of conditional probability of certain word combinations[5]. Words frequently co-occur, consequently the presence of one word may imply the presence of another word. In the case study another measure was developed, the gapped bi-gram where the middle word from a tri-gram was removed. The gapped bi-gram assisted in the identification of conditional probability of words separated by a stop-word, for example "agua da pedras", where there is a semantic relation between "agua" and "pedras". To generate the n-gram dictionaries the whole corpus was parsed. The n-grams were listed by frequency and the top 2,000 n-grams were selected for their relevant dictionary.

3.1 *Selection of Word Candidates*

Word candidates were selected from the customized dictionary as described in the above section. Although the dictionary had been "thinned", it still contained thousands of possible word candidates. It was not possible to assess each word in the dictionary for each error because the application would have been too slow. Consequently, a reduction of the number possible word candidates would improve the efficiency of the application. A common method is to use n-grams [4] to retrieve word candidates for a given error. Popular letter n-grams however, can lead to large numbers of word candidates being retrieved for a single error. It is possible to reduce the number of word candidates without removing any highly probable replacement through the use of skip grams.[1] Skip

Table 1. character error & replacement character & probability

Character Error	Character Replacement	Probability
c	ç	6.5%
a	ã	5.8%

grams are formed from letters which occupy either odd or even numbered positions in a word, for example the word "teste" would have the following 2 letter skip gram "ts et se". The popular letter n-grams were broken up into less popular skip grams and consequently when the word candidates were returned through the application of a skip gram distance a smaller and more relevant set was returned.

The use of skip grams highlighted a "quirk" of the OCR system. The OCR software frequently failed to recognize the Portuguese characters ’ç’ and ’ã’. It frequently replaced them with the characters ’c’ and ’a’. This was a significant error as ’ç’ and ’ã’ frequently appear together in Portuguese. This mistake would result in two incorrect skip grams, which may have excluded a valid word candidate from being selected.

The frequency of this mistake is shown in Table 1.

Note: These figures understate how often the OCR software made these mistakes as these figure were taken after the pre-processors had corrected the common character errors.

3.2 *Alignment of Word Candidate and Error*

The calculation of the transformation probability of error to word candidate required alignment of the word candidate and the error. This was a trivial task, if the word candidate and error were the same length. When the word candidate and error were different lengths it was necessary to return the most probable alignment with a ’#’ representing the missing character(s). There were two considerations for the algorithm design, which were accuracy and efficiency. Two algorithms were developed, one algorithm was for when the difference in length between the error and word candidate was 2 or less and the other was when the difference in length was 3 or more. The first algorithm calculated every alignment permutation and returned the most probable. The second algorithm was a compromise between accuracy and efficiency this was because the larger

the difference the more the total permutations and consequently there would be a drop in performance of the algorithm. A "sliding alignment" algorithm[8] was used where the shorter word would be moved across the longer word one character at a time. At each stage the alignment would be verified for successfully aligned characters. The word form with the most correctly aligned characters was returned.

3.3 *Calculating the Word Candidate Scores*

The scoring process initially applied a transformation probability for each of the word candidates. Word candidates were eliminated if they had less than a 0.5 transformation probability. This was because through experimentation with a statistically significant sample it was determined that word candidates with a score of less than 0.5 were unlikely to be correct. Elimination was necessary to improve the efficiency of the application. The remaining word candidates were scored for their co-occurrence probabilities with existing word n grams and gapped n-grams and the log frequency of the word candidate in the corpus was calculated as follows:

$$S = P(E \rightarrow W_c) \times (log(W_cF) + 50) \times (1 + P(X, W_c)) \times (1 + P(Y, W_c))$$

S =Score W_c =Word Candidate E =Error
W_cF =Word Candidate Frequency
X =Word which has position $\pm$ 1 of E
Y =Word which has position $\pm$ 2 of E

3.4 *Excluding Documents*

Automatic processing of the whole collection was not possible because the document collection was not of a uniformly high quality. It was possible to automatically process a large number of documents, which reduced the number of documents which needed to be processed manually. This reduced the time that was required to process the documents, but also reduced the costs involved.

The documents were classified into three categories:poor quality (no manual processing possible), low-medium (manual processing only) and medium-high quality (automatic processing possible). The quality borders where set by operators who analysed a statistically significant sample of documents at varying quality levels. The quality measures are shown in Table 2.

Table 2. Document Classification

Category	Quality measure	Action
Poor quality	$0 < q < 0.5$	no processing possible
Low - medium quality	$0.5 \leq q < 0.7$	manual processing possible
Medium - high quality	$0.7 \leq q < 1$	automatic processing possible

Poor / low quality documents had surface contamination, degraded document media and obscure or unclear typefaces which provoked an erratic response from the OCR software. In some circumstances the document was too degraded to perform any form of manual correction or re-keying. There were other documents where Tong's assumption[9] that OCR software is semi-determinate system no longer held, but were of sufficient quality to be re-keyed or manually corrected. The exclusion of documents on which probabilistic methods would function poorly allowed the algorithm to process "good quality" documents where there was sufficient certainty that the results would be acceptable. The operators worked on the remaining documents.

4 RESULTS

The probabilistic approach worked well on word errors which were not the result of errant splitting and joining and had a small number of character errors. The probabilistic approach functioned adequately on words with a larger number of character errors, however there were a significant number of incorrect choices which declined with the increasing length of the error. The same results were gained with split and joined word errors. The probabilistic approach functioned well on good quality documents because they contained more errors with a small number of character errors. Accuracy declined rapidly with decreasing document quality because of the increased number of split and joined words errors as well errors with increased number of incorrect characters. The probabilistic approach was tested on a documents which were earmarked for manual processing only and for a large number of errors there were no suggested replacements.

The rule based pre-processors corrected a larger proportion of errors than the probabilistic technique because error frequency followed a Zipf

distribution, consequently common errors constituted a very large proportion of the total error count.

This approach reduced the number of documents that needed to be processed from 9 million to 2.2 million. Approximately 5 million documents were processed by probabilistic methods and 1.8 million were rejected as too poor to process. This allowed the reduction of time required to transfer the remaining documents to electronic storage from 9 years to 2 years. It had taken the previous three years to process three million documents manually. If the project had been approached in this manner from the beginning it is estimated that the total project length would have been less than 3 years, which was the original project estimate. The project was two years late.

5 CONCLUSION

Transferring large numbers of less than pristine documents to a digital library / storage with a high degree of accuracy is a time consuming process. Manual correction / re-keying is only feasible if there are sufficiently large numbers of staff or the document count is reasonably small. Probabilistic methods work well on pristine documents with errors which have a low number of character errors, but their performance declines dramatically as media quality drops. Rule based methods are more robust as quality declines. Error frequency follows a Zipf distribution, consequently correcting common errors will have disproportionate effect on document quality. Portuguese has it's own unique challenges with accents and the "ce de cedilha (ç)" which OCR software frequently misinterprets.

Companies which attempt to transfer large numbers of documents to electronic storage via OCR software with the text requiring certain degree of accuracy should consider automatic methods of correcting text. The reduction of the time required for manual processing equates to a saving in costs which will pay a programmers time in constructing the text correct algorithms. Automatic text correction should be considered from the beginning of the project, not when the project is in obvious trouble. The economic case of automatic text correction methods increases with the size of the document collection.

REFERENCES

1. Pirkola A, Keskustalo H, Leppanen E, Kansala A., and Jarvelin K. Targeted s-gram matching: a novel n-gram matching technique for cross- and monolingual word form variants. Information Research, 2002.
2. J.J. Almeida and Ulisses Pinto. Jspell – um módulo para análise léxica genérica de linguagem natural. In Actas do X Encontro da Associão Portuguesa de Linguística, pages 1-15, 1994, 1995.
3. Eric Brill and Robert More. An improved error model for noisy channel spelling correction. In Annual Meeting of the ACL.
4. Ethan Miller Dan, Dan Shen, Junli Liu, Charles Nicholas, and Ting Chen. Techniques for gigabyte-scale n-gram based information retrieval on personal computers. In Personal Computers, Proceedings of the 1999 International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA 99), Las Vegas, NV, 1999.
5. S. M. Harding, W. B. Croft, and C. Weir. Probabilistic retrieval of ocr degraded text using n-grams. In Research and Advanced Technology for Digital Libraries.
6. JSTOR. JSTOR OCR rates. fsearch-sandbox.jstor.org/about/images.html, consulted in 2008.
7. W. Li. Random texts exhibit zipf's-law-like word frequency distribution. In Information Theory, IEEE Transactions on.
8. Lasko TA and Hauser SE. Approximate string matching algorithms for limited-vocabulary ocr output correction. In Proceedings of SPIE.
9. Xiang Tong and David A. Evans. A statistical approach to automatic ocr error correction in context. In Proceedings of the Fourth Workshop on Very Large Corpora (WVLC-4, pages 88–100, 1996.
10. I. Written and D. Bainbridge. How to Build a Digital Library. Morgan Kaufmann, 2003.

BRETT DRURY
LIAAD-INESC, PORTUGAL
E-MAIL: <BRETT.DRURY@GMAIL.COM>

JOSE JOÃO ALMEIDA
UNIVERSITY OF MINHO, PORTUGAL
E-MAIL: <JJ@DI.UMINHO.PT>

Reviewing Committee of the Volume

Eneko Agirre
Sivaji Bandyopadhyay
Roberto Basili
Christian Boitet
Nicoletta Calzolari
Dan Cristea
Alexander Gelbukh
Gregory Grefenstette
Eva Hajičová
Yasunari Harada
Graeme Hirst
Eduard Hovy
Nancy Ide
Diana Inkpen
Alma Kharrat
Adam Kilgarri
Igor Mel'čuk
Rada Mihalcea
Ruslan Mitkov
Dunja Mladeníc
Masaki Murata
Vivi Nastase
Nicolas Nicolov
Kemal Oflazer
Constantin Orasan
Maria Teresa Pazienza
Ted Pedersen
Viktor Pekar
Anselmo Peñas
Stelios Piperidis
James Pustejovsky
Fuji Ren
Fabio Rinaldi
Roracio Rodriguez
Vasile Rus
Franco Salvetti
Serge Sharo
Grigori Sidorov
Thamar Solorio
Juan Manuel Torres-Moreno
Hans Uszkoreit
Manuel Vilares Ferro
Leo Wanner
Yorick Wilks
Annie Zaenen

Additional Referees

Rodrigo Agerri
Muath Alzghool
Javier Artiles
Bernd Bohnet
Ondřej Bojar
Nadjet Bouayad-Agha
Luka Bradesko
Janez Brank
Julian Brooke
Miranda Chong
Silviu Cucerzan
Lorand Dali
Víctor Manuel Darriba Bilbao
Amitava Das
Dipankar Das
Arantza Díaz de Ilarraza
Kohji Dohsaka
Iustin Dornescu
Asif Ekbal
Santiago Fernández Lanza
Robert Foster
Oana Frunza

René Arnulfo García Hernández
Ana García-Serrano
Byron Georgantopoulos
Chikara Hashimoto
Laura Hasler
William Headden
Maria Husarciuc
Adrian Iftene
Iustina Ilisei
Ikumi Imani
Aminul Islam
Toshiyuki Kanamaru
Fazel Keshtkar
Jason Kessler
Michael Kohlhase
Olga Kolesnikova
Natalia Konstantinova
Valia Kordoni
Hans-Ulrich Krieger
Geert-Jan Kruij
Yulia Ledeneva
Yang Liu
Oier Lopez de Lacalle
Fernando Magán-Muñoz
Aurora Marsye
Kazuyuki Matsumoto
Alex Moruz
Sudip Kumar Naskar
Peyman Nojoumian
Blaz Novak
Inna Novalija
Tomoko Ohkuma
Bostjan Pajntar
Partha Pakray
Pavel Pecina
Ionut Cristian Pistol
Natalia Ponomareva
Marius Raschip
Luz Rello
Francisco Ribadas
German Rigau
Alvaro Rodrigo
Franco Salvetti
Kepa Sarasola
GeroldSchneider
Marc Schroeder
Ivonne Skalban
Simon Smith
Mohammad G. Sorba
Tadej Štajner
Sanja Štajner
Jan Strakova
Xiao Sun
Masanori Suzuki
Motoyuki Suzuki
Motoyuki Takaai
Irina Temnikova
Zhi Teng
Nenad Tomasev
Eiji Tomida
Sulema Torres
Mitra Trampas
Diana Trandabat
Stephen Tratz
Yasushi Tsubota
Hiroshi Umemoto
Masao Utiyama
Andrea Varga
Tong Wang
Ye Wu
Keiji Yasuda
Zhang Yi
Daisuke Yokomori
Caixia Yuan
Zdeně Zabokrtský
Venta Zapirain
Daniel Zeman
Hendrik Zender