A Distributed Database System for Developing Ontological and Lexical Resources in Harmony

Aleš Horák[1], Piek Vossen[2], and Adam Rambousek[1]

[1] Faculty of Informatics
Masaryk University
Botanická 68a, 602 00 Brno
Czech Republic
hales@fi.muni.cz, xrambous@fi.muni.cz
[2] Faculteit der Letteren
Vrije Universiteit van Amsterdam
e Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
Piek.Vossen@irion.nl

Abstract. In this article, we present the basic ideas of creating a new information-rich lexical database of Dutch, called Cornetto, that is interconnected with corresponding English synsets and a formal ontology. The Cornetto database is based on two existing electronic dictionaries - the Referentie Bestand Nederlands (RBN) and the Dutch wordnet (DWN). The former holds FrameNet-like information for Dutch and the latter is structured as the English wordnet. In Cornetto, three different collections are maintained for lexical units, synsets and ontology terms. The database interlinks the three collections and aims at clarifying the relations between them. The organization and work processes of the project are briefly introduced.
We also describe the design and implementation of new tools prepared for the lexicographic work on the Cornetto project. The tools are based on the DEB development platform and behave as special dictionary clients for the well-known DEBVisDic wordnet editor and browser.

1 Introduction

Lexical data and knowledge resources has rapidly developed in recent years both in complexity and size. The maintenance and development of such resources require powerful database systems with specific demands. In this paper, we present an extension of the DEBVisDic environment [1] for the development of a lexical semantic database system for Dutch that is built in the Cornetto project. The system holds 3 different types of databases that are traditionally studied from different paradigms: lexical units from a lexicological tradition, synsets within the wordnet framework and an ontology from a formal point of view. Each of these databases represents a different view on meaning. The database system is specifically designed to create relations between these databases and to allow to

Learning finite state transducers using bilingual phrases[1]

Jorge González, Germán Sanchis, and Francisco Casacuberta

Instituto Tecnológico de Informática
Universidad Politécnica de Valencia
{jgonzalez, gsanchis, fcn}@dsic.upv.es

Abstract. Statistical Machine Translation is receiving more and more attention every day due to the success that the phrase-based alignment models are obtaining. However, despite their power, state-of-the-art systems using these models present a series of disadvantages that lessen their effectiveness in working environments where temporal or spacial computational resources are limited. A finite-state framework represents an interesting alternative because it constitutes an efficient paradigm where quality and realtime factors are properly integrated in order to build translation devices that may be of help for their potential users. Here, we describe a way to use the bilingual information in a phrase-based model in order to implement a phrase-based ngram model using finite state transducers. It will be worth the trouble due to the notable decrease in computational requirements that finite state transducers present in practice with respect to the use of some well-known stack-decoding algorithms. Results for the French-English EuroParl benchmark corpus from the 2006 Workshop on Machine Translation of the ACL are reported.

1 Introduction

Machine translation (MT) is an important area of Information Society Technologies in different research frameworks of the European Union. While the development of a classical MT system requires a great human effort, *Statistical machine translation* (SMT) has proved to be an interesting framework due to being able to automatically build MT systems if adequate parallel corpora are provided [1].

Given a sentence $\mathbf{s}$ from a source language, it is commonly accepted that a convenient way to express the SMT problem is through the Bayes' rule [1]:

$$\hat{\mathbf{t}} = \operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}|\mathbf{s}) = \operatorname*{argmax}_{\mathbf{t}} \Pr(\mathbf{t}) \cdot \Pr(\mathbf{s}|\mathbf{t}) \tag{1}$$

where $\hat{\mathbf{t}}$ stands for the most likely hypothesis, according to the model, from all the possible output sentences $\mathbf{t}$. $\Pr(\mathbf{t})$ is frequently approached by a *language model*, which assigns high probability to well formed target sentences, and $\Pr(\mathbf{s}|\mathbf{t})$ is modelled by a *translation model* that is based on stochastic dictionaries and alignment models [2, 3].

[1] This work has been partially supported by the EC (FEDER) and the Spanish projects TIN2006-15694-C02-01 and the Consolider Ingenio 2010 CSD2007-00018

SIGNUM
A graph algorithm for terminology extraction

Axel-Cyrille Ngonga Ngomo[1]

University of Leipzig, Johannisgasse 26, Leipzig D-04103, Germany,
ngonga@informatik.uni-leipzig.de,
WWW home page: http://bis.uni-leipzig.de/AxelNgonga

Abstract. Terminology extraction is an essential step in several fields of natural language processing such as dictionary and ontology extraction. In this paper, we present a novel graph-based approach to terminology extraction. We use SIGNUM, a general purpose graph-based algorithm for binary clustering on directed weighted graphs generated using a metric for multi-word extraction. Our approach is totally knowledge-free and can thus be used on corpora written in any language. Furthermore it is unsupervised, making it suitable for use by non-experts. Our approach is evaluated on the TREC-9 corpus for filtering against the MESH and the UMLS vocabularies.

1 Introduction

Terminology extraction is an essential step in many fields of natural language processing, especially when processing domain-specific corpora. Current algorithms for terminology extraction are most commonly knowledge-driven, using differential analysis and statistical measures for the extraction of domain specific termini. These methods work well, when a large, well-balance reference corpus for the language to process exists. Yet such datasets exist only for a few of the more than 6,000 languages currently in use on the planet. The need is thus for knowledge-free approaches to terminology extraction. In this work, we propose the use of a graph-based clustering algorithm on graphs generated using techniques for the extraction of multi-word units (MWUs). After presenting work related to MWU extraction, we present the metric for MWU extraction used: SRE. This metric is used to generate a directed graph on which SIGNUM is utilized. We present the results achieved using several graph configurations and sizes and show that SIGNUM improves terminology extraction. In order to evaluate our approach, we used the Medical Subject Headings (MESH), with which the TREC-9 collection was tagged, and the Unified Medical Language System (UMLS) vocabularies as gold standards. Last, we discuss some further possible applications of SIGNUM and the results generated using it.

XTM: A Robust Temporal Text Processor

Caroline Hagège[1], Xavier Tannier[2]

[1] Xerox Research Centre Europe, 6 Chemin de Maupertuis, 38240 Meylan ,France
Caroline.Hagege@xrce.xerox.com
[2] LIMSI, 91403 Orsay, France
Xavier.Tannier@limsi.fr

Abstract. We present in this paper the work that has been developed at [hidden name] to build a robust temporal text processor. The aim of this processor is to extract events described in texts and to link them, when possible, to a temporal anchor. Another goal is to be able to establish temporal ordering between the events expressed in texts. One of the originalities of this work is that the temporal processor is coupled with a syntactico-semantic analyzer. The temporal module takes then advantage of syntactic and semantic information extracted from text and at the same time, syntactic and semantic processing benefits from the temporal processing performed. As a result, analysis and management of temporal information is combined with other kinds of syntactic and semantic information, making possible a more refined text understanding processor that takes into account the temporal dimension.

1 Motivation

Although interest in temporal and aspectual phenomena is not new in NLP and AI, temporal processing of real texts is a topic that has been of growing interest in recent years (see [5]). The usefulness of temporal information has become clear for a wide range of applications like multi-document summarization, question/answering systems (see for instance [10]) and information extraction applications. For presenting search results, Google also offers now, in an experimental way, a timeline view to provide results of a search (see www.google.com/experimental). Temporal taggers and annotated resources such as TimeBank ([7]) have been developed. An evaluation campaign for temporal processing has also been organized recently (see [11]).

Arabic/English Multi-Document Summarization with CLASSY—The Past and the Future

Judith D. Schlesinger[1], Dianne P. O'Leary[2], and John M. Conroy[1]

[1] IDA/Center for Computing Sciences, Bowie MD 20715, USA,
[judith,conroy]@super.org
[2] University of Maryland, CS Dept. and Inst. for Advanced Computer Studies,
College Park MD 20742, USA,
oleary@cs.umd.edu

Abstract. Automatic document summarization has become increasingly important due to the quantity of written material generated world-wide. Generating good quality summaries enables users to cope with larger amounts of information.

English-document summarization is a difficult task. Yet it is not sufficient. Environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events.

CLASSY (Clustering, Linguistics, And Statistics for Summarization Yield) is an automatic, extract-generating, summarization system that uses linguistic trimming and statistical methods to generate generic or topic(/query)-driven summaries for single documents or clusters of documents. CLASSY has performed well in the Document Understanding Conference (DUC) evaluations and the Multi-lingual (Arabic/English) Summarization Evaluations (MSE).

We present a description of CLASSY. We follow this with experiments and results from the MSE evaluations and conclude with a discussion of on-going work to improve the quality of the summaries–both English-only and multi-lingual–that CLASSY generates.

1 Introduction

Automatic multi-document summarization poses interesting challenges to the Natural Language Processing (NLP) community. In addition to addressing single document summarization issues such as determining the relevant information, pronoun resolution, and coherency of the generated summary, multi-document summary-generating systems must be capable of drawing the "best" information from a set of documents.

Automatic single document text summarization [11] has long been a field of interest, beginning in the 1950s, with a recent renaissance of activity beginning in the 1990s. System generated single document summaries for English are generally of good quality. Therefore, NIST ended single document summarization evaluation after the 2002 Document Understanding Conference (DUC). See [17] for DUC research papers and results over the years.

In contrast to the single document task, summarization of multiple documents written in English remains an ongoing research effort. A wide range of strategies to analyze documents in a collection and then synthesize/condense information to produce a multi-document summary have been explored by various research groups. System performance has improved but still lags behind human performance.

Nevertheless, environmental, economic, and other global issues make it imperative for English speakers to understand how other countries and cultures perceive and react to important events. Thus it is vital that English speakers be able to access documents in a variety of languages.

Learning Spanish-Galician Translation Equivalents using a Comparable Corpus and a Bilingual Dictionary

Pablo Gamallo Otero[1] and José Ramom Pichel Campos[2]

[1] Departamento de Língua Espanhola, Faculdade de Filologia
Universidade de Santiago de Compostela, Galiza, Spain
[2] Departamento de Tecnologia Linguística da Imaxin|Software
Santiago de Compostela, Galiza

Abstract. So far, research on extraction of translation equivalents from comparable, non-parallel corpora has not been very popular. The main reason was the poor results when compared to those obtained from aligned parallel corpora. The method proposed in this paper, relying on *seed patterns* generated from external bilingual dictionaries, allows us to achieve similar results to those from parallel corpus. In this way, the huge amount of comparable corpora available via Web can be viewed as a never-ending source of lexicographic information. In this paper, we describe the experiments performed on a comparable, Spanish-Galician corpus.

1 Introduction

There exist many approaches to extract bilingual lexicons from parallel corpora [8, 16, 1, 22, 14]. These approaches share the same basic strategy: first, bitexts are aligned in pairs of segments and, second, word co-ocurrences are computed on the basis of that alignment. They usually reach high score values, namely about 90% precision with 90% recall. Unfortunately, parallel texts are not easily available, in particular for minority languages. To overcome this drawback, different methods to extract bilingual lexicons have been implemented lately using non-parallel, comparable corpora. These methods take up with the idea of using the Web as a huge resource of multilingual texts which can be easily organized as a collection of non-parallel, comparable corpora. A non-parallel and comparable corpus (hereafter "comparable corpus") consists of documents in two or more languages which are not translation of each other and deal with similar topics. However, the accuracy scores of such methods are not as good as those reached by the strategies based on aligned parallel corpora. So far, the highest values have not improved an 72% accuracy [18], and that's without taking into consideration the coverage of the extracted lexicon over the corpus.

This paper proposes a new method to extract bilingual lexicons from a POS tagged comparable corpus. Our method relies on the use of a bilingual dictionary to identify bilingual correlations between pairs of lexico-syntactic patterns. Such

Bilingual Segmentation for Alignment and Translation

Chung-Chi Huang, Wei-Teh Chen, and Jason S. Chang

Information Systems and Applications, NTHU, HsingChu, Taiwan 300 R.O.C
{u901571, weitehchen, jason.jschang}@gmail.com

Abstract. We propose a method that bilingually segments sentences in languages with no clear delimiter for word boundaries. In our model, we first convert the search for the segmentation into a sequential tagging problem, allowing for a polynomial-time dynamic-programming solution, and incorporate a control to balance monolingual and bilingual information at hand. Our bilingual segmentation algorithm, the integration of a monolingual language model and a statistical translation model, is devised to tokenize sentences more suitably for bilingual applications such as word alignment and machine translation. Empirical results show that bilingually-motivated segmenters outperform pure monolingual one in both the word-aligning (12% reduction in error rate) and the translating (5% improvement in BLEU) tasks, suggesting monolingual segmentation is useful in some aspects but, in a sense, not built for bilingual researches.

1. Introduction

A statistical translation model (STM) is a model that, relied on lexical information or syntactic structures of languages involved, decodes the process of human translation and that, in turn, detects most appropriate word correspondences in parallel sentences. Ever since the pioneer work of (Brown et al., 1993), the field of STMs has drawn myriads of attention. Some researchers exploited Hidden Markov models to approach relatively monotonic word-aligning problems in similarly-structured language pairs

Non-Interactive OCR Post-Correction for Giga-Scale Digitization Projects

Martin Reynaert

Induction of Linguistic Knowledge, Tilburg University, The Netherlands

Abstract. This paper proposes a non-interactive system for reducing the level of OCR-induced typographical variation in large text collections, contemporary and historical. Text-Induced Corpus Clean-up or TICCL (pronounce 'tickle') focuses on high-frequency words derived from the corpus to be cleaned and gathers all typographical variants for any particular focus word that lie within the predefined Levenshtein distance (henceforth: LD). Simple text-induced filtering techniques help to retain as many as possible of the true positives and to discard as many as possible of the false positives. TICCL has been evaluated on a contemporary OCR-ed Dutch text corpus and on a corpus of historical newspaper articles, whose OCR-quality is far lower and which is in an older Dutch spelling. Representative samples of typographical variants from both corpora have allowed us not only to properly evaluate our system, but also to draw effective conclusions towards the adaptation of the adopted correction mechanism to OCR-error resolution. The performance scores obtained up to LD 2 mean that the bulk of undesirable OCR-induced typographical variation present can fully automatically be removed.

1 Introduction

This paper reports on efforts to reduce the massive amounts of non-word word forms created by OCRing large collections of printed text in order to bring down the type-token ratios of the collections to the levels observed in contemporary 'born-digital' collections of text. We report on post-correction of OCR-errors in large corpora of the Cultural Heritage. On invitation by the National Library of The Netherlands (Koninklijke Bibliotheek - Den Haag) we have worked on contemporary and historical text collections. The contemporary collection comprises the published Acts of Parliament (1989-1995) of The Netherlands, referred to as 'Staten-Generaal Digitaal' (henceforth: SGD)[1]. The historical collection is referred to as 'Database Digital Daily Newspapers' (henceforth: DDD)[2], which comprises a selection of daily newspapers published between 1918 and 1946 in the Netherlands. The historical collection was written in the Dutch spelling 'De Vries-Te Winkel', which in 1954 was replaced by the more contemporary spelling

[1] URL: http://www.statengeneraaldigitaal.nl/

[2] URL: http://kranten.kb.nl/ In actual fact, this collection represents the result of a pilot project which is to be incorporated into the far more comprehensive DDD.

Domain Information for Fine-grained Person Name Categorization

Zornitsa Kozareva, Sonia Vazquez and Andres Montoyo

Departamento de Lenguajes y Sistemas Informaticos
Universidad de Alicante
{zkozareva,svazquez,montoyo}@dlsi.ua.es

Abstract. Named Entity Recognition became the basis of many Natural Language Processing applications. However, the existing coarse-grained named entity recognizers are insufficient for complex applications such as Question Answering, Internet Search engines or Ontology population. In this paper, we propose a domain distribution approach according to which names which occur in the same domains belong to the same fine-grained category. For our study, we generate a relevant domain resource by mapping and ranking the words from the WordNet glosses to their WordNetDomains. This approach allows us to capture the semantic information of the context around the named entity and thus to discover the corresponding fine-grained name category. The presented approach is evaluated with six different person names and it reaches 73% f-score. The obtained results are encouraging and perform significantly better than a majority baseline.

1 Introduction

The Named Entity Recognition (NER) task was first introduced in the Message Understanding Conference (MUC) as it was discovered that most of the elements needed for the template filling processes in Information Extraction systems are related to names of people, organizations, locations, monetary, date, time and percentage expressions.

There are two main paradigms for NER. In the first one, NEs are recognized on the basis of a set of rules and gazetteer lists [6], [1]. The coverage of these systems is very high, however they depend on the knowledge of their human creator, the number of hand-crafted rules and the kind of entries in the gazetteer lists. In addition, NER rule-based systems are domain and language dependent, therefore they need lots of time in order to be developed and to be adapted.

The other paradigm is machine learning (ML) based NER. Given a set of feature vectors characterizing a named entity, a machine learning algorithm learns these properties and then assigns automatically NE categories to unseen entities. These systems are easily adaptable to different domains, they can function with language-independent characteristics [16], [17], however, their main drawback is related to the number of hand-labeled examples from which the ML system

A Probabilistic Model for Guessing Base Forms of New Words by Analogy

Krister Lindén

Department of General Linguistics, P.O. Box 9, FIN-00014 University of Helsinki
Krister.Linden@Helsinki.fi

Abstract. Language software applications encounter new words, e.g., acronyms, technical terminology, loan words, names or compounds of such words. Looking at English, one might assume that they appear in base form, i.e., the lexical look-up form. However, in more highly inflecting languages like Finnish or Swahili only 40-50 % of new words appear in base form. In order to index documents or discover translations for these languages, it would be useful to reduce new words to their base forms as well. We often have access to analyzes for more frequent words which shape our intuition for how new words will inflect. We formalize this into a probabilistic model for lemmatization of new words using analogy, i.e., guessing base forms, and test the model on English, Finnish, Swedish and Swahili demonstrating that we get a recall of 89-99 % with an average precision of 76-94 % depending on language and the amount of training material.

1 Introduction

New words and new usages of old words are constantly finding their way into daily language use. This is particularly prominent in quickly developing domains such as biomedicine and technology. Humans deal with new words based on previous experience: we treat them by analogy to known words. The new words are typically acronyms, technical terminology, loan words, names or compounds containing such words. They are likely to be unknown by most hand-made morphological analyzers. In some applications, hand-made guessers are used for covering this low-frequency vocabulary.

Unsupervised acquisition of morphologies from scratch has been studied as a general problem of morphology induction in order to automate the morphology building procedure. For overviews, see [8] and [2]. The problem is alleviated by the fact that there often are dictionaries available with common base forms or word roots for the most frequent words. If the inflectional patterns can be learned approximately from a corpus, the most common base forms can be checked against a dictionary in order to boost the performance of the methods. However, when we approach the other end of the spectrum, we have very rare words for which there are no ready base forms available in dictionaries and for heavily inflecting languages only 40-50 % of the words appear in base form in a

Linguistic Support for Revising and Editing

Cerstin Mahlow and Michael Piotrowski

University of Zurich
Institute of Computational Linguistics
Binzmühlestrasse 14
8050 Zürich, Switzerland
{mahlow, mxp}@cl.uzh.ch

Abstract. Revising and editing are important parts of the writing process. In fact, multiple revision and editing cycles are crucial for the production of high-quality texts. However, revising and editing are also tedious and error-prone, since changes may introduce new errors.

Grammar checkers, as offered by some word processors, are not a solution. Besides the fact that they are only available for few languages, and regardless of the questionable quality, their conceptual approach is not suitable for experienced writers, who actively create their texts. Word processors offer few, if any, functions for handling text on the same cognitive level as the author: While the author is thinking in high-level linguistic terms, editors and word processors mostly provide low-level character oriented functions. Mapping the intended outcome to these low-level operations is distracting for the author, who now has to focus for a long time on small parts of the text. This results in a loss of global overview of the text and in typical revision errors (duplicate verbs, extraneous conjunctions, etc.).

We therefore propose functions for text processors that work on the conceptual level of writers. These functions operate on linguistic elements, not on lines and characters. We describe how these functions can be implemented by making use of NLP methods and linguistic resources.

1 Introduction

Writing a text involves several steps and various tasks, starting from planning activities to writing a first draft and then revising and editing[1] to get to the final version. Revising and editing are typically recursive processes, continuing until an acceptable state is achieved.

Writing means creating a coherent text from linguistic elements, such as words, phrases, clauses and sentences. When revising and editing texts, authors are working with these elements, arranging and rearranging them, exchanging them for others, maybe even "playing" with them.

In this paper we will try to develop the idea of tools based on linguistics to support writers in the writing process, especially during revising and editing.

First, to get an idea of the abstraction level on which writers are thinking about their texts, we will have a look at recommendations for writers and editors: What are the

[1] In composition research, a distinction is typically made between *revising*, which takes place on the discourse level, and *editing*, which takes place at the sentence and word level (see [1] for a discussion).

Discovering Word Senses from Text Using Random Indexing

Niladri Chatterjee [1] and Shiwali Mohan [2]

[1] Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India 110016
niladri@maths.iitd.ac.in
[2] Yahoo! Research and Development India, Bangalore, India 560 071
shiwali@yahoo-inc.com

Abstract. Random Indexing is a novel technique for dimensionality reduction while creating Word Space model from a given text. This paper explores the possible application of Random Indexing in discovering word senses from the text. The words appearing in the text are plotted onto a multi-dimensional Word Space using Random Indexing. The geometric distance between words is used as an indicative of their semantic similarity. Soft Clustering by Committee algorithm (CBC) has been used to constellate similar words. The present work shows that the Word Space model can be used effectively to determine the similarity index required for clustering. The approach does not require parsers, lexicons or any other resources which are traditionally used in sense disambiguation of words. The proposed approach has been applied to TASA corpus and encouraging results have been obtained.

1 Introduction

Automatic disambiguation of word senses has been an interesting challenge since the very beginning of computational linguistics in 1950s [1]. Various clustering techniques, such as bisecting K-means [2], Buckshot [3], UNICON [4], Chameleon [5], are being used to discover different senses of words. These techniques constellate words that have been used in similar contexts in the text. For example, when the word ‘plant’ is used in the living sense, it is clustered with words like ‘tree’, ‘shrub’, ‘grass’ etc. But when it is used in the non-living sense, it is clustered with ‘factory’, ‘refinery’ etc. Similarity between words is generally defined with the help of existing lexicons, such as WordNet [6], or parsers (e.g. Minipar [7]).

Word Space model [8] has long been in use for semantic indexing of text. The key idea of Word Space model is to assign vectors to the words in high dimensional vector spaces, whose relative directions are assumed to indicate semantic similarity. The Word Space model has several disadvantages: *sparseness* of the data and *high dimensionality* of the semantic space when dealing with real world applications and large size data sets. Random Indexing [9] is an approach developed to deal with the problem of high dimensionality in Word Space model.

EFL Learner Reading Time Model for Evaluating Reading Proficiency

Katsunori Kotani[1/2], Takehiko Yoshimi[3/2], Takeshi Kutsumi[4],
Ichiko Sata[4], and Hitoshi Isahara[2]

[1] Kansai Gaidai University
16-1 Nakamiya Higashino-cho, Hirakata, Osaka, Japan
kat@khn.nict.go.jp
[2] National Institute of Information and Communications Technology
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
[3] Ryukoku University
1-5 Yokotani, Seta Oe-cho, Otsu, Shiga, Japan
[4] Sharp Corporation
492 Minosho-cho, Yamatokoriyama, Nara, Japan

Abstract. We propose a reading time model for learners of English as a foreign language (EFL) that is based on a learner's reading proficiency and the linguistic properties of sentences. Reading proficiency here refers to a learner's reading score on the Test of English for International Communications (TOEIC), and the linguistic properties are the lexical, syntactic and discourse complexities of a sentence. We used natural language processing technology to automatically extract these linguistic properties, and developed a model using multiple regression analysis as a learning algorithm in combining the learner's proficiency and linguistic properties. Experimental results showed that our reading time model predicted sentence-reading time with a 22.9% error rate, which is lower than the models constructed based on linguistic properties proposed in previous studies.

1 Introduction

One of the critical issues in learning or teaching a foreign language is learners' individual differences in proficiency. Unlike first language acquisition, proficiencies in acquiring a foreign language vary greatly. Thus, a language teacher has to understand each learner's problems and help the learner contend with them. The learners' problems principally arise from lack of lexical or syntactic knowledge. For instance, if a learner encounters a lexical item the meaning of which the learner does not know, he or she has to guess the meaning based on contextual information. Reading such a sentence should take more time than reading a sentence without unknown lexical items. Given this, some learners' problems can be identified by measuring his or her

Portuguese Pronoun Resolution: Resources and Evaluation

Ramon Ré Moya Cuevas, Willian Yukio Honda, Diego Jesus de Lucena,
Ivandré Paraboni and Patrícia Rufino Oliveira [1]

[1] Escola de Artes, Ciências e Humanidades – Universidade de São Paulo (EACH / USP)
Av.Arlindo Bettio, 1000 - 03828-000, São Paulo, Brazil
{fusion, kio, diego.si, ivandre, proliveira}@usp.br

Abstract. Despite being one of the most widely-spoken languages in the world, Portuguese remains a relatively resource-poor language, for which only in recently years NLP tools such as parsers, taggers and (fairly) large corpora have become available. In this work we describe the task of pronominal co-reference annotation and resolution in Portuguese texts, in which we take advantage of information provided by a tagged corpus and a simple annotation tool that has been developed for this purpose. Besides developing some of these basic resources from scratch, our ultimate goal is to investigate the multilingual resolution of Portuguese personal pronouns to improve the accuracy of their translations to both Spanish and English in an underlying MT project.

1 Introduction

Pronoun resolution – the task of identifying the antecedent of a pronoun in discourse - is often crucial to a variety of NLP applications, ranging from Text Summarization to Machine Translation (MT), and it has long been recognised as a challenging computational problem for which existing approaches - either making use of learning techniques or otherwise - often need to resort to large amounts of knowledge produced by standard NLP tools such as parsers and POS taggers applied on annotated corpora [1].

The challenge is however considerably increased if we speak of languages for which some of these basic resources are still under development, or may have only recently become available. This is the case, for instance, of Portuguese, one of the most widely-spoken languages in the world, and which still lacks somewhat behind as a relatively resource-poor language in NLP.

Our own work focuses on pronoun resolution as required by a Portuguese-Spanish-English MT project under development. Our present choice - Portuguese third person plural pronouns (“Eles/Elas”) - is based on the assumption that these (as well as their Spanish counterparts) are less prone to ambiguity, and arguably easier to resolve than the English equivalent (“They”), which may suggest an interesting multilingual approach to anaphora resolution not unlike [7].

A Preliminary Study on the Robustness and Generalization of Role Sets for Semantic Role Labeling

Beñat Zapirain[1], Eneko Agirre[1], and Lluís Màrquez[2]

[1] IXA NLP Group
University of The Basque Country
{benat.zapirain,e.agirre}@ehu.es
[2] TALP Research Center
Technical University of Catalonia
lluism@lsi.upc.edu

Abstract. Most Semantic Role Labeling (SRL) systems rely on available annotated corpora, being PropBank the most widely used corpus so far. Propbank role set is based on theory-neutral numbered arguments, which are linked to fine grained verb-dependant semantic roles through the verb framesets. Recently, thematic roles from the computational verb lexicon VerbNet have been suggested to be more adequate for generalization and portability of SRL systems, since they represent a compact set of verb-independent general roles widely used in linguistic theory. Such thematic roles could also put SRL systems closer to application needs. This paper presents a comparative study of the behavior of a state-of-the-art SRL system on both role role sets based on the SemEval-2007 English dataset, which comprises the 50 most frequent verbs in PropBank.

1 Introduction

Semantic Role Labeling is the problem of analyzing clause predicates in open text by identifying arguments and tagging them with semantic labels indicating the role they play with respect to the verb. Such sentence–level semantic analysis allows to determine "who" did "what" to "whom", "when" and "where", and, thus, characterize the participants and properties of the *events* established by the predicates. This kind of semantic analysis is very interesting for a broad spectrum of NLP applications (information extraction, summarization, question answering, machine translation, etc.), since it opens the avenue for exploiting the semantic relations among linguistic constituents.

The increasing availability of large semantically annotated corpora, like PropBank and FrameNet, has contributed to increase the interest on the automatic development of Semantic Role Labeling systems in the last five years. Since Gildea and Jurafsky's initial work "Automatic Labeling of Semantic Roles" [3] on FrameNet-based SRL, many researchers have devoted their efforts on this exciting and relatively new task. Two evaluation exercises on SRL were conducted by the 'shared tasks' of CoNLL-2004 and CoNLL-2005 conferences [1, 2], bringing to scene a comparative analysis of almost 30 competitive systems trained on the PropBank corpus. From there, PropBank became the most widely used corpus for training SRL systems.

Lexical Cohesion Based Topic Modeling for Summarization

Gonenc Ercan and Ilyas Cicekli

Dept. of Computer Engineering
Bilkent University, Ankara, Turkey
ercangu@cs.bilkent.edu.tr, ilyas@cs.bilkent.edu.tr

Abstract. In this paper, we attack the problem of forming extracts for text summarization. Forming extracts involves selecting the most representative and significant sentences from the text. Our method takes advantage of the lexical cohesion structure in the text in order to evaluate significance of sentences. Lexical chains have been used in summarization research to analyze the lexical cohesion structure and represent topics in a text. Our algorithm represents topics by sets of co-located lexical chains to take advantage of more lexical cohesion clues. Our algorithm segments the text with respect to each topic and finds the most important topic segments. Our summarization algorithm has achieved better results, compared to some other lexical chain based algorithms.

Keywords: text summarization, lexical cohesion, lexical chains.

1 Introduction

Summary is the condensed representation of a document's content. For this reason, they are low cost indicators of relevance. Summaries could be used in different applications both as informative tools for humans and as similarity functions for information retrieval applications. Summaries could be displayed in search results as an informative tool for the user. The user can measure the relevance of a document that he gets as a result of a search on Internet by just looking its summary. In order to measure similarities between documents, their summaries can be used instead of whole documents, and indexing algorithms can index their summaries instead of whole documents.

Depending on its content, summaries can be categorized into two groups: *extract* and *abstract*. If a summary is formed of sentences that appear in the original text, it is called as an *extract*. A summarization system targeting extracts should evaluate each sentence for its importance. Abstracts are the summaries that are formed from paraphrased or generated sentences. Building abstracts has additional challenges.

Different clues can be exploited to evaluate the importance of sentences. There are extractive summarization systems that take advantage of surface level features like word repetition, position in text, cue phrases and similar features that are easy to compute. Ideally, a summarization system should perform full understanding, which is very difficult and only domain dependant solutions are currently available.

Some summarization algorithms including ours, rely on more sophisticated clues that require deeper analyses of the text. A meaningful text is not a random sequence of

Hybrid Method for Personalized Search in Scientific Digital Libraries

Thanh-Trung Van and Michel Beigbeder

Centre G2I/Département RIM
Ecole Nationale Supérieure des Mines de Saint Etienne
158 Cours Fauriel, 42023 Saint Etienne, France
{van,mbeig}@emse.fr

Abstract. Users of information retrieval systems usually have to repeat the tedious process of searching, browsing, and refining queries until they find relevant documents. This is because different users have different information needs, but user queries are often short and, hence, ambiguous. In this paper we study personalized search in digital libraries using user profile. The search results could be re-ranked by taking into account specific information needs of different people. We study many methods for this purpose: citation-based method, content-based method and hybrid method. We conducted experiments to compare performances of these methods. Experimental results show that our approaches are promising and applicable in digital libraries.

1 Introduction

Search in digital libraries is usually a boring task. Users have to repeat the tedious process of searching, browsing, and refining queries until they find relevant documents. This is because different users have different information needs, but user queries are often short and, hence, ambiguous. For example, the same query "java" could be issued by a person who is interested in geographical information about the Java island or by another person who is interested in the Java programming language. Even with a longer query like "java programming language", we still do not know which kind of document this user wants to find. If she/he is a programmer, perhaps she/he is interested in technical documents about the Java language; however, if she/he is a teacher, perhaps she/he wants to find tutorials about Java programming for her/his course.

From these examples, we can see that different users of an information retrieval system have different information needs. Furthermore, a person can have different interests at different times. A good information retrieval system have to take into account these differences to satisfy its users. This problem could be solved if the system can learn some information about the interests and the preferences of the users and use this information to improve its search results. This information is gathered in *user profiles*. Generally, a user profile is a set of information that represent interests and/or preferences of a user. This information could be collected by implicitly monitoring the user's activities [1, 2] or by

Alignment-based Expansion of Textual Database Fields

Piroska Lendvai

ILK / Communication and Information Sciences
Tilburg University
P.O. Box 90153, 5000 LE, Tilburg
The Netherlands

Abstract. Our study describes the induction of a secondary metadata layer from textual databases in the cultural heritage domain. Metadata concept candidates are detected and extracted from complex fields of a database so that content can be linked to new, finer-grained labels. Candidate labels are mined drawing on the output of Alignment-Based Learning, an unsupervised grammatical inference algorithm, by identifying head - modifier dependency relations in the constituent hypothesis space. The extracted metadata explicitly represent hidden semantic properties, derived from syntactic properties. Candidates validated by a domain expert constitute a seed list for acquiring a partial ontology.

1 Introduction

The data model underlying the structure of a database is often manually constructed, with the risk of becoming out-of-date over time: records that are arranged according to this structure outgrow it as the number of data attributes increases when resources of various formats get merged, updated, and new concepts emerge. These tendencies sometimes result in a mix of attributes joined in an ad-hoc way in loosely defined (free text) columns of the database, typically labelled as *Special remarks*. Such columns are of lower semantic coherence, and are suboptimal for effective database querying as they may contain several identical data types, for example numbers, that implicitly describe different, perhaps idiosyncratic properties, of a record.

Consider the following example from the specialRemarks column of a museum collection database:

```
Slides MSH 1975-xviii-27/29, 1975-xix-20/25; tape recording 1975 II
B 297-304. Acquired as gift from the British Museum (Nat. Hist.),
BMNH 1975. 1348
```

If a researcher is searching this column for tape recordings of a certain year, he needs to browse through all slide identification numbers and other ID numbers as well, because retrieving numbers cannot be securely narrowed down any further than to accessing the entire field. A query would be more efficient if the various ID numbers of slides, tape recordings, registration numbers, etc. would be separately

Detecting Expected Answer Relations through Textual Entailment

Matteo Negri, Milen Kouylekov, Bernardo Magnini

Fondazione Bruno Kessler
Via Sommarive, 18 - Povo, Trento, Italy
{negri,kouylekov,magnini}@fbk.eu

Abstract. This paper presents a novel approach to Question Answering over structured data, which is based on Textual Entailment recognition. The main idea is that the QA problem can be recast as an entailment problem, where the text (T) is the question and the hypothesis (H) is a relational pattern, which is associated to "instructions" for retrieving the answer to the question. In this framework, given a question Q and a set of answer patterns P, the basic operation is to select those patterns in P that are entailed by Q. We report on a number of experiments which show the great potentialities of the proposed approach.

1 Introduction

Question Answering (QA) over structured data has been traditionally addressed through a deep analysis of the question in order to reconstruct its logical form, which is then translated in the query language of the target data ([1], [2]). This approach implies a complex mapping between linguistic objects (*e.g.* lexical items, syntactic structures) and data objects (*e.g.* concepts and relations in a knowledge base). Several experiences, however, have shown that such a mapping requires intensive manual work, which represents a bottleneck in the realization of large scale and portable natural language interfaces to structured data.

More recently, Textual Entailment (TE) has been proposed as a unifying framework for applied semantics ([3]), where the need for an explicit representation of a mapping between linguistic objects and data objects can be, at least partially, bypassed through the definition of semantic inferences at the textual level. In this framework, a text (T) is said to entail a hypothesis (H) if the meaning of H can be derived from the meaning of T.

According to the TE framework, in this paper we propose that QA can be approached as an entailment problem, where the text (T) is the question, and the hypothesis (H) is a relational pattern, which is associated to instructions for retrieving the answer to the question. In this framework, given a question Q and a set of relational patterns P=$\{p_1, ..., p_n\}$, the basic operation is to select those patterns in P that are entailed by Q. Instructions associated to patters may be viewed as high precision procedures for answer extraction, which are dependent on the specific data source accessed for answer extraction. In case of QA over

n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation

Lucia Specia[1,2], Baskaran Sankaran[2] and
Maria das Graças Volpe Nunes[1]

[1] NILC/ICMC - Universidade de São Paulo
Trabalhador São-Carlense, 400, São Carlos, 13560-970, Brazil
lspecia@icmc.usp.br, gracan@icmc.usp.br
[2] Microsoft Research India
"Scientia", 196/36, 2nd Main, Sadashivanagar, Bangalore-560080, India
baskaran@microsoft.com

Abstract. Although it has been always thought that Word Sense Disambiguation (WSD) can be useful for Machine Translation, only recently efforts have been made towards integrating both tasks to prove that this assumption is valid, particularly for Statistical Machine Translation (SMT). While different approaches have been proposed and results started to converge in a positive way, it is not clear yet how these applications should be integrated to allow the strengths of both to be exploited. This paper aims to contribute to the recent investigation on the usefulness of WSD for SMT by using n-best reranking to efficiently integrate WSD with SMT. This allows using rich contextual WSD features, which is otherwise not done in current SMT systems. Experiments with English-Portuguese translation in a syntactically motivated phrase-based SMT system and both symbolic and probabilistic WSD models showed significant improvements in BLEU scores.

1 Introduction

The need for Word Sense Disambiguation (WSD) in Machine Translation (MT) systems has been discussed since the early research on MT back in the 1960's. While MT was primarily addressed by rule-based approaches, the consequences of the lack of semantic disambiguation was already emphasized in DARPA's report by Bar-Hillel [2], which resulted in a considerable reduction in the funding for research on MT that time. Meanwhile, WSD grew as an independent research area, without focusing on any particular application.

With the introduction in the 1990's of the Statistical Machine Translation (SMT) approach [3], it has been taken for granted that SMT systems can implicitly address the sense disambiguation problem, especially by their word alignment and target language models. However, the SMT systems normally consider a very short window as context and therefore lack richer information coming from larger contexts or other knowledge sources.

German decompounding in a difficult corpus

Enrique Alfonseca, Slaven Bilac and Stefan Pharies

Google, Inc.
{ealfonseca,slaven,stefanp@google.com}

Abstract. Splitting compound words has proved to be useful in areas such as Machine Translation, Speech Recognition or Information Retrieval (IR). In the case of IR systems, they usually have to cope with noisy data, as user queries are usually written quickly and submitted without review. This work attempts at improving the current approaches for German decompounding when applied to query keywords. The results show an increase of more than 10% in accuracy compared to other state-of-the-art methods.

1 Introduction

The so-called compounding languages, such as German, Dutch, Danish, Norwegian or Swedish, allow the generation of complex words by merging together simpler ones. So, for instance, the German word *Blumensträuße* (flower bouquet) is made up of *Blumen* (flower) and *sträuße* (bouquet). This allows speakers of these languages to easily create new words to refer to complex concepts by combining existing ones, whereas in non-compounding languages these complex concepts would normally be expressed using multiword syntactic constituents.

For many language processing tools that rely on lexicons or language models it is very useful to be able to decompose compounds to increase their coverage. In the case of German, the amount of compounds in medium-size corpora (tens of millions of words) is large enough that they deserve special handling: 5-7% of the tokens and 43-47% of the word forms in German newswire articles are compounds [1, 2]. When decompounding tools are not available, language processing systems for compounding languages must use comparatively much larger lexicons [3]. German decompounders have been used successfully in Information Retrieval [4, 5], Machine Translation [6–8], word prediction systems [1] and Speech Recognition [3, 9].

When decompounding German words from well-formed documents that have undergone editorial review, one can assume that most of the words or compound parts can be found in dictionaries and thesauri and the number of misspellings is low, which greatly simplifies the problem of decompounding. For example, Marek [10] observed that only 2.8% of the compounds in texts from the German computer magazine called *c't* contain at least one unknown part.

On the other hand, when working with web data, which has not necessarily been reviewed for correctness, many of the words are more difficult to analyze. This includes words with spelling mistakes, and texts that, being mostly written

Sense annotation in the Penn Discourse Treebank

Eleni Miltsakaki*, Livio Robaldo+, Alan Lee*, Aravind Joshi*

*Institute for Research in Cognitive Science, University of Pennsylvania
{elenimi, aleewk, joshi}@linc.cis.upenn.edu
+Department of Computer Science, University of Turin
robaldo@di.unito.it

Abstract. An important aspect of discourse understanding and generation involves the recognition and processing of discourse relations. These are conveyed by discourse connectives, i.e., lexical items like *because* and *as a result* or implicit connectives expressing an inferred discourse relation. The Penn Discourse TreeBank (PDTB) provides annotations of the argument structure, attribution and semantics of discourse connectives. In this paper, we provide the rationale of the tagset, detailed descriptions of the senses with corpus examples, simple semantic definitions of each type of sense tags as well as informal descriptions of the inferences allowed at each level.

1 Introduction

Large scale annotated corpora have played and continue to play a critical role in natural language processing. The continuously growing demand for more powerful and sophisticated NLP applications is evident in recent efforts to produce corpora with richer annotations [6], including annotations at the discourse level[2], [8], [4]. The Penn Discourse Treebank is, to date, the largest annotation effort at the discourse level, providing annotations of explicit and implicit connectives. The design of this annotation effort is based on the view that discource connectives are predicates taking clausal arguments. In Spring 2006, the first version of the Penn Discourse Treebank was released, making availalble thousands annotations of discourse connectives and the textual spans that they relate.

Discourse connectives, however, like verbs, can have more than one meaning. Being able to correctly identify the intended sense of connectives is crucial for every natural language task which relies on understanding relationships between events or situations in the discourse. The accuracy of information retrieval from text can be significantly impaired if, for example, a temporal relation anchored on the connective *since* is interpreted as causal.

A well-known issue in sense annotations is identifying the appropriate level of granularity and meaning refinement as well as identifying consistent criteria for making sense distinctions. Even if an 'appropriate' level of granularity can be identified responding to the demands of a specific application, creating a flat set of sense tag is limiting in many ways. Our approach to the annotation of sense

The Role of PP Attachment in Preposition Generation

John Lee[1] and Ola Knutsson[2]

[1] Spoken Language Systems
MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, USA
jsylee@csail.mit.edu
[2] School of Computer Science and Communication
Royal Institute of Technology, Stockholm, Sweden
knutsson@csc.kth.se

Abstract. This paper is concerned with the task of preposition generation in the context of a grammar checker. Relevant features for this task can range from lexical features, such as words and their part-of-speech tags in the vicinity of the preposition, to syntactic features that take into account the attachment site of the prepositional phrase (PP), as well as its argument/adjunct distinction. We compare the performance of these different kinds of features in a memory-based learning framework. Experiments show that using PP attachment information can improve preposition generation accuracy on Wall Street Journal texts.

1 Introduction

Preposition usage is among the more frequent types of errors made by non-native speakers of English. In an analysis of texts [1], written by students in English-as-a-Second-Language classes, errors involving prepositions form the largest category, at about 29%[3]. A system that can automatically detect and correct preposition usage would be of much practical and educational value. Research efforts towards building such a grammar checking system have been described in [2], [3], and [4].

When dealing with preposition errors, the system typically makes two decisions. First, a *preposition generation* model needs to determine the best preposition to use, given its context in the input sentence. It should, for example, predict the preposition "*in*" to be the most likely choice for the input sentence:

Input: *He participated* **at?** *the competition.*
Corrected: *He participated* **in** *the competition.*

If the predicted preposition differs from the original one, a *confidence* model would then need to decide whether to suggest the correction to the user. In this case, confidence in the predicted preposition "*in*" should be much higher than the original "*at*", and correction would be warranted.

[3] As cited in [2].

Growing TreeLex

Anna Kupść[1] and Anne Abeillé[2]

[1] Université de Bordeaux, ERSSàB/SIGNES and IPIPAN;
Université Michel de Montaigne, Domaine Universitaire, UFRL
33607 Pessac Cedex, France
[2] Université Paris7, LLF/CNRS
UMR 7110, CNRS-Université Paris 7, Case 7031, 2, pl. Jussieu
75251 Paris Cedex 05, France
akupsc@u-bordeaux3.fr, anne.abeille@linguist.jussieu.fr

Abstract. TreeLex is a subcategorization lexicon of French, automatically extracted from a syntactically annotated corpus. The lexicon comprises 2006 verbs (25076 occurrences). The goal of the project is to obtain a list of subcategorization frames of contemporary French verbs and to estimate the number of different verb frames available in French in general. A few more frames are discovered when the corpus size changes, but the average number of frames per verb remains relatively stable (about 1.91–2.09 frames per verb).

Key words: Verb valence, subcategorization, treebank

1 Introduction

The paper presents TreeLex, a subcategorization lexicon for French, automatically extracted from a syntactically annotated corpus.

Information about the combinatory potential of a predicate, i.e., the number and the type of its arguments, is called a subcategorization frame or valence. For example, the verb *embrasser* 'kiss' requires two arguments (the subject and an object), both of them realized as a noun phrase, whereas the predicative adjective *fier* 'proud' selects a prepositional complement introduced by the preposition *de*. This kind of syntactic properties is individually associated with every predicate, both within a single language and cross-linguistically. For example, the English verb *miss* has two NP arguments but the second argument of its French equivalent *manquer* is a PP (and semantic roles of the two arguments are reversed). This implies that subcategorization lexicons which store such syntactic information have to be developed for each language individually.[3] In addition to their importance in language learning, they play a crucial role in many NLP

[3] Work on mapping theory has revealed partial correlations between lexical semantics and subcategorization frames, see for example [10] for linking relations of verbs' arguments. We are not aware of any similar work done for other types of predicates, e.g., adjectives or adverbs.

Layer Structures and Conceptual Hierarchies in Semantic Representations for NLP

Hermann Helbig, Ingo Glöckner, and Rainer Osswald

Intelligent Information and Communication Systems
FernUniversität in Hagen, Germany

Abstract. Knowledge representation systems aiming at full natural language understanding need to cover a wide range of semantic phenomena including lexical ambiguities, coreference, modalities, counterfactuals, and generic sentences. In order to achieve this goal, we argue for a multidimensional view on the representation of natural language semantics. The proposed approach, which has been successfully applied to various NLP tasks including text retrieval and question answering, tries to keep the balance between expressiveness and manageability by introducing separate semantic layers for capturing dimensions such as facticity, degree of generalization, and determination of reference. Layer specifications are also used to express the distinction between categorical and situational knowledge and the encapsulation of knowledge needed e.g. for a proper modeling of propositional attitudes. The paper describes the role of these classificational means for natural language understanding, knowledge representation, and reasoning, and exemplifies their use in NLP applications.

1 Introduction

Although envisaged since the early days of automated natural language processing (NLP), there are currently only a few implemented systems that aim at a full semantic analysis of unrestricted natural language. One of the reasons for this situation may be the diversification in formal semantics with its highly elaborate but specialized theories focusing on specific semantic phenomena such as presuppositions, generics, pluralities or modalities (see e.g. [1]), whereas building language understanding systems calls for a uniform and concise formalism covering all aspects of natural language semantics up to a certain degree of granularity. A second reason may be the current dominance of shallow NLP even in areas where traditionally deep semantic analysis has been taken as a *sine qua non*. For instance, many approaches to open-domain textual question answering rest mainly on text retrieval techniques with no or little analysis of the underlying documents. However, this approach will fail if the relevant information is mentioned only once in the text collection, has no lexical overlap with the question and uses other syntactic constructions, or is distributed over several sentences linked by anaphora.

In this paper, we argue for a concept-centered representation of natural language semantics that employs a moderate amount of reification and represents semantic aspects like plurality or facticity by ontological features whenever appropriate. The proposed approach is explicated by a slightly simplified version of the *MultiNet* knowledge representation formalism, which is described in detail in [2]. MultiNet has been designed

Improving Question Answering by Combining Multiple Systems via Answer Validation

Alberto Téllez-Valero[1], Manuel Montes-y-Gómez[1],
Luis Villaseñor-Pineda[1], and Anselmo Peñas[2]

[1] Instituto Nacional de Astrofísica, Óptica y Electrónica
Grupo de Tecnologías del Lenguaje
Luis Enrrique Erro no. 1, Sta. María Tonantzintla, Pue.; 72840; Mexico
{albertotellezv,mmontesg,villasen}@inaoep.mx
[2] Universidad Nacional de Educación a Distancia
Depto. Lenguajes y Sistemas Informáticos
Juan del Rosal, 16; 28040 Madrid; Spain
anselmo@lsi.uned.es

Abstract. Nowadays there exist several kinds of question answering systems. According to recent evaluation results, most of these systems are complementary (i.e., each one is better than the others in answering some specific type of questions). This fact indicates that a pertinent combination of various systems may allow improving the best individual result. This paper focuses on this problem. It proposes using an answer validation method to handle this combination. The main advantage of this approach is that it does not rely on internal system's features nor depend on external answer's redundancies. Experimental results confirm the appropriateness of our proposal. They mainly show that it outperforms individual system's results as well as the precision obtained by a redundancy-based combination strategy.

1 Introduction

Question Answering (QA) systems are a kind of search engines that allow responding to questions written in unrestricted natural language. Different to traditional IR systems that focus on finding relevant documents for general user queries, this kind of systems are especially suited to resolve very specific information needs.

Currently, given the great number of its potential applications, QA has become a promising research field. As a result, several QA methods have been developed and different evaluation forums have emerged (such as those at TREC[3] and CLEF[4]). Latest results from these forums evidenced two important facts about the state of the art in QA. On the one hand, they indicated that it already does not exist any method capable of answering all types of questions with

[3] Text REtrieval Conference. http://trec.nist.gov/

[4] Cross Language Evaluation Forum. http://www.clef-campaign.org/

On Ontology Based Abduction For Text Interpretation

Irma Sofia Espinosa Peraldi, Atila Kaya, Sylvia Melzer, Ralf Möller

Hamburg University of Technology,
Institute for Software Systems,
Hamburg, Germany
sofia.espinosa@tuhh.de, at.kaya@tuhh.de,
sylvia.melzer@tuhh.de, r.f.moeller@tuhh.de

Abstract. Text interpretation can be considered as the process of extracting deep-level semantics from unstructured text documents. Deep-level semantics represent abstract index structures that enhance the precision and recall of information retrieval tasks. In this work we discuss the use of ontologies as valuable assets to support the extraction of deep-level semantics in the context of a generic architecture for text interpretation.

1 Introduction

The growing amount of unstructured electronic documents is a problem found in proprietary as well as in public repositories. In this context, the web is a representative example where the need of logic-based information retrieval (IR) to enhance precision and recall is evident. Logic-based IR means the retrieval of unstructured documents with the use of abstract terms that are not directly readable from the surface of the text, but only between its lines. For example, *Chocolate Cake Recipe* is an abstract term for the following text:

> *Yield: 10 Servings, 5 oz. semisweet chocolate (chopped), 3 oz. unsweetened chocolate (chopped), 1/4 lb. (8 Tbs.) unsalted butter, 1/4 cup all-purpose flour, 4 eggs at room temperature, ...*

Relational index structures are crucial for IR. Therefore, the task of defining the necessary index structures for abstract terms to allow logic-based IR is unavoidable. In our work, the necessary structures for logic-based IR are called *deep-level semantics* and the process of extracting deep-level semantics from unstructured text documents is understood as *text interpretation*. In the course of the work presented here, we will highlight that a feasible architecture (see Figure 1) to enable the automatic extraction of deep-level semantics from large-scale corpora can be achieved through:

- A two phase process of information extraction (IE), where the first phase exploits state-of-the-art shallow text processing mechanisms to extract surface-level structures as input for the second phase. The second phase called deep-level interpretation, exploits reasoning techniques over ontologies to extract deep-level semantics.

Word Distribution Analysis for Relevance Ranking and Query Expansion

Patricio Galeas and Bernd Freisleben*

Dept. of Mathematics and Computer Science, University of Marburg,
Hans-Meerwein-Str. 3, D-35032 Marburg, Germany
{galeas,freisleb}@informatik.uni-marburg.de

Abstract. Apart from the frequency of terms in a document collection, the distribution of words plays an important role in determining the relevance of documents for a given search query. In this paper, *word distribution analysis* as a novel approach for using descriptive statistics to calculate a compressed representation of word positions in a document corpus is introduced. Based on this statistical approximation, two methods for improving the evaluation of document relevance are proposed: (a) a relevance ranking procedure based on how query terms are distributed over initially retrieved documents, and (b) a query expansion technique based on overlapping the distributions of terms in the top-ranked documents. Experimental results obtained for the TREC-8 document collection demonstrate that the proposed approach leads to an improvement of about 6.6% over the term frequency/inverse document frequency weighting scheme without applying query reformulation or relevance feedback techniques.

1 Introduction

In a typical information search process, results are obtained by literally matching terms in documents with those of a query. However, due to *synonymy* and *polysemy,* lexical matching methods are likely to be inaccurate when they are used to meet a user's information need [1].

One way to address this problem is to consider contextual information [2]. In fact, several search engines make use of contextual information to disambiguate query terms [3]. Contextual information is either derived from the user, the document structure or from the text itself by performing some form of statistical analysis, such as counting the frequency and/of distance of words.

In this paper, we present an information retrieval approach that incorporates novel contextual analysis and document ranking methods. The proposed approach, called *word distribution analysis*, is based on a compressed statistical description of the word positions in a document collection, represented through their measures of *center* and *spread.* As a complement to the term frequency/inverse document frequency (*tfidf*) metric, we propose the *term density*

* This work is partially supported by the Deutsche Forschungsgemeinschaft (DFG, SFB/FK 615)

Context-Based Sentence Alignment in Parallel Corpora

Ergun Biçici

Koç University
Rumelifeneri Yolu 34450
Sariyer, Istanbul, Turkey
ebicici@ku.edu.tr

Abstract. This paper presents a language-independent context-based sentence alignment technique given parallel corpora. We can view the problem of aligning sentences as finding translations of sentences chosen from different sources. Unlike current approaches which rely on pre-defined features and models, our algorithm employs features derived from the distributional properties of words and does not use any language dependent knowledge. We make use of the context of sentences and the notion of Zipfian word vectors which effectively models the distributional properties of words in a given sentence. We accept the context to be the frame in which the reasoning about sentence alignment is done. We evaluate the performance of our system based on two different measures: sentence alignment accuracy and sentence alignment coverage. We compare the performance of our system with commonly used sentence alignment systems and show that our system performs 1.2149 to 1.6022 times better in reducing the error rate in alignment accuracy and coverage for moderately sized corpora.

Keywords
sentence alignment, context, Zipfian word vectors, multilingual

1 Introduction

Sentence alignment is the task of mapping the sentences of two given parallel corpora which are known to be translations of each other to find the translations of corresponding sentences. Sentence alignment has two main burdens: solving the problems incurred by a previous erroneous sentence splitting step and aligning parallel sentences which can later be used for machine translation tasks. The mappings need not necessarily be 1-to-1, monotonic, or continuous. Sentence alignment is an important preprocessing step that affects the quality of parallel text.

A simple approach to the problem of sentence alignment would look at the lengths of each sentence taken from parallel corpora and see if they are likely to be translations of each other. In fact, it was shown that paragraph lengths for the English-German parallel corpus from the economic reports of Union Bank of Switzerland (UBS) are highly correlated with a correlation value of 0.991 [1]. A more complex approach would look at the neighboring sentence lengths as well. Our approach is based on this knowledge of context for given sentences from each corpus and the knowledge of distributional features of words, which we name Zipfian word vectors, for alignment purposes. A Zipfian

Language Independent First and Last Name Identification in Person Names

Octavian Popescu[1] and Bernardo Magnini[1]

[1] FBK-Trento, Italy
{popescu, magnini}@fbk.eu

Abstract. In this paper we address the problem of first name and last name identification in a news collection. The approach presented is based on corpus investigation and is language independent. At the core of the system there is a name classifier based on the values of different parameters. In its most general form, the name category identification is not an easy task. The hardest problems are raised by ambiguous tokens – those that can be either a first or a last name and/or by tokens with just one occurrence. However, the system is able to predict the name category with high accuracy. The experiments have been run on an Italian newspaper and the evaluation has been carried on I-CAB.

1 Introduction

Knowing whether a token composing the name of a person refers either to her/his first or last name is an important task in several respects. It is probably one of the first things a reader would like to know about a person, especially if she/he has no native intuitions. The name category (first vs. last) is also important for further processing of textual information. It plays an important role in enhancing the overall accuracy of a cross document coreference system (Popescu&Magnini, 2007). Many name mentions consist of only one token, but they definitely stand for two-token names; knowing the category of the token that is missing may give important clues about the person that carries that name.

The task consists in determining for each name occurrence in a large corpus the name category of each individual token composing it. For example, "*George W. Bush*" should be analyzed like "*George*<first name> *W.*<first name> *Bush*<last name>". In its most general form the name category identification is not an easy task. The hardest problems are raised by ambiguous tokens – those that can be either first or last names and/or by tokens with just one occurrence.

In this paper we address the problem of first, last name identification in a news collection. While relying on gazetteers or name dictionaries seems to be an easy way out, we show that this is not enough. The approach we are going to present is based on

Identification of Transliterated Foreign Words in Hebrew Script

Yoav Goldberg and Michael Elhadad

Computer Science Department
Ben Gurion University of the Negev
P.O.B 653 Be'er Sheva 84105, Israel
{yoavg,elhadad}@cs.bgu.ac.il

Abstract. We present a loosely-supervised method for context-free identification of transliterated foreign names and borrowed words in Hebrew text. The method is purely statistical and does not require the use of any lexicons or linguistic analysis tool for the source languages (Hebrew, in our case). It also does not require any manually annotated data for training – we learn from noisy data acquired by over-generation. We report precision/recall results of 80/82 for a corpus of 4044 unique words, containing 368 foreign words.

1 Introduction

Increasingly, native speakers tend to use borrowed foreign terms and foreign names in written texts. In sample data, we found genres with as many as 5% of the word instances borrowed from foreign languages. Such borrowed words appear in a transliterated version. Transliteration is the process of writing the phonetic equivalent of a word of language A in the alphabet of language B. Borrowed words can be either foreign loan words with no equivalent in language B, or words from language A used as slang in language B. Identifying foreign words is not a problem in languages with very similar alphabets and sound systems, as the words just stay the same. But this is not the case in words borrowed from languages that have different writing and sound systems, such as English words in Japanese, Hebrew and Arabic texts.

Transliterated words require special treatment in NLP and IR systems. For example, in IR, query expansion requires special treatment for foreign words; when tagging text for parts of speech, foreign words appear as unknown words and the capability to identify them is critical for high-precision PoS tagging; in Machine Translation, back transliteration from the borrowed language to the source language requires the capability to perform the inverse operation of transliteration; in Named Entity Recognition and Information Extraction, the fact that a word is transliterated from a foreign language is an important feature to identify proper names.

We focus in this paper on the task of identifying whether a word is a transliteration from a foreign language – and not on the task of mapping back the

Deep Lexical Semantics

Jerry R. Hobbs

Information Sciences Institute
University of Southern California
Marina del Rey, California

Abstract. In the project we describe, we have taken a basic core of about 5000 synsets in WordNet that are the most frequently used, and we have categorized these into sixteen broad categories, including, for example, time, space, scalar notions, composite entities, and event structure. We have sketched out the structure of some of the underlying abstract core theories of commonsense knowledge, including those for the mentioned areas. These theories explicate the basic predicates in terms of which the most common word senses need to be defined or characterized. We are now encoding axioms that link the word senses to the core theories. This may be thought of as a kind of "advanced lexical decomposition", where the "primitives" into which words are "decomposed" are elements in coherently worked-out theories. In this paper we focus on our work on the 450 of these synsets that are concerned with events and their structure.

1 Introduction

Words describe the world, so if we are going to draw the appropriate inferences in understanding a text, we must have underlying theories of aspects of the world and we must have axioms that link these to words. This includes domain-dependent knowledge, of course, but 70-80% of the words in most texts, even technical texts, are words in ordinary English used with their ordinary meanings. For example, so far in this paragraph, only the words "theories" and "axioms" and possibly "domain-dependent" have been domain-dependent.

Domain-independent words have such wide utility because their basic meanings tend to be very abstract, and they acquire more specific meanings in combination with their context. Therefore, the underlying theories required for explicating the meanings of these words are going to be very abstract.

For example, a core theory of scales will provide axioms involving predicates such as *scale*, $<$, *subscale*, *top*, *bottom*, and *at*. These are abstract notions that apply to partial orderings as diverse as heights, money, and degrees of happiness. Then, at the "lexical periphery" we will be able to define the rather complex word "range" by the following axiom:

$$\begin{aligned}
&(\forall\, x,y,z)range(x,y,z) \;\equiv\\
&\quad(\exists\, s,s_1,u_1,u_2)scale(s) \,\wedge\, subscale(s_1,s) \,\wedge\, bottom(y,s_1)\\
&\qquad\wedge\, top(z,s_1) \,\wedge\, u_1 \in x \,\wedge\, at(u_1,y) \,\wedge\, u_2 \in x \,\wedge\, at(u_2,z)\\
&\qquad\wedge\, (\forall\, u \in x)(\exists\, v \in s_1)at(u,v)
\end{aligned}$$

Translation Paraphrases in Phrase-Based Machine Translation

Francisco Guzmán and Leonardo Garrido

Center for Intelligent Systems
ITESM Campus Monterrey, Mexico
guzmanhe@gmail.com leonardo.garrido@itesm.mx

Abstract. In this paper we present an analysis of a phrase-based machine translation methodology that integrates paraphrases obtained from an intermediary language (French) for translations between Spanish and English. The purpose of the research presented in this document is to find out how much extra information (i.e. improvements in translation quality) can be found when using Translation Paraphrases (TPs). In this document we present an extensive statistical analysis to support conclusions.

1 Introduction

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation [1]. There have been several attempts to improve the performance of such systems. Non-syntactic phrase-based translation systems[2] certainly outperform word-based systems[3].

Nevertheless, Statistical Machine Translation (STMT) effectiveness is limited to situations where large amounts of data are available. Such a condition, limits the performance of SMT systems over"low density" language pairs [4]. Scarce training data, often leads to a low coverage problem, that is, a low amount of learned translations for a language pair.

There are several efforts trying to improve translation quality of STMT systems. Many state-of-the-art systems involve the introduction of syntactic information to phrase-based machine translations [5,6,7,8,9].

On the other hand, we find several efforts which do not use syntactic information. One main topic of discussion is the usage of paraphrases. For example Callison [4] improves translation quality by giving alternatives to broaden coverage of a phrase-based machine translation system through the use of paraphrases. They use paraphrases in cases when a phrase is not found in their phrase-tables. Other effort is conducted by Guzman and Garrido [10] who obtain what they call "translation paraphrases" from pivoting through an intermediary language.

In this paper we analyze their methodology to assess whether the inclusion of Translation Paraphrases (TP) in a STMT system are useful to improve translation quality, in comparison to systems that do not include such features.

Trusting Politicians' Words (for Persuasive NLP)

Marco Guerini, Carlo Strapparava, and Oliviero Stock
FBK-irst, I-38050, Povo, Trento, ITALY
{guerini, strappa, stock}@itc.it

Abstract. This paper presents resources and lexical strategies for persuasive natural language processing. After the introduction of a specifically tagged corpus of political speeches, some forms of affective language processing in persuasive communication and prospects for application scenarios are provided. In particular *Valentino*, a prototype for valence shifting of existing texts, is described.

1 Introduction

In order to automatically produce and analyze persuasive communication, specific resources and methodologies are needed. For persuasive NLP we built a resource called CORPS that contains political speeches tagged with audience reactions. A key role in persuasive communication is played by affects: we have focused on lexical choice and we present here a tool for modifying existing textual expressions towards more positively or negatively valenced versions, as an element of a persuasive system.

The paper is structured as follows: Section 2 gives an overview of key concepts connected to persuasion and briefly describes the state of the art in related areas. Section 3 describes the resources we built for statistical acquisition of persuasive expressions. Finally, Section 4 describes how this approach can be used for various persuasive NLP tasks, while Section 5 presents the *Valentino* prototype, built upon the resources we presented.

2 Persuasion, affect and NLP

According to Perelman and Olbrechts-Tyteca [1], persuasion is a skill that human beings use - in communication - in order to make their partners perform certain actions or collaborate in various activities. Here below we introduce some related key concepts.

Argumentation and Persuasion. In AI the main approaches focus on the argumentative aspects of persuasion. Still, argumentation is considered as a process that involves "rational elements", while persuasion includes also elements like emotions. In our view, a better distinction can be drawn considering their different foci of attention: while the former focuses on message correctness (its being a valid argument) the latter is concerned with its effectiveness.

Dynamic Translation Memory: Using Statistical Machine Translation to improve Translation Memory Fuzzy Matches

Ergun Biçici[1] and Marc Dymetman[2]

[1] Koç University (Istanbul, Turkey)
ebicici@ku.edu.tr
[2] Xerox Research Centre Europe (Grenoble, France)
marc.dymetman@xrce.xerox.com

Abstract. Professional translators of technical documents often use Translation Memory (TM) systems in order to capitalize on the repetitions frequently observed in these documents. TM systems typically exploit not only complete matches between the source sentence to be translated and some previously translated sentence, but also so-called *fuzzy matches*, where the source sentence has some substantial commonality with a previously translated sentence. These fuzzy matches can be very worthwhile as a starting point for the human translator, but the translator then needs to manually edit the associated TM-based translation to accommodate the differences with the source sentence to be translated. If part of this process could be automated, the cost of human translation could be significantly reduced. The paper proposes to perform this automation in the following way: a phrase-based Statistical Machine Translation (SMT) system (trained on a bilingual corpus in the same domain as the TM) is combined with the TM fuzzy match, by extracting from the fuzzy-match a large (possibly gapped) bi-phrase that is dynamically added to the usual set of "static" bi-phrases used for decoding the source. We report experiments that show significant improvements in terms of BLEU and NIST scores over both the translations produced by the stand-alone SMT system and the fuzzy-match translations proposed by the stand-alone TM system.

1 Introduction

Translation Memory (TM) systems [1, 2] have become indispensable tools for professional translators working with technical documentation. Such documentation tends to be highly repetitive, due to several factors, such as multiple versioning of similar products, importance of maintaining consistent terminology and phraseology, and last but not least, simplification of the translation process itself. TM systems typically exploit not only *complete matches* between the source sentence to be translated and some previously translated sentence, but also so-called *fuzzy matches* [3], where the source sentence has some substantial commonality with a previously translated sentence. These fuzzy matches can be

A Semantics-Enhanced Language Model for Unsupervised Word Sense Disambiguation

Shou-de Lin[1] and Karin Verspoor[2]

[1] National Taiwan University, sdlin@csie.ntu.edu.tw
[2] Los Alamos National Laboratory, verspoor@lanl.gov

Abstract. An N-gram language model aims at capturing statistical word order dependency information from corpora. Although the concept of language models has been applied extensively to handle a variety of NLP problems with reasonable success, the standard model does not incorporate semantic information, and consequently limits its applicability to semantic problems such as word sense disambiguation. We propose a framework that integrates semantic information into the language model schema, allowing a system to exploit both syntactic and semantic information to address NLP problems. Furthermore, acknowledging the limited availability of semantically annotated data, we discuss how the proposed model can be learned without annotated training examples. Finally, we report on a case study showing how the semantics-enhanced language model can be applied to unsupervised word sense disambiguation with promising results.

1 Introduction

Syntax and semantics both play an important role in language use. Syntax refers to the grammatical structure of a language whereas semantics refers to the meaning of the symbols arranged with that structure. To fully comprehend a language, a human must understand its syntactic structure, the meaning each symbol represents, and the interaction between the two. In most languages, syntactic structure conveys something about the semantics of the symbols, and the semantics of symbols may constrain valid syntactic realizations. As a simple example: when we see a noun following a number in English (e.g. "one book"), we can infer that the noun is countable. Conversely, if it is known that a noun is countable, a speaker of English knows that it can plausibly be preceded by a numeral. It is therefore reasonable to assume that for a computer system to successfully process natural language, it has to be equipped with capabilities to represent and utilize both the syntactic and semantic information of the language simultaneously.

The n-gram language model (LM) is a powerful and popular framework for capturing the word order information of language, or fundamentally syntactic information. It has been applied successfully to a variety of NLP problems such as machine translation, speech recognition, and optical character recognition. As described in equation (1), an n-gram language model utilizes conditional probabilities to capture word order information, and the validity of a sentence

Mixing Statistical and Symbolic Approaches for Chemical Names Recognition

Florian Boudin[♮], Juan Manuel Torres-Moreno[♮,♭] and Marc El-Bèze[♮]

[♮]Laboratoire Informatique d'Avignon
339 chemin des Meinajaries, BP1228
84911 Avignon Cedex 9, France
[♭] École Polytechnique de Montréal - Département de génie informatique
CP 6079 Succ. Centre Ville H3C 3A7
Montréal (Québec), Canada.
{florian.boudin,juan-manuel.torres,marc.elbeze}@univ-avignon.fr
http://www.lia.univ-avignon.fr

Abstract. This paper investigates the problem of automatic chemical Term Recognition (TR) and proposes to tackle the problem by fusing Symbolic and statistical techniques. Unlike other solutions described in the literature, which only use complex and costly human made ruled-based matching algorithms, we show that the combination of a seven rules matching algorithm and a naïve Bayes classifier achieves high performances. Through experiments performed on different kind of available Organic Chemistry texts, we show that our hybrid approach is also consistent across different data sets.

Key words: Term Recognition, Text Mining, Chemical Informatics.

1 Introduction

Over one million new chemical compounds are discovered and published annually. As in many scientific domains, the Organic Chemistry (OC) data are not published coherently but scattered through thousands of different journal articles. Identifiying and extracting chemical compounds is a critical task for chemical information retrieval. Information extraction technology arose in response to the need for efficient processing of documents in specialized domains. Classical Natural Language Processing (NLP) tools such as parsers, taggers or chunkers achieve very poor on OC documents. This is due to the specificity of the domain, a very wide vocabulary, long sentences containing a high quantity of "hapax legomen"[1]. Scientists, especially chemists, want to be able to search for articles related to particular chemical compounds. Nowadays, search engines mainly depend on the "classical" title, author(s) and keywords scheme searching. Extracting chemicals from texts and using them to classify, organize and accelerate the information access fit to a wide range of possible applications.

[1] Terms which only appears once in a text.

Acquisition of elementary synonym relations from biological structured terminology

Thierry Hamon[1] and Natalia Grabar[2]

[1] LIPN – UMR 7030, Université Paris 13 – CNRS, 99 av. J-B Clément,
F-93430 Villetaneuse, France
thierry.hamon@lipn.univ-paris13.fr
[2] Université Paris Descartes, UMR_S 872, Paris, F-75006 France;
INSERM, U872, Paris, F-75006, France
natalia.grabar@spim.jussieu.fr

Abstract. Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Nevertheless, we notice that such resources, particularly in specialised domains, are missing. However, specialised domains, *i.e.* biomedicine, propose several structured terminologies. In this paper, we propose a high-quality method for exploiting a structured terminology and inferring a specialised elementary synonym lexicon. The method is based on the analysis of syntactic structure of complex terms. We evaluate the approach on the biomedical domain by using the terminological resource `Gene Ontology`. It provides results with over 93% precision. Comparison with an existing synonym resource (the general-language resource `WordNet`) shows that there is a very small overlap between the induced lexicon of synonyms and the `WordNet` synsets.

1 Background

Acquisition and enrichment of lexical resources have long been acknowledged as an important research in the area of computational linguistics. Indeed, such resources are often helpful for the deciphering and computing semantic similarity between words and terms within tasks like information retrieval (especially query expansions), knowledge extraction or terminology matching.

We make the distinction between terminological and lexical resources. The aim of terminological resources is collecting terms used in a specialised area, describing and organizing them. Within terminologies, terms can be simple (*reproduction*) but mostly complex (*formation of catalytic spliceosome for first transesterification step*; *cell wall mannoprotein synthesis*). They can be linked between them with semantic relations (hierarchical, synonymous, ...). Other features of terms (*i.e.*, definitions, areas of usage) can be precised. As for lexical resources, they gather mostly simple lexical units (*i.e.*, synonyms like *formation*, *synthesis* and *biosynthesis*). These units can belong to common language or be specific to some specialised languages. They can receive descriptions (syntactic, phonetic, morphological, ...) or propose relations between them. Our observation is that

Arabic Morphology Parsing Revisited

Suhel Jaber[1] and Rodolfo Delmonte[1]

[1] University Ca' Foscari, Dept. Language Sciences, Laboratory Computational Linguistics,
Ca' Bembo, Dorsoduro 1705, 30123 Venezia, Italy
{jaber, delmont}@unive.it

Abstract. In this paper we propose a new approach to the description of Arabic morphology using 2-tape finite state transducers, based on a particular and systematic use of the operation of composition in a way that allows for incremental substitutions of concatenated lexical morpheme specifications with their surface realization for non-concatenative processes (the case of Arabic templatic interdigitation and non-templatic circumfixation).

Keywords: Arabic, morphology, non-concatenative, finite state, composition.

1 Introduction

In this paper we propose a new approach to the description of Arabic morphology using 2-tape finite state transducers, based on a particular and systematic use of the operation of composition in a way that allows for incremental substitutions of concatenated lexical morpheme specifications with their surface realization for non-concatenative processes (the case of Arabic templatic interdigitation and non-templatic circumfixation). Then we compare it with what in our opinion represents the state-of-the-art among the 2-tape finite-state implementations, that of Xerox [1], which is mainly based on the operation of intersection. We intentionally limit ourselves to the evaluation of 2-tape strictly finite-state implementations for this paper, leaving out n-tape implementations such as [2] and [3], and those based on extended finite-state automata, such as [4]. In any case we believe that our approach could be trivially adapted to n-tape implementations as well.

In this paper we argue that:

1. the use of composition allows to overcome certain technical problems inherent to the use of intersection;
2. the method of incremental substitutions through compositions allows for an elegant description of all main morphological processes present in natural languages including non-concatenative ones in strict finite-state terms, without the need to resort to extensions of any sort;
3. our approach allows for the most logical encoding of every kind of dependency, including traditional long-distance ones (mutual exclusiveness), circumfixations and idiosyncratic root and pattern combinations;

Various Criteria of Collocation Cohesion in Internet: Comparison of Resolving Power[⋆]

Igor A. Bolshakov[1], Elena I. Bolshakova[2], Alexey P. Kotlyarov[1], and Alexander Gelbukh[2]

[1]Center for Computing Research (CIC)
National Polytechnic Institute (IPN), Mexico City, Mexico
{igor,gelbukh}@cic.ipn.mx
[2]Moscow State Lomonosov University
Faculty of Computational Mathematics and Cybernetics, Moscow, Russia
bolsh@cs.msu.su, koterpillar@gmail.com

Abstract. For extracting collocations from the Internet, it is necessary to numerically estimate the cohesion between potential collocates. Mutual Information cohesion measure (MI) based on numbers of collocate occurring closely together (N_{12}) and apart (N_1, N_2) is well known, but the Web page statistics deprives MI of its statistical validity. We propose a family of different measures that depend on N_1, N_2 and N_{12} in a similar monotonic way and possess the scalability feature of MI. We apply the new criteria for a collection of N_1, N_2, and N_{12} obtained from AltaVista for links between a few tens of English nouns and several hundreds of their modifiers taken from Oxford Collocations Dictionary. The nounits own adjective pairs are true collocations and their measure values form one distribution. The nounalien adjective pairs are false collocations and their measure values form another distribution. The discriminating threshold is searched for to minimize the sum of probabilities for errors of two possible types. The resolving power of a criterion is equal to the minimum of the sum. The best criterion delivering minimum minimorum is found.

1 Introduction

During the two recent decades, the vital role of collocationsin any their definitionwas fully acknowledged in NLP. Thus great effort was made to develop methods of collocation extraction from texts and text corpora. As pilot works we can mention [3, 6, 17, 18]. However, up to date we have no large and humanly verified collocation databases for any language, including English. The only good exception is Oxford Collocations Dictionary for Students of English (OCDSE) [11], but even in its electronic version it is oriented to human use rather than to NLP. So the development of the methods of collocation extraction continues [4, 5, 9, 12–16, 19].

[⋆] Work done under partial support of Mexican Government (CONACyT, SNI, CGEPI-IPN) and Russian Foundation of Fundamental Research (grant 06-01-00571).

Evaluation of Internal Validity Measures in Short-Text Corpora*

Diego Ingaramo[1], David Pinto[2,3], Paolo Rosso[2], Marcelo Errecalde[1]

[1]Development and Research Laboratory in Computacional Intelligence (LIDIC),
UNSL, Argentina
[2]Natural Language Engineering Lab.,
Department of Information Systems and Computation,
Polytechnic University of Valencia, Spain
[3]Faculty of Computer Science (FCC),
BUAP, Mexico
{daingara,merreca}@unsl.edu.ar, {prosso,dpinto}@dsic.upv.es

Abstract. Short texts clustering is one of the most difficult tasks in natural language processing due to the low frequencies of the document terms. We are interested in analysing these kind of corpora in order to develop novel techniques that may be used to improve results obtained by classical clustering algorithms. In this paper we are presenting an evaluation of different internal clustering validity measures in order to determine the possible correlation between these measures and that of the F-Measure, a well-known external clustering measure used to calculate the performance of clustering algorithms. We have used several short-text corpora in the experiments carried out. The obtained correlation with a particular set of internal validity measures let us to conclude that some of them may be used to improve the performance of text clustering algorithms.

1 Introduction

Document clustering consists in the assignment of documents to unknown categories. This task is more difficult than supervised text categorization [13,8] because the information about categories and correctly categorized documents is not provided in advance. An important consequence of this lack of information is that clustering results cannot be evaluated with typical external measures like F-Measure and, therefore, the quality of the resulting groups is evaluated with respect to *structural properties* or *internal measures*. Classical internal measures used as cluster validity measures include the *Dunn* and *Davies-Bouldin* indexes, new graph-based measures like *Density Expected Measure* and Λ-Measure as well as some measures based on the corpus vocabulary overlapping.

When clustering techniques are applied to collections containing *very short* documents, additional difficulties are introduced due to the low frequencies of

* This work has been partially supported by the MCyT TIN2006-15265-C06-04 project, as well as by the BUAP-701 PROMEP/103.5/05/1536 grant

Analysis of Joint Inference Strategies for the Semantic Role Labeling of Spanish and Catalan

Mihai Surdeanu[1], Roser Morante[2], Lluís Màrquez[3]

[1]Barcelona Media Innovation Center
[2]Tilburg University
[3]Technical University of Catalonia

Abstract. This paper analyzes two joint inference approaches for semantic role labeling: re-ranking of candidate semantic frames generated by one local model and combination of two distinct models at argument-level using meta learning. We perform an empirical analysis on two recently released corpora of annotated semantic roles in Spanish and Catalan. This work yields several novel conclusions: (a) the proposed joint inference strategies yield good results even under adverse conditions: small training corpora, only two individual models available for combination, minimal output available from the individual models; (b) stacking of the two joint inference approaches is successful, which indicates that the two inference models provide complementary benefits. Our results are currently the best for the identification of semantic role for Spanish and Catalan.

1 Introduction

Semantic Role Labeling (SRL) is the task of analyzing clause predicates in open text by identifying arguments and tagging them with semantic labels indicating the role they play with respect to the verb, as in:

[Mr. Smith]$_{Agent}$ *sent* [the report]$_{Object}$ to [me]$_{Recipient}$ [this morning]$_{Temporal}$
Such sentence–level semantic analysis allows to determine "who" did "what" to "whom", "when" and "where", and, thus, characterize the participants and properties of the *events* established by the predicates. This semantic analysis in the form of event structures is very interesting for a broad spectrum of NLP applications.

The work proposed in this paper fits in the framework of supervised learning with joint inference for SRL. We introduce a stacking architecture that exploits several levels of global learning: in the first level we deploy two base SRL models that exploit only information local to each individual candidate argument; in the second level we perform re-ranking of the candidate frames generated by the base models; and lastly, we combine the outputs of the two individual models (after re-ranking) using meta-learning and sentence-level information.

The combination/joint inference models we introduce are not novel in themselves: all state-of-the-art SRL systems (see, e.g., [1–4]) include some kind of combination to increase robustness and to gain coverage and independence from parse errors. One may combine: 1) the output of several independent SRL basic systems [2, 5], or 2) several outputs from the same SRL system obtained by changing input annotations or other internal parameters [4, 3]. The combination can be as simple as selecting the best among

Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model

Amber Wilcox-O'Hearn, Graeme Hirst, and Alexander Budanitsky*

Department of Computer Science, University of Toronto
Toronto, Ontario, Canada M5S 3G4
amber, gh, abm@cs.toronto.edu

Abstract. The trigram-based noisy-channel model of real-word spelling-error correction that was presented by Mays, Damerau, and Mercer in 1991 has never been adequately evaluated or compared with other methods. We analyze the advantages and limitations of the method, and present a new evaluation that enables a meaningful comparison with the WordNet-based method of Hirst and Budanitsky. The trigram method is found to be superior, even on content words. We then show that optimizing over sentences gives better results than variants of the algorithm that optimize over fixed-length windows.

1 Introduction

Real-word spelling errors are words in a text that, although correctly spelled words in the dictionary, are not the words that the writer intended. Such errors may be caused by typing mistakes or by the writer's ignorance of the correct spelling of the intended word. Ironically, such errors are also caused by spelling checkers in the correction of non-word spelling errors: the "auto-correct" feature in popular word-processing software will sometimes silently change a non-word to the wrong real word (Hirst and Budanitsky 2005), and sometimes when correcting a flagged error, the user will inadvertently make the wrong selection from the alternatives offered. The problem that we address in this paper is the automatic detection and correction of real-word errors.

Methods developed in previous research on this topic fall into two basic categories: those based on human-made lexical or other resources and those based on machine-learning or statistical methods. An example of a resource-based method is that of Hirst and Budanitsky (2005), who use semantic distance measures in WordNet to detect words that are potentially anomalous in context — that is, semantically distant from nearby words; if a variation in spelling[1] results in a word that was semantically closer to the context, it is hypothesized that the original word is an error (a "*malapropism*")

* This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada. We are grateful to Bryce Wilcox-O'Hearn for his assistance.

[1] In this method, as in the trigram method that we discuss later, any consistent definition, narrow or broad, of what counts as the spelling variations of a word may be used. Typically it would be based on edit distance, and might also take phonetic similarity into account; see our remarks on Brill and Moore (2000) and Toutanova and Moore (2002) in section 5 below.

A Comparison of Co-occurrence and Similarity Measures as Simulations of Context

Stefan Bordag

Natural Language Processing Department, University of Leipzig
sbordag@informatik.uni-leipzig.de

Abstract. Observations of word co-occurrences and similarity computations are often used as a straightforward way to represent the global contexts of words and achieve a simulation of semantic word similarity for applications such as word or document clustering and collocation extraction. Despite the simplicity of the underlying model, it is necessary to select a proper significance, a similarity measure and a similarity computation algorithm. However, it is often unclear how the measures are related and additionally often dimensionality reduction is applied to enable the efficient computation of the word similarity. This work presents a linear time complexity approximative algorithm for computing word similarity without any dimensionality reduction. It then introduces a large-scale evaluation based on two languages and two knowledge sources and discusses the underlying reasons for the relative performance of each measure.

1 Introduction

One way to simulate associative and semantic relations between words is to view each word as a distinct entity. That entity may occur in a linear stream of sentences or other easily observable linguistic units. It is then possible to measure the statistical correlation between the common co-occurrence of such entities (i.e. words) within these units [1, 2]. If additional knowledge such as word classes or morphological relatedness is available, this model allows to construct a variety of applications that depend on knowledge about word relatedness, but do not necessarily need this knowledge to be precise. For example, it is sufficient to know the most significant co-occurring word pairs in a corpus to enable the creation of a helpful tool for extraction of collocations, idioms or multi-word-expressions [3–5]. Similarly, knowledge about contextual similarity modeled as co-occurrence vector comparisons helps to build thesaurus construction tools such as the Sketch engine [6] or to design specific semi-automatic algorithms that create approximations of a thesaurus [7–10].

Assuming the simple vector-space model where each word defines a new dimension, the question arises how exactly significant co-occurrence or word similarity is to be modeled. Several variations of the same underlying vector space model were proposed. One is to apply Latent Semantic Indexing (LSI) to the matrix containing the raw co-occurrence counts of words [11]. However, it is

Verb Class Discovery from Rich Syntactic Data

Lin Sun[1], Anna Korhonen[1] and Yuval Krymolowski[2]

[1] Computer Laboratory, University of Cambridge
15 JJ Thomson Avenue, Cambridge CB3 0FD, UK
alk23@cam.ac.uk,ls418@cam.ac.uk
[2] Department of Computer Science, University of Haifa
31905, Haifa, Israel
yuvalkry@gmail.com

Abstract. Previous research has shown that syntactic features are the most informative features in automatic verb classification. We investigate their optimal characteristics by comparing a range of feature sets extracted from data where the proportion of verbal arguments and adjuncts is controlled. The data are obtained from different versions of VALEX [1] – a large SCF lexicon for English which was acquired automatically from several corpora and the Web. We evaluate the feature sets thoroughly using four supervised classifiers and one unsupervised method. The best performing feature set includes rich syntactic information about both arguments and adjuncts of verbs. When combined with our best performing classifier (a novel Gaussian classifier), it yields the promising accuracy of 64.2% in classifying 204 verbs to 17 Levin (1993) classes. We discuss the impact of our results on the state-or-art and propose avenues for future work.

1 Introduction

Recent research shows that it is possible, using current natural language processing (NLP) and machine learning technology, to automatically induce lexical classes from corpus data with promising accuracy [2–5]. This research is interesting, since lexical classifications, when tailored to the application and domain in question, can provide an effective means to deal with a number of important NLP tasks (e.g. parsing, word sense disambiguation, semantic role labeling), as well as enhance performance in applications, (e.g. information extraction, question-answering, machine translation) [6–10].

Lexical classes are useful for NLP because they capture generalizations over a range of (cross-)linguistic properties. Being defined in terms of similar meaning components and (morpho-)syntactic behaviour of words [11, 12] they generally incorporate a wider range of properties than e.g. classes defined solely on semantic grounds [13]. For example, verbs which share the meaning component of 'manner of motion' (such as *travel*, *run*, *walk*), behave similarly also in terms of subcategorization (*I traveled/ran/walked to London*) and usually have zero-related nominals (*a run*, *a walk*).

NLP systems can benefit from lexical classes in many ways. For example, such classes can be used i) to define a mapping from surface realization of arguments to predicate-argument structure, ii) as a means to abstract away from individual words when required, or (iii) to build a lexical organization which predicts much of the syntax and semantics of a new word by associating it with an appropriate class.

Clause Boundary Identification Using Conditional Random Fields

Vijay Sundar Ram. R and Sobha Lalitha Devi

AU-KBC Research Centre,
MIT Campus Anna University,
Chromepet, Chennai -44,
India
(Sobha,sundar)@au-kbc.org

Abstract. This paper discusses about the detection of clause boundaries using a hybrid approach. The Conditional Random fields (CRFs), which have linguistic rules as features, identifies the boundaries initially. The boundary marked is checked for false boundary marking using Error Pattern Analyser. The false boundary markings are re-analysed using linguistic rules. The experiments done with our approach shows encouraging results and are comparable with the other approaches

1 Introduction

The clause identification is one of the shallow parsing tasks, which is important in various NLP applications such as translation, parallel corpora alignment, information extraction, machine translation and text-to-speech. The clause identification task aims at identifying the start and end boundaries of the clauses in a sentence, where clauses are word sequences which contain a subject and a predicate. The subject can be explicit or implied. For the clause identification task we have come up with a hybrid approach, where conditional random fields (CRFs), a machine learning technique and rule-based technique are used. The CRFs module with linguistic rules as features identifies the clause boundaries initially. The erroneous clause boundary detections are identified using an error analyzer and those sentences are processed using the rule-based module.

The clause identification was a shared task in CoNLL 2001. The task of identifying the clause boundaries is non-trivial. More research has been done in this task. The initial approaches to this task were using rule-based technique, which was followed by machine learning and hybrid techniques.

Early experiments in the clause boundary detection are Eva Ejeuhed's basic clause identification system for improving AT&T text to speech system [7], Papergeorgiou's rule based clause boundary system as preprocessing tool for bilingual alignment parallel text [15]. Leffa's rule based system reduces clauses to noun, adjective or an adverb, which was used in English/Portuguese machine translation system [10].

Comparison of Different Modeling Units for Language Model Adaptation for Inflected Languages

Tanel Alumäe

Institute of Cybernetics at Tallinn University of Technology,
Akadeemia tee 21, Tallinn, 12618, Estonia
tanel.alumae@phon.ioc.ee

Abstract. This paper presents a language model adaptation framework for highly inflected languages that use sub-word units as basic units in a language model for large vocabulary speech recognition. The proposed adaptation method uses latent semantic analysis based information retrieval to find documents similar to a tiny adaptation corpus. The approach enables to use different language units for modeling document similarity. The method is tested on an Estonian broadcast news transcription task. We compare words, lemmas and morphemes as basic units for similarity modeling. We observe a drop in speech recognition error rate after building adapted language model for each news story. Morpheme-based adaptation is found to give significantly larger improvement than word and lemma-based adaptation.

1 Introduction

Language model adaptation is a task of building a language model (LM) for speech recognition that is better suited for the given domain than a general background model, given a small adaptation corpus. In recent years, *latent semantic analysis* (LSA) has been successfully used for integrating long-term semantic dependencies into statistical language models [1]. The LSA-based approach gradually adapts the background language model based on the recognized words by boosting the unigram probabilities of semantically related words, using co-occurrence analysis of words and documents.

However, this approach cannot be efficiently directly used for highly inflective and/or agglutinative languages, such as Estonian, Finnish, Turkish, Korean and many others. In such languages, each word-phrase can occur in a large number of inflected forms, depending on its syntactic and semantic role in the sentence. In addition, many such languages are also so-called compounding languages, i.e., compound words can be formed from shorter particles to express complex concepts as single words. The compound words can again occur in different inflections. As a result, the lexical variety of such languages is very high and it is not possible to achieve a good vocabulary coverage when using words as basic units for language modeling. In order to increase coverage, subword units,

Unsupervised and Knowledge-free Learning of Compound Splits and Periphrases

Florian Holz, Chris Biemann

NLP Group, Department of Computer Science, University of Leipzig
{holz|biem}@informatik.uni-leipzig.de

Abstract. We present an approach for knowledge-free and unsupervised recognition of compound nouns for languages that use one-word-compounds such as Germanic and Scandinavian languages. Our approach works by creating a candidate list of compound splits based on the word list of a large corpus. Then, we filter this list using the following criteria:
(a) frequencies of compounds and parts,
(b) length of parts.
In a second step, we search the corpus for periphrases, that is a reformulation of the (single-word) compound using the parts and very high frequency words (which are usually prepositions or determiners). This step excludes spurious candidate splits at cost of recall. To increase recall again, we train a trie-based classifier that also allows splitting multi-part-compounds iteratively.
We evaluate our method for both steps and with various parameter settings for German against a manually created gold standard, showing promising results above 80% precision for the splits and about half of the compounds periphrased correctly. Our method is language independent to a large extent, since we use neither knowledge about the language nor other language-dependent preprocessing tools.
For compounding languages, this method can drastically alleviate the lexicon acquisition bottleneck, since even rare or yet unseen compounds can now be periphrased: the analysis then only needs to have the parts described in the lexicon, not the compound itself.

1 Introduction

A number of languages extensively use compounding as an instrument of combining several word stems into one (long) tokens, e.g. Germanic languages, Korean, Greek and Finnish. Compared to languages such as English, where (noun) compounds are expressed using several tokens, this leads to a tremendous increase in vocabulary size. In applications, this results in sparse data, challenging a number of NLP applications. For IR experiments with German, Braschler et al. report that decompounding results in higher text retrieval improvements than stemming [1].

As an example, consider the German compound "Prüfungsvorbereitungsstress" (stress occurring when preparing for an the exam) - without an analysis,

Semantic and Syntactic Features for Dutch Coreference Resolution

Iris Hendrickx[1], Veronique Hoste[2], and Walter Daelemans[1]

[1] CNTS - Language Technology Group,
University of Antwerp, prinsstraat 13, Antwerp
Belgium
iris.hendrickx@ua.ac.be, walter.daelemans@ua.ac.be
[2] LT3 - Language and Translation Technology Team,
University College Ghent, Groot-Brittaniëlaan 45, Ghent,
Belgium
veronique.hoste@hogent.be

Abstract. We investigate the effect of encoding additional semantic and syntactic information sources in a classification-based machine learning approach to the task of coreference resolution for Dutch. We experiment both with a memory-based learning approach and a maximum entropy modeling method.

As an alternative to using external lexical resources, such as the low-coverage Dutch EuroWordNet, we evaluate the effect of automatically generated semantic clusters as information source. We compare these clusters, which group together semantically similar nouns, to two semantic features based on EuroWordNet encoding synonym and hypernym relations between nouns.

The syntactic function of the anaphor and antecedent in the sentence can be an important clue for resolving coreferential relations. As baseline approach, we encode syntactic information as predicted by a memory-based shallow parser in a set of features. We contrast these shallow parse based features with features encoding richer syntactic information from a dependency parser. We show that using both the additional semantic information and syntactic information lead to small but significant performance improvement of our coreference resolution approach.

1 Introduction

Coreference resolution is the task of resolving different descriptions of the same underlying entity in a given text. Written and spoken texts contain a large number of coreferential relations and a good text understanding largely depends on the correct resolution of these relations. Resolving ambiguous referents in a text can be a helpful preprocessing step for many NLP applications such as text summarization or question answering.

As an alternative to the knowledge-based approaches, in which there has been an evolution from the systems which require an extensive amount of linguistic and non-linguistic information (e.g. [1]) toward more knowledge-poor approaches

Why Don't Romanians Have a Five O'clock Tea, Nor Halloween, but Have a Kind of Valentines Day?

Corina Forăscu

University Al.I. Cuza of Iaşi, Faculty of Computer Science
Research Institute for Artificial Intelligence, Romanian Academy
16, Gen. Berthelot, Iaşi – 700483, Romania
corinfor@info.uaic.ro

Abstract. Recently the focus on temporal information in NLP applications has increased. Based on general temporal theories, annotations and standards, the paper presents the steps performed towards obtaining a parallel English-Romanian corpus, with the temporal information marked in both languages. The automatic import from English to Romanian of the TimeML markup has a success rate of 96.53%. The paper analyzes the main situations that appeared during the automatic import: perfect or impossible transfer, transfer with amendments or for the language specific phenomena. This corpus study permits to decide how import techniques can be used on the temporal domain.

1 Introduction

The temporal information is expressed in natural language through:

- Time-denoting temporal expressions – references to a calendar or clock system, expressed by NPs, PPs, or AdvPs, as in *Friday; yesterday; the previous month*.
- Event-denoting temporal expressions – explicit/implicit/vague references to an event; syntactically they are realized through:
  - sentences – more precisely their syntactic head, the main verb, as in *She <u>flew</u> as the first ever co-pilot.*
  - noun phrases, as in *She followed a normal <u>progression</u> within NASA*.
  - adjectives, predicative clauses or prepositional phrases, as in : *Many experts thought was once <u>invincible</u>*.

Recent work in document analysis started focusing on the temporal information in documents, mainly for their use in many practical Natural Language Processing (NLP) applications such as:.

- linguistic investigation, lexicon induction, and translation using very large annotated corpora;
- question answering (questions like "when", "how often" or "how long");
- information extraction or information retrieval;
- machine translation (translated and normalized temporal references; mappings between different behavior of tenses from language to language);

Innovative Approach for Engineering NLG Systems: the Content Determination Case Study

Marco Fonseca, Leonardo Junior, Alexandre Melo, Hendrik Macedo

Departamento de Computação, Universidade Federal de Sergipe, 49100-000, São Cristóvão/SE, Brazil
marcos_ufs@yahoo.com.br, leonardobsjr@yahoo.com.br, asmelo10@hotmail.com, hendrik@ufs.br

Abstract. The purpose of Natural Language Generation (NLG) systems is that of automating the production of linguistically correct texts from a data source. Generators are usually built using ad-hoc software engineering practices, lacking a well-defined development process, standard software architecture, and the use of worldwide programming languages. This paper describes a new development approach that leverages the most recent programming languages and standards of modern software engineering to enhance the practical use of NLG applications. To show the practicability of the proposal, a content determination system, which accepts as input wrapped Web data regarding soccer championship results, was developed.

1 Introduction

Natural Language Generation (NLG) [13] is a conceptually consolidated technology. Past research has clarified many fundamentals issues and conceived solutions that are robust and scalable enough for practical use. Furthermore, opportunities for practical applications have multiplied with the information inundation from relevant Web content sources.

Unfortunately, NLG techniques remain virtually unknown and unused by mainstream and professional computing. This situation is probably due mainly to the fact that until recently, NLG was built using *ad-hoc* software engineering practices with no explicit development process and no standard software architecture. Reliance on special-purpose esoteric modeling and implementation languages and tools is another NLG issue. Every system is designed and implemented following specific domain complexities and needs and little has been done to change the portrayed situation. A good example is surface realization activity. Many realization components have been built based on different grammatical formalisms and theories used to describe NLG [8].

This work proposes an innovative approach to the development of NLG systems, in which the pipeline of text generation tasks work as a set of consecutive rule base for model transformation. Such methodology for building applications by applying transformations on models in different levels of abstraction was recently popularized

Terms Derived from Frequent Sequences for Extractive Text Summarization*

Yulia Ledeneva,[1] Alexander Gelbukh,[1] René Arnulfo García-Hernández [2]

[1] Natural Language and Text Processing Laboratory,
Center for Computing Research, National Polytechnic Institute, DF 07738, Mexico
yledeneva@yahoo.com, www.Gelbukh.com

[2] Instituto Tecnologico de Toluca, Mexico
renearnulfo@hotmail.com

Abstract. Automatic text summarization helps the user to quickly understand large volumes of information. We present a language- and domain-independent statistical-based method for single-document extractive summarization, i.e., to produce a text summary by extracting some sentences from the given text. We show experimentally that words that are parts of bigrams that repeat more than once in the text are good terms to describe the text's contents, and so are also so-called maximal frequent sentences. We also show that the frequency of the term as term weight gives good results (while we only count the occurrences of a term in repeating bigrams).

1 Introduction

A summary of a document is a (much) shorter text that conveys the most important information from the source document. There are a number of scenarios where automatic construction of such summaries is useful. For example, an information retrieval system could present an automatically built summary in its list of retrieval results, for the user to quickly decide which documents are interesting and worth opening for a closer look—this is what Google models to some degree with the snippets shown in its search results. Other examples include automatic construction of summaries of news articles or email messages to be sent to mobile devices as SMS; summarization of information for government officials, businessmen, researches, etc., and summarization of web pages to be shown on the screen of a mobile device, among many others.

The text summarization tasks can be classified into single-document and multi-document summarization. In single-document summarization, the summary of only one document is to be built, while in multi-document summarization the summary of a whole collection of documents (such as all today's news or all search results for

* Work done under partial support of Mexican Government (CONACyT, SNI, SIP-IPN, COTEPABE-IPN, COFAA-IPN). The authors thank Rada Mihalcea for useful discussion.

Stat-XFER: A General Search-based Syntax-driven Framework for Machine Translation

Alon Lavie

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 USA
alavie@cs.cmu.edu.edu

Abstract. The CMU Statistical Transfer Framework (Stat-XFER) is a general framework for developing search-based syntax-driven machine translation (MT) systems. The framework consists of an underlying syntax-based transfer formalism along with a collection of software components designed to facilitate the development of a broad range of MT research systems. The main components are a general language-independent run-time transfer engine and decoder, along with several different tools for creating the various underlying language-pair-specific resources that are required for building a specific MT system for any given language pair. We describe the general framework, its unique properties and features, and its application to the construction of MT research prototype systems for a diverse collection of language pairs.

1 Introduction

The field of Machine Translation (MT) has dramatically shifted in the course of the past decade. Modern state-of-the-art approaches to MT rely on machine learning methods of increasing complexity and sophistication in order to automatically acquire their underlying translation models from available data resources. Phrase-based Statistical MT (PB-SMT) [1–3] has become the predominant approach in recent years. In PB-SMT, simple statistical modeling methods are used to acquire likely phrase-to-phrase translation equivalents from large volumes of sentence-parallel text corpora. In the absence of large sentence-parallel data, the statistical estimation methods break down, and the approach becomes ineffective. Vast sentence-parallel corpora exist only for a limited number of language pairs (primarily pairs of European languages, Chinese, Japanese and Arabic), severely limiting the applicability of this approach. While the amount of online resources for many languages will undoubtedly grow over time, many of the languages spoken by smaller ethnic groups and populations in the world will not have such resources within the foreseeable future. Corpus-based MT approaches will therefore not be effective for such languages for some time to come.

Statistical Machine Translation into a Morphologically Complex Language

Kemal Oflazer

Faculty of Engineering and Natural Sciences
Sabancı University
Istanbul, Tuzla, 34956, Turkey
oflazer@sabanciuniv.edu

Abstract. In this paper, we present the results of our investigation into phrase-based statistical machine translation from English into Turkish – an agglutinative language with very productive inflectional and derivational word-formation processes. We investigate different representational granularities for morphological structure and find that (i) representing both Turkish and English at the morpheme-level but with some selective morpheme-grouping on the Turkish side of the training data, (ii) augmenting the training data with "sentences" comprising only the content words of the original training data to bias root word alignment, and with highly-reliable phrase-pairs from an earlier corpus-alignment (iii) re-ranking the n-best morpheme-sequence outputs of the decoder with a word-based language model, and (iv) "repairing" translated words with incorrect morphological structure and words which are out-of-vocabulary relative to the training and the language model corpus, provide an non-trivial improvement over a word-based baseline despite our very limited training data. We improve from 19.77 BLEU points for our word-based baseline model to 26.87 BLEU points for an improvement of 7.10 points or about 36% relative. We briefly discuss the applicability of BLEU to morphologically complex languages like Turkish and present a simple extension to compare tokens not in a all-or-none fashion but taking lexical-semantic and morpho-semantic similarities into account, implemented in our BLEU+ tool.

1 Introduction

Statistical machine translation from English-to-Turkish poses a number of difficulties. Typologically English and Turkish are rather distant languages: while English has very limited morphology and rather fixed SVO constituent order, Turkish is an agglutinative language with a very rich and productive derivational and inflectional morphology, and a very flexible (but SOV dominant) constituent order. One implication of complex morphology is that, in parallel texts, Turkish words usually align to multiple words on the English side. When done at the word level, this is very noisy and masks the more (statistically) meaningful alignments at the sub-lexical level. Another issue of practical significance is the lack of large scale parallel text resources, with no substantial improvement expected

Natural Language as the Basis for Meaning Representation and Inference

Ido Dagan[1], Roy Bar-Haim[1], Idan Szpektor[1], Iddo Greental[2], and Eyal Shnarch[1]

[1] Bar Ilan University, Ramat Gan 52900, Israel,
dagan@cs.biu.ac.il
[2] Tel Aviv University, Tel Aviv 69978, Israel

Abstract. Semantic inference is an important component in many natural language understanding applications. Classical approaches to semantic inference rely on logical representations for meaning, which may be viewed as being "external" to the natural language itself. However, practical applications usually adopt shallower lexical or lexical-syntactic representations, which correspond closely to language structure. In many cases, such approaches lack a principled meaning representation and inference framework. We describe a generic semantic inference framework that operates directly on language-based structures, particularly syntactic trees. New trees are inferred by applying entailment rules, which provide a unified representation for varying types of inferences. Rules were generated by manual and automatic methods, covering generic linguistic structures as well as specific lexical-based inferences. Initial empirical evaluation in a Relation Extraction setting supports the validity and potential of our approach. Additionally, such inference is shown to improve the critical step of unsupervised learning of entailment rules, which in turn enhances the scope of the inference system.
This paper corresponds to the invited talk of the first author at CICLING 2008.

1 Introduction

It has been a common assumption that the structure of natural language is not suitable to formally represent meanings and to conduct inferences over them. Indeed, according to the traditional formal semantics approach inference is conducted at a logical level. Texts are first translated, or *interpreted*, into some logical form and then new propositions are inferred from interpreted texts by a logical theorem prover. Meaning and inference are thus captured by representations that are "external" to the language itself, and are typically independent of the structure of any particular natural language.

However, practical text understanding systems usually employ shallower lexical and lexical-syntactic representations, which clearly correspond to the structure of the particular natural language being processed. Such representations are sometimes augmented with partial semantic annotations like word senses,

What we are talking about and what we are saying about it

Eva Hajičová

Charles University
Institute of Formal and Applied Linguistics
Malostranské nám. 25, 118 00 Prague, Czech Republic
hajicova@ufal.mff.cuni.cz

Abstract. In view of the relationships between theoretical, computational and corpus linguistics, their mutual contributions are discussed and illustrated on the issue of the aspect of language related to the information structure of the sentence, distinguishing ”what we are talking about” and ”what we are saying about it”.

1 Introduction

The name of the research domain of Computational Linguistics seems to be self-explanatory; however, there has always been a dispute what exactly ‘computational’ means (especially from the point of view of the relation between its theoretical and applied aspects and from the point of view of its supposedly narrowing scope due to the prevalent use of statistical methods). In addition, with the expansion of the use of computers for linguistic studies based on very large empirical language material, and, consequently, with the appearance of an allegedly new domain, corpus linguistics, a question has emerged what is the position of corpus linguistics with regard to computational linguistics.

After a summary of some of the issues related to the problem of ‘how many linguistics there are’ (Sect. 2), we briefly sketch in which respects the different ‘linguistics’ can mutually contribute to each other (Sect. 3). The main objective of our paper is to illustrate on an example of a linguistically based multi-layered annotation scenario (Sect. 4) and of a selected linguistic phenomenon, namely the information structure of the sentence (Sect. 5.1), how linguistic theory can contribute to a build-up of an integrated scenario of corpus annotation (Sect. 5.2) and, in the other direction, how a consistent application of such a scenario on a large corpus of continuous texts can provide a useful feedback for the theory (Sect. 5.3). In Section 6, some conclusions will be drawn from the personal experience with working with the given theory and scenario.

2 How many linguistics?

If the terms *computational linguistics* and *corpus linguistics* are understood rather broadly, as covering those domains of linguistics that are based on the