IJCLA **ISSN 0976-0962**

International Journal of Computational Linguistics and Applications

Vol. 4 **No. 1** **Jan-Jun 2013**

Guest Editor
Ajith Abraham

Editor-in-Chief
Alexander Gelbukh

ISSN 0976-0962

International Journal of Computational Linguistics and Applications

Vol. 4 **No. 1** **Jan-Jun 2013**

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

The views expressed herein are those of the authors. The journal reserves the right to edit the material.

Indexing: Cabell's Directory of Publishing Opportunities.

Editor-in-Chief: Alexander Gelbukh

Subscription: India: Rs. 2699
Rest of the world: US$ 249

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

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

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

International Journal of Computational Linguistics and Applications

Vol. 4 **No. 1** **Jan-Jun 2013**

CONTENTS

Editorial

It is my pleasure to present to the readers a new issue of IJCLA. This issue presents papers on four topics: lexical resources and specifically WordNet; grammar, semantics, and dialogue; information extraction; and sentiment analysis and social networks.

Lexical resources are the heart of most natural language processing technologies. Specifically, WordNet has traditionally been the most widely used lexical resource. It groups words with the same meaning together (such groups are called synsets and represent specific lexical meanings that exist in a given language) and separates different senses of the same word into different synsets. In addition, it specifies a wide variety of relationships between such lexical meanings, such as genus–species, part–whole, etc.

R. Amaro *et al.* (Portugal and Spain) report new developments in building a Portuguese WordNet. The original WordNet was built for English; development of WordNet-like dictionaries for other languages is a priority task for the corresponding communities. In addition, such development sheds light on the commonalities between languages and differences that require adjustments in the structure of WordNet. For Portuguese recently a number of rich high-quality lexical resources have been recently developed (a comparative analysis of some of them is given in another paper in this volume), which makes Portuguese an attractive alternative to English for language-independent and multilingual natural language processing experiments. Amaro et al. describe their efforts on increasing the density of relationships represented in Portuguese WordNet.

A. A. Freihat *et al.* (Italy and India) address the phenomenon of specialization polysemy and study it on the material found in WordNet. Specialization polysemy is a phenomenon observed when a word has two senses, one of which can be considered in a certain way more specific than another, that is, included in the other. Detection of this phenomenon is important in natural language processing applications such as machine translation or information retrieval. Freihat *et al.*

describe and classify different situations in which specialization polysemy appears in WordNet.

The next section is devoted to classical problems of natural language processing: grammar, semantics, and dialogue.

T. Seraku (UK) presents a solution to a class of problems related with incremental syntactic parsing. In some languages the syntactic structure of a sentence is defined very late in the analysis process, practically only when the whole sentence has been read by the parser. This poses efficiency and complexity problems to the parsing algorithm. Seraku proposes a solution based on dynamic addressing of the parsing tree nodes in the parser's internal memory. Examples are given for the Japanese language.

Y. Haralambous and **V. Klyuev** (France, Japan) consider the task of semantic analysis of text, which can also be called text understanding. Text understanding is in a way the philosopher's stone, the ultimate goal of natural language processing with which, when it is achieved, all other tasks would be easily solved. In its turn semantic analysis requires deep and broad knowledge about language and about the world and human life. Probably the wider available single source of such knowledge is Wikipedia. Haralambous and Klyuev improve a particular technique of semantic analysis, known as Explicit Semantic Analysis, using knowledge that can be extracted from the structure of Wikipedia.

N. Vadász *et al*. (Hungary) continues the topic of understanding text, addressing the problem of understanding the intentions of the speakers in monologs or dialogs. They present a formal framework for representing people's beliefs, desires, and intentions in a logical form. Their framework allows for analysis of joking, lying, fibbing, bluffing, expressing polarity and opinions, etc. They illustrate their formal ideas with numerous examples.

Z. Wei *et al*. (HK, UK, Quatar, China) show how to identify exploratory dialogs in text. Exploratory dialog is a type of communication between persons with deep understanding of each other's ideas, which implies proactive, positive, and creative participation in the communication. Such dialogs are especially important in learning environments and can occur between students or between the teacher and a student. It is important to detect and encourage this kind of dialog, and discourage non-explorative dialogs in learning and academic environments. Wei *et al*. use machine learning techniques to classify dialogs into explorative and non-explorative. I guess the authors themselves have mastered well the important art of

explorative communication, showing an impressive example of successful academic cooperation between teams from four different countries!

The next paper is devoted to the field of information extraction: identifying specific facts or relations in a given thematic domain expressed in the text.

R. Nawaz *et al*. (UK) present their system for determining descriptions of new biological events in biomedical scientific papers. The amount of published scientific literature nowadays does not allow the researcher to read or even look through all published literature on the topic of their research. Instead, automatic or semi-automatic methods have to be used to locate relevant pieces of information. This problem is especially observed in biomedical literature with its huge and rapidly growing body of published experimental data. Reports about newly observed events are intermixed in the texts with mentions of already known events; however, it is important to identify the novel contents of a scientific paper and the new biological events communicated in this paper. Nawaz *et al*. report more than 99% accuracy of their system in classifying the mentions of bio-events into new and previously known.

Finally, the last three papers are devoted to sentiment analysis, opinion mining, and analysis of the phenomena in the blogosphere and social networks. This is a very hot topic nowadays, with a lot of attention from private companies and governmental bodies drawn to it.

L.A. de Freitas and **R. Vieira** (Brazil) compare a number of lexical resources available for opinion mining and sentiment analysis in Portuguese language. As I have mentioned above, Portuguese natural language analysis community has developed good infrastructure with rich and high-quality lexical resources, which are not only useful for development of accurate applications for this language, but also for testing language-independent or otherwise non-English-oriented methods. Freitas and Vieira give an overview of such lexical resources available for Portuguese in the area of emotion, sentiment, and opinion analysis.

B.-K.H. Vo and N. Collier (Japan) analyze the emotions that Twitter users expressed during the tragic 2011 Great East Japan earthquake and tsunami and its aftermath that included leak of radioactivity from the Fukushima nuclear reactors and uncertainty about possible nation-wide nuclear catastrophe. They argue that

automatic analysis of emotions expressed in social networks in critical situations can help the government to quickly make correct decisions on social help and overall control of the situation. Specifically for the earthquake situations, Vo and Collier present a selection of the corresponding emotions to be tracked and two classification methods for these emotions to be automatically identified in massive Twitter flows.

L. Jia *et al*. (USA) address the task of blog retrieval with an additional requirement: the retrieved blog posts should not only correspond to the user query but also be of a specified facet: opinionated or factual, personal or official, and in-depth or shallow. Obviously, for this the blog posts are to be classified along these dimensions. Jia *et al*. propose the corresponding classifiers and show experimental results that confirm the effectiveness of their proposed methods.

This issue of IJCLA will be useful for researchers, students, and general public interested in various aspects of natural language processing.

GUEST EDITOR:

AJITH ABRAHAM
DIRECTOR,
MACHINE INTELLIGENCE RESEARCH LABS
(MIR LABS), USA
E-MAIL: <AJITH.ABRAHAM@IEEE.ORG>

Lexical Resources

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 11–27
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 14/06/13

Increasing Density through New Relations and PoS Encoding in WordNet.PT

RAQUEL AMARO,[1] SARA MENDES,[1,2] AND PALMIRA MARRAFA[1]

[1] *Universidade de Lisboa, Portugal*
[2]*Universitat Pompeu Fabra*

ABSTRACT

This paper reports research developed in the scope of building a wordnet for Portuguese (WordNet.PT), particularly focusing on the impact the results obtained have in the density of the network of relations and, thus, on its usability for NLP tasks. Following from basic research on different linguistic phenomena and on strategies for modeling them in relational models of the lexicon, the implementation of these results amounts to a richer resource, with new cross-PoS relations and information on event and argument structures, thus crucially contributing to accurately modeling all the main PoS in the database. We also define a way to integrate prepositions in wordnets and discuss the motivations and modeling strategies used to do so. Based on this work, we show how our contributions augment the coverage and the accuracy of WordNet.PT, by increasing the density of the network of relations, thus making it more usable for NLP applications.

KEYWORDS**:** *wordnets; cross-PoS lexical semantic relations; network density; linguistic coverage*

1 Introduction

WordNet.PT [1] (WN.PT) ([1],[2]), a wordnet for Portuguese developed according to the approach of EuroWordNet (EWN) ([3]), presents distinctive properties concerning the extension of the set of relations used and the strategies employed for attaining lexical coverage.

The initial strategies employed for building WN.PT had as main concern the accuracy of the resulting resource, rather than its extension. This, together with a strong focus on research, motivated the option for the manual selection, description and encoding of all WN.PT data, resulting in a smaller but much more reliable lexical resource, compared with automatically and semi-automatically constructed databases. The enlargement of the database has followed the semantic domains approach, involving the integration of lexical items from different PoS, which motivated the need for enriching the model with more information, namely information on selectional properties and new PoS, and for encoding new relations, in particular cross-PoS relations.

In this paper we present research developed in the scope of building WN.PT, particularly focusing on the impact the results obtained have in the density of the network and, thus, on its usability for NLP tasks. In Section 2 we present and discuss research on different linguistic phenomena, particularly regarding new relations, with a special focus on cross-PoS relations, introduced in WN.PT to model all the main PoS in the database and to encode information on argument structures. Section 3 is dedicated to the impact the contributions and modeling strategies implemented in WN.PT have on the density of the network. Section 4 concludes this paper with our final remarks and considerations regarding future work.

2 WordNet.PT Relations: Innovation and Coverage

WN.PT adopts almost entirely the set of relations defined in EWN, exception being the DERIVED, PERTAINS and BE IN STATE relations. The first two, besides being morphological relations, are somewhat complementary to the set of relations used in EWN (see [3]:37): the relation DERIVED is only used when there is a morphological link between two synsets and a lexical-conceptual relation already stands; the relation PERTAINS fulfills a void,

[1] http://www.clul.ul.pt/clg/wordnetpt

whenever there is a clear morphological link between a given noun and a given adjective and no other relation clearly stands. Given the lack of a clear stable conceptual relation holding between word forms linked via these relations, we do not use them. BE IN STATE relation is addressed further below.

Interesting enough, these relations are mostly used to relate nouns and adjectives, and can be seen as a way of linking adjectives in the lexicon, given that hyperonymy is not a structuring relation in the case of this PoS and that it does not hold for many adjectival synsets. Fundamental research on event structure and on adjectives developed within WN.PT ([4],[5]) has led to the definition of new semantic relations that further support discarding the ones mentioned above.

2.1 *Adjectives in WordNet.PT*

Following research on adjectives and their modeling in relational lexica ([6],[4]), in WN.PT we defined the following set of relations – CHARACTERIZES WITH REGARD TO, SETS VALUE TO, IS BY DEFINITION RELATED TO, IS A CHARACTERISTIC OF and IS TELIC SUBEVENT OF –, dealing with various complex lexical semantics phenomena regarding adjectives in a general and systematic way.

Although HYPERONYMY is the main structuring relation in wordnets, the semantic organization of adjectives is considerably different ([7]): nothing like the hierarchies of hyponymic relations is available for adjectives. Also, descriptive and relational adjectives[2] differ in terms of intrinsic meaning and of syntactic and semantic behavior (see [4]:53-76 for a detailed discussion on this issue). In WordNet ([8], [9]), descriptive adjectives are organized in clusters of synsets, an organization that mirrors psychological principles of the organization of the lexicon ([7]).

As argued in detail in [6] and [4], descriptive adjectives typically apply an incidence relation of a single property to the denotation of the noun they are related to in context. Put somewhat simplistically, they assign a value of an attribute to a noun. These values can be of different types: Boolean values, scalar values, and values that are neither one nor the other. Encoding this information in wordnets contributes to a more accurate lexical

[2] Wordnets leave out non-restricting adjectives. This option is based on the fact that, as pointed out by different authors ([10], [11], [4], etc.), non-restricting adjectives are a small class with a very particular semantic contribution, closer to semantic operators than to other adjectives.

representation of this PoS. In view of these properties, in WN.PT we use a small set of conceptual relations to represent descriptive adjectives, some of which inherited from the general EWN framework.

In WN.PT we use a semantic relation corresponding to the ATTRIBUTE relation of WordNet to encode the relation between adjectives and attributes, which we label as CHARACTERIZES WITH REGARD TO/IS CHARACTERIZABLE BY for the sake of transparency for non-specialist users:

1. $\{\text{tall}\}_{\text{Adj}}$ CHARACTERIZES W.R.T $\{\text{height}\}_{\text{N}}$/$\{\text{height}\}_{\text{N}}$ IS CHARACTERIZABLE BY $\{\text{tall}\}_{\text{Adj}}$

Naturally, our claims regarding this semantic relation are not related to the label used to encode it, but rather to the way it is used in WN.PT. In WordNet 3.0[3], in each adjective cluster, only focal adjectives are linked to an attribute. This is counter-intuitive, since the relation holding between *cold* and *temperature* is just as strong as the relation linking *gelid* and *temperature*, for instance. Moreover, the information regarding which attribute is associated to a given adjective – which is just as relevant for focal adjectives as for any other adjective in the cluster – can only be obtained in WordNet 3.0 if a mechanism for navigating the network of relations is developed in order to extract information expressed for focal adjectives and assign it to non-focal adjectives, where appropriate. Another crucial difference regards the relations used for the definition of adjective clusters: in WordNet 3.0 adjectives are associated by semantic similarity to a focal adjective to form clusters, and linked to a contrasting cluster through ANTONYMY. Instead of using a similarity relation that clearly poses problems (see [4]:95 and ff.), we claim that all adjectives ascribing values of the same attribute are linked to this attribute and thus related amongst themselves. This way, without having to encode it directly and somewhat artificially in the network, the clusters argued to be on the basis of the organization of adjectives are obtained: not around pairs of opposite adjectives, but around a common attribute, overcoming the need to define focal adjectives for each cluster.

At the same time that it overcomes the shortcomings mentioned above, it can be argued that this strategy results in loss of information, as the relation between adjectives associated to close values of a given attribute is not explicitly encoded in the network. This is particularly relevant in the case of scalar adjectives, as these adjectives determine values that are organized

[3] http://wordnet.princeton.edu/wordnet/

relatively to each other[4]. [12] state that gradation is in fact a semantic relation organizing lexical memory for adjectives. However, it is not encoded in WordNet because it is rarely lexicalized in English. But besides this individual organization relatively to each other, scalar adjectives are also organized around areas of a scale: typically two extremes and a middle value. Despite the relevance of continuing to develop research on how to model adjective scales ([13]), we start with a coarser modeling of this adjective subclass, which encodes the area of the appropriate scale to which the attribute value assigned by a given adjective belongs. To accomplish this we use a new semantic relation to link the adjective and the lexicalization of the value it assigns, typically an adverb: SETS VALUE TO/IS THE VALUE SET BY.

2. $\{tall\}_{Adj}$ SETS VALUE TO $\{plus\}_{Adv}$/$\{plus\}_{Adv}$ IS THE VALUE SET BY $\{tall\}_{Adj}$

This relation overcomes the information loss mentioned above: through the combination of the CHARACTERIZES WITH REGARD TO and the SETS VALUE TO relations we are able to obtain the cluster organization of adjectives, without the need for using fuzzy similarity relations or for defining *a priori* pairs of focal adjectives. Moreover, we can use the same strategy for encoding descriptive adjectives which do not assign scalar values. Adjectives like *dead* and *alive*, for instance, assign Boolean values, associated to the presence or absence of an attribute in the modified noun, i.e. a **yes** or **no** value of the relevant attribute. To encode this, we also use the SETS VALUE TO relation, linking such adjectives to $\{yes\}_{Adv}$ or $\{no\}_{Adv}$.

With regard to relational adjectives, things are considerably different, as these adjectives are not organized in opposite clusters. The meaning of relational adjectives is something like 'of, relating/pertaining to, associated with' some noun. In WordNet and EWN, relational adjectives are encoded as pertainyms of the nouns they are morphologically associated to. In EWN the PERTAINS relation is basically a morphological link (which is not always the case: e.g. *water* and *aquatic*), associated to a fuzzy semantic relation: it holds when no other relation clearly stands. In contrast to what is claimed in [3]:37, we argue that far from being meaningless 'themselves', relational adjectives involve sets of properties and introduce a relation between these sets of properties and the noun modified ([6], [4]). These adjectives establish an underspecified relation, which is specified in con-

[4] For a discussion on adjective scales and WordNet adjective clusters see [13].

text, between the modified noun and a domain that is exterior to it. In WN.PT, we use a very underspecified link to encode this relation: the relation IS BY DEFINITION RELATED TO. The salience of the semantic relation holding between relational adjectives and the lexicalization of the set of properties they are associated with, independently of any morphological link between them, motivates the creation of this new relation, which is exactly the opposite of what is stated about the PERTAINS relation in EWN, which focuses on the morphological link. Also, this broader relation allows for linking relational adjectives even when the set of properties involved is not lexicalized by a noun, but by a lexical item from another PoS, like in the case of *sedative*$_{\text{Adj}}$ and *sedate*$_{\text{V}}$, for instance.

This way the main relations used for encoding descriptive and relational adjectives in WN.PT are: ANTONYMY, CHARACTERIZES WITH REGARD TO, and SETS VALUE TO, for the former; and IS BY DEFINITION RELATED TO, for the latter. These semantic relations allow us to encode the basic definitional characteristics of these adjectives in a linguistically motivated way, at the same time making it possible for membership to these classes to emerge from the network of relations encoded.

But adjectives are also relevant for the codification of salient properties of other lexical items. EWN uses the BE IN STATE relation to encode "links between nouns that refer to anything in a particular state expressed by an adjective" ([3]:37), recognizing the role adjectives can play in the characterization of nominal synsets. However, the definition and scope of application of this relation is too narrow: it cannot be used with relational adjectives, which are associated to sets of properties and not to a single state. Inspired by these observations and in order to broaden the domain of application of the link used in EWN, we introduce the new relation IS A CHARACTERISTIC OF/HAS AS A CHARACTERISTIC, in (3).

3. {carnivorous}$_{\text{Adj}}$ IS A CHARACTERISTIC OF {shark}$_{\text{N}}$ *reversed*
   {shark}$_{\text{N}}$ HAS AS A CHARACTERISTIC {carnivorous}$_{\text{Adj}}$

This relation allows us to express the most salient – and definitional – features of nouns in the network, contributing to richer and more clearly defined synsets. The possibility of ascribing, but also of negating this relation allows us to encode contrasting definitional features of certain nouns, in a similar way to the features encoded by some meronymy relations[5].

[5] One of the prototypical features of *shark*, in (3), is *carnivorous*. In contrast, one of the distinctive features of *whale shark*, hyponym of *shark*, is the fact that

Being able to express this is therefore very relevant, not only because it mirrors speakers' lexical knowledge, but also because it can provide crucial information to wordnet-based applications using inference systems.

Finally, in WN.PT adjectives are also used to encode definitional properties of verbs. Following [14] and further work on the representation of complex predicates in wordnets, verb telicity is also encoded in WN.PT.

4. {sadden}$_{V}$ HAS TELIC SUBEVENT {sad}$_{Adj}$/{sad}$_{Adj}$ IS TELIC SUBEVENT {sadden}$_{V}$ *reversed*
5. [$_{T}$ [$_{P}$ act(x,y) and ~ Q(y)], [$_{e}$Q(y)]]
   T: transition, P: process, e: event, Q: atomic event
6. a. He made Mary sad./b. *He made Mary.
7. a. *He saddened Mary sad./b. He saddened Mary.

The semantics of telic verbs involves a change of state of their theme argument, i.e. the subevent that closes the whole event is an atomic event, (a state) that affects its theme and is different from its initial state. By default, these verbs are associated to an LCS (Lexical Conceptual Structure) like the one in (5).

When syntactically realized, in contexts with LCS deficitary telic verbs ([14]), for instance, the telic subevent generally corresponds to an adjectival constituent (see 6a), whereas in the general case the telic state is incorporated in the verb, hence the ill-formation in (7a). In WN.PT the telicity of these verbs is captured through the relation HAS TELIC SUBEVENT/IS TELIC SUBEVENT (see (4)). This relation is different from the SUBEVENT relation in EWN as the latter only stands for lexical entailment involving temporal proper inclusion, therefore not accounting for the geometry of the event (see (5)). This is not the case of the TELIC SUBEVENT relation which regards the atomic subevent that is the ending point of the global event denoted by the verb, thus not properly included.

---

it is not. Moreover, this is the specific difference that distinguishes it from its sisters. This example makes apparent that this relation between nouns and adjectives expresses information just as crucial as the one encoded by some MERONYMY relations: *caffeine* IS A MERONYM OF *coffee*, and the negation of this MERONYMY relation is the specific difference of *decaf*, for instance.

2.2 *Prepositions in WordNet.PT*

Besides being syntactic markers, prepositions are also regarded as a kind of relation operator, relating concepts such as space, temporality or causality, and have been described according to their conceptual properties ([15], [16], [17], among others). Studies such as these, along with the identification of the need to account for arguments introduced by prepositions for a fine-grained codification of predicates in relational models of the lexicon, motivated the integration of prepositions in WN.PT ([5]).

As other PoS, prepositions can be related by SYNONYMY [6], HYPERONYMY and ANTONYMY relations, although the criteria for establishing whether these relations hold or not between two prepositions require slight adjustments of the test formulae used for pinpointing these relations, in order to consider the preposition plus the element with reference potential it combines with ((8), (9), (10)).

8. $Prep_1$ IS SYNONYM OF $Prep_2$ in a given Context iff: if $Prep_1$ then $Prep_2$ and if $Prep_2$ then $Prep_1$ (*over* IS SYNONYM OF *on top of*)
9. $Prep_1$ IS HYPERONYM OF $Prep_2$ iff: $Prep_2$ is $Prep_1$+ (space/time/direction…) but not the converse ({toward} IS HYPERONYM OF {downward} (*toward* + direction))
10. $Prep_1$ IS ANTONYM OF $Prep_2$ iff: i) $Prep_1$ and $Prep_2$ are hyponyms of $Prep_3$; ii) $Prep_1+XP_i$ is the opposite of $Prep_2+XP_i$ and $Prep_2+XP_i$ is the opposite of $Prep_1+XP_i$; therefore if $Prep_1+XP_i$ then not $Prep_2+XP_i$ and if $Prep_2+XP_i$ then not $Prep_1+XP_i$ ({above} IS ANTONYM OF {below})

Interestingly, the linguistic tests for HYPONYMY show that prepositions denoting source and goal locations, for instance, are not hyponyms of a preposition denoting location ([5]). In fact, there is a strong semantic relation between the concepts of location, and source and goal locations, but it is a causality relation rather than a specification relation: moving something to a goal location causes that something to be in that location (see (11)), just as moving something from a source location causes that some-

[6] Although it exists, SYNONYMY is not very productive for this PoS. This fact is probably not independent of prepositions being a closed-class, and seems to be conversely proportional to the highly polysemic behavior of prepositional expressions.

thing not to be in that location[7]. This way, prepositional nodes can also be related by CAUSE relations.

11. $Prep_1$ CAUSES $Prep_2$ iff: $Prep_1$+XP_i CAUSES/HAS AS CONSEQUENCE $Prep_2$+XP_i, but not the converse ({to} CAUSES/HAS AS CONSEQUENCE {at})

The integration of prepositions in wordnets, besides allowing to explicitly state subcategorization properties of predicates, contributes to compensate some shortcomings of mainstream wordnets, namely in terms of distinguishing word senses based on the relations encoded in the database. In section 2.3, we discuss these aspects in detail, in relation with a proposal for encoding selectional properties of predicates in wordnets.

2.3 *Encoding Selection Information*

Among the cross-PoS relations available in EWN, there is a set of relations concerning the role (or function) of entities in events. As stated in [3]:29, ROLE relations are based on thematic role assignment, and are correlated with the argument structure of verbs. However, the nodes related by ROLE relations often are not coincident with the selection restrictions of verbs. In addition, in many cases, ROLE relations are only definitional to the meaning of the participant. For instance, a *passenger*/*customer* is defined as one that *travels/buys*, but the event denoted by *travel*/*buy* is not defined as an event having a *passenger/customer* as an agent.

Following research on verbal predicates ([5]), we define three new relations to account for selection information, based on the argument structure as defined in the Generative Lexicon (GL) ([18]). Argument structure in GL allows for specifying the number and semantic type of arguments of a given predicate, also including information on how these arguments behave syntactically in general, namely with regard to specific restrictions on their overt realization in context, distinguishing between true, shadow and default arguments ([18]:63 and ff.).

---

[7] PPs introduced by the preposition *at*, indicator of location, correspond to the resulting state of the movement *from* or *to* to a given location. Prepositional phrases headed by this item can replace state denoting items such as adjectives, providing evidence for this claim (see [5]:155): *John is* ***tired***./*John is* ***at the door***.

Briefly, the relation SELECTS/IS SELECTED BY refers to true arguments, i.e arguments that have to be syntactically realized (or whose omission has to be licensed by syntactic or pragmatic contexts); the INCORPORATES/IS INCORPORATED relation refers to shadow arguments, strictly incorporated in the lexical predicate, which means they cannot be overt arguments unless they are further specified; and the HAS AS DEFAULT ARGUMENT/IS DEFAULT ARGUMENT OF refers to participants in the event structure of the predicate that are mostly null, since the semantics of the predicate allows for a default interpretation (for further discussion on these relations, see [5]). Also, taking advantage of the inheritance mechanism in the WordNet model, the relation SELECTS accounts for the overt realization of the target node of this relation or any of its direct or indirect hyponyms, see (12).

12. {die}$_V$ SELECTS {living being}$_N$: All living beings / birds / men / insects / ... die.

The implementation of these relations in WN.PT takes advantage of the possibility of relating either variants or synsets, and from the conjunction operator, available in the EWN framework. The first allows for stating different selection restrictions for the members of a synset, in (13) below. If nothing is stated, the relation applies to all the elements. Otherwise, the variant-to-variant restriction has to be activated, and the two elements related explicitly identified. As to the conjunction operator, it allows for simultaneously linking the elements of complex arguments, as it is the case, for instance, of arguments introduced by a preposition, illustrated in (14):

13. {voltar, regressar}$_V$ [≅ return, come back]
    SELECTS {para}$_P$ (**variant to variant**: *voltar* - *para*);
    SELECTS {a}$_P$
14. {engarrafar}$_V$ [≅ bottle; put in a bottle]
    INCORPORATES {em}$_P$ [≅ in] (**conjunctive 1**)
    INCORPORATES {garrafa}$_N$ [≅ bottle] (**conjunctive 2**)

An inheritance mechanism drastically reduces the work involved in specifying this information, since selection information relations are inherited through hyponymic chains, as mentioned above. However, selectional information is not always completely inherited by hyponyms, as made apparent by the case of incorporated arguments, further motivating a mechanism of lexical inheritance by default: hyponyms inherit all the information that characterizes their hyperonyms if nothing is stated otherwise.

ROLE and selection information relations are not always coincident, even when considering definitional properties of predicates only. WN.PT data shows that ROLE and selection information relations are typically coincident in the case of agents ({dress}$_V$ INVOLVED_AGENT/SELECTS {person}$_N$), the same not being necessarily true when other participants are at stake. In (15), the instrument used in an event like *selar* (seal) is identified through a ROLE relation, but this relation does not allow us to know that, in the specific case of this verb, this is an incorporated argument, and, as such, only syntactically realized under strict constraints.

15. {selar}$_V$ [seal, ≅ close with a seal]
    INVOLVED_INSTRUMENT {selo}$_N$ [≅ seal]
    INCORPORATES {com}$_P$ [≅ with] (conj. 1); INCORPORATES {selo}$_N$ [≅ seal] (conj.2)

According to the literature ([9]), the specification of the manner in which events occur has a special significance in the determination of verbal meaning. This specification, when a lexicalization of the manner is available, is encoded through the IN_MANNER relation ([3]:36), linking verb and adverb synsets, such as {run}$_V$ IN_MANNER {fast}$_{Adj}$. In a similar way, when no lexicalization of manner is available, but this information is incorporated in the verbal predicate, we claim that the INCORPORATES relation can be used, as shown in (16).

16. {puxar}$_V$ [≅ move with traction, pull]
    INCORPORATES {com}$_P$[≅ with] (conj. 1); INCORPORATES {tracção}$_N$ [≅ traction] (conj. 2)

The introduction of selection information relations allows for distinguishing and representing different levels of information in the WordNet model, increasing the amount of information that can be expressed in it: ROLE and IN_MANNER relations (existing in the EWN framework) – conceptual properties; SELECTS, INCORPORATES and HAS AS DEFAULT ARGUMENT relations – selectional properties and syntactic restrictions. Selection information relations coherently complement the existing relations, resulting in a more accurate description of lexical items and linking synsets which otherwise would not be associated. Our goal is not to provide complete syntactic frames for each synset, but to make available richer descriptions of lexical-conceptual units, following the assumption that selection information reflects semantic and syntactic relations ([19]). Consider, for instance, the verb *pôr* (≅ put; to move to a location), that selects

an argument introduced by the preposition *em* (≅ indicator of location). On the one hand, the specific information that distinguishes this verb from its hyperonym is directly related to the determination of a final location. On the other hand, there is no node in the lexicon suitable to be linked through INVOLVED_LOCATION to the verb *pôr*, since almost any lexicalization of a surface, object, area, body part, etc., can occur with *pôr*. The integration of prepositions in WN.PT, together with the selection information relations described, result in a richer and more accurate characterization of the semantic properties of lexical items, explicitly modeling semantic content and co-occurrence information reflecting semantic properties, and enabling the establishment of word senses based on the information expressed in the network ([19]):

17. {pôr}$_V$ [≅ put] HAS_HYPERONYM {mover}$_V$ [≅ move]
    INCORPORATES {para}$_P$ [≅ indicator of goal location]
    SELECTS {em}$_P$ [≅ indicator of location]

Although the preposition *em* (≅ in) does not impose strong selection constraints on the NP it combines with (e.g. *in the table / bedroom / fridge / field / shoe / air…*), the incorporated meaning component of *pôr* referring a goal location is now accounted for through the INCORPORATES relation, and the concept of location is indicated by the prepositional nodes selected.

These more precise and richer descriptions of synsets allow for contributing to the distinction of word senses based on the relations encoded in the network. In fact, selection information alone can be used to straightforwardly distinguish between word senses, as illustrated in (18). This is particularly relevant since, as argued in [20], information on the co-occurrence of words is easily available in texts – and thus easily accessible for NLP tasks based on corpora analysis and statistic oriented models –, whereas information on the co-occurrence of meanings is harder to extract from raw data and requires complex strategies involving text disambiguation.

18. {tratar}$_V$ [≅ treat, heal]
    SELECTS {animal}$_N$ [≅ animal]
    {tratar}$_V$ [≅ take care]
    SELECTS {de}$_P$ [≅ of] (conj.1); SELECTS {animal}$_N$ [≅ animal] (conj. 2)

3 WordNet.PT Data: Informational Richness and Density

The density of wordnets is specifically significant considering that in this model word senses are represented in terms of relationships between synsets. Also, WordNet has been used to solve primary barriers in the development of reliable information retrieval, machine translation, summarization and language generation systems, or word-sense disambiguation applications, for which rich language resources are crucial.

Particularly in the case of word sense disambiguation, the density of WordNet has been considered limited ([21]). Thus, augmenting the density of relational language resources is decisive from the point of view of their usability, whether we consider inference-based applications, where the richer the connectivity in the database, the more inference is possible ([22]), or applications that draw on measurements of semantic relatedness between concepts, since higher relational density provides shorter average paths between lexical objects ([23]). For these reasons, several strategies have been put forth in order to augment the density of wordnets, such as those depicted in [19], [20], [21] or [23], to name a few, and further developments resulting in the increase of network density continue to be welcomed. In this section, we show how the new relations proposed under the scope of our work accomplish just that.

Table 1 compares the density (number of relations per synset) of WN.PT, after the implementation of the new lexical-conceptual relations described, with the density of WordNet 1.5[8], regarding adjectives and a subset of verbs.

Comparing WN.PT with WordNet 3.0 instead of WordNet 1.5 (see footnote 8), particularly considering PoS differentiation, could provide different numbers. Nonetheless, it is possible to show how density increases as a result of using the new selection information relations with regard to the verbs tested. The same occurs when comparing EWN relations and

[8] Developed in the general framework of EWN, WN.PT was originally implemented with Polaris, which determined its mapping with WordNet 1.5 data, the mapping of Inter-Lingual links to WordNet 3.0 being still ongoing. For this reason, and given the fact that WordNet 3.0 statistics do not cover the number of relations in total or by PoS, the data considered here for purposes of comparison are those of WordNet 1.5. Based on the statistics available (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html) and data offered by [24]:374, the density of WordNet 3.0 considering the total number of synsets and relations is 2.0.

Table 1. Network density for adjectives and a sample of verbs (change of location verbs) in WordNet 1.5, WN.PT only with EWN relations, and with its new specific relations

	WordNet 1.5	EWN relations	WN.PT relations
Adjectives	1.48	–	**4.48**
Change of location verbs	2.13	4.12	**5.68**

new relations implemented in WN.PT in general. The increase in density of WN.PT with regard to WordNet 1.5 is quite substantial: about 200% for adjectives and more than 165% for the verbs tested. This increase is not as high when comparing WN.PT using EWN relations only with WN.PT using the full set of relations proposed. However, it is still quite significant (37%).

Adding to the linguistic motivation, these results further sustain the use of these new relations in wordnets. Besides the importance of having a denser network from the point of view of wordnet-based applications, increasing the density of wordnets is a crucial aspect for relational models of the lexicon themselves since the meaning of each unit is determined by the set of relations it holds with other units: a denser network of relations results in richer and more appropriately defined synsets.

4 Final Remarks

The implementation of new relations and the integration of new PoS in WN.PT decisively contribute to enhancing its density, consistency and coverage. The new relations allow for more accurate and motivated descriptions but also for the integration of new PoS, enhancing the usability of the database in different types of computational applications. This has been tested in several applications, both in terms of the contribution to the treatment of different linguistic phenomena (such as co-occurrence restrictions of co-hyponyms and contrasts in Aktionsart values within troponymic chains ([5]), word-sense disambiguation ([25], [26]), or usability in Language Engineering applications ([27])).

However, several issues require further attention. First, some promising results emerge from the work already developed. With regard to adverbs, studying to which extent the comprehensive treatment of event modifying

adjectives can contribute to the treatment of this category and how the set of properties identified is mirrored in wordnets are open questions that can contribute to a deeper treatment of this PoS. Further research and testing on the selection properties of predicates is also due, specifically in the case of underspecified arguments that correspond to high nodes in the hierarchy.

Regarding this, using the information available in WN.PT to establish selection features might provide a solution. Consider, for instance, the following Portuguese pair of verbs: *enjaular* (≅ cage; put inside a large and resistant cage, typically made of metal, animals of considerable size), as in *the hunter caged the lion*; and *engaiolar* (≅ cage; put inside a cage, small animals, typically birds or small mammals), as in *the child caged the canary*. Setting *animal* as argument of these two verbs overgenerates, since many hyponyms of *animal* cannot be arguments of either one or the other of the two verbs. The solution might be to consider features expressed by other available relations, such as HAS AS CHARACTERISTIC {grande}$_{Adj}$ (big) or {pequeno}$_{Adj}$ (small), for the arguments of *enjaular* and *engaiolar*, respectively.

References

1. Marrafa, P.: The Portuguese WordNet: General Architecture and Semantic Internal Relations, *DELTA* (2002).
2. Marrafa, P.: WordNet do Português - Uma base de dados de conhecimento linguístico. Lisbon: Instituto Camões (2001).
3. Vossen, P.: EuroWordNet General Document. EuroWordNet Project LE2-4003 & LE4-8328 report, University of Amsterdam (2002)
4. Mendes, S.: Syntax and Semantics of Adjectives in Portuguese: analysis and modeling. PhD dissertation. University of Lisbon (2009)
5. Amaro, R.: Computation of Verbal Predicates in Portuguese: relational network, lexical-conceptual structure and context. PhD dissertation. University of Lisbon (2009)
6. Mendes, S.: Adjectives in WordNet.PT. 3rd Global WordNet Association Conference, pp. 225–230. Jeju Island, Korea (2006)
7. Miller, K. J.: Modifiers in WordNet. In: Fellbaum, C. (ed.) WordNet: an electronic lexical database, pp. 47–68. Cambridge, MA: The MIT Press (1998)
8. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.J.: Introduction to WordNet: an On-line Lexical Database. International Journal of Lexicography, vol. 3, number 4 (1990).

9. Fellbaum, C.: A Semantic Network of English Verbs. In: Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, pp. 69–104. Cambridge, MA: The MIT Press (1998)
10. Kamp, H.: Two theories about adjectives. In: Keenan, E. (ed.) Formal Semantics of Natural Language, pp. 123-155. Cambridge: Cambridge University Press (1975)
11. Chierchia, G., McConnel-Ginet, S.: Meaning and Grammar: an Introduction to Semantics. Cambridge, MA: The MIT Press (1990)
12. Fellbaum, C., Gross, D., Miller, K. J.: Ajectives in WordNet. In Five Papers on WordNet. Princeton, USA (1993).
13. Sheinman, V., Tokunaga,T.: AdjScale: Differentiating between similar adjectives for language learners. 1st International Conference on Computer Supported Education, pp. 229–235 (2009)
14. Marrafa, P.: Modelling constituency and predication in Portuguese. Revista PaLavra, vol. 12 (special issue: Linguística Computacional), pp. 106–118 (2004)
15. Jensen, P., Nilsson J. F.: Ontology-Based Semantics for Prepositions. ACL-SIGSEM workshop. Institut de Recherche en Informatique de Toulouse (2003)
16. Saint-Dizier, P.: PrepNet: a framework for describing prepositions: preliminary investigation results. 6th International Workshop on Computational Semantics ITK, pp. 25–34. Tilburg (2005)
17. Mcshane, M, Beale S., Nirenburg S.: Disambiguating Homographous Prepositions and Verbal Particles In An Implemented Ontological Semantic Analyzer. Working Paper 01-05. ILIT, University of Maryland Baltimore County (2005)
18. Pustejovsky, J.: The Generative Lexicon. Cambridge, MA: The MIT Press (1995)
19. Agirre, E., Martinez, D.: Integrating selectional preferences in WordNet. 1st International WordNet Conference, pp. 1–9. Mysore (2002)
20. Bentivogli, L., Pianta E.: Extending WordNet with Syntagmatic Information. 2nd International WordNet Conference, pp. 47–53. Brno, Czech Republic (2004)
21. Boyd-Graber, J., Fellbaum, C., Osherson D., Schapire R.: Adding dense, weighted connections to WordNet. 3rd Global WordNet Meeting, pp. 29–35. Jeju Island, Korea (2006)
22. Harabagiu, S., Moldovan D.: Knowledge Processing on an Extended WordNet. In: Fellbaum, C. (ed.) WordNet. An Electronic Lexical Database, pp. 353–378. Cambridge, MA: The MIT Press (1998)
23. Lemnitzer, L., Wunsch H., Gupta P.: Enriching GermaNet with Verb-noun Relations - a Case Study of Lexical Acquisition. LREC 2008, pp.156–160. Marrakech, Morocco (2008)

24. Agirre, E., Montse, C., German R., Soroa, A.: Exploring Knowledge Bases for Similarity. LREC 2010, pp. 373–377. Valletta, Malta (2010)
25. Marrafa, P., Mendes, S.: Using WordNet.PT for translation: disambiguation and lexical selection decisions. International Journal of Translation, vol. 19. Bahri Publications (2007)
26. Marrafa, P., Amaro, R., Freire N., Mendes S.: Controlled Portuguese: coping with ambiguity. In: Kuhn, T., Fuchs N. E. (eds.) CNL 2012, LNCS 7427, pp. 152–166. Springer-Verlag Berlin Heidelberg (2012)
27. Marrafa, P., Ribeiro, C., Santos R., Correia, J.: Gathering Information from a Relational Lexical-Conceptual Database: A Natural Language Question-Answering System. 8th World Multi-Conference on Systemics, Cybernetics and Informatics. Orlando (2004)

RAQUEL AMARO
CENTRO DE LINGUÍSTICA,
UNIVERSIDADE DE LISBOA,
AVENIDA PROFESSOR GAMA PINTO, 2, 1649-003 LISBOA, PORTUGAL
E-MAIL: < RAMARO@CLUL.UL.PT >

SARA MENDES
CENTRO DE LINGUÍSTICA,
UNIVERSIDADE DE LISBOA,
AVENIDA PROFESSOR GAMA PINTO, 2, 1649-003 LISBOA, PORTUGAL
AND
UNIVERSITAT POMPEU FABRA
ROC BORONAT 138, BARCELONA, SPAIN
E-MAIL: < SARA.MENDES @CLUL.UL.PT >

PALMIRA MARRAFA
CENTRO DE LINGUÍSTICA,
UNIVERSIDADE DE LISBOA,
AVENIDA PROFESSOR GAMA PINTO, 2, 1649-003 LISBOA, PORTUGAL
E-MAIL: < PALMIRA.MARRAFA@NETCABO.PT >

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 29–52
RECEIVED 13/12/12 ACCEPTED 11/01/13 FINAL 21/03/13

Solving Specialization Polysemy in WordNet

ABED ALHAKIM FREIHAT,[1] FAUSTO GIUNCHIGLIA,[1] AND BISWANATH DUTTA[2]

[1]*University of Trento, Italy*
[2]*Indian Statistical Institute (ISI), India*

ABSTRACT

Specialization polysemy refers to the type of polysemy, when a term is used to refer to either a more general meaning or to a more specific meaning. Although specialization polysemy represents a large set of the polysemous terms in WordNet, no comprehensive solution has been introduced yet. In this paper we present a novel approach that discovers all specialization polysemy patterns in WordNet and introduces new operations for solving all the instances of the problem.

KEYWORDS: *WordNet, Polysemy, Specialization Polysemy, regular Polysemy, Polysemy Reduction*

1 Introduction

Solving the polysemy problem in WordNet [1] is very crucial in many research fields including Machine translation, information retrieval and semantic search [13]. Several approaches have been introduced to solve the polysemy problem, but no approach gives a comprehensive solution to the problem. Solving the polysemy problem is very important because the high polysemous nature of WordNet leads to insufficient quality of natural language processing (NLP) and semantic applications.

Current polysemy approaches classify the polysemy problem in *contrastive polysemy* which corresponds to the polysemous terms that have unrelated meanings and *complementary polysemy* which corresponds to the polysemous terms with related meanings. This classification is correct but in general, it is not sufficient to solve the problem. We need further analysis of the different types of related polysemy and introduce a solution that solves the problem according to the specific nature of each of these related polysemy types [12]. For example, the methods for solving the *metonymy polysemy* (described in section 2) cannot be applied for solving the *specialization polysemy*, although both polysemy types belong to the complementary polysemy.

Specialization polysemy is a type of complementary polysemy that refers to the cases, when a term is used to refer to either a more general meaning or a more specific meaning [5]. The more general/ more specific meaning relation between the senses of specialization polysemy terms reflects a hierarchical relation between the senses that is encoded implicitly at lexical level rather than the semantic level. For instance, in the following example, sense 2 is a more general meaning than sense 1:

1. **correctness**, rightness: conformity to fact or truth.
2. **correctness**: the quality of conformity to social expectations.

Although Specialization polysemy represents a large set of the polysemous terms in WordNet, no comprehensive solution has been introduced yet. Systematic polysemy approaches such as CORELEX [4] did not provide a solution for specialization polysemy. Regular polysemy approaches such as the work presented in [5] discovered some patterns of specialization polysemy cases without offering a solution. On the other hand, polysemy reduction approaches tried to solve a subset of the specialization polysemy cases through merging the similar meanings of polysemous terms [3].

In this paper, we present a novel approach to solve the specialization polysemy in WordNet. The presented solution solves the specialization polysemy problem by providing a semi automatic method for discovering the specialization polysemy cases by means of regular structural patterns. It also provides criteria for determining the nature of the hierarchical relation between the senses of a specialization polysemy cases and new operations that solve the specialization polysemy problem by transforming the implicit relations between the

synsets at lexical level into explicit relations at the semantic level. The advantages of our approach are that it improves the ontological structure of specialization polysemy cases and increases the knowledge in WordNet by adding new missing senses and relations rather than merely decreasing knowledge as it is suggested in polysemy reduction [3] and sense clustering approaches [10, 11].

This paper is organized as follows: in Section two, we give an overview of polysemy types in WordNet and make a comparison between specialization polysemy and other polysemy types. In Section three, we present the structural patterns of specialization polysemy and an algorithm for discovering these patterns. In Section four, we introduce the synset patterns in the case of specialization polysemy and show how we use these patterns to solve the specialization polysemy problem. In Section five, we discuss the results and evaluation of our approach. In Section six, we conclude the paper and describe our future research work.

2 Specialization Polysemy

WordNet [1, 2] is a lexical database that organizes synonyms of English words into sets called synsets where each synset is described through a gloss. WordNet organizes the relations between synsets through semantic relations where each grammatical category has a number of relations that are used to organize the relations between the synsets of that grammatical category. For example, the hyponymy relation (X is a type of Y) is used to organize the ontological structure of nouns. WordNet 2.1 contains 147,257 words, 117,597 synsets and 207,019 word-sense pairs. Among these words there are 27,006 polysemous words, where 15776 of them are nouns. The number of senses of polysemous nouns may range from 2 senses to 33 senses. Nevertheless, 90% of these nouns have less than 5 senses. WordNet uses sense ranking to order the synsets of the polysemous words. This order reflects the familiarity of the senses. The sense number 1 is the most familiar or *the common sense* of the synset. Another important ranking is the synset synonyms ranking. This ranking reflects which term is usually used to express a synset, where the first synonym is the most used term and so on. The first synonym of a synset is also called the *preferred term* of the synset. Note that, in this paper, we use the

notion term to refer for a word and its part of speech. For example, the word *love* has two terms: love as a noun, and love as a verb. We use the notion sense(s) to refer to synset(s) of a term. Notice that, in this paper, we are concerned with polysemous nouns only.

Polysemy approaches differentiate between *contrastive polysemy*, i.e. terms with completely different and unrelated meanings—also called homonyms; and *complementary polysemy*, i.e. terms with different but related meanings. Complementary polysemy is classified in three sub types: *Metonymy*, *specialization polysemy*, and *metaphoric polysemy*. Polysemy approaches did not offer a solution for the polysemy problem that takes into account the different nature of each of these types. For example, regular polysemy approaches dealt with metonymy and metaphoric cases only. Classifying polysemy types and providing a solution for each type is a very important improvement towards making WordNet a suitable resource for NLP applications. In the following we explain the different polysemy types and discuss the difference between specialization polysemy and metonymy and metaphors.

2.1 *Specialization Polysemy*

Specialization polysemy is a type of complementary polysemy which denotes a hierarchical relation between the senses of a polysemous term. In case of abstract senses, we say that a sense A is a more general meaning of a sense B. In this case we say also that the sense B is a more specific meaning of the sense A. In the cases, where the senses denote physical entities, we may also use the taxonomic notations type and subtype instead of more general meaning and more specific meaning respectively. In the following examples, sense 2 denotes a subtype of the type denoted by sense 1 for the term turtledove:

1. Australian turtledove, **turtledove**, Stictopelia cuneata: small Australian dove
2. **turtledove**: any of several Old World wild doves.

A very important characteristic of specialization polysemy terms that differentiate it from contrastive polysemy and metonymy terms is the *type compatibility* of the term senses. By type compatibility, we mean that the term senses belong to the same type. For example both types of

turtledove belong to the type dove. Some metaphoric cases as we shall see later, have the type compatibility property also.

2.1.1 Metonymy Polysemy

Metonymy polysemy happens when we substitute the name of an attribute or a feature for the name of the thing itself such as the second sense in the following example.

1. fox: alert carnivorous mammal with pointed muzzle and ears and a bushy tail.
2. fox: the grey or reddish-brown fur of a fox.

In metonymy, there is always a *base meaning* of the term and other *derived meanings* that express different aspects of the base meaning [8]. Sense 1 of the term fox in the previous example is the base meaning and sense 2 is a derived meaning of the term. Metonymy is different from specialization polysemy in the following way: The senses of metonymy terms belong to different types and thus the relation more general meaning/ more specific meaning is not applicable for metonymy. For example, the base meaning of the term fox belongs to the **animal** category while derived meaning belongs to **artifact**. This means, the relation between the derived meanings and the base meaning of a metonymy term cannot be hierarchical as it is the case in specialization polysemy. It is possible to find type compatible metonymy cases. The point here is that in such cases it is very difficult to distinguish between metonymy and specialization polysemy. We think that treating such cases as specialization polysemy is better since the hierarchical relation is stronger than the metonymic relation.

2.1.2 Metaphoric Polysemy

Metaphoric polysemous terms are the terms that have literal and figurative meanings. In the following example, the first sense of the term **honey** is the literal meaning and the second sense is the figurative:

1. **honey**: a sweet yellow liquid produced by bees.
2. beloved, dear, dearest, loved one, **honey**, love: a beloved person.

The metaphoric relation between the literal sense and the metaphoric sense may disappear or it may become difficult to understand the metaphoric link between the metaphoric and literal sense of the term. We

call such cases dead metaphors. For example, the senses of animator indicate a dead metaphor:

1. energizer, vitalizer, **animator**: someone who imparts energy and vitality to others.
2. **animator**: the technician who produces animated cartoons.

From a hierarchical point of view metaphors can be divided into two groups:

a. *Type compatible metaphors*: the cases, where the literal meaning and the figurative meaning belong to the same type. Consider the term role player for example:

   1. pretender, **role player**: a person who makes deceitful pretenses.
   2. actor, **role player**: a theatrical performer.

b. *Type incompatible metaphors*: the cases, where the literal meaning and the figurative meaning belong to the different types. The literal meaning of the term **honey** for example belongs to the **food** category, while the figurative meaning belongs to **person**.

The metaphoric relation is not hierarchical. The metaphoric link between the senses is raised usually through inconsistency between the literal and the metaphoric sense. Although both senses of the term **role player** belong to the type person, these senses are inconsistent and cannot be generalized to a common type. In the case of dead metaphors and/or the cases, where it is difficult to grasp the metaphoric link between the senses, compatible metaphors can be treated as specialization polysemy while incompatible metaphors can be categorized as homonyms.

2.1.3 Contrastive Polysemy

The senses of a contrastive polysemous term have different etymological origins and they are not related. These senses are also said to be *homographs*. For example, the origin of sense 1 of the term bank is Italian, while the second sense is Norwegian.

1. depository financial institution, **bank**,: a financial institution.

2. **bank**: sloping land (especially the slope beside a body of water).

Although, there is no relation between the senses of contrastive terms, it is possible to find cases with related senses. For example, the two senses of the term animator can be considered homonyms. The link between the senses can be ignored in some cases as in the following example:

1. **Pascal**, Pa: a unit of pressure equal to one newton per square meter.
2. **Pascal**: a programing language designed to teach programming.

Although both senses share the same term that refers to the famous French mathematician **Pascal**, they are in fact homonyms since they belong to two different categories: unit of measurement and programming language, respectively.

3 Structural Patterns

In defining regular structural patterns, our approach relies on Apresjan's definition of regular polysemy: *"A polysemous Term T is considered to be regular if there exists at least another polysemous T' that is semantically distinguished in the same way as T"* [8].

In the following, we describe type compatible structural patterns, and how we use these patterns to discover specialization polysemy terms.

3.1 Types of Structures

Structural patterns in WordNet are found at three levels of the ontological structure of WordNet. In general, the patterns at the upper level ontology correspond to metonymy and incompatible metaphoric cases. The patterns at the middle level and lower level correspond to specialization polysemy and compatible metaphoric cases. Homonyms do not follow any pattern and can be found at any level of the ontological structure of WordNet. Accordingly, we consider homonyms found in specialization polysemy patterns as false positives. In the following, we

define a subset of the of the structural patterns in wordNet, namely the type compatible patterns, where we consider the type within its subtypes as a pattern to capture type compatible patterns.

Definition 1: Type Compatible Pattern Let T be a polysemous term that has n meanings, $n > 1$. Let S be the set of the synsets of T. Let R be a subset of S. Let Q an ordered sequence of R, where $|R| = m$, $2 \le m \le n$, and $Q = < s_1,..,s_m >, s_i \in R, s_i \ne s_j$, for $i \ne j$. A pattern *ptrn* of T is defined as $p\# < p_1,..,p_m >$, such that each p_i is a direct hyponym of p and subsumes $s_i, 1 \le i \le m$. We call p the type (the category) of the pattern and p_i the subtypes of the pattern. For example, **vascular plant** is the type of the pattern *vascular plant#<herbaceous plan, bulbous plant>* that has the subtypes **herbaceous plant** and **bulbous plant**.

The previous definition is suitable for capturing type compatible patterns in the upper and middle level ontology of WordNet. However, this definition is not suitable to capture patterns at the lower level ontology since polysemous terms at the lower level ontology correspond usually to the cases, where the senses of each polysemous term share a common parent. To be able to capture the structural regularity at the lower level ontology, we define the common parent class:

Definition 2: Common parent class Let T be a term that has n meanings, $n > 1$. Let S be the set of the senses of T. T belongs to the common parent class if the following occurs:

$$\exists R(R \subseteq S \wedge |R| > 1 \wedge \forall s(s \in R \Rightarrow \exists p\,(hypernym(s, p) \wedge \neg\exists s'(s' \in R \wedge \neg hypernym(s', p))))$$

In Figure 1, the sense of croaker is a hypernym of the two senses of white croaker and is therefore an example of common parent class.

Not all polysemous terms at the lower level ontology share the same parent. There are cases, where the direct parent of one synset is an indirect parent of the other. In some cases, the distance between the indirect synset and the common root is two. We consider these terms as members of the common parent class.

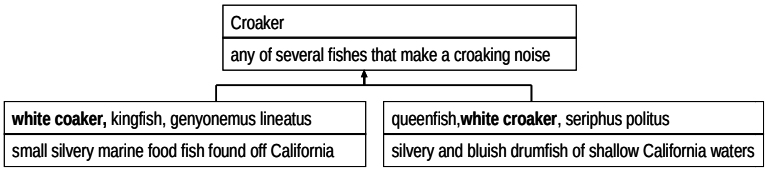


Fig. 1: An example for common parent class

Definition 3: Regular Type Compatible Pattern Let T be a polysemous term that has *n* meanings. Let S be the set of the synsets of T, Let *ptrn* be a pattern of *T*. *T* is considered to belong to a regular patterns if the following occurs: There exists at least another Term T' such that T and T' are not synonyms and *T'* belongs to *ptrn* or $\exists Q(Q \subseteq S \wedge Q \in common\ parent)$.

The pattern *vascular plant#<herbaceous plan, bulbous plant>* is regular since there are 6 terms that belong to it. In addition to regular patterns we are also interested in sub patterns. Our hypothesis is that the sub patterns of a specialization polysemy pattern belong also to specialization polysemy.

Definition 4: Sub pattern For a regular pattern $ptrn = p\#< p_1,..,p_m >$, A *ptrn'* is considered to be a sub pattern of *ptrn* if $ptrn' = p\#< p_1',..,p'_k >$ and $\exists p_i, p'_j (p_i = p'_j), 1 \le i \le m, 1 \le j \le k$.

For example, the regular pattern *vascular plant#<herbaceous plant, bulbous plant>* has the following sub pattern: *vascular plant#<bulbous plant, hydrophytic plant>*.

3.2 Discovering Specialization Polysemous Terms via Structural Patterns

The basic idea of our solution is to find all terms in WordNet, where the senses of these terms fulfill the type compatibility criterion since this criterion is the main characteristic of all specialization polysemy terms. At the lower level ontology, the terms that belong to the common parent class automatically fulfill this criterion. The patterns at the top level ontology including CORELEX patterns do not fulfill this criterion. In the middle level ontology, we have patterns that corre-

spond to specialization polysemy and other patterns that correspond to compatible metaphoric terms. Both polysemy types fulfill the type compatibility criterion. Thus our task is to classify those patterns into specialization polysemy patterns and metaphoric patterns. Notice here that there are patterns that include both polysemy types. These patterns require further step to identify the specialization polysemy terms. Our approach works in four phases as follows:

A. Patterns Identification
B. Patterns Classification
C. Polysemy Type Assignment
D. Validation

The first and the third phases are automatic, while the second and fourth are manual. In the following we discuss the four phases of our approach that we applied on the nouns that have exactly two senses.

A. Patterns Identification

In this phase, we used the following algorithm to identify the regular type compatible patterns.

Algorithm: Regular Type Compatible Patterns Extraction
Input:
PNOUNS = Polysemous nouns in WordNet
UNIQUEBEGINNERS = list of the unique beginners in WordNet
SENSENUMBER = the number of the term synsets,
Output:
N = an associative array to store the regular patterns.
M = an associative array to store the sub patterns
P = a list to store the elements of the common parent class
O = a list of singleton patterns
1. *poly_nouns* = retrieve_polysemous_nouns(*SENSENUMBER*)
2. For each *noun* in *poly_nouns*
3. *S* = retrieve_synsets(*noun*)
4. *ptrns* = get_patterns(S)
5. For each $Q \subseteq S$
6. If $Q \in$ *Common Parent*
7. add <*noun*, *Q*> to P.
8. For each pattern *ptrn* = $p\#<p_1,..,p_m>$ in *ptrns*
9. If $p \notin$ *UNIQUEBEGINNERS*

10. Add *noun* to the list under *ptrn* in N.
11. For each *ptrn* in N
12. If |N[*ptrn*]| > 1
13. M[ptrn] = sub_patterns(*ptrn*)
14. Remove sub_patterns(*ptrn*) from N
15. For each *ptrn* in N
16. If |N[*ptrn*]| < 2
17. Add *ptrn* to O
18. Remove *ptrn* from N
19. return <N,M,P,O>

The presented algorithm works in three phases:

1. *Patterns and common parent terms identification* (lines 1 to 10): We retrieve the list of all nouns that have the sense number given in the algorithm input. We check, whether the term belongs to the common parent class and also whether it has regular patterns. We exclude the top level ontology patterns such as *physical entity<physical object, physical process>*. Such patterns correspond usually to CORELEX patterns and they are not specialization polysemy patterns. Notice also that it is possible for terms that have more than 2 senses to have more than one pattern.
2. *Sub patterns identification* (lines 11 to 14): If more than one term belong to a pattern, thus it is a regular pattern, then we search all singleton patterns to identify possible sub patterns of that pattern. Identified sub patterns are removed from the patterns list and added to the sub patterns list.
3. *Singleton patterns identification* (lines 15 to 18): After identifying the sub patterns, the remaining singleton patterns are removed from the patterns list and added to the list of the singleton patterns.

The results of applying the algorithm on the terms that have two synsets are as follows: the total number of the nouns in WordNet that have two senses is 9328 nouns. 2899 nouns of them were identified by the algorithm to belong to type compatible patterns. The algorithm returned four lists: a pattern list that contains 333 patterns, a sub patterns list that contains 344 sub patterns, the list of the common parents that contains1002 terms, and a list that contains 358 singleton patterns.

B. Patterns Classification

Our task in this phase is to classify the patterns in specialization polysemy and metaphoric polysemy. First of all, the terms that belong to the common parent are considered as specialization polysemy candidates. We consider also the polysemy type of the sub patterns as the polysemy type of the pattern, they belong to. To classify the patterns, we have arranged them into hierarchies. The roots of the hierarchies are shown in the following table. The numbers right to the types correspond to the number of patterns that belong to that type.

Table 1. The roots of type compatible patterns in WordNet

Patterns under physical entity		Patterns under abstract entity	
Type	#patterns	Type	#patterns
substance	6	psychological feature	2
organism	4		
person	106	cognition	12
animal	20	attribute	26
plant	18	communication	18
artifact	73	measure	14
process	9	group	9
location	4	time period	4
thing	5	relation	3

Analyzing the patterns under these types shows that these patterns can be classified into four groups:

1. Specialization polysemy patterns
2. Metaphoric patterns
3. Homonymy patterns
4. Mixed patterns

In the following, we explain our criteria by classifying the patterns.

1. *Specialization Polysemy patterns*: the type of some specialization polysemy patterns can be determined directly by considering the type of the pattern only. For example, it is clear that the patterns whose type belongs to *animal*, and the types under *animal* are specialization polysemy or at least it is not common at all to find a metaphoric link between the types under *animal*. The criteria for determining other

specialization polysemy patterns is the *consistency* of the pattern subtypes.

2. *Metaphoric patterns*: to determine metaphoric patterns, we followed the idea that metaphors are human centric in the sense that we use metaphors to express our feelings, judgments, situations, irony and so on. For example, when we use sponger to refer to some one, we are making a judgment upon that person. This gives us a hint, where to search for metaphoric patterns, namely under the person type or the types whose subtypes indicate meaning transfer from their literal meaning to a (metaphoric) human centric meaning as discussed below. Here, the type attribute is an example of such cases.

a. Metaphoric patterns under person: we found under the type person 106 patterns. Some of these patterns are specialization polysemy patterns and others are metaphoric. To determine metaphoric patterns under the type person, we searched for inconsistency between the subtypes of the patterns. We find such inconsistency for example in the pattern person#<bad person, worker>, the sub type bad person is not consistent with the type worker and therefore a specialization polysemy is totally excluded in this pattern. The term iceman is an example of terms that belong to this pattern:

1. iceman: someone who cuts and delivers ice.
2. hatchet man, iceman: a professional killer.

On the other hand the subtypes of the pattern person#<expert, worker> are consistent and is considered as a specialization polysemy pattern. The term technician is an example for this pattern:

1. **technician**: someone whose occupation involves training in a technical process.
2. **technician**: someone known for high skill in some intellectual or artistic technique.

b. Metaphoric patterns under attribute: Our criteria here was to find meaning transfer between the sub types. Attribute has the following four patterns: attribute#<property, trait>, attribute#<property, state>, attribute#<property, quality>, and attribute#<quality, trait>, with the following meanings:

Property: a basic or essential attribute shared by all members of a class.
State: a state of depression or agitation.
Quality: an essential and distinguishing attribute of something or someone.
Trait: a distinguishing feature of your personal nature.

The meaning transfer from property to human centric meaning is clear in the first three patterns. For example, in the term *chilliness*:

1. **chilliness**, coolness, nip: the property of being moderately cold.
2. coldness, frigidness, iciness, **chilliness**: a lack of affection or enthusiasm.

In the fourth pattern, the relation between quality and trait depends on whether the term under the quality subtype refers to *an attribute of something* or *an attribute of someone*. The first case corresponds to metaphoric polysemy while the second corresponds to specialization polysemy.

3. *Homonymy Patterns:* In general, homonymy can not be considered as a type of regular polysemy. Nevertheless, we cannot exclude the existence of homonymy patterns. WordNet contains few homonymy patterns such as the following pattern: *organism#<animal, plant>*, where we find type mismatch between the subtypes. Specialization or metaphoric polysemy in such patterns is totally excluded.

4. *Mixed patterns*: This group contains the patterns that were identified to have more than one polysemy type. For example, the pattern *attribute#<quality, trait>* belongs to this group.

In summary: there are some patterns whose sub types indicate type inconsistency. After excluding these patterns, all patterns under the *physical entity* are candidates for specialization polysemy except the patterns under *person* which contains both polysemy types. In the case of *abstract entity*, most of the patterns under attribute are candidates for metaphoric polysemy. The patterns under *cognition* and *communication* contain both polysemy types, and the rest types are candidates for specialization polysemy.

C. Polysemy Type Assignment

In this phase, each of the nouns, that were determined to belong to type compatible polysemy patterns, is assigned to specialization polysemy or metaphoric according to the pattern of the term. The terms that belong to both polysemy types and the terms that belong to the singleton patterns are not assigned and they are subject to manual treatment in the validation phase.

D. Validation

In this phase, we manually validate the assigned polysemy type. Our criterion is to determine the relation between the senses of a term and thus the polysemy type, is the synset gloss. In difficult cases, we also consider the hierarchical properties of the term synsets. We have three tasks in this phase:

1. *Validation of the assigned polysemy types*: we check whether each of the nouns belong to its assigned polysemy type.
2. *Assigning the polysemy type*: for the terms that belong to the mixed patterns and singleton patterns.
3. *Excluding of false positives*: we exclude the false positives from the terms of the 4 groups.

Our judgments during the validation process are based on knowledge organization in such a way that word etymology and linguistic relatedness have secondary role in our judgments. The primary criterion is:

1. In case of specialization polysemy: Is it possible for both senses to be generalized to a common type? If the answer is no or we don't know, then we consider the term to be a homonymy case. The term cardholder is an example for such cases:

    1. **cardholder**: a person who holds a credit card or debit card.
    2. **cardholder**: a player who holds a card or cards in a card game.

2. In case of metaphoric polysemy: Is it easy to discover the metaphoric link between the senses? If the answer is no or we don't know, then we consider the term to be specialization polysemy candidate. The term agreeableness that belongs to the metaphoric pattern attribute#<quality, trait> is an example for such cases:

1. **agreeableness**, amenity: pleasantness resulting from agreeable conditions.
2. **agreeableness**, agreeability: a temperamental disposition to be agreeable.

4 Synset patterns

The structural patterns served as a criterion for identifying specialization polysemy candidates. The next step is how to solve the polysemy problem for the identified candidates. The *more general meaning/more specific meaning* relation between the senses of the specialization polysemy terms reflects a hierarchical relation between the senses. Thus, the solution should reflect this relatedness. In the following, we explain how the synonyms of the specialization polysemy synsets are used to organize the hierarchical relation between the senses.

4.1 Types of Synsets

In our approach, we have analyzed the relation between the synset synonyms and the possible relation between the synsets of specialization polysemy cases. The idea here is that the nature of the relation between the synsets of specialization polysemy terms can be determined based on the synonyms of such terms. Based on the synset synonyms , we divided the specialization polysemy terms in three groups:

1. Twin synsets
2. Type – sub type synsets
3. General meaning – example meaning synsets

1. *Twin synsets*: both synsets of such terms contain other synonyms beside the polysemous terms. Analyzing these cases shows that the *is a* relation does not hold between the synsets themselves. In fact both synsets are more specific in meanings of some (non existing) third synset as in the following example:

1. **white croaker**, queenfish, Seriphus politus: silvery and bluish fish of California.
2. **white croaker**, kingfish, Genyonemus lineatus: silvery fish of California.

2. *Type - sub type synsets*: One synset contains the polysemous terms only, the other contains the polysemous terms and other synonyms. Analyzing these cases shows that the gloss of the synset that contains the polysemous terms only usually begins with the following phrase: "*any of several "* which reflects that this synset encodes a more general meaning while the synset with additional synonyms describes a specific type that belong to the type of both synsets. For example, sense 1 describes a specific type, while sense 2 is a general description of turtledove.

1. Australian turtledove, **turtledove**, Stictopelia cuneata: small Australian dove.
2. **turtledove**: any of several Old World wild doves.

3. *General meaning - example meaning synsets*: both synsets contain the polysemous terms only. Analyzing these cases shows that there is a synset which denotes the meaning of the term in general while the other synset denotes an example of that general meaning. According to our analysis, the synset with the general meaning has usually sense rank 1. For example sense 1 denotes the general meaning of the term timetable while sense 2 is an example of the term. Notice that, there are many other examples of timetables such as *schedule of lessons in the school*. We think that sense 2 is an example for unnecessary sense enumeration in WordNet and we consider the senses as candidates to be merged.

1. **timetable**: a schedule listing events and the times at which they will take place.
2. **timetable**: a schedule of times of arrivals and departures.

4.2 Organizing Specialization Polysemous Terms via Synset Patterns

According to the above analysis, we suggest to solve the specialization polysemy by reorganizing the ontological structure of the synsets, where the implicit hierarchical relation between the synsets at lexical level is transformed into explicit hierarchical relation at semantic level. This requires adding missing synsets, is a relations and removing redundant is a relations.

1. *Solution for Twins synsets*: We add a new (missing) parent in cases, where the polysemous meanings of a term T can be seen more specific

meanings of an absent more general meaning: Let s_1, s_2 be two synsets of a term belonging to the missing parent cases. Let $\{T_1,..,T_n\}$ be the set intersection of s_1 and s_2. Let T' be the preferred term in s_1 and s_2 or the term with the highest rank in both synsets. Let T be the preferred term of the type of s_1 and s_2. We create a common parent S_p of s_1 and s_2 as follows:

i) Create a new synset S_p such that:

The lemmas are the intersection of the lemmas of s_1 and s_2;
The gloss of S_p = T' is a T.

ii) Remove the common lemmas from s_1 and s_2
iii) Connect S_p to T via the *is-a* relation
iv) Connect the senses s1 and s2 to S via the *is-a* relation
v) Remove redundant relations

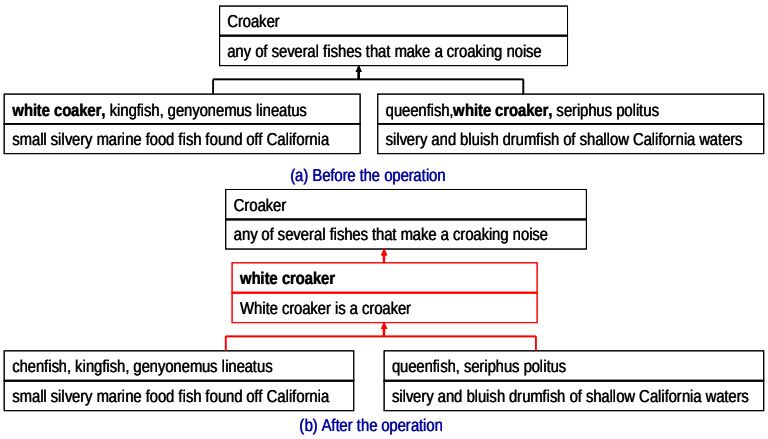


Fig. 2. Example for adding a new missing parent

2. *Solution for type – sub type synsets*: In such cases we establish a missing *is_a* relation to denote that a sense of a polysemous term T is more specific than another more general meaning of T: Let s_1, s_2 be two synsets of a term belonging to the missing relation cases. Let s_2 be the synset that has the polysemous terms and additional terms. Let s_1 be the synset that contains the polysemous terms only.

i) Connect s1 to s2 via the *is-a* relation: s_2 *is-a* s_1.
ii) Remove redundant relations.

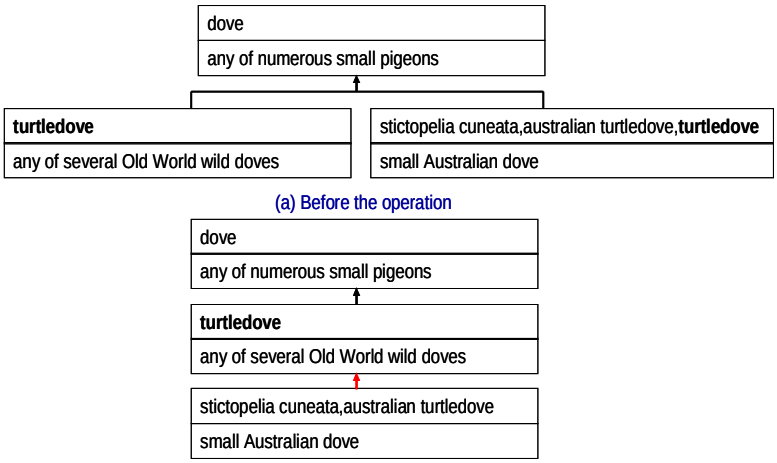


Fig. 3. Example for adding missing relation

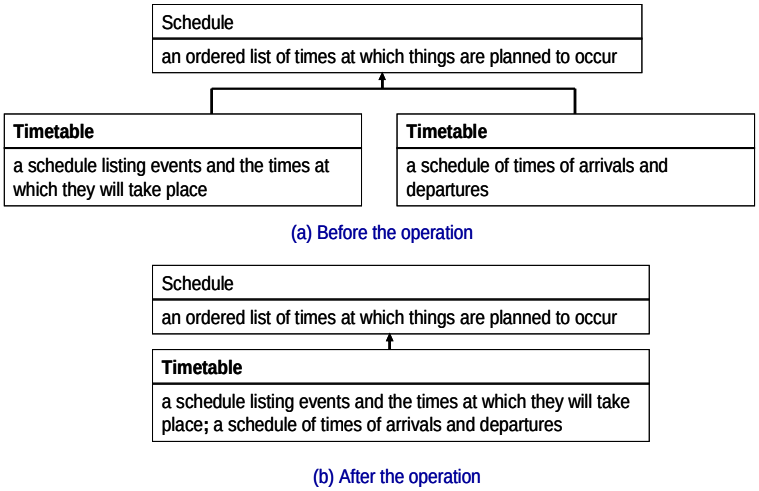


Fig. 4. Example for merge operation

3. *Solution for general - example synsets*: In such cases, we merge the senses of the terms as follows**.**

Let s_1, s_2 be two synsets of a term belonging to the merge cases. We keep the synset with sense rank 1 and remove the other one as follows:

i) The lemmas of s_1 are the same as before since both synsets share the same lemmas.
ii) The gloss of s_1 = the gloss of s_1; the gloss of s_2
iii) The relations of s_1 are the union of the relations of both synsets
iv) Remove redundant relations

5 Results and Evaluation

In the following, we describe the results and the evaluation of our approach. Table 2 presents the results of the four pattern groups and common parent group after the validation.

Table 2. Validated results of the approach

Patterns group	#total cases	#Specialization Polysemy	#Metaphors	#Homonyms
Spec. Polysemy Patterns	807	673	26	108
Metaphoric Patterns	221	28	170	23
Homonyms Patterns	56	0	0	56
Mixed Patterns	111	41	39	31
Common Parent	1002	927	40	35
Sub patterns and singleton patterns	702	455	90	157
Total nouns	2899	2124	365	410

In Table 3, we present the pattern groups that have been identified.

Table 3. Distribution of type compatible patterns

#Patterns	#Spec. Polysemy Patterns	#Metaphoric Patterns	#Homonym Patterns	#Mixed Patterns
333	225	79	15	14

As we can see in Table 2, 73% of the identified terms belong to specialization polysemy. In table 3, we find that 67.5% of the identified patterns are specialization polysemy patterns. In Table 2, we can also see that not all terms that belong to the common parent group are specialization polysemy terms. About 4% of these terms are in fact homo-

graphs. Consider for instance, term *apprehender* that belongs to the common parent group.

1. knower, **apprehender**: a person who knows or apprehends.
2. **apprehender**: a person who seizes or arrests.

Although both senses belong to the type person, they are in fact homographs. Also, about 3.5% of the common parent group were identified as metaphors. Consider for example, the following senses of the term *moment of truth*.

1. **moment of truth**: the moment in a bullfight when the matador kills the bull.
2. **moment of truth**: a crucial moment on which much depends.

We have examined CORELEX patterns to find overlap between CORELEX patterns and the patterns identified in our approach. We did not find any overlap between them. This was expected, since CORELEX patterns belong to the top level ontology, where as the specialization polysemy patterns were found at the middle and lower level ontology. An important thing to note here is that none of the terms that belong to CORELEX were identified as specialization polysemy terms. They belong mainly to metonymy.

In Table 4, we list the distribution of specialization polysemy operations.

Table 4. Specialization polysemy operations

Operation	Adding missing parent	Adding missing relation	Merge	Total
#cases	1045	685	409	2124

The total number of reduced polysemous words is 2124 words. The total number of merged synsets represents about 14% of the total processed cases. We have added 1045 new synsets and 1730 new relations, while deleted 409 synsets and 409 relations. Compared to polysemy reduction approaches, 86% of the cases were not merged. Instead of merging, we have reorganized the ontological structure of the terms. It is important here to notice that our approach improves the ontological structure of WordNet by increasing knowledge rather than decreasing knowledge as it is suggested by other approaches.

To evaluate our approach, 834 cases have been evaluated by two evaluators. In Table 5, we report the evaluation statistics, where the column polysemy type refers to homonymy, metaphoric, or specialization polysemy and polysemy operation refers to creating missing parent, adding missing relation, or merging operation. Note that, polysemy operation is applicable in case of specialization polysemy. The table presents the agreement between the evaluators and our approach. The third row represents the number of cases, where at least one evaluator agrees with our approach.

Table 5. Evaluation results

	Polysemy type agreement	Polysemy operation agreement
Evaluator 1	803 ≈ 96.2%	750 ≈ 89.9%
Evaluator 2	775 ≈ 92.9%	686 ≈ 82.2%
Partial agreement	824 ≈ 98.8%	796 ≈ 95.4%

6 Conclusion

In this paper, we have introduced an approach for solving the specialization polysemy problem based on type compatible regular patterns. This approach decreases polysemy, but at the same time knowledge is increased. It improves the ontological structure of WordNet, where the implicit relations between the synsets of polysemous terms which is encoded at lexical level are transformed into explicit semantic relations.

In the current paper, we presented the result of our approach applied on nouns that have two senses. Our future plan is to apply the approach on all other nouns in WordNet as a first step towards solving the other polysemy types.

References

1. Miller, G.A.: WordNet: a lexical database for English. In: Communications of the ACM, 38 (11), pp. 39–41 (1995)
2. Miller, G. A., Beckwith, R., Fellbaum, Ch., Gross, D. and Miller, K.: Introduction to WordNet: An on-line lexical database. In: International Journal of Lexicography (1990)

3. Mihalcea, R., Moldovan, D.I.: EZ.WordNet: Principles for Automatic Generation of a Coarse Grained WordNet. In: FLAIRS, pp. 454–458 (2001)
4. Buitelaar, P. P.: CORELEX: Systematic Polysemy and Underspecification. In: PhD thesis, Brandeis University, Department of Computer Science (1998)
5. Bouquet, P., Giunchiglia, F.: Reasoning About Theory Adequacy. A New Solution to the Qualification Problem. In: Don Perlis (editor): Fundamenta Informaticae, vol. 2/3/4, pp. 247-262 (1995)
6. Barque, L., Chaumartin, F.R.: Regular Polysemy in WordNet. In: JLCL 24(2), pp. 5–18 (2009)
7. Giunchiglia, F., Yatskevich, M., Avesani, P., Shvaiko, P.: A Large Scale Dataset for the Evaluation of Ontology Matching Systems. In: The Knowledge Engineering Review Journal (KER), Cambridge University Press, 24 (2) (2009)
8. Apresjan, J.: Regular Polysemy. In: Linguistics, 142, pp. 5–32 (1974).
9. Peters, W., Peters, I.: Lexicalized systematic polysemy in WordNet, Language Resources and Evaluation (2000)
10. Snow, R., Prakash, S., Jurafsky, D., Ng, A.Y.: Learning to Merge Word Senses. EMNLP-CoNLL 2007: 1005–1014 (2007)
11. Navigli, R.: Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. ACL (2006)
12. Freihat, A.A., Giunchiglia, F., Dutta, B.: Approaching Regular Polysemy in WordNet. In: proceedings of 5th International Conference on Information, Process, and Knowledge Management (eKNOW), Nice, France (2013)
13. Giunchiglia, F., Kharkevich, I., Zaihrayeu, I.: Concept search. In Processing of ESWC'09, Lecture Notes in Computer Science. Springer (2009)
14. Peters, W., Peters, I.: Lexicalised Systematic Polysemy in WordNet, Department of Computer Science,University of Sheffield, U.K.
15. Giunchiglia, F., Dutta, B., Maltese, V.: Faceted lightweight ontologies. In: Conceptual Modeling: Foundations and Applications, Alex Borgida, Vinay Chaudhri, Paolo Giorgini, Eric Yu (Eds.) LNCS 5600 Springer (2009)
16. Pustejovsky, J.: The Generative Lexicon, MIT Press, Cambridge, MA (1995)
17. Navigli, R.: Word sense disambiguation: a survey, ACM Computing Surveys 41(2):1–69 (2009)
18. Nerlich, B., Clarke, D.D.: Polysemy and flexibility: introduction and overview. Berlin, New York: Mouton de Gruyter, 3–29 (2003)
19. Zaihrayeu, I., Su, L., Giunchiglia, F., Pan, W., Ju, Q., Chi, M., Huang, X.: From Web Directories to Ontologies: Natural Language Processing Challenger, 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, ISWC + ASWC 2007, Busan, Korea. (2007)

Abed Alhakim Freihat
Dept. of Information Engineering and Computer Science,
University of Trento,
Trento, Italy
E-mail: <fraihat@disi.unitn.it>

Fausto Giunchiglia
Dept. of Information Engineering and Computer Science,
University of Trento,
Trento, Italy
E-mail: <fausto@disi.unitn.it>

Biswanath Dutta
Documentation Research and Training Centre (DRTC),
Indian Statistical Institute (ISI),
Bangalore, India
E-mail: <bisu@drtc.isibang.ac.in>

Grammar, Semantics, Dialog

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 55–77
RECEIVED 26/11/12 ACCEPTED 11/01/13 FINAL 29/03/13

Structural Underspecification and Resolution within a Processing-oriented Grammar Formalism

TOHRU SERAKU

University of Oxford, UK

ABSTRACT

A challenge to modeling incrementality in language processing is posed by complex NPs in some verb-final languages, where a parser does not see whether a clause that a parser currently processes is part of a complex NP and how deeply it is embedded. These indeterminacies are handled by structural underspecification and resolution within Dynamic Syntax. This article points out that the previous implementation of the mechanism faces a formal problem of introducing indistinguishable nodes into the tree, and proposes a solution by letting a parser determine node-addresses flexibly. Concrete analyses are given to Japanese relatives as a case of complex NPs in verb-final languages.

KEYWORDS**:** *Dynamic Syntax, incrementality, Japanese, relative clauses*

1 Introduction

A central issue in recent processing studies is whether the incremental parsing thesis holds of verb-final languages. Despite initial negative suggestions [14], there has been a growing body of research pointing to a conclusion in which the answer is positive [6]. From a parser's point of view, particularly challenging are complex NPs (e.g. NP with a relative clause, NP with an appositive clause) in some verb-final

languages such as Japanese and Korean: a complex NP in these languages consists of a clause ending with a verb and a head noun following the clause. So, in processing a clause, a parser does not see in advance (a) whether the current clause is a main clause or part of a complex NP and, if it is part of a complex NP, (b) how deeply it is embedded.

These two indeterminacies are illustrated by the Japanese strings (1, 2, 3). First, as shown in (1), argument NPs in Japanese may be dropped when they are identifiable contextually. The parentheses in (1) indicate that *Mary-ga* and *hon-o* may be dropped.

(1) (*Mary-ga*) (*hon-o*) *ka-tta.*
(Mary-NOM) (book-ACC) buy-PAST
'Mary bought a book.'

In Japanese, a relative clause precedes a head noun. Thus, the relative clause *Mary-ga ka-tta* in (2) is identical to the string (1) if *hon-o* is dropped in (1).

(2) [[*Mary-ga* *ka-tta*] *hon*]-*wa* *omoshiroi.*
[[Mary-NOM buy-PAST] book]-TOP interesting
'A book which Mary bought is interesting.'

Note that the string (2) contains no morpheme that marks a relative clause.[1] Thus, a parser, which processes *Mary*, cannot see whether *Mary* belongs to a relative clause as in (2) or a matrix clause as in (1). Further, as demonstrated in (3), a parser, which has processed the complex NP string *Nai-ta otoko*, is still unable to see whether this complex NP belongs to a matrix clause or, as in (3), it is part of a larger complex NP.

(3) [[*Nai-ta* *otoko*]-*o* *nagusame-ta* *hito*]-*ga* *nige-ta*
[[cry-PAST man]-ACC comfort-PAST person]-NOM run.away-PAST
'A person who comforted a man who cried ran away.'

[1] It is reported that a verb in a relative clause in Japanese has a special intonation [13]. This intonational cue, however, is not available until a parser processes the verb *kau* (= 'buy') in (2). In Korean, the verbal suffix -*u*(*n*) indicates a relative clause [17], but, once again, this morphological cue is not available until a parser processes a verb.

An appropriate parser for Japanese must be flexible enough to accommodate these two indeterminacies.

A reasonable method of handling such indeterminacies is to introduce structural indeterminacies to trees. This idea is implemented within Dynamic Syntax (DS) [2, 8, 10] as structural underspecification and resolution. This is intuitively plausible, but, as will be pointed out, the previous analysis [2, 9, 13] ends up inducing indistinguishable nodes into the tree. This constitutes a rather serious problem because it overturns a principal basis for explaining diverse linguistic data (Greek clitics [3], Japanese clefts [16]) and it prevents the DS modeling of English dialogue [15] from being applied to Japanese dialogue. In short, complex NPs in verb-final languages such as Japanese offer a good test case for evaluating the DS formalism.

The aim of this article is to point out a formal problem that the extant DS treatment of complex NPs suffers from and to propose a solution by letting a parser determine node-addresses flexibly. The refined DS parser, it is argued, provides a more realistic model of language understanding in that a "look ahead" mechanism may be avoided and that intonational cues are more effectively utilized. To illustrate this point, the article examines Japanese relatives as a case of complex NPs in verb-final languages.

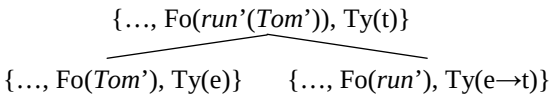
2 Dynamic Syntax

Dynamic Syntax (DS) is a grammar formalism that models knowledge of language; thus, DS is a theory of competence and regarded as generative grammar in the sense explicated by Noam Chomsky [4]. Unlike mainstream generative grammar, however, knowledge of language, or competence, is defined as a set of constraints on language performance, more specifically, the building-up of interpretation in context [2, 8, 10]. With such constraints, a parser processes a string of words left-to-right, and builds up semantic representation incrementally, without a separate level of syntactic structure: "syntax" within DS is no more than a set of constraints on how to build up a semantic tree progressively in context.

2.1 *Trees and Tree Descriptions*

The aim of a parser is to construct a semantic tree that represents an interpretation of a string in context on the basis of word-by-word processing. Trees in DS are binary, an argument node being on the left and a functor node being on the right. Each node is decorated with a declarative unit, consisting of a formula and labels.[2] A formula is semantic content at a node, and labels indicate various properties of the content; one example of labels is a logical type, which indicates the combinatorial property of the content. A formula is represented with the predicate Fo, whose argument comes from D_{Fo} = {*Tom*', *run*', …}. Content of some lexical items is not an element in D_{Fo}; for instance, the content of *she* is a place-holding variable U, called "meta-variable", whose value is supplied contextually. A logical type is represented with the predicate Ty, whose argument comes from D_{Ty} = {e, t, e→t, …}. D_{Ty} is a finite set (for instance, it does not include a type for five-place predicates), and no operations are stipulated to generate types, such as type-lifting and composition of functors. For example, the parse of *Tom runs* gives rise to the semantic tree (4); for the sake of simplicity, tense is ignored throughout this article.

(4)



The notation "…" in each declarative unit indicates additional labels which are not explicitly shown here. Another example of labels is a decoration in LOFT (Logic Of Finite Trees [1]). This is a language to talk about trees, which enables a parser to describe the other nodes in the tree from the perspective of a current node. LOFT-operators are defined as follows. First, there are operators to model an immediate dominance relation: $\langle\downarrow_0\rangle$ is for argument daughters and $\langle\downarrow_1\rangle$ for functor daughters. For instance, $\langle\downarrow_0\rangle$Ty(e) indicates that the argument daughter is of type-e; this label holds at the top node in (4). The inverses, $\langle\uparrow_0\rangle$ and $\langle\uparrow_1\rangle$, describe a mother node from the perspective of an argument node and from the perspective of a functor node, respectively. Second, operators with the Kleene star * model a dominance relation. $\langle\downarrow_*\rangle$ describes a node somewhere below the current node, together with its inverse, $\langle\uparrow_*\rangle$. These operators may describe a node at an arbitrary distance, but not across a "LINK" relation. Third, the "down" operator <D> and the "up" operator <U> model the weakest relation and may describe a node across a "LINK" relation. Finally, <L> and its inverse $\langle L^{-1}\rangle$ describe a node within

[2] Formally, DS structure is represented by a set of declarative units, where their relations are governed by LOFT (Logic Of Finite Tree) [1].

another structure that is LINKed from/to a current node. (For LINK relations, see Section 2.4.)

Another type of label is a node identifier, Tn(a), where Tn is a tree-node predicate. If a node is annotated with Tn(a), Tn(a0) indicates its argument daughter, and Tn(a1) indicates its functor daughter. A root node is marked by Tn(0), its argument daughter being by Tn(00) and its functor daughter being by Tn(01). Thus, the declarative unit at the root node in (4) is more precisely as in (5).

(5) {…, Tn(0), $<\downarrow_0>$Tn(00), $<\downarrow_1>$Tn(01), Fo(*run*'(*Tom*')), Ty(t), ◇}

This declarative unit contains a pointer ◇. In a DS tree, there always exists a single node that is under development. Such an active node is marked by a pointer ◇.

In non-final states, a tree is a "partial" structure in the sense that there exists a node decorated with a set of "requirements". A tree is said to be well-formed iff there are no outstanding requirements, and a string is said to be grammatical iff there exists a tree update that leads to a well-formed tree. A requirement is notated as the label ?α at a node, which requires that α will hold at the node. For instance, ?Ty(e) requires that the node will be decorated with Ty(e). Every node is introduced with requirements and every single tree-update is driven by some form of requirements. A parser runs a set of actions in order to satisfy requirements, as we shall see in the next sub-section.

2.2 *Actions for Tree Updates*

Trees grow progressively on the basis of left-to-right processing of a string in context without postulating an independent level of syntactic structure. The starting point of tree update is determined by the AXIOM, which introduces an initial node with the following declarative unit:

(6) {?Ty(t), ◇}

?Ty(t) requires that this node will be of type-t. This requirement corresponds to the parser's goal to build up an interpretation of a string: in this sense, tree growth is goal-directed. As a string is processed word-by-word, the initial node becomes increasingly richer: it is updated gradually and monotonically by a combination of general, lexical, and pragmatic actions.[3]

[3] In earlier works [10], the initial node is also annotated with Tn(a), an arbitrary node-address. Tn(a) is not articulated in recent works [2, 8], the

First, general actions are a set of actions that are stored in the DS system and that are not lexicalized. Each general action is formulated as a program, or a sequence of instructions to update a tree. Instructions are in the conditional format (7).

(7)
	IF	… ("…" is a condition to be met by a node highlighted by ◇)
	THEN	… ("…" is an action to be run if the condition is met)
	ELSE	… ("…" is an action to be run if the condition is not met)

The application of general actions is optional: a parser may run general actions at any time as long as the IF block is met by an active node. Examples of general actions will be presented in the next sub-section.

Second, lexical actions are a set of actions that are stored in the DS system and that are lexicalized. Lexical items also encode a sequence of instructions to update a tree, but lexical actions differ from general actions in terms of optionality: a package of actions encoded in a lexical item α must be run every time α is parsed. For instance, *inu* (= 'dog') encodes the macro of actions (8), where `put`(α) is a primitive action to decorate a node with α.

(8)
	IF	?Ty(e)
	THEN	`put`(Fo(ε, x, *inu*'(x)), Ty(e))
	ELSE	ABORT

Thus, (8) declares that if a current node is decorated with ?Ty(e), a parser annotates the node with Fo(ε, x, *inu*'(x)) and Ty(e). ABORT in the ELSE block ensures that this action cannot be executed unless the IF block is met. In (8), (ε, x, *inu*'(x)) is a type-e term that denotes a dog, expressed in Epsilon Calculus.[4] As shown in (1), argument NPs in

---

assumption being that the node introduced by the AXIOM is a root node of the whole tree. In Section 4, I shall modify the AXIOM so that it introduces a node that is underspecified for a node-address.

[4] Epsilon Calculus is a formal study of arbitrary names in natural deduction in Predicate Logic, proposed by David Hilbert. Every quantified NP is mapped onto an epsilon term, a type-e term defined as a triple: an operator, a variable, and a restrictor. In the case of (ε, x, *inu*'(x)), the existential operator ε binds the variable x that is restricted by the predicate *inu*'. This term stands for an arbitrary witness of the Predicate Logic formula ∃x.*inu*'(x). Since quantified NPs are uniformly analyzed as type-e terms, a quantified NP at an object position is handled without assuming type-shifting or quantifier movement [5]. A scope relation is expressed in a scope statement, where each term is in a dependency relation to others. This statement is constructed gradually as quantified NPs are parsed. Once a complete statement arises,

Japanese may be dropped. Thus, verbs encode a macro of actions to build up a propositional skeleton with argument slots. If NPs are dropped, such slots are contextually assigned content; if NPs have been processed, such slots collapse with the nodes that have been created by the parse of these NPs (cf. Section 3).

Third, pragmatic actions are a set of actions whose schematic rule-structures are stored in the DS system but whose execution involves pragmatic inference. A case of pragmatic actions pertinent to the present article is SUBSTITUTION, which saturates a meta-variable. For instance, the parse of *he* puts a meta-variable Fo(U_{MALE}) at a node, with a requirement that the node will be annotated with a formula denoting a male. This requirement drives SUBSTITUTION, replacing the variable with a content denoting a male with reference to contextual factors. SUBSTITUTION resolves underspecification in content. This is a quite familiar process in linguistics, but DS assumes another, less familiar form of underspecification: underspecification of structural relation.

2.3 *Structural Underspecification and Resolution*

Within DS, a node may be initially unfixed and resolved later. There are three types of general actions to induce unfixed relations with different locality restrictions:

(9)
a. LOCAL *ADJUNCTION: to induce a node that is "locally" unfixed
b. *ADJUNCTION: to induce a node that is "non-locally" unfixed
c. GENERALIZED ADJUNCTION: to induce a node that is "globally" unfixed

These general actions may be run only if a pointer ◇ is at a type-t-requiring node; so, unfixed nodes are always hung from a type-t-requiring node.

First, LOCAL *ADJUNCTION induces an unfixed node that must be fixed within a local proposition. This node is decorated with $<\uparrow_0><\uparrow_{1*}>$?Ty(t). This means that if a pointer ◇ moves up from an argument node (and possibly keeps going through functor nodes), then a parser finds a type-t-requiring node. For instance, $<\uparrow_0><\uparrow_{1*}>$?Ty(t) may be $<\uparrow_0>$?Ty(t), $<\uparrow_0><\uparrow_1>$?Ty(t), $<\uparrow_0><\uparrow_1><\uparrow_1>$?Ty(t), and so on.

every term in a proposition is "evaluated": it reflects the full scope relation into the restrictor of that term. Since this evaluation process is not pertinent, it is disregarded in this article.

Given this restricted dominance relation, the node is fixed under the closest type-t-requiring node. If a pointer crosses a type-t-requiring node, the relation includes more than one $\langle\uparrow_0\rangle$, as in $\langle\uparrow_0\rangle\langle\uparrow_1\rangle\langle\uparrow_1\rangle\langle\uparrow_0\rangle$?Ty(t), which contradicts $\langle\uparrow_0\rangle\langle\uparrow_{1*}\rangle$?Ty(t). This unfixed relation is resolved by a case particle. For instance, the lexical action encoded in the nominative-case particle *ga* puts the label $\langle\uparrow_0\rangle$?Ty(t) at an unfixed node, fixing it as a subject node under the closest type-t-requiring node.

Second, *ADJUNCTION induces an unfixed node that may be resolved at any node as long as the unfixed relation does not cross a LINK relation. Such nodes are marked by $\langle\uparrow_*\rangle$?Ty(t), which ensures that a pointer may cross a type-t-requiring node. This non-local unfixed relation cannot be resolved lexically. For instance, the accusative-case particle *o* narrows down possible fixed positions to a set of object nodes, each under some type-t-requiring node, but it does not specify a unique position. However, this unfixed relation may be resolved by the general action UNIFICATION: ?Ty(α)-unfixed node unifies with a Ty(α)-fixed node, as a result of which the fixed node is annotated with the union of the two declarative units.

Third, GENERALIZED ADJUNCTION induces a node that is wholly unfixed (i.e. may be across a LINK boundary). This globally unfixed relation is modeled by decorating the unfixed node with <U>?Ty(t), where the "up" operator <U> models a dominance relation across a LINK relation, allowing a pointer ◊ to move up and to cross a LINK boundary (cf. Section 2.1). An unfixed node induced by GENERALIZED ADJUNCTION may not be resolved by the parse of case particles for the same reason as stated in the last paragraph.

2.4 *LINK Relations*

Within DS, two structures may be built up in tandem, one of which is LINKed to the other. LINK is a relation between two structures that share a formula, and it is used for modeling, among other things, relatives in the following manner: a parser builds up an adjunct structure and LINKs the top node of the adjunct structure to a fresh node in an emergent main structure; a parser enriches this fresh node with the content of the adjunct structure. In this course of LINK transitions there are two crucial steps.

First, the general action LINK INTRODUCTION induces a LINK relation between a top node in an adjunct structure and a new type-e-requiring node in an emergent main structure. From the perspective of a node in a main structure, the top node of an adjunct structure may be described by the operator <L> (cf. Section 2.1). So, the label <L>α at a node in a main structure declares that if a parser looks at a LINKed node in an adjunct structure, the LINKed node is annotated with α.

Given the inverse operator $<L^{-1}>$, the following relation holds: $<L^{-1}>Tn(a) \Leftrightarrow Tn(aL)$.

(10) LINK INTRODUCTION

IF Ty(t), <D>(Fo(α))
THEN `make`(<L^{-1}>); `go`(<L^{-1}>); `put`(?∃x.Fo(x[α]), ?Ty(e))
ELSE ABORT

In (10), `make` and `go` are primitive actions concerning a node creation and a pointer movement, respectively. The IF block requires that a current node be of type-t and that a node somewhere below this node be decorated with Fo(α), where α is an arbitrary type-e term.[5] The THEN block requires that, if the IF block is satisfied, a parser initiate an inverse LINK relation from the current node to a fresh node in an unfolding main structure, and decorate the node with the requirements: ?∃x.Fo(x[α]) and ?Ty(e). ?∃x.Fo(x[α]) requires that this node will be decorated with a term that contains α as a sub-term; this ensures that the two LINKed structures share a term α.

Second, the fresh node in an emergent main structure is decorated by a head noun, and enriched with the content of the adjunct structure. This enrichment process is formulated as the general action LINK EVALUATION.

(11) LINK EVALUATION

IF Ty(e), Fo(ε, y, φ(y))
THEN IF <L>(Fo(ψ[(ε, x, P(x))]))
THEN `put`(Fo(ε, y, φ(y)&ψ[y/(ε, x, P(x))]))
ELSE ABORT
ELSE ABORT

(ε, y, φ(y)) is the content of a head noun, and ψ is the content of a relative clause, where (ε, x, P(x)) is the content of a gap in the relative clause. A parser reflects ψ into the term (ε, y, φ(y)) as an additional restrictor by re-binding (ε, x, P(x)) in ψ with the variable y, as in (ε, y, φ(y)&ψ[y/(ε, x, P(x))]). As a consequence, this composite term denotes an entity that satisfies not only the description of the head noun but also the description of the relative clause.

[5] In the previous work [9], the operator with the Kleene-star $\downarrow_*$ (instead of the "down" operator <D>) was used. This article presents LINK INTRODUCTION by replacing $\downarrow_*$ with <D>. This is because Japanese relatives are not sensitive to islands, as will be pointed out in Section 4.4. The next section shows that, even if this modification is made, the present version of LINK INTRODUCTION is not adequate.

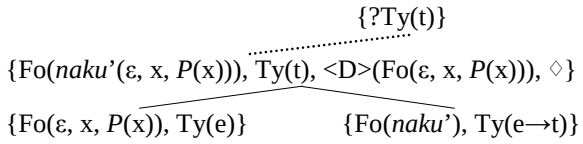
3 The Problem

Let us outline the previous DS account of Japanese relatives [2, 9, 13]. Consider (12), where the head noun *otoko* (= 'man') is preceded by the relative clause *Nai-ta*.

(12) [*Nai-ta* *otoko*]-*ga* *nige-ta.*
[cry-PAST man]-NOM run.away-PAST
'A man who cried ran away.'

In this earlier view, the AXIOM induced the initial node (6). Since *naku* (= 'cry') may belong to an embedded structure of an arbitrary depth, a parser introduced a globally unfixed type-t-requiring node by running GENERALIZED ADJUNCTION. This unfixed relation is shown by the dotted line in (13). Under this node, a parser ran the lexical actions encoded in *naku*, constructing a propositional template with a subject slot. Since no argument NPs had been parsed, a parser annotated this subject slot with the term (ε, x, *P*(x)), where *P* is an abstract predicate.[6]

(13) Parsing *Nai-ta*[7]

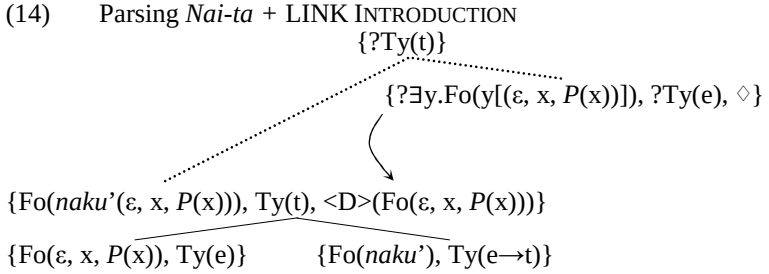


Then, in order to parse the head noun *otoko*, a parser executed LINK INTRODUCTION, initiating an inverse LINK relation from the type-t node to a new type-e-requiring node in an unfolding main structure, as shown by the curved arrow in (14). This node was also globally unfixed with respect to the root node since it might turn out to be part of a larger structure.

[6] In some previous accounts [2, 13], the node for a gap is notated as a variable. But this article follows a more recent account [9] in decorating the node with a term involving an abstract predicate P. However, this is just for expository purposes, and the analysis to be proposed in Section 4 may be recast in line with the previous accounts [2, 13].

[7] In this and subsequent trees, only relevant labels are expressed in declarative units.

(14) Parsing *Nai-ta* + LINK INTRODUCTION



The current node in (14) was then decorated by the parse of the head noun *otoko*, and enriched by LINK EVALUATION. The resulting declarative unit is shown in (15).

(15) {Fo(ε, y, *otoko'*(y)&*naku'*(y)), Ty(e), ◊}

This type-e node was fixed as a subject by the parse of the nominative-case particle *ga*. The parse of *nigeru* (= 'run away') then created a main structure, where the type-e node decorated with the declarative unit (15) was identified as a subject node.

Notice that the previous DS account ends up with two unfixed nodes of the same type hung from the same node, as shown by the two dotted lines in (14). That is, the AXIOM set out an initial node as the root node of the whole tree, and with respect to this root node two unfixed nodes were introduced for the relative clause and for the head noun. But multiplication of unfixed relations is not licit: in Logic Of Finite Trees [1], each node must be uniquely identifiable with respect to the other nodes in a tree; but if two unfixed nodes with the same locality restriction were hung from the same node, they would be indistinguishable and cannot be uniquely defined in the tree.[8]

More than one unfixed node, however, may be hung from the same node if they are of different sorts. Recall that there are three types of locality restrictions on unfixed relations and that they differ in terms of where an unfixed node may be resolved (cf. Section 2.3). This means that if two unfixed nodes have different locality restrictions, they are distinguishable and may be introduced from the same node. In (14), however, the two unfixed relations are both globally unfixed and cannot be distinguished. Thus, the tree (14) is formally illegitimate, and

[8] "Structural underspecification and resolution" is formally similar to "functional uncertainty" within LFG [7], but it seems there is no LFG analogue of the unique-unfixed-node constraint; a functional uncertainty for FOCUS may have more than one solution if the value is a set, each member of the set being associated with different values [11] (Mary Dalrymple p.c.).

it is concluded that the previous DS account of Japanese relatives [2, 9, 13] is inadequate.

The problem of multiplying unfixed relations occurs generally in the DS treatment of complex NPs in verb-final languages such as Japanese and Korean. This is because a modifier (e.g. relative clause) in these languages precedes a head noun, and the head noun could be part of a larger complex NP. The challenge is how a parser processes complex NPs incrementally in these languages without multiplying unfixed relations with the same locality restriction.

4 Solution

This section proposes a solution to the problem raised in the last section. The heart of the proposal is to let a parser determine node-addresses flexibly. To this end, I shall drop the assumption that the AXIOM introduces a root node of the whole tree and that a head noun is processed with respect to this root node.

Firstly, the AXIOM is modified so that it introduces a node decorated with not only the type requirement ?Ty(t) but also the node-address requirement ?∃x.Tn(x), together with a place-holding variable for a node-address, as in Tn(U).

(16) AXIOM (modified)

{Tn(U), ?∃x.Tn(x), ?Ty(t), ◇}

The meta-variable U may be substituted with 0, in which case the node is identified as a root node. Alternatively, it may be substituted with an arbitrary constant "a", whose actual manifestation will be determined at a later step (cf. Section 4.1).[9]

Second, LINK INTRODUCTION is modified as in (17), where the essential point is that a node for a head noun is structurally underspecified with respect to a new type-t-requiring node. In plain English, (17) declares the following: if a node is of type-t and decorated with a proposition involving a term α, a parser initiates an inverse LINK relation from this propositional node to a type-e-requiring node; this type-e-requiring node is annotated with the requirement that this node will be annotated with a term containing α as a sub-term; a parser

[9] The use of meta-variables in modeling an underspecification of node-address is inspired by Ronnie Cann, and the use of arbitrary constants to saturate such meta-variables is suggested by Ruth Kempson. I am grateful for insightful discussions I have had with them.

structurally underspecifies this type-e-requiring node with respect to a new type-t-requiring node.[10]

(17) LINK INTRODUCTION (modified)

IF Ty(t), <D>(Fo(α))
THEN make(<L^{-1}>); go(<L^{-1}>); put(?∃x.Fo(x[α]), ?Ty(e));
make(<↑*>); go(<↑*>);
put(Tn(U), ?∃x.Tn(x), ?Ty(t)); go(<↓*>)
ELSE ABORT

4.1 *Illustration One: Simple Cases of Relatives*

For illustration, let us consider the simple case of relatives (12), repeated here as (18).

(18) [*Nai-ta* *otoko*]-*ga* *nige-ta.*
[cry-PAST man]-NOM run.away-PAST
'A man who cried ran away.'

An initial node is set out by the modified AXIOM. Unlike the previous DS account [2, 9, 13], a parser may process the relative clause *Nai-ta* directly under this initial node.

(19) Parsing *Nai-ta*

{Tn(U), ?∃x.Tn(x), Fo(*naku'*(ε, x, *P*(x))), Ty(t), <D>(Fo(ε, x, *P*(x))), ◇}

{Fo(ε, x, *P*(x)), Ty(e)} {Fo(*naku'*), Ty(e→t)}

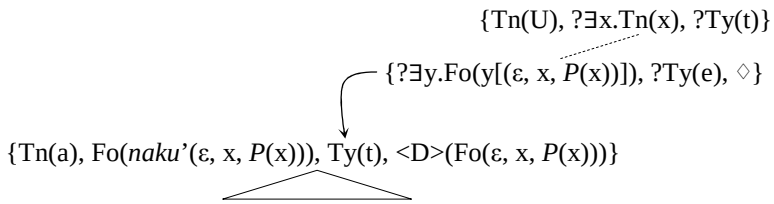
If the string ended here, a parser would identify the top node as a root node of the tree by saturating Tn(U) as Tn(0). In (18), however, *Nai-ta* is a relative clause.[11] Further, it is unknown at this point how deeply

[10] The locality restriction on this type-e-requiring unfixed node is the same as that imposed by *ADJUNCTION. This is because a head noun may be long-distance scrambled; for the detail, see a DS account of long-distance scrambling [2].

[11] When the verb *naku* appears in a relative clause, it has a special intonation [13] (cf. Section 1). This intonational cue cannot be made use of in the previous analysis [2, 9, 13], where GENERALIZED ADJUNCTION had to fire before the parse of relative clauses. By contrast, in my analysis, a parser does not run GENERALIZED ADJUNCTION, and may process a relative clause

the relative clause is embedded. Thus, a parser substitutes Tn(0) with Tn(a), where "a" is an arbitrary constant whose manifestation is worked out at a later step. A parser then runs LINK INTRODUCTION (17) in order to create a node for the head noun *otoko* (= 'man'). At this stage, the tree has been updated as in (20), where a triangle schematizes the internal structure.

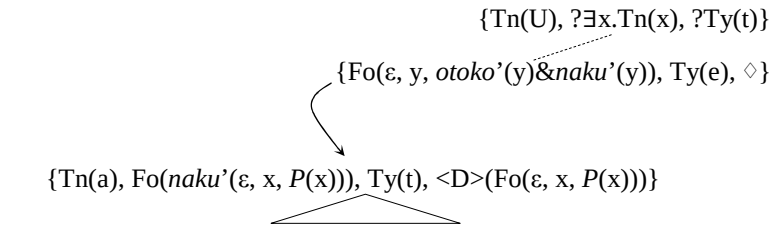
(20) Parsing *Nai-ta* + LINK INTRODUCTION



In (20), the current node is non-locally unfixed with respect to a new type-t-requiring node (cf. footnote 10). This non-local unfixed relation is shown by the dashed line.

Now that a type-e-requiring node is present, a parser may process the head noun *otoko* (= 'man'), decorating the node with content and type, and LINK EVALUATION then incorporates the content of the relative clause into the node.

(21) Parsing *Nai-ta otoko* + LINK EVALUATION

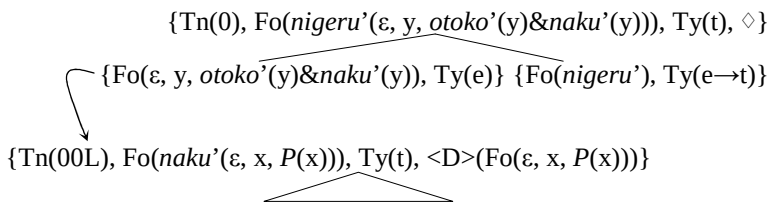


The rest of the process is as usual: the nominative-case particle *ga* marks the current node in (21) as a subject node, and the matrix verb *nigeru* (= 'run away') fleshes out a main propositional structure, where the subject node is identified as the node for the head noun. Finally, a parser saturates Tn(U) at the top node as Tn(0), ensuring that this is a root node. Once this node-address is specified, the actual manifestation

directly under an initial node set out by the AXIOM. The intonational cue then helps the parser to saturate Tn(U) at the initial node as Tn(a).

of the arbitrary constant "a" in Tn(a) is automatically explicated as Tn(00L).

(22) Parsing [*Nai-ta otoko*]-*ga nige-ta*



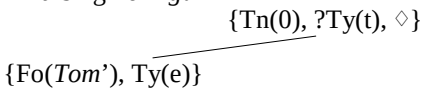
Notice that in the tree update above, no multiple unfixed nodes have been induced. This is because a node for a head noun is structurally underspecified with respect to a new type-t-requiring node that may be distinct from the root node of the whole tree.

The account is also applicable to (23), where, unlike (18), part of the matrix clause (i.e. *Tom-ga*) is processed before the relative clause *nai-ta*.

(23) *Tom-ga* [*nai-ta* *otoko*]-*o* *nagusame-ta.*
Tom-NOM [cry-PAST man]-ACC comfort-PAST
'Tom comforted a man who cried.'

Again, an initial node is set out by the AXIOM (16), and after LOCAL *ADJUNCTION creates a type-e-requiring unfixed node, *Tom* decorates the node with content and type and *ga* fixes it as a subject node. Since *Tom-ga* is part of a matrix clause, Tn(U) may be saturated as Tn(0), a node-address for a root node of the whole tree.

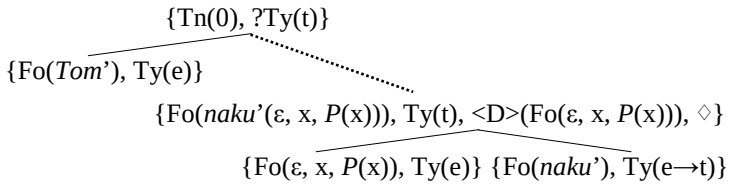
(24) Parsing *Tom-ga*



What comes next is *naku* (= 'cry'). A parser would develop the current propositional structure if *naku* were a matrix verb. In (23), an intonational break between *Tom-ga* and *nai-ta* signals that *naku* is an embedded verb, and a parser runs GENERALIZED ADJUNCTION to induce

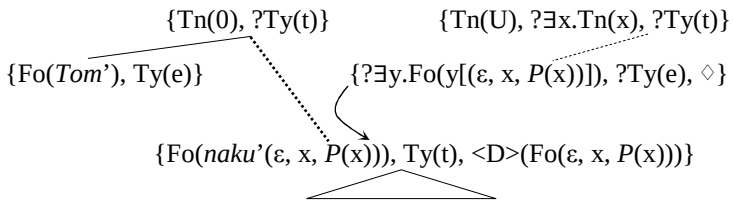
a globally unfixed type-t-requiring node.[12] The lexical actions encoded in *naku* flesh out this type-t-requiring node, providing a propositional template where a subject slot is decorated with the term (ε, x, *P*(x)), as usual.

(25) Parsing *Tom-ga nai-ta* + GENERALIZED ADJUNCTION



A parser runs LINK INTRODUCTION, initiating an inverse LINK relation to a type-e-requiring node that is unfixed with respect to a fresh type-t-requiring node.

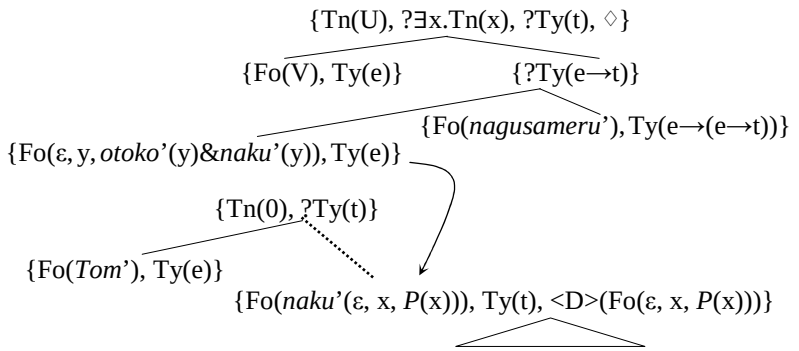
(26) Parsing *Tom-ga nai-ta* + LINK INTRODUCTION



The rest of the process is as usual: (a) the head noun *otoko* decorates the current node with content and type; (b) LINK EVALUATION incorporates the content of the relative clause into the node for the head noun; (c) the accusative-case particle *o* marks this node as an object under the type-t-requiring node; (d) the matrix verb *nagusameru* (= 'comfort') develops this type-t-requiring node by providing a propositional schema, where the object slot collapses with the node for the head noun and the subject slot is decorated with a meta-variable as in Fo(V).

[12] Here, *ADJUNCTION cannot fire because this general action requires that a current node not have any dominated node. In the present case, the current node has a dominated node (i.e. the node decorated with Fo(*Tom*')).

(27) Parsing *Tom-ga* [*nai-ta otoko*]-*o nagusame*



Now, a parser may saturate Tn(U) at the current node as Tn(0). As a result, this node is identified with the node set out by the AXIOM. Concomitantly, the node decorated with Fo(V) collapses with the node decorated with Fo(*Tom*'). (Recall that the dotted line indicates a globally unfixed relation, which may cross a LINK boundary.) For reasons of space, only the declarative unit at the root node is provided here as (28), which correctly represents the truth-conditional content of the string (23).

(28) {Tn(0), Fo(*nagusameru*'(ε, y, *otoko*'(y)&*naku*'(y))(*Tom*')), Ty(t), ◊}

4.2 *Illustration Two: Relative Clause Nesting*

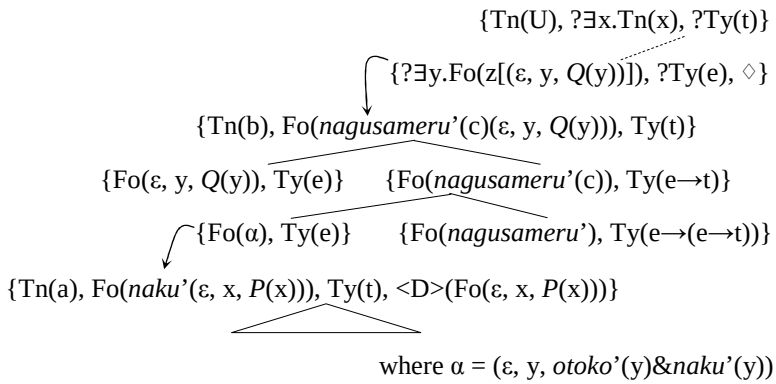
In the present account, a node for a head noun is structurally underspecified within a new propositional structure, and once this propositional structure is fully developed a parser may run LINK INTRODUCTION to induce another inverse LINK relation. Thus, the account naturally models successive relative clause embedding without failing to capture the left-to-right processing of the sequence. To illustrate, consider the case of relative clause nesting as in (29), where the complex NP *Nai-ta otoko* (= 'a man who cried') is part of the relative clause that modifies the head noun *hito* (= 'person').

(29) [[*Nai-ta otoko*]-*o nagusame-ta hito*]-*ga nige-ta*

[[cry-PAST man]-ACC comfort-PAST person]-NOM run.away-PAST
'A person who comforted a man who cried ran away.'

The parse of this string up to *otoko* (= 'man') gives rise to the same tree as (21). The unfixed node for *otoko* is marked as an object node by the accusative-case particle *o*. Then, *nagusameru* (= 'comfort') provides a propositional template; an object slot collapses with the node for *otoko*, and a subject slot is decorated with Fo(ε, y, Q(y)). At this stage, a parser may run LINK INTRODUCTION once again in order to parse the head noun *hito*, initiating another inverse LINK relation from the propositional node decorated with Fo(*nagusameru*'(c)(ε, y, *Q*(y))) to a type-e-requiring node.

(30) Parsing [*Nai-ta otoko*]-*o nagusame-ta* + LINK INTRODUCTION



where α = (ε, y, *otoko*'(y)&*naku*'(y))

The rest of the process is as usual: (a) the current node is decorated by the head noun *hito*; (b) LINK EVALUATION reflects the content of the relative clause into the node for *hito*; (c) the node for *hito* is marked as a subject by the nominative-case particle *ga* under a new type-t-requiring node; (d) this type-t-requiring node is fleshed out by *nigeru* (= 'run away'), where the subject slot collapses with the node for *hito*; (e) finally, Tn(U) at the top node is saturated as Tn(0), a node-address for a root node of the whole tree. The declarative unit at the root node is shown in (31).

(31) {Fo(*nigeru'*(ε, z, *hito*'(z)&*nagusameru*'(ε, y, *otoko*'(y)&*naku*'(y))(z))), Tn(0), Ty(t), ◇}

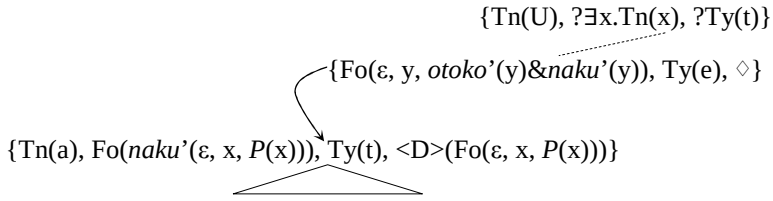
4.3 *Illustration Three: Scrambling of Complex NPs*

Japanese allows the permutation of arguments, so-called "scrambling". Thus, a head noun modified by a relative clause may be fronted: compare (23) with (32).

(32) [*Nai-ta* *otoko*]-*o* *Tom-ga* *nagusame-ta.*
[cry-PAST man]-ACC Tom-NOM comfort-PAST
'Tom comforted a man who cried.'

Scrambling is also dealt with by the present account. In (32), the parse of the relative clause *Nai-ta* provides a propositional template, where a subject slot is decorated with Fo(ε, x, *P*(x)), and LINK INTRODUCTION initiates an inverse LINK relation from this type-t node to a type-e-requiring unfixed node. This unfixed node is decorated by the head noun *otoko* (= 'man') and enriched by LINK EVALUATION.

(33) Parsing *Nai-ta otoko* + LINK EVALUATION



The current node is marked as an object node by the accusative-case particle *o*. Then, a pointer ◇ goes up to the type-t-requiring node, where *Tom-ga* induces a subject node and *nagusameru* (= 'comfort') creates a propositional template, where a subject slot collapses with the node for *Tom*. Finally, Tn(U) is saturated as Tn(0). The root node is decorated with (34); this declarative unit is exactly the same as the one in (28), which predicts that the string (32) is truth-conditionally equivalent to the string (23).

(34) {Tn(0), Fo(*nagusameru*'(ε, y, *otoko*'(y)&*naku*'(y))(*Tom*')), Ty(t), ◇}

4.4 *Illustration Four: Unbounded-Dependency and Island-Insensitivity*

Japanese relatives exhibit "unbounded-dependency": a head noun may be associated with a gap in a relative clause across a clause boundary.

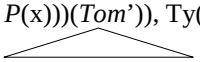
Thus, in (35), the head noun *otoko* (= 'man') is associated with the subject gap of *naku* (= 'cry') across the clause boundary *Tom-ga … i-tta*.

(35) [[*Tom-ga* [*nai-ta* *to*] *i-tta*] *otoko*]-*ga* *nige-ta.*
[[Tom-NOM [cry-PAST COMP] say-PAST] man]-NOM run.away-PAST
'A man who Tom said cried ran away.'

Prior to the head noun *otoko*, the parse of (35) leads to the semantic tree (36).

(36) Parsing *Tom-ga nai-ta to i-tta*

{Tn(a), Fo(*iu*'(*naku*'(ε, x, *P*(x)))(*Tom*')), Ty(t), <D>(Fo(ε, x, *P*(x))), ◊}



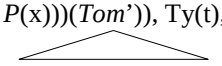
<D>(Fo(ε, x, *P*(x))) declares that the term (ε, x, *P*(x)) is found somewhere below the current node (possibly, across a LINK boundary; cf. Section 2.4.) Thus, the IF block of LINK INTRODUCTION is met and a parser initiates an inverse LINK relation to a type-e-requiring node, imposing a requirement that this node will be annotated with a term containing (ε, x, *P*(x)) as a sub-term. This type-e-requiring node is decorated by the head noun *otoko* (= 'man') and enriched by LINK EVALUATION.

(37) Parsing *Tom-ga nai-ta to i-tta otoko* + LINK EVALUATION

{Tn(U), ?∃x.Tn(x), ?Ty(t)}

{Fo(ε, y, *otoko*'(y)&*iu*'(*naku*'(y))(*Tom*')), Ty(e), ◊}

{Tn(a), Fo(*iu*'(*naku*'(ε, x, *P*(x)))(*Tom*')), Ty(t), <D>(Fo(ε, x, *P*(x)))}



The current node is marked as a subject by the nominative-case particle *ga*, and the matrix verb *nigeru* (= 'run away') creates a propositional schema where a subject slot collapses with the node for *otoko*. The root node in the final state is decorated with the declarative unit in (38).

(38) {Tn(0), Fo(*nigeru*'(ε, y, *otoko*'(y)&*iu*'(*naku*'(y))(*Tom*'))), Ty(t), ◊}

Given the lack of restrictions on where the term to be shared in the pair of LINKed structures is to be detected, it is predicted that Japanese relatives are not sensitive to "islands": that is, a head noun may be associated with a gap across an island boundary [12]. Thus, as shown in (39), the head noun *hito* (= 'man') may be associated with the subject gap of *kau* (= 'buy') even though this association crosses a complex NP island boundary that is formed by *Ka-tta tokei*.

(39) [[*Ka-tta* *tokei*]-*ga* *nisemonoda-tta* *hito*]-*ga* *nai-ta.*
[[buy-PAST watch]-NOM fake-PAST man]-NOM cry-PAST
'A man such that a watch he bought was a fake cried.'

One may wonder whether the use of the operator <D> in LINK INTRODUCTION is a stipulation, but there is a rationale. As has been assumed, verbs in Japanese provide a propositional skeleton where argument slots are decorated with meta-variables, and saturation of meta-variables is not structurally constrained. So, LINK INTRODUCTION is defined with the operator <D>, which models the weakest dominance relation, so that the label <D>(Fo(α)) and the primitive action put(?∃x.Fo(x[α]), ?Ty(e)) ensure that a term which will inhabit a node for a head noun may be found "deep inside" the relative clause structure (i.e. across a LINK relation).

5 Conclusion

This article has pointed out that the extant DS account of complex NPs in verb-final languages, especially Japanese relatives, is not adequate in that it multiplies unfixed relations with the same locality restriction. This formal problem disappears if node-addresses are specified flexibly. To this end, the AXIOM and LINK INTRODUCTION are modified and tested against a range of data posed by Japanese relatives.

In closing, it should be noted that the refined DS parser is more realistic than the past DS parser [2, 9, 13]. In the previous account, some sort of "look ahead" device needs to be assumed: that is, a parser must foresee that an incoming string has an embedded clause and run GENERALIZED ADJUNCTION before it starts to process the string. Although it was suggested that intonational cues were available to the parser, such cues would not obtain until a verb within a relative clause is parsed. By contrast, the parser proposed in this article may start to process a string without executing GENERALIZED ADJUNCTION in advance because an initial node set out by the AXIOM does not have to

be a root node of the whole tree and it may be developed by the parse of an embedded clause. This account makes use of intonational cues more effectively in order to saturate Tn(U), an underspecified node-address.[13]

ACKNOWLEDGMENTS. I have benefitted from constructive and encouraging exchanges with Ronnie Cann, David Cram, Mary Dalrymple, Ruth Kempson, and Jieun Kiaer. My deepest gratitude goes to Ruth Kempson, who provided me with a number of valuable comments on this work. Needless to say, the author alone is responsible for any inadequacies in the present article. This research was supported by the Clarendon Fund Scholarship, the Oxford-Kobe Scholarship, and the Sasakawa Fund Scholarship.

References

1. Blackburn, P., Meyer-Viol, W.: Linguistics, Logic, and Finite Trees. Bulletin of Interest Group of Pure and Applied Logics 2, 2-39 (1994)
2. Cann, R., Kempson, R., Marten, L.: The Dynamics of Language. Elsevier, Oxford (2005)
3. Chatzikyriakidis, S., Kempson, R.: Standard Modern and Pontic Greek person restrictions. Journal of Greek Linguistics 11, 127-166 (2011)
4. Chomsky, N.: Aspects of the Theory of Syntax. MIT Press, MA, Cambridge (1965)
5. Heim, I., Kratzer, A.: Semantics in Generative Grammar. Blackwell, Oxford (1998)
6. Kamide, Y.: Incrementality in Japanese sentence processing. In: Nakayama, M. et al. (eds.) The Handbook of East Asian Psycholinguistics, Vol. 2, Japanese. Cambridge University Press, Cambridge (2006)
7. Kaplan, R. M., Zaenen, A.: Long-distance dependencies, constituent structure, and functional uncertainty. In: Baltin, M., Kroch, A. (eds.)

[13] The mechanism proposed in this article has an implication for a cross-linguistic modeling of relatives in verb-final languages. The AXIOM remains invariant across languages, but it is possible that LINK INTRODUCTION is realized in a different form in different languages. For instance, in Korean, where the suffix -(*u*)*n* serves as a marker for relative clauses [17], the macro of actions in LINK INTRODUCTION may be lexically encoded in the relative clause marker. Further, the essential idea of the modified LINK INTRODUCTION is to underspecify a node for a head noun within a new propositional structure. This insight is quite general and may be made use of when we define general actions for other types of complex NPs.

Alternative Conceptions of Phrase Structure. University of Chicago Press, Chicago (1989)
8. Kempson, R., Gregoromichelaki, E., Howes, C.: The Dynamics of Lexical Interfaces. CSLI, Stanford (2011)
9. Kempson, R., Kurosawa, A.: At the syntax-pragmatics interface. In: Hoshi, H. (ed.) The Dynamics of Language Faculty. Kuroshio, Tokyo (2009)
10. Kempson, R., Meyer-Viol, W., Gabbay, D.: Dynamic Syntax. Blackwell, Oxford (2001)
11. King, T. H.: Focus domains and information structure. In: Butt, M., King, T. H. (eds.) The Proceedings of the LFG 97 Conference. CSLI, Stanford (1997)
12. Kuno, S.: The Structure of the Japanese Language. MIT Press, MA, Cambridge (1973)
13. Kurosawa, A.: On the Interaction of Syntax and Pragmatics. Ph.D. thesis, King's College London (2003)
14. Pritchett, B. L.: Grammatical Competence and Parsing Performance. University of Chicago Press, Chicago (1992)
15. Purver, M., Cann, R., Kempson, R.: Grammars as parsers. Research on Language and Computation 4, 289-326 (2006)
16. Seraku, T.: Multiple foci in Japanese clefts and the growth of semantic representation. In: Aloni, M. et al. (eds.) Lecture Notes in Computer Science 7218. Springer, Berlin (2012)
17. Sohn, H.: The Korean Language. Cambridge University Press, Cambridge (1999)

Tohru Seraku
St. Catherine's College,
University of Oxford,
Manor Road, Oxford, OX1 3UJ, UK.
E-mail: <tohru.seraku@stcatz.ox.ac.uk>

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 79–94
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 17/06/13

Thematically Reinforced Explicit Semantic Analysis

YANNIS HARALAMBOUS[1] AND VITALY KLYUEV[2]

[1] *Institut Mines-Télécom - Télécom Bretagne and Lab-STICC UMR CNRS 6285, France*
[2] *University of Aizu, Japan*

ABSTRACT

We present an extended, thematically reinforced version of Gabrilovich and Markovitch's Explicit Semantic Analysis (ESA), where we obtain thematic information through the category structure of Wikipedia. For this we first define a notion of categorical tfidf which measures the relevance of terms in categories. Using this measure as a weight we calculate a maximal spanning tree of the Wikipedia corpus considered as a directed graph of pages and categories. This tree provides us with a unique path of "most related categories" between each page and the top of the hierarchy. We reinforce tfidf of words in a page by aggregating it with categorical tfidfs of the nodes of these paths, and define a thematically reinforced ESA semantic relatedness measure *which is more robust than standard ESA and less sensitive to noise caused by out-of-context words. We apply our method to the French Wikipedia corpus, evaluate it through a text classification on a 37.5 MB corpus of 20 French newsgroups and obtain a precision increase of 9–10% compared with standard ESA.*

1 INTRODUCTION

1.1 *Explicit Semantic Analysis*

Unlike semantic similarity measures, which are limited to ontological relations such as synonymy, hyponymy, meronymy, etc., *semantic relatedness*

measures detect and quantify semantic relations of a more general kind. The typical example is the one involving the concepts CAR, VEHICLE and GASOLINE. A car is a special kind of vehicle, so we have an hyperonym relation between the concepts, which can easily be quantified by a semantic similarity measure (for example, by taking the inverse of the length of the shortest path between the corresponding synsets in WordNet). But between CAR and GASOLINE, there is no semantic *similarity*, since a car is a solid object and fuel is a liquid. Nevertheless, there is an obvious semantic relation between them since most cars use gasoline as their energy source, and such a relation can be quantified by a semantic *relatedness* measure.

Gabrilovich & Markovitch [1] introduce the semantic relatedness measure ESA (= Explicit Semantic Analysis, as opposed to the classical method of Latent Semantic Analysis [2]). ESA is based on the Wikipedia corpus. Here is the method: after cleaning and filtering Wikipedia pages (keeping only those with a sufficient amount of text and a given minimal number of incoming and outgoing links), they remove stop words, stem all words and calculate their tfidfs. Wikipedia pages can then be represented as vectors in the space of (nonempty, stemmed, distinct) words, the vector coordinates being normalized tfidf values. By the encyclopedic nature of Wikipedia, one can consider that every page corresponds to a concept. We thus have a matrix whose columns are concepts and whose lines are words. By transposing it we obtain a representation of words in the space of concepts. The ESA measure of two words is simply the cosine of their vectors in this space.

Roughly, two words are closely ESA-related *if they appear frequently in the same Wikipedia pages* (so that their tfs are high), *and rarely in the corpus as a whole* (for their dfs to be low).

Despite the good results obtained by this method, it has given rise to some criticism. Thus, Haralambous & Klyuev [3] note that ESA has poor performance when the relation between words is mainly ontological. As an example, in the English corpus, the word “mile” (length unit) does not appear in the page of the word “kilometer” and the latter appears only once in the page of the former: this is hardly sufficient to establish a nonzero semantic relatedness value; however, such a relation is obvious, since both words refer to units of length measurement. As pointed out in [3], an ontological component, obtained from a WordNet-based measure, can, at least partially, fill this gap.

Another, more fundamental, criticism is that of Gottron et al. [4], who argue that the choice of Wikipedia is irrelevant, and that any corpus of comparable size would give the same results. To prove it, they base ESA not on Wikipedia, but on the Reuters news corpus, and get even better results

than with standard ESA. According to the authors, the semantic relatedness value depends only on the collocational frequency of the terms, and this whether documents correspond to concepts or not. In other words they deny the "concept hypothesis," namely that ESA specifically uses the correspondence between concepts and Wikipedia pages. Also they state that while "the application of ESA in a specific domain benefits from taking an index collection from the same topic domain while, on the other hand, a "general topic corpus" such as Wikipedia introduces noise," and this has precisely been our motivation for strengthening the thematic robustness of ESA. Indeed, in this article we will enhance ESA by adopting a different approach: the persistence of tfidfs of terms when leaving pages and entering the category graph.

1.2 *Wikipedia Categories*

A Wikipedia page can belong to one or more categories. Categories are represented by specific pages using the "Category:" prefix; these pages can again belong to other categories, so that we obtain a directed graph structure, the nodes of which can be standard pages (only outgoing edges) or categories (in- and outgoing edges). A page can belong to several categories and there is no ranking of their semantic relevance. For this reason, to be able to use categories, we first need an algorithm to determine the single semantically most relevant category, and for this we use, once again, ESA.

Wikipedia's category graph has been studied thoroughly in [5] (for the English corpus).

1.3 *Related Work*

Scholl et al. [6] also enhance the performance of ESA using categories. They proceed as follows: let T be the matrix whose rows represent the Wikipedia pages and whose columns represent words. The value $t_{i,j}$ of cell (i, j) is the normalized tfidf of the jth word in the ith page. For each word m there is therefore a vector $\boldsymbol{v}_m$ whose dimension is equal to the number of pages. Now let C be the matrix whose columns are pages and whose lines are categories. The value of a cell $c_{i,j}$ is 1 when page j belongs to category i and 0 otherwise. They take the product of matrices $\boldsymbol{v}_m \cdot C$ which provides a vector whose jth component is $\sum_{i|D_i \in c_j} t_{i,j}$, that is the sum of tfidfs of word m for all pages belonging to the jth category. They use the concatenation of vector $\boldsymbol{v}_m$ and of the transpose of $\boldsymbol{v}_m \cdot C$ to improve

system performance on the text classification task. They call this method XESA (eXtended ESA).

We see that in this attempt, page tfidf is extended to categories by simply taking the sum of tfidfs of all pages belonging to a given category. This approach has a disadvantage when it comes to high-level categories: instead of being a way to find the words that characterize a given category, the tfidf of a word tends to become nothing more than the average density of the word in the corpus, since for large categories, tf tends to be the total number of occurrences of the word in the corpus, while the denominator idf remains constant and equal to the number of documents containing the given word. Thus, this type of tfidf loses its power of discrimination for high-level categories. As we will see in Section 2.2, we propose another extension of tfidf to categories, which we call *categorical tfidf*. The difference lies in the denominator, where we take the number, not of all documents containing the term, but only of those *not belonging* to the category. Thus our categorical tfidf (which is equal to the usual tfidf in the case of pages) is high when the term is common in the category and *rare elsewhere* (as opposed to *rare on the entire corpus* of Scholl et al.).

In [7], the authors examine the problem of inconsistency of Wikipedia's category graph and propose a shortest path approach (based on the number of edges) between a page and the category "*Article*," which is at the top of the hierarchy. The shortest path provides them with a semantic and thematic hierarchy and they calculate similarity as shortest length between vertices on these paths, a technique already used in WordNet [8]. However, as observed in [8, p. 275], the length (in number of edges) of the shortest path can vary randomly, depending on the density of pages (synsets, in the case of WordNet) in a given domain of knowledge. On the other hand, the distance (in number of edges) between a leaf and the top of the hierarchy is often quite short, frequently requiring an arbitrary choice between paths of equal length.

What is common with our approach is the intention to simplify Wikipedia's category graph. But instead of counting edges, we weight the graph using ESA measure and use this weight, which is based on the statistical presence of words on pages belonging to a given category, to calculate a maximum spanning tree. The result of this operation is that any page (or category other than "*Article*") has exactly one parent category that is semantically closest to it. This calculation is global, in the sense that the total weight of the tree is maximum.

We use this tree to define *thematically reinforced ESA*. Our goal is to avoid words which, by accident, have a high tfidf in a given page despite the fact that they thematically do not really belong to it. This happens in

the very frequent case where words have low frequencies (in the order of 1–3) so that the presence of an unsuitable word in a page results in a tfidf value as high (or even higher, if the word is seldom elsewhere) as the one of relevant words. Our hypothesis is that a word having an unduly high tfidf will disappear when we calculate its (categorical) tfidf in categories above the page, while, on the contrary, relevant words will be shared by other pages under the same category and their tfidfs will continue to be nonzero when switching to them. Such words will "survive" when we move away from leaves of the page-and-category tree and towards the root.

2 THEMATIC REINFORCEMENT

2.1 *Standard Tfidf, Concept Vector and ESA Measure*

Let us first formalize the standard ESA model.[3]

Let $\mathcal{W}$ be the Wikipedia corpus pruned by the standard ESA method, $p \in \mathcal{W}$ a Wikipedia page, and $w \in p$ a word.[4] The tfidf $t_p(w)$ of the word w on page p is defined as:

$$t_p(w) := (1 + \log(f_p(w))) \cdot \log\left(\frac{\#\mathcal{W}}{\sum_{\substack{p\in\mathcal{W}\\ w\in p}} 1}\right),$$

where $f_p(w)$ is the frequency of w on page p, $\#\mathcal{W}$ the cardinal of $\mathcal{W}$ and $\sum_{\substack{p\in\mathcal{W}\\ w\in p}} 1$, also known as the df (= document frequency) of w, is the number of Wikipedia pages containing w.

Consider the space $\mathbb{R}^{\#\mathcal{W}}$, where dimensions correspond to pages p of $\mathcal{W}$. Then we define the "concept vector" $\boldsymbol{w}$ of word w as

$$\boldsymbol{w} := \sum_{p\in\mathcal{W}} t_p(w) \cdot 1_p \in \mathbb{R}^{\#\mathcal{W}}$$

where 1_p is the unitary vector of $\mathbb{R}^{\#\mathcal{W}}$ corresponding to page p.

Let w and w' be words appearing in Wikipedia (and hence the Euclidean norms $\|\boldsymbol{w}\|$ and $\|\boldsymbol{w'}\|$ of their concept vectors are nonzero). The ESA semantic relatedness measure μ is defined as follows:

$$\mu(w, w') := \frac{\langle \boldsymbol{w}, \boldsymbol{w'}\rangle}{\|\boldsymbol{w}\| \cdot \|\boldsymbol{w'}\|}.$$

[3] All definitions in Section 2.1 are from [1].

[4] By "word" we mean an element of the set of character strings remaining after removing stopwords and stemming the Wikipedia corpus.

2.2 *Categorical Tfidf*

Let c be a Wikipedia category. We define $\mathcal{F}(c)$ as the set of all pages p such that

- either p belongs to c,
- or p belongs to c_1, and there a sequence of subcategory relations $c_1 \to c_2 \to \cdots \to c$, ending with c.

Definition 1 *Let $w \in p$ be a word of $p \in \mathcal{W}$, $t_p(w)$ its standard tfidf in p, and c a category of $\mathcal{W}$. We define the* categorical tfidf $t_c(w)$ *of w for category c as follows:*

$$t_c(w) := \left(1 + \log\left(\sum_{p \in \mathcal{F}(c)} f_p(w)\right)\right) \cdot \left(\log\left(\frac{\#\mathcal{W}}{1 + \sum_{\substack{p \in \mathcal{W} \setminus \mathcal{F}(c) \\ w \in p}} 1}\right)\right).$$

The difference with the tfidf defined by [6] is in the calculation of df: instead of $\sum_{\substack{p \in \mathcal{W} \\ w \in p}} 1$, that is the amount of pages containing w in the entire Wikipedia corpus, we focus on those in $\mathcal{W} \setminus \mathcal{F}(c)$, namely the set difference between the whole corpus and pages that are ancestors of c in the category graph, and we use $1 + \sum_{\substack{p \in \mathcal{W} \setminus \mathcal{F}(c) \\ w \in p}} 1$ instead (the unit is added to prevent a zero df in the case where the word does not appear outside $\mathcal{F}(c)$). We believe that this extension of tfidf to categories improves discriminatory potential, even when the sets of pages become large (see discussion in Section 1.3).

2.3 *Vectors of Pages and Categories*

Let $p \in \mathcal{W}$ be a page. We define the *page vector* $\boldsymbol{p}$ as the normalized sum of concept vectors of its words, weighted by their tfidfs:

$$\boldsymbol{d} := \frac{\sum_{w \in p} t_p(w) \cdot \boldsymbol{w}}{\| \sum_{w \in p} t_p(w) \cdot \boldsymbol{w} \|}.$$

Similarly let c be a category of Wikipedia, we define the *category vector* $\boldsymbol{c}$ as

$$\boldsymbol{c} := \frac{\sum_{w \in \mathcal{F}(c)} t_c(w) \cdot \boldsymbol{w}}{\| \sum_{w \in \mathcal{F}(c)} t_c(w) \cdot \boldsymbol{w} \|}.$$

where $w \in \mathcal{F}(c)$ means that there exists a page p such that $p \in \mathcal{F}(c)$ and $w \in p$.

2.4 *Wikipedia Arborification*

Definition 2 *Let p be a Wikipedia page and c, c' Wikipedia categories. Let $p \to c$ be the membership of page d to category c, and $c \to c'$ the subcategory relation between c and c. We define the* weight of semantic relatedness *of these relations as*

$$p(p \to c) = \langle \boldsymbol{p}, \boldsymbol{c} \rangle.$$
$$p(c \to c') = \langle \boldsymbol{c}, \boldsymbol{c'} \rangle,$$

where $\langle \, . \, , \, . \, \rangle$ is the Euclidean scalar product of two vectors.

This product is equal to the cosine metric since the vectors are all unitary. By this property we also have $\mathrm{Im}(p) \subset [0, 1]$.

The relations considered in Definition 2 correspond to vertices of the Wikipedia category graph. Let $\mathcal{W}'$ be the weighted Wikipedia digraph; its vertices are pages and categories, its edges are memberships of pages and inclusions of categories, and its weight is the weight of semantic relatedness.

At this point we can already reinforce the standard tfidf of words on pages, by the categorical tfidf of the same words in related categories. But how can we choose these categories? Taking all those containing a page would result in cacophony since categories can be more or less relevant and sometimes have no semantic relation whatsoever. Not to mention the fact that the Wikipedia category graph is quite complex, and using it as such would be computationally prohibiting.

The solution we present to this problem is to simplify $\mathcal{W}'$ by extracting a maximal spanning tree. It should be noted that standard minimal/maximal spanning tree algorithms such as Kruskal or Prim cannot be applied because $\mathcal{W}'$ is directed, has a global sink, namely the "Article" page, and we want the orientation of the directed spanning tree to be compatible with the one of the directed graph[5].

To obtain the maximal spanning tree, we utilized Chu-Liu & Edmonds' algorithm [9, p. 113-119], published for the first time in 1965. This semi-linear algorithm returns a minimum weight forest of rooted trees covering the digraph. The orientation of these rooted trees is compatible with the one of the graph. In the general case, connectivity is not guaranteed (even though the graph may be connected). But in the case of a digraph containing a global sink, the forest becomes a single tree, and we get a true *directed*

[5] It is a known fact that every rooted tree has exactly two possible orientations: one going from the root to the leaves and one in the opposite direction.

maximal spanning tree of the graph. If our case, the global sink is obviously the category that is hierarchically at the top, namely "*Article*."[6]

Let $\mathcal{T}$ be the maximal spanning tree of $\mathcal{W}'$ obtained by our method. As in any tree, there is a unique path between any two nodes. In particular, there is a unique path between any page-node and the root; we call it the *sequence of ancestors* of the page.

2.5 *Thematically Reinforced ESA*

We will use the page ancestors in the maximal spanning tree to update tfidf values of words in the page vectors. Indeed, a word in a given page may have a high tfidf value simply because it occurred one or two times, this does not guarantee a significant semantic proximity between the word and the page. But if the word appears also in ancestor categories (and hence, in other pages belonging to the same category), then we have stronger chances for semantic pertinence.

Definition 3 *Let p be a Wikipedia page, w a word $w \in p$, $t_p(w)$ the standard tfidf of w in p, $(\pi^i(p))_i$ the sequence of ancestors of p, and $(\lambda_i)_i$ a decreasing sequence of positive real numbers converging to 0. We define the* thematically reinforced tfidf $t_{p,\lambda_*}(w)$ *as*

$$t_{p,\lambda_*}(w) = t_p(w) + \sum_{i \geq 0} \lambda_i t_{\pi^i(p)}(w).$$

The sum is finite because the Wikipedia maximal spanning tree is finite and hence there is a maximal distance from the root, after which the π^i become vacuous.

Definition 4 *With the notations of Definition 3, we define the* thematically reinforced concept vector $\boldsymbol{w}_{\lambda_*}$ *as*

$$\boldsymbol{w}_{\lambda_*} := \sum_{p \in \mathcal{W}} t_{p,\lambda_*}(w) \cdot 1_p \in \mathbb{R}^{\#\mathcal{W}}.$$

[6] It should be noted, however, that the path between a page and the root on the maximal spanning tree is not a maximal path per se, since the importance is given to the global maximality of weight, for the whole tree. If our goal were to find the most appropriate taxonomy for a specific page, i.e., the most relevant path from this page to the top, then it would be more appropriate to use a shortest/longest path algorithm, such as Dijkstra. This has already been proposed in [7], but for the metric of the number of edges; in our case we would rather use the measure given by the weight of the graph.

In other words, it is the usual concept vector definition, but using thematically reinforced tfidf.

With these tools we can define our extended version of ESA, as follows:

Definition 5 *With the notations of Definition 3 and $w, w' \in \mathcal{W}$, we define the* thematically reinforced ESA semantic relatedness measure μ_{λ_*} *as:*

$$\mu_{\lambda_*}(w, w') := \frac{\langle \boldsymbol{w}_{\lambda_*}, \boldsymbol{w}'_{\lambda_*} \rangle}{\|\boldsymbol{w}_{\lambda_*}\| \cdot \|\boldsymbol{w}'_{\lambda_*}\|}.$$

In other words, it is the usual ESA measure definition, but using thematically reinforced concept vectors and tfidf.

3 CORPUS

We have chosen to work on the French Wikipedia corpus (version of December 31, 2011), which is smaller than the English one and, to our knowledge, has not yet been used for ESA. To adapt ESA to French Wikipedia, we followed the same steps as [1] and [10] except for one: we have preceded stemming by lemmatization, to avoid loss of information due to poor stemming of inflected words. (In English, inflection is negligible, so that stemming can be performed directly.)

Originally, the authors of [1] pruned the 2005 English Wikipedia corpus down to 132,689 pages. In our case, by limiting the minimum size of pages to 125 (nonstop, lemmatized, stemmed and distinct) words, 15 incoming and 15 outgoing links, we obtained a number of Wikipedia pages comparable to that of the original ESA implementation, namely 128,701 pages (out of 2,782,242 in total) containing 1,446,559 distinct words (only 339,679 of which appear more than three times in the corpus).

Furthermore, the French corpus contains 293,244 categories, 680,912 edges between categories and 12,935,688 edges between pages and categories. As can be seen in Fig. 1, by the logarithmic distribution of incoming and outgoing degrees, this graph follows a power distribution $p^{-\alpha}$ with $\alpha = 2.08$ for incoming degrees and $\alpha = 7.51$ for outgoing degrees. According to [11, p. 248], the former value is typical, while the latter can be considered very high, and this was another motivation for simplifying the Wikipedia graph by extracting the maximal spanning tree, instead of performing heavy calculations on the entire graph.

The French Wikipedia category graph is fairly complex and, in particular, contains cycles. Indeed, according to [12], "cycles are not encouraged

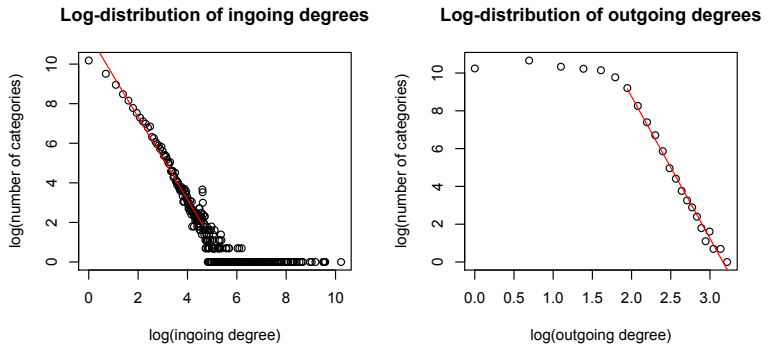


Fig. 1. Ingoing and outgoing degree distribution of the French Wikipedia categories.

but may be tolerated in rare cases." The very simple example of categories "*Zoologie*" (= Zoology) and "*Animal*" (in French Wikipedia) pointing to each other, shows that the semantic relation underlying subcategories is not always hyperonymy. Here ANIMAL is the object of study of the discipline ZOOLOGY. We attempted the following experiment: starting from the 2,782,242 (unfiltered) French Wikipedia pages, we followed random paths formed by the category links. The choice of each subsequent category was made at random, but did not change during the experiment. 78% of these paths contained cycles, but it turned out that it was always the same 50 cycles, 12 of which were of length 3 (triangles) and all others of length 2 (categories pointing to each other, as in the example above, which was detected by this method). Hence, we were able to turn this directed graph acyclic by merely removing 50 edges.

4 EVALUATION

Gabrilovich and Markovitch [1] evaluate their method on WS-353, a set of 352 English word pairs, the semantic relatedness of which has been evaluated by 15–16 human judges. Their criterion is the Spearman correlation coefficient between the rank of pairs obtained by ESA and that obtained by taking the average of human judgments. Our first attempt was to translate these pairs into French, but the result was rather disappointing.[7]

[7] Indeed, some twenty words are untranslatable into a simple term (the current version of ESA covers only single-word terms), such as "seafood" which

We have therefore chosen to evaluate our implementation of ESA in a more traditional way, by performing a text classification task. We have extracted a total of 20,000 French language messages from the 20 most popular French newsgroups. The characteristics of our evaluation corpus can be seen on Table 1, where the second column represents the number of messages for a given newsgroup, the third the number of words, and the fourth, the number of distinct stemmed nonstop words that also occur in Wikipedia.

Table 1. Characteristics of the evaluation corpus

Theme	Newsgroup	# mess.	# words	# terms
Medicine	fr.bio.medecine	1,000	738,258	14.785
Writing	fr.lettres.ecriture	1,000	688,849	14,948
French language	fr.lettres.langue.francaise	1,000	594,143	14,956
Animals	fr.rec.animaux	1,000	391,270	10,726
Classical music	fr.rec.arts.musique.classique	1,000	379,794	15,056
Rock music	fr.rec.arts.musique.rock	1,000	318,434	12,764
Do-it-yourself	fr.rec.bricolage	1,000	358,220	8,349
Movies	fr.rec.cinema.discussion	1,000	680,480	18,284
Gardening	fr.rec.jardinage	1,000	495,465	12,042
Photography	fr.rec.photo	1,000	415,767	10,931
Diving	fr.rec.plongee	1,000	485,059	11,326
Soccer	fr.rec.sport.football	1,000	612,842	13,548
Astronomy	fr.sci.astronomie	1,000	444,576	10,781
Physics	fr.sci.physique	1,000	598,079	13,916
Economics	fr.soc.economie	1,000	737,795	14,797
Environment	fr.soc.environnement	1,000	683,806	15,756
Feminism	fr.soc.feminisme	1,000	612,844	16,716
History	fr.soc.histoire	1,000	675,957	16,458
Religion	fr.soc.religion	1,000	763,477	16,124
Sects	fr.soc.sectes	1,000	738,327	16,732
Global		20,000	11,413,442	67,902

can be translated only as "*fruits de mer*." Furthermore there are ambiguities of translation resulting from word polysemy: When we translate the pair "flight/car" by "*vol*/*voiture*," we obtain a high semantic relatedness due to the criminal sense of "*vol*" (= theft) while the sense of the English word "flight" is mainly confined to the domain of aviation. Finally, some obvious collocations disappear when translating word for word, such as "soap/opera" which is unfortunately not comparable to "*savon*/*opéra*"...

Table 2. Evaluation results (ordered by decreasing precision)

λ_1	λ_2	λ_3	λ_4	λ_5	C	# SVs	Precision	λ_1	λ_2	λ_3	λ_4	λ_5	C	# SVs	Precision
1.5	0	0.5	0.25	0.125	3.0	786	**75.015%**	0	1	0.5	0.25	0.125	3.0	710	74.716%
1	0	0.5	0.25	0.125	3.0	709	74.978%	2	1	0.5	0.25	0.125	3.0	899	74.705%
1.5	1	0.5	0.25	0.125	3.0	827	74.899%	2	0	0.5	0.25	0.125	3.0	852	74.675%
0.25	1.5	0.5	0.25	0.125	3.0	761	74.87%	0.5	0.25	0.125	0.0625	0.0312	3.0	653	74.67%
0.5	0	0.5	0.25	0.125	3.0	698	74.867%	2	0.5	0.5	0.25	0.125	3.0	899	74.641%
1	0.5	0.25	0.125	0.0625	3.0	736	74.845%	0.25	0.125	0.0625	0.0312	0.015	3.0	615	74.613%
0.5	1	0.5	0.25	0.125	3.0	736	74.795%	1	1	1	0.5	0.25	3.0	796	74.61%
1	1.5	0.5	0.25	0.125	3.0	865	74.791%	0	1.5	1	0.5	0.25	3.0	792	74.548%
0.5	0.5	0.5	0.25	0.125	3.0	682	74.789%	1.5	1.5	1	0.75	0.25	3.0	900	74.471%
0.5	1.5	0.5	0.25	0.125	3.0	778	74.814%	2	1.5	1	0.5	0.25	3.0	**995**	74.36%
1.5	0.5	0.2	0.1	0.05	3.0	775	74.780%	0	0	0	0	0	3.0	324	65.58%

To perform text classification we need to extend the definitions of tfidf and document vector to the evaluation corpus. Let $\mathcal{C}$ be the evaluation corpus and d a document $d \in \mathcal{C}$. We define the tfidf $t_d(w)$ of a word $w \in d$ in $\mathcal{C}$ as

$$t_d(w) := (1 + \log(f_d(w))) \cdot \log\left(\frac{\#\mathcal{C}}{\mathrm{df}(w)}\right),$$

where f_d is the frequency of w in d; $\#\mathcal{C}$ the total number of documents; $\mathrm{df}(w)$ the number of documents in $\mathcal{C}$, containing w.

Furthermore, our ESA implementation provides us with a concept vector $\boldsymbol{w}$ for every word w. We define the *document vector* $\boldsymbol{d}$ as:

$$\boldsymbol{d} := \frac{\sum_{w \in d} t_d(w) \cdot \boldsymbol{w}}{\| \sum_{w \in d} t_d(w) \cdot \boldsymbol{w} \|}.$$

where the denominator is used for normalization.

Using these vectors, text classification becomes standard classification in $\mathbb{R}^{\#\mathcal{W}}$ for the cosine metric. We applied the linear multi-class SVM classifier SVM$^{\text{multiclass}}$ [13] to the set of these vectors and the corresponding document classes, and after a tenfold cross-validation, we obtained an average precision of 65.58% for a C coefficient of 3.0. The classification required 324 support vectors. Admittedly the precision obtained is rather low, which is partly due to the thematic proximity of some classes (like, for example, Religion and Sects, or Writing and French language). However, our goal is not to compare ESA to other classification methods, but to show that our approach improves ESA. So, this result is our starting point and we intend to improve it.

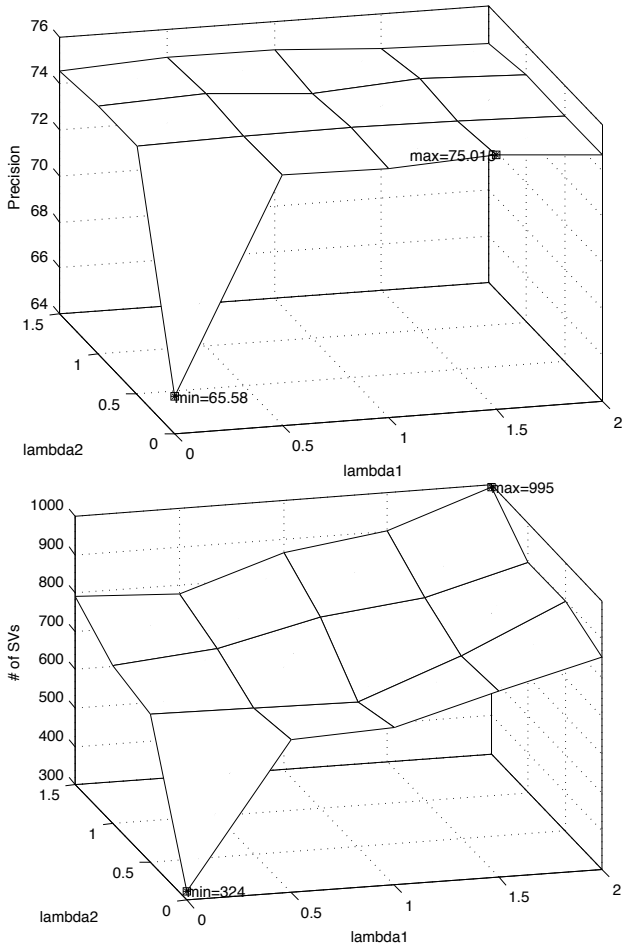


Fig. 2. Precision (to the left) and number of support vectors used (to the right), as functions of the parameters λ_1 and λ_2.

We followed the same modus operandi using thematically reinforced methods and obtained the results displayed on Table 2. The results show a significant improvement over the standard ESA version (that corresponds to $\lambda_i = 0$ for all i. This confirms our approach. In Fig. 2 the reader can see the precision obtained as function of the two first parameters λ_1 and λ_2, as well the number of support vectors used. We notice that the precision

varies slightly (between 74.36% and 75.015%, that is less than 1%) as long as λ_1 or λ_2 are nonzero, and abruptly goes down to 65.58% when they are both zero. For nonzero values of λ_i the variation of precision follows no recognizable pattern. On the other hand, the number of support vectors shows a pattern: it is clearly correlated with λ_1 and λ_2, the highest value being 995, number of support vectors used when both λ_1 and λ_2 take their highest values. Since CPU time is roughly proportional to the number of support vectors, it is most interesting to take small (but nonzero) values of λ_i so that, at the same time, precision is high and the number of support vectors (and hence CPU time) is kept small.

5 Conclusion and Hints for Further Research

By reinforcing the thematic context of words in Wikipedia pages, context obtained through the category structure, we claim to be able to improve the performance of the ESA measure.

We evaluated our method on a text classification task based on messages from the 20 most popular French language newsgroups: thematic reinforcement allowed us to improve the classification precision by 9–10%.

Here are some hints for research to be done:

1. propose the notion of the "most relevant category" to Wikipedia users and use their feedback to improve the system;
2. when we take the "most relevant category" for each page, we don't consider by how much it is better than the others. For small differences of semantic relevance weight between categories one could imagine alternative "slightly worse" spanning trees and compare the results;
3. by comparing relevance between alternative "most relevant" categories for the same page one could quantify a "global potential" of the Wikipedia corpus. Compare with Wikipedia corpora in other languages;
4. aggregate the thematically reinforced measure with collocational and ontological components, as in [3];
5. define another measure, based on links between pages (or categories), proportional to the number of links (or link paths) between pages and inversely proportional to the length of these paths. Compare it to ESA (which uses the number of links between pages to filter Wikipedia, but does not include it in semantic relatedness calculations) and thematically reinforced ESA;
6. and, more generally, explore the applications of graph theory to the formidable mathematical-linguistic objects represented by the different graphs extracted from Wikipedia.

REFERENCES

1. Gabrilovich, E., Markovitch, S.: Computing semantic relatedness using Wikipedia-based explicit semantic analysis. IJCAI'07: Proceedings of the 20th international joint conference on Artifical intelligence (2007)
2. Deerwester, S.C., Dumais, S.T., Furnas, G.W., Harshman, R.A., Landauer, T.K., Lochbaum, K.E., Streeter, L.A.: Computer information retrieval using latent semantic structure (1989) US Patent 4,839,853 of June 13, 1989.
3. Haralambous, Y., Klyuev, V.: A Semantic Relatedness Measure Based on Combined Encyclopedic, Ontological and Collocational Knowledge. In: International Joint Conference on Natural Language Processing, Chiang-Mai, Thailand (2011)
4. Gottron, T., Anderka, M., Stein, B.: Insights into explicit semantic analysis. In: CIKM'11: Proceedings of the 20th ACM international conference on Information and knowledge management. (2011)
5. Zesch, T., Gurevych, I.: Analysis of the Wikipedia category graph for NLP applications. In: Workshop TextGraphs-2 : Graph-Based Algorithms for Natural Language Processing. (2007) 1–8
6. Scholl, P., Böhnstedt, D., García, R.D., Rensing, C., Steinmetz, R.: Extended explicit semantic analysis for calculating semantic relatedness of web resources. In: EC-TEL'10: Proceedings of the 5th European conference on Technology enhanced learning conference on Sustaining TEL: from innovation to learning and practice, Springer (2010)
7. Collin, O., Gaillard, B., Bouraoui, J.L.: Constitution d'une ressource sémantique issue du treillis des catégories de wikipedia. In: TALN 2010. (2010)
8. Leacock, C., Chodorow, M.: Combining local context and WordNet similarity for word sense identification. In Fellbaum, C., ed.: WordNet, an electronic lexical database, The MIT Press (1998) 266–283
9. Gabow, H.N., Galil, Z., Spencer, T., Tarjan, R.E.: Efficient algorithms for finding minimum spanning trees in undirected and directed graphs. Combinatorica **6** (1986) 109–122
10. Çallı, Ç.: Improving search result clustering by integrating semantic information from Wikipedia. Master's thesis, Middle East Technical University, Ankara (2010)
11. Newman, M.: Networks. An Introduction. Oxford University Press (2010)
12. Medelyan, O., Legg, C., Milne, D., Witten, I.H.: Mining meaning from Wikipedia. International Journal of Human-Computer Studies **67**(9) (2009) 716–754
13. Joachims, T.: Making large-scale SVM learning practical. In Schölkopf, B., Burges, C., Smola, A., eds.: Advances in Kernel Methods - Support Vector Learning, MIT Press (1999)

YANNIS HARALAMBOUS
INSTITUT MINES-TÉLÉCOM - TÉLÉCOM BRETAGNE,
AND LAB-STICC UMR CNRS 6285,
TECHNOPÔLE BREST-IROISE, CS 83818,
29238 BREST CEDEX 3, FRANCE
E-MAIL: <YANNIS.HARALAMBOUS@TELECOM-BRETAGNE.EU>

VITALY KLYUEV
SOFTWARE ENGINEERING LABORATORY,
UNIVERSITY OF AIZU,
AIZU-WAKAMATSU, FUKUSHIMA-KEN 965-8580, JAPAN
E-MAIL: <VKLUEV@U-AIZU.AC.JP>

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 95–110
RECEIVED 30/11/12 ACCEPTED 11/01/13 FINAL 11/03/13

The Matrix of Beliefs, Desires and Intentions – Sentence by Sentence

NOÉMI VADÁSZ, JUDIT KLEIBER, AND GÁBOR ALBERTI

University of Pécs, Hungary

ABSTRACT

This paper is grounded in the dynamic semantic [7] model ℜeALIS [1] about human interpreting 'minds' as they are in communication with each other. Following in the footsteps of studies [3-4], here we offer a text analysis method which proceeds from sentence to sentence and thus gradually opens up the intensional status of the information as it is obtained by the hearer. 'Matrix' here refers to a combination of a pragmatic text analysis (e.g. through the formalization of Grice's approach [5]) and the intensional messages of linguistic clues [3–4]. Within the matrix, the elements of intensionality cease to exist as sporadic 'specialties'. Rather, an inherent part of the semantic content of each given sentence is the information concerning what beliefs (and each with what level of certainty), desires and/or intentions the speaker has, as well as what he/she thinks in the same respect about his/her conversational partner, and also what the partner thinks of him/her correspondingly, and so on. In an implementation of ℜeALIS, numerical matrices were developed [2], which produce the truth-conditional interpretation of the sentences that are attributed to particular agents as speakers at certain moments. This method makes it possible to interpret various opinions connected to the sentences – opinions like "This has been a (white) lie / a bluff," or "The speaker has killed the joke."

KEYWORDS: *representational dynamic discourse semantics, intensionality, information state*

1 Introduction

This paper is grounded in the dynamic semantic [7] model ℜeALIS [1] about human interpreting 'minds' as they are in communication with each other. Following in the footsteps of studies [3–4], here we offer a text analysis method which proceeds from sentence to sentence and thus gradually opens up the intensional status of the information as it is obtained by the hearer. 'Matrix' here refers to a combination of a pragmatic text analysis (e.g. through the formalization of Grice's approach [5]) and the intensional messages of linguistic clues [3–4].

Within the matrix, the elements of intensionality cease to exist as sporadic 'specialties'. Rather, an inherent part of the semantic content of each given sentence is the information concerning what beliefs (and each with what level of certainty), desires and/or intentions the speaker has, as well as what he/she thinks in the same respect about his/her conversational partner, and also what the partner thinks of him/her correspondingly, and so on.

2 Formalization

Let us take a simple example to evoke the theory put forth in [1] and the technical apparatus presented in our 2012 CICLing publication [4]: *Mary is at home*. Here the 'primary' segment of the information state (Γ_s^0) in Grice's ideal speaker holds that eventuality *e*, registering Mary's being at home, is thought honestly true by him/her (speaker *s*). With the formal apparatus of ℜeALIS, this piece of knowledge can be captured in the representation of the speaker's mind as a 'worldlet', which can be characterized by the following five-item label: ⟨BEL, max, s, τ, +⟩. The first parameter (in this case 'BEL') shows modality. 'MAX' indicates a higher level of belief or belief with the power of "knowledge". Symbol *s* refers to the speaker; τrefers to time; while + refers to a possible polarity (π_1). (In a later phase of the research, we will introduce further parameters for emotion and style.) Compared to the above, the relevant segment of the information state (Γ_i^0) in the interpreter who enters the conversation in an "ideal" manner can be described as follows:

(1) $\Gamma_i^0=\{\langle BEL,max,i,\tau,0\rangle, \langle DES,great,i,\tau,+\rangle\langle BEL,max,i,\tau',\theta\rangle\}$

Conversational partner *i* is therefore not aware of the eventuality being true or not (π=0) but has a strong desire (DES) for it to turn out.

This is what symbol '0' (drawn zero) refers to. It should be noted that the gratification of the above desire in a latter phase of the communication is represented in the hearer's mind by the appearance of a + or a – in the place of the drawn zero. It would be a mistake, however, to infer from this a "dangerously four-rated" background-logical calculus: we do not suggest that an operational chart can be directly assigned to these four rates. Pragmatic rules can only be set up for complex lists of label-series.

We note here for those well-acquainted with logics that the rules to be provided here concern, in general terms, those sub-structures present in the cognitive network of information states which can be formalized. This way, we aim to sidestep the logical approaches which to a linguist might seem too "sterile", idealized or simplified.

The next level of representation shows that the information states of the conversational parties contain a great number of assumptions (of different states) about their partner's internal worlds. The speaker intends to alter the hearer's information state by letting him/her know that *e* is true. In addition, he/she makes it probable ('great') that his/her hearer is an "ideal" one in the earlier sense of the word (namely that he/she can be described with the start-out information state (Γ_i^0):

$$(2) \qquad \Gamma_s^1=\{\langle \mathrm{INT,max,s,\tau,(\pi_2=)+}\rangle\langle \mathrm{BEL,max,i,\tau',+}\rangle\} \cup \{\langle \mathrm{BEL,great,s,\tau,+}\rangle\} \wedge \Gamma_i^0$$

The second segment of the hearer's information state can be represented in a similar way:

$$(3) \qquad \Gamma_i^1=\{\langle \mathrm{BEL,great,i,\tau,+}\rangle\} \wedge (\Gamma_s^1 \cup \Gamma_s^0)$$

Formulas are applied as formal representations of the n^{th} segment of both the speaker's and the hearer's information states, in a most general way:

$$(4) \qquad \Gamma_s^n=\{\langle \mathrm{BEL},1/(n+1),\mathrm{s,\tau,+}\rangle\} \wedge \Gamma_i^{n-1},$$
$$\Gamma_i^n=\{\langle \mathrm{BEL},1/(n+1),\mathrm{i,\tau,+}\rangle\} \wedge \Gamma_s^{n-1}.$$

The deeper the recursion is, the smaller the fraction. This suggests a decrease in the intensity of knowledge – namely that we have increasingly vague ideas about information contained by segments of information states which are farther and farther away from the initial segments. In an actual communication situation the participants can barely rely on $n > 2$ cases. In formal (generative) linguistics, nevertheless, it is not

proper to exclude competence to a deeper recursion already at the start. The potentially unlimited union of the unrestrained (as for *n*) appropriate segments evokes the absolute sum of the speaker and the hearer's relevant information in the actual communication situation: $\cup\Gamma_s^n$ and $\cup\Gamma_i^n$.

So far we have described an ideal communication situation – ideal in the Gricean or in a post-Gricean sense. The chart below demonstrates that changes to the parameters of certain 'worldlet' labels can also capture non-ideal communication situations such as in the cases of misleading or lying.

Although the concept of an 'ideal communication situation' is widely used in pragmatics, it is very difficult to pinpoint. For a communication situation to be ideal it takes ideal partners (a hearer and a speaker) and ideal circumstances.

In the present paper, the meaning of 'ideal' shall be extended beyond the Gricean sense. 'Ideal' here means some kind of smoothness when nothing disturbs the smooth flow of conversation. Course books and foreign language books, for example, typically feature ideal speech situations to illustrate humorless but easy-to-process discourse. The reason why it is crucial to mark off the case of an ideal communication situation is because all other (deviant) cases can be correlated to it; this, then, makes it possible to allocate all the different situations in one system. Grice's theory and maxims come handy when one wants to demonstrate what the ideal situation is like since they provide a good enough definition for the "obligations" of the speaker who does not wish to upset the flow of this more or less humorless conversation in any way.

Earlier it has been said that the conversational parties aim to keep to a common goal. Now this may be misleading if taken in the strict sense: it may well be that it takes a certain degree of non-ideality for human conversations to be diverse in nature. It can be well presumed that most of the conversations one encounters day by day do not conform to the genuinely, *per definition* ideal standard. Also, most of the ideal communication situations are to be found in formal contexts (which are not devoid of misunderstandings, either). When one is talking to his/her immediate friends, he/she economizes on very little information in order to save time for both parties – almost to an extent of breaching Grice's maxim of quantity. Very probably, many of us have had the feeling of hardly being capable to provide answers to our partners that would be long and detailed enough. Oftentimes, we may have the feeling of only being capable of hurling fragments of information

on the other one (while breaking the maxim of manner) in the hope that every situation becomes clear at some point. Again, our answers to certain questions may not satisfy the needs of the questioner at all; we use them in order to dissuade the questioner from further questioning (e.g. “What did you have for lunch at school?” “A first and a second course.”).

Yet, every discourse features the common goal above somehow – except that in most of them this goal is not reached in a straightforward way. It has been mentioned earlier that the concept of an ideal speaker and hearer is necessary to allocate ideal and various non-ideal situations in one system. Although it is a daring idea to divide all communication situations into ideal and non-ideal, this is a necessary step to take here. Cases on the vague borderline between the two types of situations will not be addressed here; ‘ideal’ in this paper shall refer to a speaker and a hearer as they were specified above, while all other behavior of the speaker and the hearer shall be perceived as ‘non-ideal’. The present paper focuses on cases where the speaker improperly or poorly identifies the desires of his/her conversational partner and where he/she misleads their partner on purpose. In addition, the paper will also attempt to account for the mistakes of the hearer.

3 Polarity

In what follows, Π (marker of polarity) is replaced by +, –, 0 or ϴ. These changes enable the system of formalization to handle different non-ideal communication situations.

(5) $<BEL,max,s,\tau,\Pi^1>$

The speaker’s knowledge of e

(6) $<BEL,great,s,\tau,\Pi^2><BEL,i,\tau,\Pi^3>$

The speaker’s knowledge about the hearer’s knowledge of e

(7) $<BEL,great,s,\tau,\Pi^4><DES,i,\tau,\Pi^5><BEL,i,\tau',\Pi^6>$

The speaker’s knowledge about the hearer’s desire of e

(8) $<INT,max,s,\tau,\Pi^7><BEL,i,\tau',\Pi^8>$

The speaker’s intension of the hearer’s knowledge of e

The chart below (9) shows the differences between the various formalized situations as regards changes in polarity. The chart makes it easy to assess the differences between the polarity adjustments of various situations. It can be seen, for instance, that all cases of misleading (from concealing information to lying) share the fact that there is a difference in the polarities of their parameters Π^1 and Π^8. This indicates a difference between what the speaker knows and what the speaker desires the hearer to (not) know. In other words, the speaker is expected to pass on information to the hearer about which he/she is convinced as being not true or about which he/she is not convinced as true. It may also happen that the speaker provides the hearer information that the latter one does not desire to have.

(9) Situations and ℜeALIS Polarity Values:

Π^1	Π^2	Π^3	Π^4	Π^5	Π^6	Π^7	Π^8	SITUATIONS
+	+	0	+	+	~~0~~	+	+	Ideal
+	+	+	+	0	~~0~~	+	+	Clarification
+	+	–	+	0	~~0~~	+	+	Correction
+	+	0	+	+	~~0~~	+	0	Concealing
0	+	0	0	+	~~0~~	+	+	Bluff
+	+	0	+	0	~~0~~	+	–	Fib
+	+	0	+	+	–	+	–	White lie
+	+	0	+	+	~~0~~	+	–	Lie

In the first two situations neither misleading nor lying takes place on the part of the speaker. By changing the variability of the polarities, however, very interesting situations can be illustrated – such as, for example, when the speaker corrects the hearer or clarifies the information that both of them have.

4 Clarifications

Making a clarification can be easily captured in the words ‘so’ and ‘OK, so’. The hearer and the speaker have the same information but the speaker finds it important to clarify this fact (to avoid later misunderstandings).

(10) Γ_s^0 ={<BEL,max,s,τ,+>}
Γ_s^1 ={<BEL,great,s,τ,+><BEL,i,τ,+><BEL,great,s,τ,+>
<DES,i,τ,0><BEL,i,τ',~~θ~~><INT,max,s,τ,+><BEL,i,τ',+>}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

5 Corrections

Making a correction is similar to making a clarification. The speaker knows that the hearer is wrong so the former corrects the latter. This is what the words 'yes, indeed' indicate in the dialogue.

(11) Γ_s^0={<BEL,max,s,τ,+>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,->,<BEL,great,s,τ,+>
<DES,i,τ,0><BEL,i,τ',~~θ~~>,<INT,MAX,S,T,+><BEL,I,T',+>}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

6 The Speaker Kills the Joke

The speaker kills a joke when he/she shares a piece of information too early with the hearer, which he/she only wanted to find out later on. The hearer wants to get a certain piece of information about e only in a later moment of time; the speaker, however, driven by an ill purpose, disrespects this want on the listener's part. A classic example of this is "spoilerism", when someone deliberately hints information on the plot of a book or film that the other one has not yet read or seen. Although no lying or misleading takes place here, the situation is far from ideal.

(12) Γ_s^0={<BEL,max,s,τ,+>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,->,<BEL,great,s,τ,+>
<DES,i,τ,-><BEL,i,τ',~~θ~~>, <BEL,great,s,τ,+><DES,i,τ,+>
<BEL,i,τ'',~~θ~~>, <INT,MAX,S,T,+><BEL,I,T',+>}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

7 Concealing

When the speaker wants to conceal something from the hearer, no lying takes place, but the speaker misleads the hearer.

(13) Γ_s^0={<BEL,max,s,τ,+>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,0>,<BEL,great,s,τ,+>
<DES,i,τ,+><BEL,i,τ',~~0~~>, <INT,max,s,τ,+><BEL,i,τ',0>}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

Another subtype of concealing is when the speaker – breaking the maxim of quantity (and relevance) – "talks the hearer's arm off". Here, since the information being passed is true, no lying takes place, but the information is such that the hearer does not want to know or which does not add to the conversation. The hearer expects certain information but instead of getting that, he/she gets another piece of information, which is true but that does not matter for the hearer. A wide range of linguistic (slang) expressions are available to describe this activity (talk someone's head/arm/pants off, talk the bark off the tree, talk a blue streak, beat about the bush…). One may sense slight semantic differences between these expressions, but those might well only result from individual language use.

8 Bluffs

When a speaker bluffs, he/she gives the hearer information whose truth he/she is not confident about. He does so to fulfill his/her (ill) purpose. This interpretation of bluffing is different from the term as it is used in

poker (where it is a simple lie without words). Concentrating instead on the colloquial use of the word 'bluff', it can be explained as follows.

(14) Γ_s^0={<BEL,max,s,τ,0>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,0>,<BEL,great,s,τ,+>
<DES,i,τ,+><BEL,i,τ',θ>, <INT,max,s,τ,+><BEL,i,τ',+>}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

We should note here though that 'bluff' is also used in the following sense: In a less formal situation of a job-interview, the applicant may bluff about his/her English language proficiency or about his/her earlier experiences in order to get the job, knowing at the same that there will be no chance for this little 'fib' to turn out later on. Children's bluffs may be considered a similar case. While parents may often catch their kids bluffing, they do not expose each and every one of their bluffs since those are perceived as a natural accompaniment (up to a certain aspect) to being a child.

9 Fibs

Telling stories or fibbing is a case of lying: the speaker gives the hearer some information while he/she is convinced that it is not true. The difference between a fib and a lie is hard to tell. It is not only the level of seriousness that makes a difference between them. In risk-free fibbing, the speaker assumes that the hearer does not need the information at all (this lends fibbing a risk-free nature). For example, a husband can tell his wife without any risks that he has paid the bills (while he has not) because he can do it the next day and with this, he can straighten out the pity lie. The fibber is not driven by anything bad; he acts to protect his face.

(15) Γ_s^0={<BEL,max,s,τ,+>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,0>,
<BEL,great,s,τ,+><DES,i,τ,0><BEL,i,τ',θ>,
<INT,max,s,τ,+><BEL,i,τ',->}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

10 White Lies

White lies also belong to lying. It is not their risk-free nature that distinguishes white lies from real lies but the intention behind them. The speaker so-called "maps" the hearer's desire about *e* and tries to satisfy it.

Speaking of white lies, doctor-patient dialogues readily come into mind. Suppose a patient has very little chance of recovering. Knowing this fact may even worsen his/her well-being, so neither the doctor nor the family communicates this information to him/her. It can also be regarded as a white lie when someone compliments on someone else on their hair style (even though it looks awful) to make them feel good. Looking at the motifs and goals behind white and non-white lies, it becomes clear what makes them different: white lies are generally governed by good intentions – as opposed to lies, which are cases of pure crime. Here, too, formal explanation comes from a reason beyond form, which at the same time makes the phenomenon ready to be formalized. The hearer wants to get hold of a piece of information and – although the opposite of the information-content is true – the speaker chooses to satisfy his/her partner's desire rather than comply with the maxim of quality and not lie.

(16) Γ_s^0={<BEL,max,s,τ,+>}

Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,0>,
<BEL,great,s,τ,+><DES,i,τ,+><BEL,i,τ',->,
<INT,max,s,τ,+><BEL,i,τ',->}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to believe, in a later moment, that Mary is not at home. The speaker intends the hearer to believe that Mary is not at home.

11 Lies

A clear-cut case of lying is when the speaker gives the hearer some information while he/she is convinced that it is not true.

(17) Γ_s^0={<BELmax,s,τ,+>}
Γ_s^1={<BEL,great,s,τ,+><BEL,i,τ,0>,
<BEL,great,s,τ,+><DES,i,τ,+><BEL,i,τ',~~0~~>,
<INT,max,s,τ,+><BEL,i,τ',->}

The speaker knows that Mary is at home. The speaker strongly believes that the hearer does not know whether Mary is at home or not. The speaker strongly believes that the hearer would like to know, in a later moment, that Mary is at home or not. The speaker intends the hearer to believe that Mary is not at home.

12 Implementation

In an implementation of ℜeALIS, numerical matrices were developed by our close colleagues [2] to produce the truth-conditional interpretation of the sentences that are attributed to particular agents as speakers at certain moments. Due to the advantageous feature of ℜeALIS to represent the interpreters' minds as maps containing the labels discussed above as guideposts [4], the method makes it possible to interpret various opinions connected to the sentences – opinions like "This has been a (white) lie / a bluff," or "The speaker has killed the joke". The program simply has to seek the guideposts for the appropriate configurations of polarity values.

In what follows, the relevant properties of the implemented interpretation system are sketched out.

Instead of input sentences, it is better to choose the model of the external world for a starting point. The model consists of relations, each of which has time intervals as one type of its arguments. The relation corresponding to the verb *snow*, for instance, is a binary one which associates time intervals with spatial entities (i.e. it is given when it snows where). The relation corresponding to the adjective *bald* associates time intervals and entities which correspond to people. *Live* (or *be somewhere*) is a tertiary relation with the following types of entities as arguments: a person, a spatial entity and a time interval. *Know* (*somebody*) is also a tertiary relation with two persons and a time interval as its argument types.

The fact that a certain *n*-tuple of entities can be found in a certain relation is defined as an *infon* ([8]:242). Truth-conditional evaluation primarily relies on infons. The sentence *It is snowing* is true, for instance, if its performance is attributed to a moment *t* of time and a person whose location is a spatial entity *s*, where there is an infon of *snowing* with a time interval *T* containing *t* and a place *S* containing *s*. Similarly, the sentence *Peter knows Mary* is true if its performance is attributed to a moment *t* of time (as well as a speaker and an addressee) so that there is an infon of *knowing* with a time interval *T* containing *t* and the entities assigned to the names in question in the appropriate order (the explanation of the complex way in which this assignment is dependent on the speaker's and the addressee's information states is beyond the scope of this paper).

In an elegant linguistic model, the truth of the sentence *Mary is at home* does not directly rely on one single infon but (at least) on two infons and a meaning postulate, saying that *x is at home if x is in a place s such that x lives in s* (the knowledge of the meaning postulate is held to belong to the selected speaker's information state).[1]

In ℜeALIS [1], each "interpreter" (human being) is an entity of the world model and has further entities ("internal" ones) at his/her disposal. Situations (1-17) illustrated some varieties (of certain parts) of an interpreter's labeled network which expresses his/her momentary information state. What is crucial is that these internal networks (are defined so that they) also belong to the system of relations of the world model.

In a realistic implementation of ℜeALIS, each interpreter's information state at point *t* of time can be regarded as a modified (partial) copy (or "photograph") of the "active" infons (whose time intervals contain

[1] All heterogeneous events (e.g. *travel home* or *lose weight*) are to be evaluated via meaning postulates based upon homogeneous events that have direct connections to infons because it is more economical to define the external world model by means of a meager ontology, in which infons correspond to homogeneous events. The definition of *travel home*, for instance, can rely on eventualities associated with "earlier" and "later" points of time: e.g. *x travels home if x is travelling at t', and x intends to be at home, and x is at home at a later moment t"*, etc. The system for the components of this meaning postulate can be broadened in a sophisticated linguistic model, but the application of certain components will depend on tense and aspect. The sentence *Mary was travelling home*, for instance, does not require satisfaction of the last component mentioned above (*x is at home at a later moment t"*) to be true; the intention also mentioned above is essentially enough ([5]:147, *Imperfective Paradox*).

t). A perfect copy would mean that an interpreter's information state is such that it contains a corresponding eventuality referent for each active infon, represented in the following trivial worldlet-label family: {⟨BEL,max,s,τ,+⟩}. This would mean a supernatural interpreter who would be aware of all current facts of the on-going external world. Users of our software can apply this "oracle"-mode but they can also choose to modify the worldlet-label families associated with eventuality referents to develop realistic interpreters. In the case of a realistic interpreter, the label family is only associated with an eventuality (except for cases with a small set of eventualities): {⟨BEL,max,s,τ,0⟩}. This means that the given interpreter knows nothing about the eventualities in question; an ordinary human being only has a *partial* snapshot of the surrounding world. In the case of a small set of eventualities, however, each eventuality in the information state of the realistic interpreter is associated with at least as complex worldlet-label families as those shown in (1-17). Instead of, or in addition to, knowing that *e* is true, the interpreter knows, for instance, that another person believes that *e* is false, and/or (s)he wants this person to believe that *e* is true, and/or (s)he wishes that a third person would intend to convince the second person that *e* is true, etc. What we argue here is that a human being has a *modified* snapshot of the surrounding world, compared to the perfect picture at a potential oracle's disposal.

This approach, thus, provides a manifold mirroring of external relations. The capriciously modified images and the genuine relations all belong to the same relational system whilst, due to what we call worldlet-labels, the internal status of each "image" is precisely defined and is detectable within a particular information state which belongs to a particular interpreter. This results in the fact that the evaluation of sentences (18a–c) – discussed above – is not significantly simpler than the evaluation of the sentences shown in (19) (whose interpretation requires an intensional apparatus in other logical systems).

(18) a. It is snowing.
b. Peter knows your brother.
c. Polly is at home.

(19) a. Peter believes that it is snowing.
b. Peter has discovered that it is snowing.
c. According to Ann, Brian believes
that Cecil wants Mary to be at home.
d. According to Brian, Mary is pretty.
e. Mary is pretty.

The evaluation of sentences in (18) required *pattern matching* which pertains to active infons of the external world, where activity can be defined on the basis of when (time), where (place) and by whom (speaker) the given sentence is performed. What (18a) illustrates excellently is that no sentence can be evaluated without attributing its performance to an interpreter with an entirely elaborated information state. (18b) shows that the addressee should also be decided on for truth-conditional evaluation. In (18c), Polly is used as a nickname for Mary; we intend to call the reader's attention to the fact that it is a prerequisite for the evaluation of sentence (18c) that (in the ideal case) both the speaker and the addressee use this nickname for a certain Mary.

The evaluation of sentences in (19) also requires pattern matching. The only difference is that the appropriate patterns should not (only) be detected in the area of infons but also in other areas of the relational model of the entire world.

In (19a), it is irrelevant if it is snowing "outside"; what matters is that a certain segment of the information state belongs to a person who is a 'unique' Peter to the speaker in the given context.

(19b) requires a more complex investigation. The external world also matters: it must be snowing outside; and Peter's two information states at two points of time should be searched. It is required that in the earlier state, but not in the later state, the eventuality of snowing is *not* associated with a label like this: $\langle$BEL,max,s,τ,+$\rangle$.

The evaluation of sentence (19c) essentially requires the discovery of a special segment of the internal network of an interpreter. We should enter the information state of a person who is known by the speaker as Ann; then we should find this Ann's beliefs concerning Brian's beliefs, especially those concerning Cecil's wishes. This is a long path but it also ends in pattern matching.[2]

(19d) is to be evaluated on the basis of the information state of the person known as Brian to the speaker and chosen by the user of our program. What we would like to illustrate here is that in the case of an intensional predicate like *pretty*, it is easier to evaluate somebody's opinion than to evaluate a seemingly objective proposition like the one in (19e). Our solution relies on the approach that no infon corresponds to *pretty*, but the interpreters' current opinions should be searched. To be pretty means to be pretty according to the majority. In a more sophisticated approach, which is easily available in our system, the in-

[2] Here again, the problem of names/nicknames arises and probably results in ambiguity. We do not enter into details here.

formation states of those assumed to be known (and respected) by the selected speaker should be looked at; some relative majority will decide on the question of prettiness.

At the end of the section, let us note that the speaker-dependent truth-conditional evaluation of sentences like those in (19) requires the same background architecture as the evaluation of certain pragmatic reactions attributed to the hearer such as in "This has been a (white) lie / a bluff" etc. Here – as in the latter example – we have a further typical case to investigate: the case of worldlet-label families that are associated with (a) given eventuality referent(s) in the speaker's information state(s) (probably in addition to infons).

13 Further Goals

Going farther in and deeper down in the levels of recursion (cf. sections 4–11), there are several further situations waiting to be formalized, beyond the scope of the present paper. One of our further goals, in fact, includes the formalization of longer dialogues with several turns.

As for real-life implementation, our system could be used, for instance, to make a judge's work easier. The information state of each party concerned in a case could be registered at any selected point of time (e.g. as regards their beliefs and intentions related to external facts and/or one another); in order to prove, for example, that a given person could not have been aware of a given fact at a given point of time

ACKNOWLEDGEMENT. We are grateful to SROP-4.2.1.B-10/2/KONV/2010/ KONV-2010-0002 (Developing Competitiveness of Universities in the Southern Transdanubian Region) for their contribution to our costs at CICLing (2013, Samos) and for supporting our Research Team ℜeALIS in 2012. In 2013 ℜeALIS is supported by SROP-4.2.2.C-11/1/KONV-2012-0005 (Well-Being in the Information Society); the final version of this paper is due to this project.

References

1. Alberti, G.: ℜeALIS: An Interpretation System which is Reciprocal and Lifelong. Workshop 'Focus on Discourse and Context-Dependence' (16.09.2009, 13.30-14.30 UvA, Amsterdam Center for Language and

Comm.), http: // www.hum.uva.nl / aclc / events.cfm / C2B8E596-1321-B0BE-6825998CFA642DB2, http://lingua.btk.pte.hu/realispapers (2009)
2. Alberti, G., Károly, M., Kilián, I., Kleiber, J., and Vadász, N.: The moment of truth – or the anchoring function α of ℜeALIS from the scope of Hungarian [in Hungarian]. In: Vincze, V, and Tanács, A. (eds.): Ninth Hungarian Conference on Computational Linguistics. Dept of Informatics, Univ. of Szeged, Hungary, pp. 236–250 (2013)
3. Alberti, G. and J. Kleiber: Where are Possible Worlds? (Arguments for ℜeALIS). Acta Linguistica Hungarica 59 (1-2) (ed. Katalin É. Kiss). 3–26 (2012)
4. Alberti, G., and M. Károly: Multiple Level of Referents in Information State. A. Gelbukh (ed.): Computational Linguistics and Intelligent Text Processing, CICLing2012, New Delhi, India. Lecture Notes in Computer Science LNCS7181. Springer Verlag, Berlin, Heidelberg, pp. 34–362 (2012)
5. Dowty, D. R.: *Word Meaning and Montague Grammar*. Reidel, Dordrecht (1979)
6. Grice, P.: Logic and conversation. In: Cole, P. & Morgan, J. (eds.): Syntax and Semantics 3. New York, Academic Press, pp. 41–58 (1975)
7. Kamp, H., van Genabith, J., Reyle, U.: Discourse Representation Theory. In Handbook of Philosophical Logic, vol. 15, pp. 125–394. Springer-Verlag, Berlin (2011)
8. Seligman, J., Moss, L. S.: Situation Theory. In van Benthem, J., and ter Meulen, A. eds.: Handbook of Logic and Language. Elsevier, Amsterdam, MIT Press, Cambridge, Mass, pp. 239–309 (1997)

Noémi Vadász
Research Team ℜeALIS for Theoretical,
Computational and Cognitive Linguistics,
Department of Linguistics, University of Pécs,
Ifjúság 6, H-7624 Pécs, Hungary
E-mail: <vadasznoemi@gmail.com>

Judit Kleiber
Research Team ℜeALIS for Theoretical,
Computational and Cognitive Linguistics,
Department of Linguistics, University of Pécs,
Ifjúság 6, H-7624 Pécs, Hungary
E-mail: <kleiber.judit@pte.hu>

Gábor Alberti
Research Team ℜeALIS for Theoretical,
Computational and Cognitive Linguistics,
Department of Linguistics, University of Pécs,
Ifjúság 6, H-7624 Pécs, Hungary
E-mail: <alberti.gabor@pte.hu>

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 111–126
RECEIVED 29/10/12 ACCEPTED 11/01/13 FINAL 07/03/13

A Self-Training Framework for Automatic Identification of Exploratory Dialogue

ZHONGYU WEI,[1] YULAN HE,[2] SIMON BUCKINGHAM SHUM,[3]
REBECCA FERGUSON,[3] WEI GAO,[4] AND KAM-FAI WONG[5]

[1] *The Chinese University of Hong Kong, Hong Kong*
[2] *Aston University, UK*
[3] *The Open University, UK*
[4] *Qatar Foundation, Qatar*
[5] *Key Laboratory of High Confidence Software Technologies, China*

ABSTRACT

The dramatic increase in online learning materials over the last decade has made it difficult for individuals to locate information they need. Until now, researchers in the field of Learning Analytics have had to rely on the use of manual approaches to identify exploratory dialogue. This type of dialogue is desirable in online learning environments, since training learners to use it has been shown to improve learning outcomes. In this paper, we frame the problem of exploratory dialogue detection as a binary classification task, classifying a given contribution to an online dialogue as exploratory or non-exploratory. We propose a self-training framework to identify exploratory dialogue. This framework combines cue-phrase matching and K-nearest neighbour (KNN) based instance selection, employing both discourse and topical features for classification. To do this, we first built a corpus from transcripts of synchronous online chat recorded at The Open University annual Learning and Technology Conference in June 2010. Experimental results from this corpus show that our proposed framework outperforms several competitive baselines.

KEYWORDS: *Exploratory dialogue identification, self-training, K-nearest neighbour, classification.*

1 INTRODUCTION

Exploratory dialogue is a form of discourse associated with deep learning and learners engaging with each other's ideas constructively. It is desirable because prompting learners to employ this type of dialogue has been shown to improve learning outcomes. [1] defined exploratory dialogue as follows:"*Exploratory dialogue* represents a joint, coordinated from of co-reasoning in language, with speakers sharing knowledge, challenging ideas, evaluating evidence and considering options in a reasoned and equitable way."

Since exploratory dialogue has been shown to be a productive type of dialogue in which knowledge is made publicly accountable and reasoning is visible, the study of exploratory dialogue identification has attracted increasing attention from learning analytics researchers. Mercer et al. [2] originally conducted research on dialogue collected in face-to-face settings and identified exploratory dialogue as a type of learner talk including elements such as evaluation, challenge, reasoning and extension. Ferguson and Buckingham Shum [3] analysed transcripts from online conferences to identify exploratory dialogue. They found that markers of exploratory dialogue can be used to distinguish meaningfully between discussions and to support evaluation of them. They manually identified 94 words and phrases that signaled the presence of elements of exploratory dialogue. Examples of cue phrases for exploratory dialogue include "*but if, my view, I think, good example, good point, that is why, next step*".

Table 1 shows an excerpt from an online discussion about distance learning. Apart from those contributed by "user3", all postings are classified as *Exploratory*. Words highlighted in italics are discourse cues indicating exploratory dialogue.

Table 1. Examples of exploratory and non-exploratory dialogue.

User Id	Postings	Label
user1	*I also think* opening up the course production and design process is the way to go, but it will be a big culture change!	Exploratory
user2	*I agree with* user1 - but there are so many drivers, not least money.	Exploratory
user3	Audio back to normal speed for me now.	Non-Exploratory
user4	*I think* the key is teachers recognising that their skills lie in Learning Design, in all its variations.	Exploratory

The obvious drawback of such a cue-phrase based approach is that it is not possible to enumerate all the possible key phrases signaling the presence of exploratory dialogue. Indeed, our preliminary experiments on the online conference dataset show that the cue-phrase based approach gives high precision but low recall. In this paper, instead of using cue-phrase based methods, we investigate machine learning approaches to the automatic identification of exploratory dialogue. The three main challenges we face are:

- Firstly, the annotated dataset is limited. Although there are abundant online discussions on a wide range of topics, there are almost no annotated corpora specifically designed for detection of exploratory dialogue. This lack of annotated data corpora makes it impractical to use supervised learning methods.
- Secondly, exploratory dialogue is a form of discourse indicating that learning is likely to be taking place and that learners are going beyond a simple accumulation of ideas. Discourse features are therefore important indicators signaling the existence of exploratory dialogue. The high precision results we obtained from our collected online conference corpus using the cue-phrase based method also reveal the significance of discourse features. Discourse-based classification is intrinsically different from traditional text classification problems which are typically topic driven.
- Thirdly, although the content of online learning discussions may cover a range of topics, knowing the discussion topics in a particular dialogue segment could help with the detection of exploratory dialogue. For example, in the case of two postings extracted from an online discussion forum on the topic of "cloud computing" as shown below, both contain cue phrases indicating the presence of exploratory dialogue (these cue phrases are highlighted in italics). However, only the first posting is a positive example of exploratory dialogue. The second posting deals with an off-topic issue. This implies that both discourse and topical features should be considered when identifying exploratory dialogue.
  Posting 1: *I disagree*. Freemind is superb to use for cloud computing.
  Posting 2: I would like to join you for dinner, *but if* my wife comes home earlier, I will not make it.

In this paper, we treat exploratory dialogue detection as a binary classification problem that is concerned with labeling a given posting as ex-

ploratory or non-exploratory. To address the three challenges outlined above, we propose a SElf-training from Labeled Features (SELF) framework to carry out automatic detection of exploratory dialogue from online content. Our proposed SELF framework makes use of a small set of annotated data and a large amount of un-annotated data. In addition, it employs both cue-phrase matching and KNN-based instance selection to incorporate discourse and topical features into classification model training. The SELF framework makes use of self-learned features instead of pseudo-labeled instances to train classifiers by constraining the models' predictions on unlabeled instances. It avoids the incestuous bias problem of traditional self-training approaches that use pseudo-labeled instances in the training loop. This problem arises when instances are consistently mislabeled, which makes the model worse instead of better in the next iteration.

2 RELATED WORK

Exploratory Dialogue Detection: Research into exploratory dialogue originates in the field of educational research, where this type of dialogue has been studied for more than a decade. In face-to-face settings, Mercer and his colleagues [4, 1] distinguished three social modes of thinking used by groups of learners: disputational, cumulative and exploratory. They proposed that exploratory dialogue is the type considered most educationally desirable [5].

Ferguson et al. [3] explored methods of detecting exploratory dialogue within online synchronous text chat. They manually identified a list of cue phrases indicative of the presence of exploratory dialogue. Despite the identification of these phrases, this manual approach cannot easily be generalised to other online texts.

Apart from detecting exploratory dialogue within online and offline discussions, there has also been research [6, 7] into different approaches to the detection of exploratory sections of texts. In particular, this research has focused on science papers and feedback reports. This context is different to that of chat because such documents are usually grammatically correct, carefully punctuated and formally structured.

Dialogue Act Detection: Since exploratory dialogue detection can be carried out using discourse cues, it is closely related to dialogue act classification, which aims to analyze the intentions of the speaker, for example instruction or explanation. Samuel et al. [8] identified a number of cue phrases automatically and showed these can be powerful indicators

of the associated dialogue acts. Webb et al. [9, 10] explored the use of cue phrases to carry out direct classification of dialogue act.

Using manually annotated datasets such as *Verbmobil* [11], many supervised machine learning approaches have been applied to dialogue act recognition, including Hidden Markov Models [12], the language model [13], Bayesian networks [14], Decision Trees [15] using features including n-grams, syntactic tags (such as dependency parse chunks or part of speech tags), and pragmatic information.

Self Training from Labeled Features: Traditional self training approaches employ *self-labeled instances* in the training loop. Although the current model might be improved by adding self-labeled examples with the highest confidence values generated at each iteration, this is not the case because instances might be mislabeled, making the model worse in the next iteration. In order to address this problem, research has been conducted to explore *labeled features* in model learning without labeled instances. Druck et al. [16] proposed training discriminative probabilistic models with labeled features and unlabeled instances using generalized expectation (GE) criteria. He and Zhou [17] also made use of the GE criteria for self training. They derived labeled features from a generic sentiment lexicon for sentiment classification.

To summarise, exploratory dialogue can be detected using either a set of pre-defined cue phrases signaling the existence of exploratory dialogue or supervised classifiers trained on an annotated corpus. Manually defining cue phrases is both time consuming and labour intensive. On the other hand, annotated corpora are difficult to obtain for practical applications. We therefore propose a feature-based self-learning framework which combines the advantages of cue-phrase based and supervised learning approaches. Further-more, integrating a KNN-based instance selection method into the framework offers an opportunity to reduce the mislabeled instances introduced through self-training.

3 SELF-TRAINING FROM LABELED FEATURES (SELF) FRAMEWORK

We propose a SElf-training from Labeled Features (SELF) framework for exploratory dialogue detection. This framework is shown in Figure 1. We first train an initial maximum entropy (MaxEnt) classifier based on generalized expectation (GE) criteria [16], using self-learned features extracted from a small set of annotated dataset. The trained classifier is then applied to a large amount of un-annotated data. We employ a cue-phrase matching method together with the classifier in order to select positive

examples (exploratory dialogue) and improve the labelling accuracy. In order to take into account topical features, a KNN-based instance selection method is used to select pseudo-labeled instances. These are added to the original annotated training set to derive self-learned features. In the next training loop, the classifier is re-trained using the self-learned features based on GE. Training iterations terminate after five iterations or when the number of label changes in the un-annotated dataset is less than 0.5% of the size of the un-annotated dataset (50 in our study).

3.1 *Classifier Training using Generalized Expectation Criteria*

For exploratory dialogue classification, we define a label set with L labels denoted by $\mathcal{L} = \{\text{exploratory}, \text{non-exploratory}\}$. In addition, we have a corpus with a collection of M postings denoted by $\mathcal{C} = \{\mathbf{d}_1, \mathbf{d}_2, ..., \mathbf{d}_M\}$ where the bold-font variables denote the vectors. Each posting in the corpus is a vector of M_d features denoted by $\mathbf{d} = \{f_1, f_2, ..., f_{M_d}\}$.

In case of a classifier parameterized by θ, the label l of a dialogue post $\mathbf{d}$ is found by maximizing Equation 1.

$$\tilde{l} = \arg\max_{l} P(l|\mathbf{d}; \theta) \tag{1}$$

Assuming we have some labeled features with probability distribution on label set $\mathcal{L}$, we can construct a set of real-valued features of the observation to express some characteristic of the empirical distribution of the training data that should also hold for the model distribution.

$$F_{jk}(\mathbf{d}, l) = \sum_{i=1}^{M} \delta(l_d = j)\delta(k \in \mathbf{d}_i), \tag{2}$$

where $\delta(x)$ is an indicator function that takes a value of 1 if x is true, 0 otherwise. Equation 2 calculates how often feature k and dialogue label j co-occur in an instance.

We define the expectation of the feature as shown in Equation 3.

$$E_\theta[\mathbf{F}(\mathbf{d}, l)] = E_{\tilde{P}(\mathbf{d})}[E_{P(l|\mathbf{d};\theta)}[\mathbf{F}(\mathbf{d}, l)]], \tag{3}$$

where $\tilde{P}(\mathbf{d})$ is the empirical distribution of $\mathbf{d}$ in the dialogue corpus $\mathcal{C}$, $P(l|\mathbf{d}; \theta)$ is a conditional model distribution parameterized at θ, and $E_\theta[\mathbf{F}(\mathbf{d}, l)]$ is a matrix of size $L \times K$, where K is the total number of features used in model learning. The jk_{th} entry denotes the expected number of instances that contain feature k and have label j.

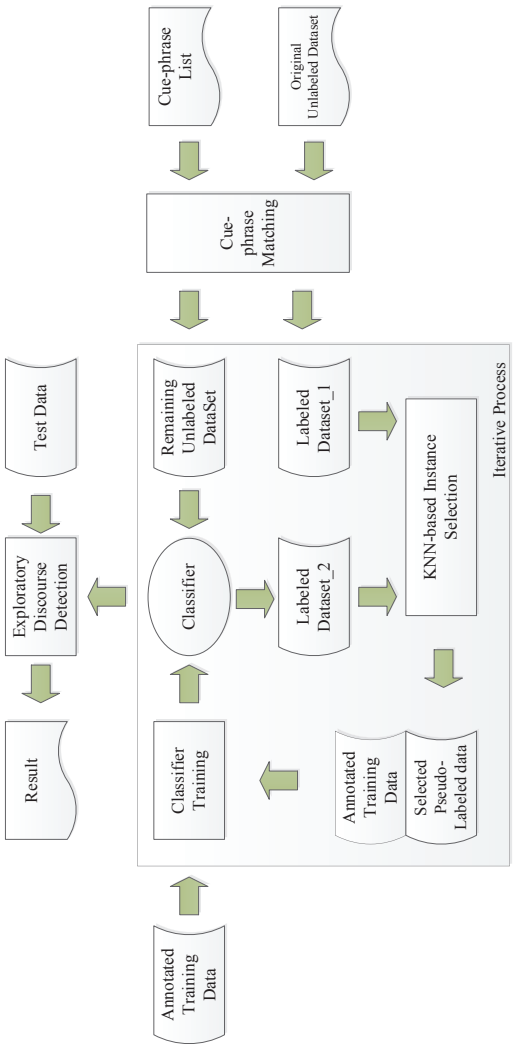


Fig. 1. A self-training framework for exploratory dialogue detection.

A criterion can be defined that minimises the KL divergence of the expected label distribution and a target expectation $\bar{\mathbf{F}}$, which is essentially an instance of generalized expectation criteria that penalizes the

divergence of a specific model expectation from a target value [16].

$$G(E_\theta|\mathbf{F}(\mathbf{d},l)) = KL(\bar{\mathbf{F}}||E_\theta[\mathbf{F}(\mathbf{d},l)]) \quad (4)$$

We can use the target expectation $\bar{\mathbf{F}}$ to encode human or task prior knowledge. For example, the feature "but-if" (bi-gram feature of combining two words "but if")typically signifies an exploratory dialogue. We thus expect this feature to appear in an exploratory dialogue posting more often than in a posting that does not contain exploratory dialogue.

In our experiments, we built a MaxEnt classifier based on GE. In order to do so, we first had to select the indicative features for each class, decide on their respective class labels, and suggest the target or reference feature-class distribution for each of them.

Given a small set of annotated training data, information gain can be used to select representative features. Features with probability higher than threshold ρ are selected. The expected feature-class distribution for a given feature f is defined as a vector $\mathbf{F}(d)$ where

$$F(f,j) = \tilde{P}(j|f;\theta) \quad (5)$$

That is, $F(f,j)$ element is the probability of a label $l = j$ being assigned given that feature *f* is present in a dialogue post. Such probabilities can be estimated directly from data.

3.2 *Incorporating Cue Phrases for Un-annotated Data Labelling*

In our preliminary experiments, the cue-phrase matching method based on the 94 cue phrases identified in [3] has been found to give a high precision over 95% when detecting exploratory dialogue. This suggests that discourse features based on cue phrases could potentially improve the accuracy of exploratory dialogue detection. In our proposed SELF framework, cue phrases can be utilised in two ways. One approach is to combine them with the features extracted from a small set of annotated data in order to train MaxEnt using GE. Another approach is to use them to select positive examples (exploratory dialogue) from un-annotated data, which can subsequently be combined with a small set of annotated data to train classifiers.

Our preliminary experimental results found that features selected from our small set of annotated data are typically in the range of thousands. Hence, merely combining 94 cue phrases with the selected features does not bring any obvious improvement in exploratory dialogue detection performance. Therefore, in this paper, we use cue phrases to identify exploratory dialogue within the un-annotated data and then add them to the originally labelled data set for subsequent classifier training.

3.3 *KNN-Based Instance Selection*

Within a self-training framework, pseudo-labeled instance selection is a crucial step, because adding consistently mislabeled instances to the training set can degrade the model in subsequent iterations. A straightforward way of selecting pseudo-labeled instances is only to select instances with confidence values generated by the current classifier that are above a certain threshold. Nevertheless, as mentioned in Section 1, we argue that topical features are also crucial to exploratory dialogue detection. Therefore, we propose a KNN-based instance selection method to utilise local topical features in order to reduce the number of mislabeled instances.

Once a classifier is trained, it is applied to the un-annotated data with a total of N postings $\mathcal{C}^U = \{\mathbf{d}_1^u, \mathbf{d}_2^u, ..., \mathbf{d}_N^u\}$, and it generates a corresponding label for each posting $\mathcal{L}^U = \{l_1^u, l_2^u, ..., l_N^u\}$ together with a confidence value $\mathcal{Z}^U = \{z_1^u, z_2^u, ..., z_N^u\}$ indicating how confident the classifier is when assigning the corresponding label.

We first select k nearest neighbors for each posting $\mathbf{d}_i^u \in \mathcal{C}^U$ based on the cosine similarity measurement as defined by Equation 6.

$$Sim(\mathbf{d}_i^u, \mathbf{d}_j^u) = \frac{\mathbf{d}_i^u \times \mathbf{d}_j^u}{||\mathbf{d}_i^u|| \times ||\mathbf{d}_j^u||} \tag{6}$$

This essentially selects postings that are topically similar to $\mathbf{d}_i^u$. We then decide whether the instance $\mathbf{d}_i^u$ should be selected for subsequent classifier training by considering the pseudo-labels of its k nearest neighbors. A support value s_i is calculated for instance selection.

$$s_i = \frac{\sum_{j=1}^{k} \delta(l_i^u = l_j^u) z_j^u}{k} \tag{7}$$

where $\delta(x)$ is an indicator function which takes a value of 1 if x is true, 0 otherwise.

A pseudo-labeled instance $\mathbf{d}_i^u$ is selected only if its corresponding support value s_i is higher than a threshold η. In our experiment, we empirically set η to 0.4 and k to 3.

4 EXPERIMENTS

4.1 *The Open University Conference 2010 Dataset*

The dataset for evaluating our proposed exploratory dialogue detection method was constructed from the Annual Learning and Technology Con-

ference: Learning in an Open World[6], run by the UK Open University (OU) in June 2010. Statistics relating to the OU Conference 2010 dataset (OUC2010) are provided in Table 2. The two-day conference was made up of four sessions - a morning session and an wvening session on each day. During the conference, 164 participants generated 2,636 postings within the synchronous text chat forum. These consisted of 6,689 distinct word tokens. These postings are typically short with a mean average of 10.14 word tokens in each one.

In addition to OUC2010, we constructed an additional un-annotated dataset from three open online courses, including 49 sessions containing 10,568 dialogue postings in total. Statistics relating to the un-annotated dataset are provided in the *Un-annotated* category of Table 2. We will make both the OUC2010 and un-annotated corpora available for public access.

Table 2. Statistics of the original OUC2010 and the un-annotated datasets.

	SessionID	Participant#	Posting#	Token#	Vocabulary#	Ave. Length
Annotated	OU_22AM	76	667	7204	2506	10.80
	OU_22PM	61	860	9073	3074	10.55
	OU_23AM	54	541	5517	2037	10.19
	OU_23PM	54	568	4937	1932	8.69
	total	164	2636	26731	6798	10.14
Un-annotated		1152	10568	97699	17268	9.244

We hired three graduate students with expertise in educational technology to annotate a subset of OUC2010. The task was to classify whether a dialogue posting was exploratory or not. The dialogue postings were presented in chronological order so that annotators could make decisions based on contextual information (i.e., postings before and after the current posting).

The Kappa coefficient [18] for inter-annotator agreement was 0.5977 for the binary classification of exploratory / non-exploratory. Statistics relating to the annotated OUC2010 dataset are presented in Table 3.

4.2 *Experimental Setup*

As shown in Table 2, the average length of each posting was relatively short. We therefore did not carry out stopwords removal or stemming.

[6] `http://cloudworks.ac.uk/cloudscape/view/2012/`

Table 3. Statistics of annotated OUC2010 dataset.

SessionID	Agreed Posting#	Exploratory#	Non-Exploratory#
OU_22AM	529	380	149
OU_22PM	661	508	153
OU_23AM	456	310	146
OU_23PM	441	219	222
total	2087	1417	670

Our preliminary experiments showed that combining unigrams with bigrams and trigrams gave better performance than using any one or two of these three features. Therefore, in the experiments reported here, we use the combination of unigrams, bigrams and trigrams as features for classifier training and testing.

We compare our proposed framework with the following approaches in order to explore the effectiveness of the framework:

- **Cue phrase labelling (CP)**. Detect exploratory dialogue using cue phrases only.
- **MaxEnt**. Train a supervised MaxEnt classifier using annotated data.
- **GE**. Train a MaxEnt model using labeled features based on Generalized Expectation (GE) criteria. We select labeled features if their association probabilities with any one of the classes exceed 0.65.
- **Self-learned features (SF)**. The feature based self-learning framework without cue phrase matching and KNN instance selection. Documents labeled by the initial classifier are taken as labeled instances. Features are selected based on the information gain (IG) of the feature with the class label and the target expectation of each feature is re-estimated from the pseudo-labeled examples. A second classifier is then trained using these self-learned features using GE.
- **Self-learned features + KNN (SF+KNN)**. At each training iteration, the KNN-based instance selection method is used to select the pseudo-labeled instances for the derivation of self-labeled features.
- **Self-learned features + Cue-phrase + KNN (SF+CP+KNN)**. Our proposed method integrating both cue-phrase matching method and KNN based instance selection method within the self-training framework.

In each run of experiment, one session of the annotated OUC2010 was selected as the test set, and all or part of the remainder was used as the training set. The un-annotated dataset was used for self-training. For performance evaluation, all possible training and testing combinations

were tested and the results were averaged over all such runs. In each of the re-training iterations, pseudo-labeled instances were selected with the same ratio of exploratory to non-exploratory as in the initial training set. We evaluated our method using metrics including accuracy, precision, recall and F-measure.

4.3 *Results*

OVERALL PERFORMANCE Table 4 shows the exploratory dialogue classification results on the OUC2010 dataset using the methods described above. We used half a session from one of the four annotated sessions for training. The total amount of training postings ranged from 220 to 330. *CP* gives the highest precision of over 95%. However, it also generates the lowest recall value, only 42%. This indicates that the manually defined cue phrases are indeed accurate indicators of exploratory dialogue. However, they missed over half the positive exploratory dialogue.

Training from labeled features only (*GE*) performs worse than the supervised classifier *MaxEnt*. The original self-learned features method, *SF*, presents a similar performance when compared to *GE*. *SF+KNN*, incorporating the KNN-based pseudo-labelled instances selection method, outperforms *SF*, showing the effectiveness of adding instances based on the labels of their k-nearest neighbours. Our proposed method, which is *SF+CP+KNN*, incorporating both cue phrase matching and KNN based instance selection, outperforms all the other baselines according to accuracy and F1 value, generating 3.4% and 4% improvement to accuracy compared to the *GE* method. Although the improvement seems modest compared to supervised learning methods such as MaxEnt, our significance test shows that the improvement is statistically significant. In addition, while supervised learning methods require annotated data for training, our proposed SELF framework only requires a small set of labelled features. This is important for exploratory dialogue detection because annotated data are scarce.

VARYING TRAINING SET SIZE To explore the influence of the amount of training data on accuracy and to investigate the effectiveness of two components within SELF, we varied the size of the annotated training set from 1/8 session to 1 session and compared the performance of different approaches. As shown in Figure 2, as the size of the training set increases, the performance of all approaches grow improves. *SF+CP+KNN* outperforms all the other methods with regard to accuracy across different sizes

Table 4. Exploratory dialogue classification results.

Approach	Accuracy	Precision	Recall	F1
CP	0.5389	**0.9523**	0.4241	0.5865
MaxEnt	0.7886	0.8262	0.8609	0.8301
GE	0.7658	0.7753	**0.8717**	0.8017
SF	0.7659	0.7572	0.8710	0.8062
SF+KNN	0.7701	0.7865	0.8539	0.8148
SF+CP+KNN	**0.7924**	0.8083	0.8688	**0.8331**

of training set. As the size of the training set increases, the accuracy of *GE* rises quickly exceeding both *SF* and *SF+KNN* when the size of the annotated data reaches 1 session. This shows that when annotated data are abundant, the effect of self-labeled feature learning and KNN-based instance selection diminishes. Nevertheless, incorporating both cue-phrase matching and KNN-based instance selection *SF+CP+KNN*, our proposed method performs significantly better than all other methods tested.

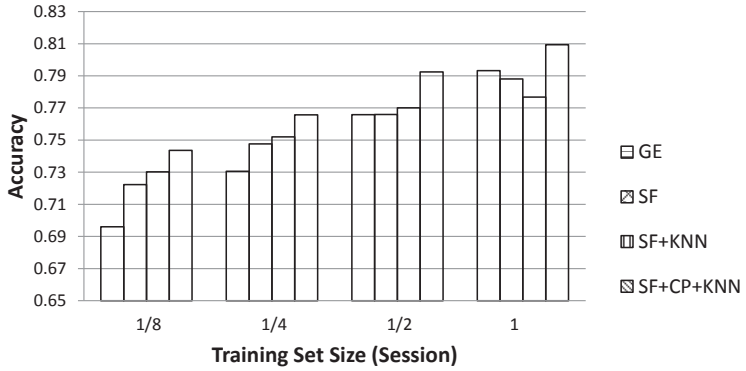


Fig. 2. Accuracy vs. training set size.

VARYING k IN KNN-BASED INSTANCE SELECTION To explore the impact of k in KNN based instance selection on the performance of our proposed SELF framework, we varied k, the number of neighbours, in *SF+CP+KNN*. Here, we only used half a session of the annotated dataset

for training. As shown in Table 5, the best performance is achieved when k is set to 3.

Table 5. Performance of proposed framework on different k

k	Accuracy	Precision	Recall	F1
1	0.7868	0.8007	0.8666	0.8282
3	**0.7924**	**0.8083**	**0.8688**	**0.8331**
5	0.7881	0.8005	0.8685	0.8292
7	0.7586	0.7505	0.8640	0.8001

5 CONCLUSIONS

In this paper, we have proposed a self-training framework for the detection of exploratory dialogue within online dialogue. Cue phrases have been employed to utilise discourse features for classification and a KNN-based instance selection method has been proposed to make use of topical features in order to reduce the erroneously-labeled instances introduced by self training. We have built the first annotated corpus for the detection of exploratory dialoge, OUC2010, from the OU Online Conference. Experimental results on OUC2010 show that our approach outperforms competitive baselines.

To the best of our knowledge, our study is the pioneer work on the automatic detection of exploratory dialogue. There are elements of this work that we would like to explore further. In the current paper, we have only focused on the use of n-grams. It would be possible to explore other features, such as the position of dialogue postings within one session. For example, dialogue exchanges at the beginning of sessions are likely to be non-exploratory because people tend to introduce themselves and greet each other when they first arrive. Moreover, if we know that one posting is exploratory, for example, if someone challenges a previous statement, then the next posting is also likely to be exploratory. Hence, contextual information such as previous and subsequent postings could be taken into account when classifying a posting. Another interesting direction will be to explore automatic ways of expanding the cue phrase list and combining it with machine learning methods for exploratory dialogue detection.

ACKNOWLEDGEMENTS This work was partially supported by General Research Fund of Hong Kong (No. 417112).

REFERENCES

1. Mercer, N., Littleton, K.: Dialogue and the development of children's thinking: A sociocultural approach. Taylor & Francis (2007)
2. Mercer, N.: Sociocultural discourse analysis. Journal of Applied Linguistics **1**(2) (2004) 137–168
3. Ferguson, R., Buckingham Shum, S.: Learning analytics to identify exploratory dialogue within synchronous text chat. In: Proceedings of the 1st International Conference on Learning Analytics and Knowledge. (2011) 99–103
4. Mercer, N.: Developing dialogues. Learning for life in the 21st century (2002) 141–153
5. Mercer, N., Wegerif, R.: Is "exploratory talk" productive talk? In Littleton, K., Light, P., eds.: Learning with computers. Routledge (1998) 79–101
6. Whitelock, D., Watt, S.: Open mentor: Supporting tutors with their feedback to students. In: 11th CAA International Computer Assisted Assessment Conference. (2007)
7. Sándor, Á., Vorndran, A.: Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In: Workshop on Text and Citation Analysis for Scholarly Digital Libraries. (2009) 36–44
8. Samuel, K., Carberry, S., Vijay-Shanker, K.: Automatically selecting useful phrases for dialogue act tagging. Arxiv preprint cs/9906016 (1999)
9. Webb, N., Hepple, M., Wilks, Y.: Dialogue act classification based on intra-utterance features. In: Proceedings of the AAAI Workshop on Spoken Language Understanding. (2005)
10. Webb, N., Liu, T.: Investigating the portability of corpus-derived cue phrases for dialogue act classification. In: 22nd International Conference on Computational Linguistics (COLING). (2008) 977–984
11. Jekat, S., Klein, A., Maier, E., Maleck, I., Mast, M., Quantz, J.J.: Dialogue acts in verbmobil. Technical report, DFKI GmbH (1995)
12. Levin, L., Langley, C., Lavie, A., Gates, D., Wallace, D., Peterson, K.: Domain specific speech acts for spoken language translation. In: Proceedings of 4th SIGdial Workshop on Discourse and Dialogue (SIGDIAL). (2003)
13. Reithinger, N., Klesen, M.: Dialogue act classification using language models. In: 5th European Conference on Speech Communication and Technology. (1997)
14. Ji, G., Bilmes, J.: Dialog act tagging using graphical models. In: ICASSP. (2005) 33–36
15. Verbree, D., Rienks, R., Heylen, D.: Dialogue-act tagging using smart feature selection; results on multiple corpora. In: IEEE Spoken Language Technology Workshop. (2006) 70–73
16. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: SIGIR. (2008) 595–602
17. He, Y., Zhou, D.: Self-training from labeled features for sentiment analysis. Information Processing & Management **47**(4) (2011) 606–616

18. Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. Computational linguistics **22**(2) (1996) 249–254

ZHONGYU WEI
THE CHINESE UNIVERSITY OF HONG KONG,
SHATIN, N.T., HONG KONG
E-MAIL: <YUTOUWEI@GMAIL.COM>

YULAN HE
SCHOOL OF ENGINEERING & APPLIED SCIENCE,
ASTON UNIVERSITY,
BIRMINGHAM, UK
E-MAIL: <Y.HE@CANTAB.NET>

SIMON BUCKINGHAM SHUM
KNOWLEDGE MEDIA INSTITUTE /
INSTITUTE OF EDUCATIONAL TECHNOLOGY,
THE OPEN UNIVERSITY,
MILTON KEYNES, UK
E-MAIL: <S.BUCKINGHAM.SHUM@OPEN.AC.UK>

REBECCA FERGUSON
KNOWLEDGE MEDIA INSTITUTE /
INSTITUTE OF EDUCATIONAL TECHNOLOGY,
THE OPEN UNIVERSITY,
MILTON KEYNES, UK
E-MAIL: <REBECCA.FERGUSON@OPEN.AC.UK>

WEI GAO
QATAR COMPUTING RESEARCH INSTITUTE,
QATAR FOUNDATION,
DOHA, QATAR
E-MAIL: <WGAO@QF.ORG.QA>

KAM-FAI WONG
KEY LABORATORY OF HIGH CONFIDENCE
SOFTWARE TECHNOLOGIES,
MINISTRY OF EDUCATION, CHINA
E-MAIL: <KFWONG@SE.CUHK.EDU.HK>

Information Extraction

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 129–144
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 14/03/13

Something Old, Something New: Identifying Knowledge Source in Bio-events

RAHEEL NAWAZ, PAUL THOMPSON, AND SOPHIA ANANIADOU

University of Manchester, UK

ABSTRACT

Locating new experimental knowledge in biomedical texts is important for several tasks undertaken by biologists. Although several systems can distinguish between new and existing knowledge, this generally happens at the text zone level. In contrast to text zones, bio-events constitute structured representations of biomedical knowledge. They bridge text with domain knowledge and can be used to develop sophisticated semantic search systems. Typically, event extraction systems locate and classify events and their arguments, but ignore interpretative information (meta-knowledge) from their textual context. Since several events (often nested) can occur in a sentence, determining which event(s) are affected by which textual clues can be complex. We have analysed knowledge source annotation in two bio-event corpora: GENIA-MK (abstracts) and FP-MK (full papers), and have developed a system to classify bio-events automatically according to their knowledge source. Our system performs with an accuracy of over 99% on both abstracts and full papers.

KEYWORDS: *knowledge source, new knowledge, meta-knowledge, event, bio-event, machine learning*

1 Introduction

In recent years, several annotation schemes, e.g., [1-4] have been developed to identify and classify textual zones (i.e., continuous spans of text, such as sentences and clauses) in scientific papers, according to their rhetorical status or general information content. In most cases, these corpora have subsequently been used as a basis for training systems to recognise this information automatically, e.g., [5-7]. Common to all of these systems is the ability to identify information about *knowledge source*. That is, whether the text zone refers to new work being described in the paper, or to work that has already been described elsewhere. Such systems can be instrumental in helping users to search for text zones that contain new experimental knowledge. The identification of such information is important for several tasks in which biologists have to search and review the literature. One such example is the maintenance of models of biological processes, such as pathways [8]. As new reactions or new evidence for reactions become available in the literature, these should be added to the corresponding pathway(s). Another area where this information is useful is in the curation of biomedical databases. One of the tasks involved in keeping such databases up-to date is to search for new evidence for a particular interaction (e.g., gene regulation) within the literature [9].

In the types of task outlined above, the biologist is likely to be looking for specific types of biological processes or reactions, and specific types of information about them, e.g., what caused the reaction to occur, where the reaction took place, etc. Although the text zone classification systems cannot help with this kind of task, another type of system, i.e., an event extraction system, can be extremely useful. Event extraction systems are usually developed through training on manually annotated bio-event corpora, e.g., GENIA [10], BioInfer [11] and GREC [12]. These corpora identify named entities, such as genes and proteins, as well as the bio-events in which these entities participate. Systems are then trained to extract bio-event structures automatically from texts. The recent BioNLP Shared Tasks on event extraction in 2009 [13] and 2011 [14] have helped to stimulate considerable advances in event extraction research.

Event extraction facilitates the development of sophisticated semantic-based search systems, e.g., [15], which allow researchers to perform structured searches over events extracted from a large body of text [16]. Although search constraints can typically be specified in terms of event type (i.e., the process or reaction of interest) and/or the types of named

entities participating in the event, the ability to specify knowledge source as a constraint is not available. Bio-events are typically contained within a single sentence, and the existing text zone identification systems would normally be able to determine knowledge source at the sentence level. However, events are not the same as text zones. Whilst text zones constitute continuous spans of text, events usually consist of several discontinuous text spans, which correspond to different elements of the event, e.g., participants, location, etc. [17]. There are also (usually) several events contained within a single sentence. This means that just because a sentence or clause may be identifiable as having a particular knowledge source, it does not follow that all events contained within that text zone will have the same knowledge source; each event may have its own interpretation, and determining which events are affected by particular textual clues can be complex. For example, consider the following sentence:

Previous studies have shown that inhibition of the MAP kinase cascade with PD98059, a specific inhibitor of MAPK kinase 1, may prevent the rapid expression of the alpha2 integrin subunit.

This sentence contains not only a speculative analysis from an *Other* source, i.e., *Inhibition of the MAP kinase may prevent the expression of the alpha2 integrin subunit*, but also a general fact, i.e., *PD98059 is a specific inhibitor of MAPK kinase 1*. The main verb in the sentence (i.e., *prevent*) describes the information that has been reported in previous studies. In a sentence-based annotation scheme, this is likely to be the only information that is encoded. However, this means that the general fact is disregarded. Some annotation schemes have attempted to overcome the fact that sentences may contain multiple types of information by annotating meta-knowledge below the sentence level, i.e., clauses [18, 19] or segments [20]. In the case of the latter scheme, a new segment is created whenever there is a change in the meta-knowledge being expressed.

In the sentence above, however, it is not possible to split the sentence into continuous segments, since the general fact is embedded within the speculative analysis. In an event-based view of the sentence, this does not matter, since events consist of structures with different "slots", each of which is filled by a different text span, drawn from anywhere within the sentence. In this way, we say that the speculative analysis is triggered by the verb *prevent*, and has the participants *Inhibition of the MAP kinase* and *the rapid expression of the alpha2 integ-*

rin subunit. Similarly, the general fact can be encoded as a separate event. Only the speculative analysis event is referring to work being carried out as part of a particular study. The general fact event is considered to be established knowledge, and so it would not be correct to attribute this event to a particular previous study.

In order to allow further information to be encoded in event extraction systems, [21] proposed a multidimensional event-based meta-knowledge annotation scheme that includes knowledge source as a dimension of event interpretation. Other dimensions included in the scheme are: knowledge type, certainty level (allowing, amongst other things, speculative analyses to be encoded), polarity, and manner. This scheme has been manually applied to a number of different corpora. Firstly, the GENIA event corpus, comprising 1000 MEDLINE abstracts, was enriched to create the GENIA-MK corpus [22]. Secondly, a corpus of 4 full papers with event annotations has been enriched to create the FP-MK corpus [23]. A third, on-going effort is the application of the scheme to a corpus of stem cell research papers [24].

This paper describes our work on analysis and automated identification of knowledge source information about bio-events, using the GENIA-MK (abstracts) and FP-MK (full papers) corpora for training and testing. In both corpora, each event is ascribed one of two knowledge source values, i.e., *Current*, for events relating to work described in the current paper (default value), or *Other*, for events relating to work originally described elsewhere. Although the analysis carried out in [23] reveals that there are significant differences in the distributions of the different knowledge source values in abstracts and full papers, and that the textual means of denoting *Other* events also varies between abstracts and full papers, our system is able to perform to an almost identical level of accuracy on both text types, i.e., 99.6% and 99.4%, for abstracts and full papers, respectively.

2 Background

2.1 *Bio-event*

In its most general form, a **textual event** can be described as an action, relation, process or state [25]. More specifically, an event is a structured semantic representation of a certain piece of information contained within the text. Events are usually anchored to particular text fragments that are central to the description of the event, e.g., *event-*

trigger, *event-participants* and *event-location*, etc. A **bio-event** is a textual event specialised for the biomedical domain, in that it constitutes a dynamic bio-relation involving one or more participants [10]. These participants can be bio-entities or (other) bio-events, and are each assigned a semantic role like *theme* and *cause*, etc. Bio-events and bio-entities are also typically assigned semantic types/classes from particular taxonomies/ontologies. Consider the sentence S1: "*It has previously been reported [12] that LTB4 augments c-jun mRNA*". This sentence contains a single bio-event of type *positive_regulation*, which is anchored to the verb *augments*. Figure 1 shows a typical structured representation of this bio-event. The event has two participants: *c-jun mRNA* and *LTB4*, which have both been assigned their respective semantic types and roles within the event.

TRIGGER:	***augmented***
TYPE:	**positive_regulation**
THEME:	***c-jun mRNA*** : RNA_molecule
CAUSE:	***LTB4*** : organic_molecule

Fig. 1. Typical representation of the bio-event contained in sentence S1

2.2 *Knowledge Source*

As mentioned above, information about knowledge source is an integral part of a number of schemes for annotating text zones and their functions. The argumentative zoning (AZ) scheme, first introduced in [1], distinguishes sentences that mention OWN work presented in the current paper and OTHER specific work presented in another paper. Later extensions based on this scheme [2, 26] recognized that different types of information about OWN work can usefully be distinguished, such as OWN_METHD (methods) and OWN_RES (results) or OWN_CONC (conclusions). Multi-dimensional schemes allow several pieces of information to be associated with a given text span, and thus provide more flexibility regarding the types of information that can be encoded. Several such schemes encode information about knowledge source as a separate dimension, e.g., the scheme of [6] includes a novelty attribute (*New* or *Old*) that is distinct from their knowledge type attribute (*Background, Method, Conclusion*, etc.) The scheme of [3] identified five dimensions of information that could reliably be identi-

fied about text fragments (mostly clauses or sentences). Their *evidence* dimension includes information about the source of knowledge expressed in the text fragment. It has four possible values, which have similarities with some of the evidence codes used during the annotation for the Gene Ontology [27]. These values are: *E0*: no indication of evidence; *E1*: mention of evidence with no explicit reference; *E2*: explicit reference is made to other papers to support the assertion; *E3*: experimental evidence is provided directly in the text.

In the event-based meta-knowledge scheme of Nawaz et al. [21], information about the knowledge source of the event is encoded using the *Source* dimension, which has two possible values. The **Other** value is assigned when the event can be attributed to a previous study. This value is normally determined through the presence of explicit clues, e.g., *previously*, *recent studies*, etc., or cited papers, in the vicinity of the event. The **Current** value is assigned when the event makes an assertion that can be attributed to the current study. This is the default category, and is assigned in the absence of explicit lexical or contextual clues, although explicit clues such as *the present study* may be encountered. As an example, the bio-event in sentence S1 (section 1.1) has been attributed to another study through the use of an in-text citation. Therefore, it will be assigned the knowledge source value of *Other*.

2.3 *Annotation of Knowledge Source in GENIA-MK and FP-MK Corpora*

The GENIA-MK corpus consists of 1000 MEDLINE abstracts, containing 36,858 events, each of which has been annotated according to the meta-knowledge scheme described in [23]. In this corpus, slightly fewer than 2% of all events are assigned a *Source* value of *Other*. This is not surprising: abstracts are meant to provide a summary of the work carried out in a given paper and, given the very limited space, there is little opportunity to discuss previous work. Indeed, the use of citations is often prohibited in abstracts.

The FP-MK corpus consists of 4 full papers, in which 1,710 events have been annotated according to the same meta-knowledge scheme. In contrast to the GENIA-MK corpus, nearly 20% of all events in the FP-MK corpus belong to the *Other* category. The analysis provided in [23] examines the distribution of *Source* annotations in the various different sections in full papers. The study reports that by far the highest concentration of *Other* events is in the *Background* sections of the papers, where over 40% of the events are attributed to other sources. This is

expected, since it is normally in the *Background* section where one encounters the highest concentration of descriptions of previous work. The *Discussion* sections of the papers also have a high concentration (over 25%) of *Other* events, since it is common to compare and contrast the outcomes of the current work with those of previous related studies as part of the discussion. The frequency of *Other* events in the remaining sections is considerably lower. For example, in the *Results* sections of the papers, less than 7% of events are annotated as *Other.*

3 Analysis of *Other* Events

3.1 *Clue Frequency*

Table 1 shows the most commonly annotated clue expressions for *Source=Other* in the GENIA-MK (abstracts) and FP-MK (full papers) corpora respectively. For abstracts, several clue expressions contain the adverbs *previously* or *recently*, or their adjectival equivalents. The phrases *have been* and *has been* have also been annotated as clues with reasonably high frequency, the reason being that the use of the passive voice with the present perfect tense (e.g. *has been studied)* is a common means to indicate that an event has previously been completed (e.g., in a previous study), but yet has relevance to the current study.

Table 1. Most frequently annotated *Other* clues in GENIA-MK and FP-MK corpora

GENIA-MK (abstracts)			FP-MK (full papers)		
Cue	Freq	%	Clue	Freq	%
previously	118	21.7%	Citation	267	78.3%
has been	89	16.3%	has been	41	12.0%
recently	67	12.3%	previously	6	1.8%
have been	39	7.2%	recently	6	1.8%
previous	38	7.0%	latter example	4	1.2%
recent	32	5.9%	studies have shown	4	1.2%
earlier	6	1.1%	we and others	4	1.2%

In contrast to abstracts, the vast majority of clue expressions in full papers correspond to citations. However, similarly to abstracts, the use

of the present perfect tense is also quite common. Other explicit markers (such as *previously* and *recently*) constitute less than 10% of the clue expressions.

3.2 *Clue Ambiguity*

The presence of an *Other* clue in a sentence is not in itself sufficient evidence for assigning the knowledge source value of *Other* to all events in the sentence. While a sentence contains, on average, 4 bio-events, the majority of *Other* clues affect only one event in the sentence, i.e., the knowledge source value for the remaining events in the sentence is *Current*. Therefore, it is highly important that the syntactic/semantic structure of the sentence is considered, in order to determine which, if any, of the events are being affected by the clue. For example, the existence/type of dependency/constituency relations between the event participants and any *Other* clue(s) present in the sentence can be considered.

Furthermore, some of the *Other* clues (e.g., the tense of the sentence) are inherently ambiguous, and only indicate an *Other* event in certain contexts. For example, the clue expression *has/have been* is a significant clue for *Other* events – it accounts for over 23% of all *Other* events in abstracts and 12% of all *Other* events in full papers. However, an analysis of events from the sentences containing the phrase *has/have been* in the GENIA-MK corpus reveals that only 8% of these events are of type *Other*. This proportion is even lower (7%) for full papers.

3.3 *Event Complexity*

We examined the distribution of events assigned the value *Source=Other* amongst **simple** and **complex** events. By simple event, we mean an event whose participants are all entities, whilst a complex event is one with at least one participant which is itself an event. In abstracts, 67% of *Other* events are complex. Conversely, 2.26% of complex events are of type *Other*, while only 0.88% of simple events are of type *Other*. This means that an arbitrary complex event is 2.6 times more likely than an arbitrary simple event to have knowledge source value of *Other*.

In full papers, an even greater proportion of *Other* events (i.e., 72%) is complex. A total of 3.32% of complex events are of type *Other*, while only 0.73% of simple events belong to this type. Therefore, in

full papers, an arbitrary complex event is 4.5 times more likely than an arbitrary simple event to have knowledge source value of *Other*.

3.4 *Relative Position within Text*

In abstracts, 74% of *Other* events appear in the 2nd, 3rd or 4th sentence. Furthermore, over 80% of the *Other* events appear in the first half of the abstract.

In full papers, the section to which the sentence containing the event belongs is more significant than the relative position of the sentence within the paper or even within a section. For example, over 60% of all *Other* events found in full papers occur within the *Background* section.

4 Classifier Design

Based on the analysis of *Other* events, we engineered 7 feature sets. We used the Enju parser [28] to obtain the lexical and syntactic information required to construct these features. A brief explanation of each feature set is as follows:

- *Syntactic features* include the tense of the sentence (since *Other* events will normally be reported using the past tense), the POS tag of the *event*-trigger, and the POS tag(s) of *Other* clue(s) found in the sentence.
- *Semantic features* include the type of the bio-event and the type and role of each participant.
- *Lexical features.* Since the presence of lexical clues is usually key to determining *Other* events, these features include whether an *Other* clue is present in the sentence, and the clue itself. The clue list was compiled by combining the clue lists extracted from the GENIA-MK and FP-MK corpora, together with regular expressions to identify citations, which are also often important for the identification of *Other* events.
- *Lexico-semantic features.* Since the presence of an *Other* clue in a sentence does not usually affect all events within the sentence, these features help to determine the likelihood that a particular lexical clue for *Other* affects a given event. The features include the proximity (surface distance) between the *Other* clue and various event compo-

nents (*event-trigger*, *event-participants* and *event-location*), whether the *Other* clue precedes or follows the *event-trigger*, etc.

– *Dependency (lexico-syntactic) features.* Proximity of *Other* clues to event components is not always sufficient to determine which events they affect. In more complex sentences, it can be important to consider syntactic structure, since the *Other* clue may not occur close to the event components, but still be structurally related. For this reason, these features are based around the presence of direct and indirect dependency relations between the *Other* clue present in the sentence and the *event-trigger*, and the length of these dependency paths.

– *Constituency (lexico-syntactic) features.* This is a further class of structural features. They are based around the *command* [29] and *scope* relations, which are derived from the constituency parse tree. The command features consider the existence of S-, VP- and NP-command relations between the *Other* clue and the *event-trigger*. The scope features consider whether the *event-trigger* falls under the syntactic scope of the *Other* clue.

– *Positional features.* As mentioned above, *Other* events are far more numerous in certain sections of full papers, while within abstracts, earlier sentences are most likely to contain such events. Therefore, we include amongst our features the section in which the sentence containing the event appears (for abstracts all events have the same value and this feature becomes redundant), and the relative position of the sentence containing the event, both within the entire text and within the section.

We used the Random Forest [30] algorithm, which develops an ensemble/forest of Decision Trees from randomly sampled subspaces of the input features. Once the forest has been created, new instances are classified by first obtaining individual classifications from each tree and then using a majority vote to attain the final classification. We used the WEKA [31] implementation of the Random Forest algorithm, which is based on [30]. Our optimization settings included: (1) setting the number of trees in the forest to 10, (2) setting the number of features used to build individual trees to log(N+1), where N is the total number of features, (3) setting no restrictions on the depth of individual trees.

5 Results and Discussion

We conducted a series of experiments using different clue lists and feature set combinations. All results were 10-fold cross validated. The best results for abstracts and full papers are shown in Table 2. In both cases, the best results were achieved by using the 7 most frequent clues (Table 1) and all feature sets.

Table 2. Best results for GENIA-MK and FP-MK

Category	GENIA-MK (abstracts)			FP-MK (full papers)		
	P	R	F	P	R	F
Current	99.6%	99.8%	99.7%	99.5%	99.2%	99.3%
Other	83.3%	70.8%	75.6%	81.3%	70.1%	75.3%
Overall	99.4%	99.4%	99.4%	95.9%	93.4%	94.6%

5.1 *Abstracts*

In abstracts, only 2% of all events are of type *Other*; therefore, the baseline accuracy (through majority-class allocation) is 98%. Our system achieves an overall accuracy of 99.6%, which is considerably higher than this baseline. Recall for the *Other* category is significantly lower than the precision (over 10%). This is mainly due to the difficulty in identifying and disambiguating *Other* clues. The overall system precision and recall are both 99.4%.

5.2 *Full Papers*

The proportion of *Other* events in full papers is almost 10 times greater than in abstracts, with just under 20% of all events belonging to the *Other* category. The baseline classification accuracy for full papers is thus 80%. Therefore, statistically, identification of knowledge source in full papers is a harder task than in abstracts. However, our system achieves a very high overall accuracy of 99.4%. The main difference between the *Other* events in abstracts and full papers is the occurrence of explicit citations as clues. Since our system also includes citation related features, it is able to perform equally well on both corpora.

Similarly to the results for abstracts, precision for full papers is significantly higher than recall. Again, this is mainly due to the difficulty in identifying/disambiguating *Other* clues. This is also reflected in overall system performance as well, where precision is 2.5% higher than recall.

5.3 *Discussion*

Our results are the first that concern the detection of knowledge source at the event level. However, some comparisons can be drawn with similar previous work at the clause, sentence, and zone level. The text zone classification system of [5] achieved a precision/recall of 51%/30% for their OTHER category and a precision/recall of 85%/86% for the OWN category. [32] achieved an overall F-score of 70% for automatic zone classification, including BACKGROUND and OWN zones. The clause classification system reported by [7] performed with F-scores of 89%, 57%, 94% and 91% for the E0, E1, E2, and E3 classes respectively. [6], whose classification is performed at the sentence level, achieved an F-score of 64% for their BACKGROUND class; however, they did not try to identify the novelty attributes separately. Although we identify knowledge source at the event level, which is more challenging than similar tasks at the clause/sentence/zone level, our results are significantly higher. This is partly because we have cast the problem as a binary classification rather than a multi-category classification.

In our system, the most common reason for misclassification was the inability of the system to identify *Other* clues. This accounted for over 52% of the misclassified events. A significant proportion (32%) of misclassified events belonged to sentences with complex syntactic structures, e.g., where the *event-trigger* and the *Other* clue belonged to different clauses. These misclassifications can be partly attributed to parsing limitations, especially in terms of identifying complex dependency relations.

6 Conclusion

The isolation of new experimental knowledge in large volumes of text is important for several tasks undertaken by biologists. Although the ability to search for events of interest can significantly reduce the biologist's workload in finding relevant information, even more time could

be saved by facilitating further refinement of the search results to include only events pertaining to reliable new experimental knowledge. This goal can be achieved through the automatic recognition of event meta-knowledge. One of the most crucial aspects of identifying new experimental knowledge is to determine the knowledge source of the event.

In this paper, we have analysed the event-level knowledge source annotations in the GENIA-MK corpus (abstracts) and the FP-MK corpus (full papers). This analysis was used to inform the process of designing a system to recognise knowledge source automatically. We have shown that the knowledge source of events can be recognised to a high degree of accuracy. In abstracts, the overall accuracy is 99.6% and the overall F-score is 99.4%. The baseline accuracy for abstracts is already extremely high (98%), given that there are few events in abstracts that refer to previous work. However, a more significant result is that the performance of the classifier on full papers is almost as high as for abstracts, even though the baseline accuracy for full papers (80%) is considerably lower than for abstracts. On full papers, the classifier performs with an overall accuracy of 99.4% and achieves an overall F-score of 94.6%. These results provide encouraging evidence that the knowledge source of biomedical events can be predicted very reliably, regardless of text type. We plan to use our system to assist in the (semi-)automatic annotation of other corpora containing bio-event or relation annotation, e.g., [11, 12, 33]. This will pave the way for a more advanced system, able to recognise source information for a wider range of event and relation types. By integrating our classification system with event extraction systems, such as [34], we will be able to develop more sophisticated systems that can extract events with associated source information fully automatically. Events are also relevant to other domains. For example, the ACE 2005 evaluation involved the recognition of events in the general language domain, including events relating to conflict, business and justice. We are in the process of adapting our meta-knowledge scheme to this domain, which will allow systems to be trained to recognise knowledge source for events in alternative domains.

ACKNOWLEDGEMENTS. The work described in this paper has been funded by the MetaNet4U project (ICT PSP Programme, Grant Agreement: No 270893).

References

1. Teufel, S., Carletta, J., Moens, M.: An annotation scheme for discourse-level argumentation in research articles. Proceedings of EACL 110–117 (1999)
2. Mizuta, Y., Korhonen, A., Mullen, T., Collier, N.: Zone analysis in biology articles as a basis for information extraction. International Journal of Medical Informatics **75**:468–487 (2006)
3. Wilbur, W.J., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotations: definitions, guidelines and corpus construction. BMC Bioinformatics **7**:356 (2006)
4. Liakata, M., Teufel, S., Siddharthan, A., Batchelor, C.: Corpora for the conceptualisation and zoning of scientific papers. Proceedings of LREC 2010, 2054-2061 (2010)
5. Teufel, S.: Argumentative Zoning. Univ. of Edinburgh, Edinburgh (1999)
6. Liakata, M., Saha, S., Dobnik, S., Batchelor, C., Rebholz-Schuhmann, D.: Automatic recognition of conceptualisation zones in scientific articles and two life science applications. Bioinformatics **28** (2012)
7. Shatkay, H., Pan, F., Rzhetsky, A., Wilbur, W.J.: Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. Bioinformatics **24**:2086–2093 (2008)
8. Oda, K., Kim, J.-D., Ohta, T., Okanohara, D., Matsuzaki, T., Tateisi, Y., Tsujii, J.i.: New challenges for text mining: mapping between text and manually curated pathways. BMC Bioinformatics **9:** S5 (2008)
9. Yeh, A.S., Hirschman, L., Morgan, A.A.: Evaluation of text data mining for database curation: lessons learned from the KDD Challenge Cup. Bioinformatics **19**: i331-i339 (2003)
10. Kim, J.-D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics **9** (2008)
11. Pyysalo, S., Ginter, F., Heimonen, J., Bjorne, J., Boberg, J., Jarvinen, J., Salakoski, T.: BioInfer: a corpus for information extraction in the biomedical domain. BMC Bioinformatics **8**: 50 (2007)
12. Thompson, P., Iqbal, S., McNaught, J., Ananiadou, S.: Construction of an annotated corpus to support biomedical information extraction. BMC Bioinformatics **10**:349 (2009)
13. Kim, J.D., Pyysalo, S., Ohta, T., Bossy, R., Nguyen, N., Tsujii, J.: Overview of BioNLP Shared Task 2011. Proceedings of BioNLP Shared Task 2011 Workshop, 1–6 (2011)
14. Kim, J.-D., Pyysalo, S., Nedellec, C., Ananiadou, S., Tsujii, J. (eds.): Selected Articles from the BioNLP Shared Task 2011, Vol. 13. BMC Bioinformatics (2012)

15. Miyao, Y., Ohta, T., Masuda, K., Tsuruoka, Y., Yoshida, K., Ninomiya, T., Tsujii, J.: Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. Proceedings of ACL, 1017–1024 (2006)
16. Ananiadou, S., Pyysalo, S., Tsujii, J., Kell, D.B.: Event extraction for systems biology by text mining the literature. Trends Biotechnol **28**: 381–390 (2010)
17. Kim, J.D., Ohta, T., Tsujii, J.: Corpus annotation for mining biomedical events from literature. BMC Bioinformatics **9**: 10 (2008)
18. Waard, A.d., Shum, B., Carusi, A., Park, J., Samwald, M., Sándor, Á.: Hypotheses, Evidence and Relationships: The HypER Approach for Representing Scientific Knowledge Claims. ISWC 2009, the 8th International Semantic Web Conference, Washington, DC., USA (2009)
19. Rubin, V.L.: Stating with certainty or stating with doubt: Intercoder reliability results for manual annotation of epistemically modalized statements. Human Language Technologies Conference: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, 141–144 (2007)
20. Wilbur, W.J., Rzhetsky, A., Shatkay, H.: New directions in biomedical text annotation: definitions, guidelines and corpus construction. BMC Bioinformatics **7**: 356 (2006)
21. Nawaz, R., Thompson, P., McNaught, J., Ananiadou, S.: Meta-Knowledge Annotation of Bio-Events. Proceedings of LREC 2010, 2498–2507 (2010)
22. Thompson, P., Nawaz, R., McNaught, J., Ananiadou, S.: Enriching a biomedical event corpus with meta-knowledge annotation. BMC Bioinformatics **12**: 393 (2011)
23. Nawaz, R., Thompson, P., Ananiadou, S.: Meta-Knowledge Annotation at the Event Level: Comparison between Abstracts and Full Papers. Proceedings of the Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012), 24-21 (2012)
24. Neves, M., Damaschun, A., Kurtz, A., Leser, U.: Annotating and evaluating text for stem cell research. Proceedings of the Third LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining (BioTxtM 2012), 16-23 (2012)
25. Sauri, R., Pustejovsky, J.: FactBank: A Corpus Annotated with Event Factuality. Language Resources and Evaluation **43**, 227–268 (2009)
26. Teufel, S., Siddharthan, A., Batchelor, C.: Towards discipline-independent argumentative zoning: Evidence from chemistry and computational linguistics. Proceedings of EMNLP 2009, 1493–1502 (2009)
27. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richard-

son, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene Ontology: tool for the unification of biology. Nature Genetics **25**: 25–29 (2000)

28. Miyao, Y., Tsujii, J.: Feature Forest Models for Probabilistic HPSG Parsing. Computational Linguistics **34**: 35–80 (2008)
29. Langacker, R.: On Pronominalization and the Chain of Command. In: Reibel, D., Schane, S. (eds.): Modern Studies in English. Prentice-Hall, Englewood Cliffs, NJ., 160–186 (1969)
30. Breiman, L.: Random forests. Machine Learning **45**: 5–32 (2001)
31. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: an update. SIGKDD Explorations **11**: 10–18 (2009)
32. Mullen, T., Mizuta, Y., Collier, N.: A baseline feature set for learning rhetorical zones using full articles in the biomedical domain. ACM SIGKDD Explorations **7**: 52-58 (2005)
33. Bunescu, R., Ge, R., Kate, R.J., Marcotte, E.M., Mooney, R.J., Ramani, A.K., Wong, Y.W.: Comparative experiments on learning information extractors for proteins and their interactions. Artif Intell Med. **33**: 139–155 (2005)
34. Miwa, M., Saetre, R., Kim, J.D., Tsujii, J.: Event extraction with complex event classification using rich features. J Bioinform Comput Biol **8**: 131–146 (2010)

Raheel Nawaz
National Centre for Text Mining,
Manchester Interdisciplinary Biocentre, University of Manchester,
131 Princess Street, Manchester, M1 7DN, UK
E-mail: <raheel.nawaz@cs.man.ac.uk>

Paul Thompson
National Centre for Text Mining,
Manchester Interdisciplinary Biocentre, University of Manchester,
131 Princess Street, Manchester, M1 7DN, UK
E-mail: <paul.thompson@manchester.ac.uk>

Sophia Ananiadou
National Centre for Text Mining,
Manchester Interdisciplinary Biocentre, University of Manchester,
131 Princess Street, Manchester, M1 7DN, UK
E-mail: <sophia.ananiadou@manchester.ac.uk>

Sentiment Analysis and Social Networks

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 147–158
RECEIVED 15/12/12 ACCEPTED 11/01/13 FINAL 27/03/13

Comparing Portuguese Opinion Lexicons in Feature-Based Sentiment Analysis

LARISSA A. DE FREITAS AND RENATA VIEIRA

PUCRS, Brazil

ABSTRACT

In this paper we evaluate different lexicons in feature level opinion mining on Brazilian Portuguese movie reviews. Research in this field often considers English data, while other languages are less explored. So we discuss and compare available resources and techniques that can be applied to Portuguese for dealing with this task. We found better results when using SentiLex adjetives. The results indicate a F-score of 0.73 for positive polarity recognition and 0.76 for negative polarity recognition.

KEYWORDS: *Opinion Mining; Sentiment Analysis; Portuguese Online Reviews; Movie Reviews*

1 INTRODUCTION

Studies about "opinion mining", also called "sentiment analysis"(SA) have been developed more intensively in the last decade. In general, research in this area focuses in detecting the holder's sentiment about a topic in a review. Opinions are important because whenever we need to make a decision, we want to know other points of view.

Nowadays, opinion mining has been investigated mainly in three levels of granularity (document, sentence or feature). According to Liu [1], both the document level and sentence level analyses do not discover what exactly people liked or not. However, feature level opinion mining, required for that, is extremely challenging.

In feature-based opinion mining, features related to an object are analysed. This technique comprises the following steps: identifying the features about the object in review, deciding whether the review is positive or negative and summarizing the information [2]. Overall, the output is a tuple containing the feature and the polarity of objects. The model of feature-based opinion mining is proposed by many researchers, such as Hu and Liu [3] and Popescu and Etzioni [4].

In the literature, recent works about ontology-based opinion mining in feature level are Zhao and Li [5] and Peñalver-Martínez et al. [6]. Both have been applied on English movie reviews, presenting high quality results.

In this context, we address the issue of feature-based opinion mining but applied on Brazilian Portuguese movie reviews. We used part-of-speech (POS) tags, movie ontology concepts and two available Portuguese opinion lexicons.

This paper is organized as follows: works about feature-based opinion mining are discussed in Section 2. Our approach is introduced in Section 3. Tests are discussed in Section 4. Finally, conclusion and future works are presented in Section 5.

2 FEATURE-BASED OPINION MINING

The works by Hu and Li [3] and Popescu and Etzinoni [4] are the most representative ones in this area of study. Hu and Li [3] use association rule mining while Popescu and Etzinoni [4] use the Pointwise Mutual Information (PMI) for feature extraction. According to Hu and Li [3] implicit features occur much less frequently than explicit ones. This paper focuses on features that appear explicitly in the reviews.

Most of the existing work on review mining and summarization is focused on product reviews [3, 4]. When people write a movie reviews, they probably comment not only on movie elements (e.g., music, vision effectcs, award, genre), but also on movie-related people (e.g., director, actor, writer, producer). Therefore, the commented features in movie reviews are much richer and more challenging than other domain, such as: hotel, restaurant and product. Zhuang et al. [7] have done a pioneer work on classifying and summarizing movie reviews by extracting high frequent opinion keywords. Feature-opinion pairs were identified by using a dependency grammar graph.

Binali et al. [8] present an overview about feature-based opinion mining. The following tasks are identified: the extraction of objects (entities

mentioned in reviews e.g., movie); the extraction of object features (components and attributes e.g., title); the detection of sentiment about object features (e.g., good title); the detection of sentiment about objects (the global sentiment expressed in relation to an entity e.g., recommended or not recommended); the comparison of two entities (e.g., movie A and movie B); the comparison of features of two entities (e.g., actors movie A and actors movie B). In our study, we intend to extract object features and detect sentiment about object features.

Feature-based opinion mining that uses ontologies, in the English language are [6, 9, 5, 10]. The literature shows that there are different levels of knowledge representation: authors using complex structures [6, 9, 10]—even if they do not use all the knowledge available—and authors using simple structures [5] for feature identification. A common point is the use of IMDb data. Unfortunately, the ontologies cited in [9, 5, 10] are not available. The only ontology we found was the Movie Ontology (MO[1]).

In this paper, we conducted the adaptation of the algorithm Polarity Recognizer in Portuguese (PIRPO) [11] applied to Brazilian Portuguese movie reviews and using MO concepts (Figure 1). PIRPO receives as input a set of reviews which are pre-processed in order to extract their sentences and detect which reviews are split into positive and negative segments. The system output is a list of sentences with polarity that reflects the polarity of the words characterising the concepts of the ontology in the reviews [11].

3 APPROACH

This approach is composed of two main steps: preprocessing and semantic orientation recogniser. These steps are described in detail below.

3.1 *Preprocessing*

The main objective of this step is to obtain the grammatical categories. For this task we used Portuguese TreeTagger[2]. The TreeTagger is a tool for annotating text with POS and lemma information. For example, the sentence “Um dos melhores filmes que já vi!” [“One of the best movies I have watched!”] and “É simplesmente o PIOR filme que vi nos últimos

[1] http://www.movieontology.org/

[2] http://gramatica.usc.es/ gamallo/tagger.htm

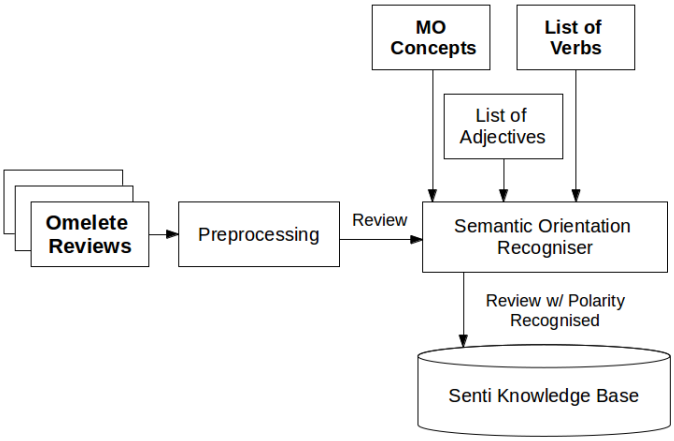


Fig. 1. PIRPO Information Architecture. Adapted from [11].

tempos." ["It is simply the WORST movie I have watched lately."] obtains the following lemmatized words accompanied by their grammatical categories:

Um DET um dos PRP+DET de melhores ADJ melhor filmes NOM filme que PR que já V <unknown> vi V ver ! SENT !

É V <unknown> simplesmente ADV simplesmente o DET o PIOR NOM pior filme NOM filme que PR que vi V ver nos P nos últimos V <unknown> tempos NOM tempo . SENT .

3.2 *Semantic Orientation Recogniser*

In this step, external resource was used, such as: ontology concepts and opinion lexicons.

The main idea is to use the opinion words around each movie concept in a review sentence to determine the opinion orientation. Still, the orientation of an opinion on a feature indicates whether the opinion is positive, negative or neutral.

In our work, features are represented by concepts the MO of ontology. Firstly, concepts are identified and extracted of pre-processed reviews.

For example:

Um DET um dos PRP+DET de melhores ADJ melhor **filmes NOM filme** que PR que já V <unknown> vi V ver ! SENT !

After, we used opinion lexicons, i.e., adjectives or verbs contained in SentiLex and OpLexicon for polarity identification. The adjectives or verbs around each movie feature identified are analysed.

For example, when we use the list of adjectives:

Um DET um dos PRP+DET de **melhores ADJ melhor filmes NOM filme** que PR que ja V <unknown> vi V ver ! SENT !

We identified the adjective "melhores" ["best"] near the word "filme" ["movie"]. In SentiLex this adjective is neutral and in OpLexicon is positive.

For example, when we use the list of verbs:

Um DET um dos PRP+DET de melhores ADJ melhor **filmes NOM filme** que PR que já V <unknown> **vi V ver** ! SENT !

We identified the verb "ver" ["watch"] around "movie" ["filme"]. In OpLexicon this verb is positive. SentiLex did not have this verb.

Finally, the output, a tuple containing the feature and polarity of objects, is stored in a database.

For example, tuple: (movie, positive).

4 TESTS

In this section, we evaluate the algorithm using the semantic orientation recogniser. We have conducted tests using the movie corpus, the MO concepts and Portuguese lexicons (SentiLex[3] and OpLexicon[4]). These resources are described below.

4.1 *Movie Corpus*

In order to build the movie corpus, initially we automatically got reviews about 1.160 movies on the website Omelete[5]. In these tests, 150 reviews were randomly selected. The corpus has only 8.999 words and 440 sentences. After that, TreeTagger is used to generate part-of-speech tags.

The manual annotation of the corpus was conducted by two people. The agreement between annotators was measured through the Kappa Statistics. The Kappa Statistics is the metric that evaluate concordance level in classification task. The value was moderate (Kappa 0.58) for agreement about opinion mining and fair (Kappa 0.39) for agreement

[3] http://xldb.fc.ul.pt/wiki/SentiLex-PT01

[4] http://ontolp.inf.pucrs.br/Recursos/downloads-OpLexicon.php

[5] http://omelete.uol.com.br/

about feature identification (see Table 1). We believe that the annotation has an acceptable value for the problem proposed in this study. In manual annotation the most frequent concepts were: *movie*, *actor*, *people* and *genre*.

Table 1. Kappa Statistics [12].

Interval	Agreement
< 0.00	Poor
0.00 to 0.20	Slight
0.21 to 0.40	**Fair**
0.41 to 0.60	**Moderate**
0.61 to 0.80	Substantial
0.81 to 1.00	Almost Perfect

4.2 *Movie Ontology*

In this study, we used the concepts of MO. MO aims at providing controlled vocabulary to describe semantically related concepts, such as a movie, genre, director, actor and individuals—for example "A Era do Gelo 3" ["Ice Age 3"], "Animação" ["Animation"], "Carlos Sandanha" and "Márcio Garcia", respectively. This ontology was described in OWL and is available in English. MO provides hierarchies of concepts and a set of instances. Only 11 out of 78 concepts (Table 2) were identified in the movie corpus, such as: *action*, *actor*, *director*, *fun*, *genre*, *kids*, *love*, *movie*, *place*, *person*, and *thrilling*.

Table 2. Movie Ontology Metrics.

Metrics	Value
Number of Concepts	78
Number of Object Properties	38
Number of Data Properties	4
Number of Individuals	282

4.3 *Portuguese Lexicons*

In the literature, many papers about opinion mining use SentiWordNet[6] [13]. SentiWordNet 3.0 [14] is a fragment of WordNet 3.0 manually annotated for positivity, negativity and neutrality. Each synset has three numeric values in the interval 0 to 1 for positive, negative and neutral. Both [5] and [6] calculated the polarity of the features using SentiWordNet. This resource has nearly 117.000 words in English.

There are languages in which this type of resource started to be built recently, as is the case of Portuguese. SentiLex and OpLexicon, Portuguese opinion lexicons, appeared in 2010.

SentiLex 2.0 [15] has 7.014 lemmas and 82.347 inflected forms (of nouns, verbs, adjectives and adverbs). SentiLex is useful for opinion mining applications involving European Portuguese, in particular for detecting and classifying sentiments and opinions. In tests we used 16.833 SentiLex adjetives and 28.989 SentiLex verbs (Table 3).

OpLexicon [16] has nearly 30.322 words and was built based on a corpus, thesaurus and translated texts. Three different opinion lexicons generated by each techniques are conjoined to create a large lexicon for Brazilian Portuguese. In tests we used 23.433 OpLexicon adjetives and 6.889 OpLexicon verbs (Table 3).

Table 3. Portuguese Lexicons.

Lexicon	Number of Words
SentiLex Adjectives	16.833
SentiLex Verbs	28.989
OpLexicon Adjectives	23.433
OpLexicon Verbs	6.889

Even though SentiLex or OpLexicon are small and new, we used this lexicon. Both have three numeric values: 1 (positive), -1 (negative) and 0 (neutral).

4.4 *Results*

The results are presented in Table 4. In the table, lines 2 and 8 give the results that uses OpLexicon adjectives and MO concepts for positive and

[6] http://sentiwordnet.isti.cnr.it/

negative polarity recognition. The results indicate that precision for negative polarity recognition is poor. Lines 3 and 9 show corresponding results that uses SentiLex adjectives and MO concepts. We can see that the f-measure is the best result. Lines 4 and 10 give the results that uses OpLexicon verbs and MO concepts for positive and negative polarity recognition. The results also indicate that precision for negative polarity recognition is poor. Lines 5 and 11 show corresponding results that uses SentiLex verbs and MO concepts. We can see that the f-measure is the same as positive polarity recognition as negative polarity recognition.

In summary, the best results are obtained when using SentiLex adjectives, the f-measure of 73% for positive polarity recognition and 76% for negative polarity recognition.

Table 4. Results for Feature-Based Opinion Mining.

		Precision	Recall	F-Measure
Positive	OpLexicon(ADJ) + MO(C)	1.0	0.45	0.62
	SentiLex(ADJ) + MO(C)	**0.87**	**0.63**	**0.73**
	OpLexicon(V) + MO(C)	1.0	0.40	0.57
	SentiLex(V) + MO(C)	1.0	0.50	0.66
	OpLexicon(ADJ and V) + MO(C)	1.0	0.43	0.61
	SentiLex(ADJ and V) + MO(C)	0.90	0.57	0.70
Negative	OpLexicon(ADJ) + MO(C)	0.08	1.0	0.15
	SentiLex(ADJ) + MO(C)	**0.66**	**0.88**	**0.76**
	OpLexicon(V) + MO(C)	0.04	1.0	0.08
	SentiLex(V) + MO(C)	0.50	1.0	0.66
	OpLexicon(ADJ and V) + MO(C)	0.11	1.0	0.20
	SentiLex(ADJ and V) + MO(C)	0.63	0.92	0.75

4.5 *Error Analysis*

In the following, we show a few examples to analyse some typical errors. Bold is used to denote feature objects and adjectives or verbs polarity indicates.

Example 1:

Sentence: “incrível o filme, me emocionei em alguns momentos, perfeitos.” [“amazing film, moved in some moments, perfect.”]

Annotated Sentence: (‘incrível’, ‘ADJ’), (‘o’, ‘DET’), (‘**filme,me**’, ‘NOM’), (‘emocionei’, ‘V’), (‘em’, ‘PRP’), (‘alguns’, ‘P’), (‘momentos,perfeito’, ‘V’), (‘.’, ‘SENT’)

Error: filme,me NOM

Expected: filme NOM

Here the word “filme” [“movie”] is grouped with comma and pronoun “me” [“me”]. In fact, there are many writting error in movie reviews. To solve the problem, a heuristic should be build.

Example 2:

Sentence: “... esse filme apesar de ruin causou ...” [“... this movie although bad cause ...”]

Annotated Sentence: ... (‘esse’, ‘DET’), (‘**filme**’, ‘NOM’), (‘apesar’, ‘L’), (‘de’, ‘PRP’), (‘**ruin**’, ‘NOM’), (‘causou’, ‘V’) ...

Error: ruin NOM

Expected: ruim ADJ

The word “ruim” [“bad”] is misspelled. Maybe phonetic algorithm or spellchecker should be used to solve the problem.

Example 3:

Sentence: “Filme excelente, elenco competente, direção fantástica, trilha sonora de Alberto Iglesias no mínimo brilhante ...” [“Excellent movie, competent cast, fantastic direction, trowel Alberto Iglesias score of at least brilliant ...”]

Annotated Sentence: (‘**Filme**’, ‘NOM’), (‘excelente’, ‘ADJ’), (‘,’, ‘VIRG’), (‘**elenco**’, ‘V’), (‘competente’,‘ADJ’), (‘,’, ‘VIRG’), (‘**direção**’, ‘V’), (‘fantástica’, ‘V’), (‘,’, ‘VIRG’), (‘**trilha**’, ‘NOM’), (‘**sonora**’, ‘ADJ’), (‘de’, ‘PRP’), (‘Alberto’, ‘NOM’), (‘Iglesias’, ‘NOM’), (‘no’, ‘PRP+DET’), (‘mínimo’, ‘NOM’), (‘brilhante’, ‘ADJ’) ...

Error: (movie, positive)

Expected: (movie, positive), (cast, positive), (direction, positive), and (soundtrack, positive)

This sentence has a (movie, positive), (cast, positive), (direction, positive), and (soundtrack, positive) tuple but the algorithm only detected a (movie, positive) tuple for review.

5 Conclusion and Future Works

In summary, the application of the adaptation of the algorithm proposed in [11] in the movie domain presented good results. In future works we intend to use the complete ontology (*concepts*, *properties*, *instances* and

hierarchies). Furthermore, we intend to redo these tests in other domains, such as: education, politics, and others.

Aiming at improving the results, the preprocessing step might be broadened. We intend to use lemmatizer in preprocessing and properties, instances and hierarchies of ontologies in identification feature.

Also, we intend to add lists of adverbs and list of nouns in polarity identification. At last, we would apply a set of linguistic rules, such as negatives and intensifiers which vary from language to language [17]. In opinion mining, the negation is a more common linguistic construction that affects the polarity. It is not only transmitted by negative words, but also by lexical units, such as diminutives and connectives. The works described in [17–19] were considered pioneers in the negation model in sentiment analysis.

Besides, we intend to study ways of solving problems such as the use of different words (e.g., *filmes* and *filmão*) that refer to the same concept.

6 Acknowledgments

We thank the Brazilian funding agency CAPES/FAPERGS for the scholarship granted.

References

1. Liu, B.: Sentiment analysis and subjectivity. In: Handbook of Natural Language Processing, Second Edition. CRC Press, Taylor and Francis Group (2010)
2. Bhuiyan, T., Xu, Y., Josang, A.: State-of-the-art review on opinion mining from online customer's feedback. In: 9th Asia-Pacific Complex Systems Conference. (2009)
3. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: 19th national conference on Artifical intelligence. (2004) 755–760
4. Popescu, A., Etzioni, O.: Extracting product features and opinions from reviews. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. (2005) 339–346
5. Zhao, L., Li, C.: Ontology based opinion mining for movie reviews. In: 3rd International Conference Knowledge, Science, Engineering and Management. (2009)
6. Peñalver Martínez, I., Valencia-Garcia, R., Garcia-Sanchez, F.: Ontology-guided approach to feature-based opinion mining. In: International Conference on Applications of Natural Language to Information Systems. (2011)

7. Zhuang, L., Jing, F., Zhu, X.: Movie review mining and summarization. In: 15th ACM international conference on Information and knowledge management. (2006)
8. Binali, H., Potdar, V., Wu, C.: A state of the art opinion mining and its application domains. In: International Conference on Industrial Technology. (2009)
9. Shein, K.P.P.: Ontology based combined approach for sentiment classification. In: 3rd International Conference on Communications and Information Technology. CIT'09, Stevens Point, Wisconsin, USA, World Scientific and Engineering Academy and Society (2009) 112–115
10. Zhou, L., Chaovalit, P.: Ontology-supported polarity mining. Journal of the American Society for Information Science and Technology **59** (2008) 98–110
11. Chaves, M., Freitas, L., Souza, M., Vieira, R.: Pirpo: An algorithm to deal with polarity in portuguese online reviews from the accommodation sector. In: 17th International Conference on Applications of Natural Language Processing to Information Systems. (2012)
12. Landis, J., Koch, G.: The measurement of observer agreement for categorical data. Biometrics **33** (1977) 159–174
13. Esuli, A., Sebastiani, F.: Sentiwordnet: A publicly available lexical resource for opinion mining. In: 5th International Conference on Language Resources and Evaluation. (2006)
14. Baccianella, S., Esuli, A., Sebastiani, F.: Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In: 7th International Conference on Language Resources and Evaluation. (2010)
15. Silva, M.J., Carvalho, P., Sarmento, L.: Building a sentiment lexicon for social judgement mining. In: 10th International Conference Computational Processing of the Portuguese Language. (2012)
16. Souza, M., Vieiras, R., Busetti, D., Chishman, R., Alves, I.M.: Construction of a portuguese opinion lexicon from multiple resources. In: 8th Brazilian Symposium in Information and Human Language Technology. (2012)
17. Polanyi, L., Zaenen, A.: Contextual valence shifters. AAAI Spring Symposium on Attitude **20** (2004) 1–10
18. Kennedy, A., Inkpen, D.: Sentiment classification of movie reviews using contextual valence shifters. Computational Intelligence **22** (2006) 110–125
19. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics **35** (2009) 399–433

LARISSA A. DE FREITAS
FACULDADE DE INFORMÁTICA,
PUCRS,
PORTO ALEGRE, BRAZIL
E-MAIL: <LARISSA.FREITAS@ACAD.PUCRS.BR>

RENATA VIEIRA
FACULDADE DE INFORMÁTICA,
PUCRS,
PORTO ALEGRE, BRAZIL
E-MAIL: <RENATA.VIEIRA@PUCRS.BR>

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 159–173
RECEIVED 29/12/12 ACCEPTED 11/01/13 FINAL 15/05/13

Twitter Emotion Analysis in Earthquake Situations

BAO-KHANH H. VO AND NIGEL COLLIER

National Institute of Informatics, Japan

ABSTRACT

Emotion keyword spotting approach can detect emotion well for explicit emotional contents while it obviously cannot compare to supervised learning approaches for detecting emotional contents of particular events. In this paper, we target earthquake situations in Japan as the particular events for emotion analysis because the affected people often show their states and emotions towards the situations via social networking sites. Additionally, tracking crowd emotions in the Internet during the earthquakes can help authorities to quickly decide appropriate assistance policies without paying the cost as the traditional public surveys. Our three main contributions in this paper are: a) the appropriate choice of emotions; b) the novel proposal of two classification methods for determining the earthquake related tweets and automatically identifying the emotions in Twitter; c) tracking crowd emotions during different earthquake situations, a completely new application of emotion analysis research. Our main analysis results show that Twitter users show their Fear and Anxiety right after the earthquakes occurred while Calm and Unpleasantness are not showed clearly during the small earthquakes but in the large tremor.

KEYWORDS: *Twitter, Social media, Emotion recognition, Sentiment analysis, Earthquake, Japan*

1 INTRODUCTION

Sentiment analysis and emotion analysis have been increasingly studied in recent years thanks to the population of big text data. Although both emotion analysis and sentiment analysis apply psychology and cognitive science to computer science applications, emotion analysis mainly targets the fine-grained emotions while sentiment analysis detects simple attitudes such as positive and negative [1]. According to Scherer's typology of affective states [2], emotion is a relatively brief episode of synchronized response to the evaluation of an external or internal event as being of major significance; attitude is relatively enduring, affectively coloured beliefs, preferences, and predispositions towards objects or persons. Hence, emotion analysis is more appropriate for particular events rather than polarity sentiment analysis. Accordingly, our work addresses emotion analysis in different earthquake events for tracking and comparing the emotion variations during these events.

Tracking crowd responses, especially opinions and emotions towards an earthquake could provide valuable situational awareness for not only authorities to manage bad situations but also for psychological scientists to understand human behaviors in such situations. When a natural disaster like an earthquake occurs, the public agencies need the up-to-the-minute the affected people's responses to tailor emergency warnings, to aid the victims, and to calm down the public anxiety. With the strong development of social media, tracking emotions becomes easier, faster and more reliable than using the traditional public surveys or polls [3]. The social media service in our research is Twitter that allows users to send and read instant text-based messages or "tweets". This work analyzes Japanese tweets in Tokyo for emotion tracking during earthquakes because of the following reasons:

- Japan is the third in the world for total Twitter usage in 2012, and Tokyo is one of the top three cities in terms of tweets.[1]
- Japan often encounters earthquakes. The Great East Japan Earthquake in 2011 affected millions of people in Japan including Tokyo.
- As Japanese does not contain white space between words like English, Twitter users can convey more information within 140 characters of a tweet in Japanese than English.

[1] See http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/

We track emotions of the four earthquake dates by the Coordinated Universal Time (UTC): March 11th, April 7th, April 11th and July 10th, 2011 for the sake of comparison between earthquakes. To analyze emotions during earthquake dates, the first important task is to identify tweets related to the earthquake. These earthquake concerned tweets can be recognized in the Twitter data by our proposed supervised classification method. For emotion analysis task, we annotated the training data to Calm, Unpleasantness, Sadness, Anxiety, Fear, and Relief emotion. We use another supervised method for emotion recognition and then plot the categorized emotions into the time interval of each earthquake. Various features and machine learning models are used for both classification methods for selecting the best features and models.

To the best of our knowledge, our work is the first research upon emotion recognition and tracking for Twitter data in earthquake situations whereas the earthquake related Twitter analysis works do not consider the emotion aspects of users. Nevertheless, the choice of emotions and the classification methods with appropriate features for Japanese social media are the first contributions for Twitter emotion analysis applications.

2 RELATED WORK

Sentiment analysis and emotion analysis are the tasks of identifying the attitude and emotion classes of the investigated document [1]. The word "document" we use here has a general meaning as it can refer to a linguistic unit including a single sentence, a paragraph, and a document of many paragraphs. There are three main approaches used for emotion and attitude identification:

- Textual keyword spotting approach: Using a set of emotion words mostly adjectives and adverbs defined by specific lexical resources like Google Profiles of Mood States [4], Linguistic Inquiry and Word Count dictionaries [5] to select documents such as tweets and Facebook statuses containing such keywords in distinct emotions. This method is used widely for social media due to the large scale and the noise of social network data because it can remove a fair amount of irrelevant documents. While this approach is directly applicable for English with no modification in adjectives and adverbs, it is difficult to be applied for Japanese, an agglutinative language.
- Rule-based linguistic approach: Each sentence is processed in stages, including symbolic cue, abbreviations, sentence parsing, and word /

phrase / sentence-level analysis [6]. This rule-based approach often has limitations due to the diversity of natural language, especially the language in social media.

– Feature-based classification approach: This empirical approach has been used from the first applications of sentiment analysis on movie reviews to current sentiment and emotion analysis applications on social media [7]. Alm et al. [8] consider determining emotion of a linguistic unit is a multi-class classification problem. This supervised learning approach generates a function that maps linguistic units to the desired emotion by looking at the features derived from linguistic unit-emotion examples of the function. The features can be n-grams, bag of words, or Twitter features such as re-tweets, hashtags, replies, punctuations, and emoticons [9].

There are a few works on sentiment analysis in crisis contexts similar to earthquake events which are hurricanes [3], and gas explosion [10]. These works also use the feature-based classification method for English sentiment analysis. Mandel et al. [3] experiment with features: two tokenizer alternatives, stop word removal, frequency pruning, worry lexicon, humor lexicon, and emoticon; and classifiers: Maximum Entropy, Decision Tree, and Naive Bayes to choose the best features and model for classifying tweets to Irene Hurricane concerned or unconcerned. Nagy and Stamberger [10] combine the available English sentiment data comprising SentiWordNet, emoticons, AFNN for classification by Bayesian Network.

Most of Japanese emotion analysis applications are for blogs [11] and Japanese emotion-provoking sentences collected in the Web [12], not Twitter data that is shorter and less information. These applications restrict the data to explicit emotive sentences that contain emotive words, not inferred emotive sentences. In another effort to analyze emotions automatically for Japanese, a Japanese WordNet Affect is being developed [13], but it is not completed yet.

For Twitter analysis in earthquake situations, the available works only concentrate on earthquake and rumor detection [14], not sentiment analysis in earthquake situations.

Due to the listed disadvantages and advantages of the related works as well as our wish to handle inferred emotive tweets, the feature-based classification method is feasible for our purpose of emotion analysis in earthquake situations.

3 METHODOLOGY

3.1 *Emotion selection for earthquake situations*

Plenty of English emotion analysis researches classify documents [1, 8] to Ekman's 6 basic emotions [15]: Surprise, Happiness, Anger, Fear, Disgust, and Sadness. Meanwhile, some Japanese works [11] base on Nakamura's Japanese emotion dictionary [16] with 10 emotion types: Excitement, Shame, Joy, Fondness, Dislike, Sorrow, Anger, Surprise, Fear, and Relief; and Tokuhisa et al. [12] use 10 emotion classes: Happiness, Pleasantness, Disappointment, Unpleasantness, Loneliness, Sadness, Anger, Anxiety, Fear, and Relief from Teramura dictionary [17]. It is clear that the emotions used by Tokuhisa et al. are more separate than Ekman's emotions. Besides, Nakamura's emotions have Excitement, Joy, Fondness are quite similar and along with Shame, they are not appropriate in earthquake contexts. Therefore, we construct our emotions from 10 emotions of Tokuhisa et al.

For the reason of emotion analysis in earthquake situations, we need to adjust the 10 emotion classes to match such situations. We remove Happiness and Pleasantness because we think Happiness and Pleasantness are too positive to fit in negative situations like earthquakes. In order to show positive emotion in such situations, Calm emotion may be the most appropriate one. Anger, Disappointment, and Unpleasantness can be grouped into Unpleasantness. Sadness can include Loneliness. To sum up, we use 6 emotion classes for data annotation and classification: Calm, Unpleasantness, Sadness, Anxiety, Fear, and Relief. This choice of emotions is also verified in the data.

3.2 *Earthquake related tweet identification and emotion analysis methods*

Before classifying emotions in the crisis situation, we need to perform the task of selecting only the tweets related to the earthquake. We call these tweets as Concerned tweets, the others are Unconcerned, and the task of filtering Concerned tweets as earthquake related tweet identification. In order to filter the Concerned tweets out of the Unconcerned ones, instead of using the simple keyword spotting approach, we apply the feature-based classification method because the word spotting approach can not cover all the tweets related to the earthquake as the empirical method does. For example, it is difficult to redefine the word list for tweets related

to the food shortage, family's safety, etc. For the purpose of emotion comparison after earthquakes, we select the tweets right after the earthquakes on the earthquake dates. Similarly, we apply the feature-based classification method for the emotion analysis task for the reasons mentioned in Section 2.

The two tasks can share the same linguistic features: n-grams, bag-of-words, stop-word removal, and emoticons. The purpose for using various features is comparing applying them in classifiers to select the best feature and classifier for the two classification tasks.

FEATURES

Emoticons Japanese emoticons or better known as kaomoji are much more complex than English emoticons, thus it is hard to fully detect emoticons in Japanese text [11]. For serving our main purpose of classification, we detect emoticons in tweets by using complex regular expressions instead of the techniques mentioned in [11]. The detected emoticons are used for two kinds of testing features: a) they are removed from tweets to become no-emoticon feature or b) they are grouped to 10 emotions of Nakamura [16] thank to the available of CAO preliminary emoticon lists mentioned in [11]; we call this feature as grouped-emoticon feature.

Bag of words Japanese words are not separated by space. Therefore, we need to use a Japanese morphological processing tool to segment words. However, because Twitter language is informal with many new words and slangs, morphological processing tools can not correctly analyze morphemes of tweets. To solve a part of this problem, we need to make a normalization dictionary for convert the wrongly segmented words, out-of-dictionary words to the correct words for adding unique words to the bag. The bag-of-word feature has two options: removing or not removing stop words beside the options of including emotions of emoticons or removing all emoticons.

N-grams Due to the bad performance probability of the Japanese morphological analysis tool for Twitter data, we think we should use n-gram features. We intend to try the uni-gram, bi-gram, and tri-gram features separately as well as the combinations of uni-gram and bi-gram; uni-gram, bi-gram and tri-gram. All of these n-gram features have the options of including emotions of emoticons or removing emoticons.

Stop-words We use the stop-word removal feature accompany with the bag-of-word feature. For the reason that there is no official stop-word list in Japanese, we need to find the most appropriate list for our purpose.

EARTHQUAKE CONCERNED/UNCONCERNED TWEET CLASSIFICATION We use different features in Section 5.1 with Support Vector Machine (SVM), Naive Bayes, Multinomial Naive Bayes (MNB), Decision Tree (J48), and Maximum Entropy (MaxEnt) models for the purpose of comparison to select the best features and models.

EMOTION CLASSIFICATION For this multi-class classification task, we use the Sequential Minimal Optimization (SMO) and Multinomial Naive Bayes (MNB) models with features in Section 5.1 for classification.

4 DATA

4.1 *Data collecting and pre-processing*

We collected Twitter data for five months, starting from March 10th 2011 to July 31st 2011 using Twitter API[2] with the geolocation feature set to track messages originating within Tokyo because Tokyo, the capital city with biggest population and Twitter users of Japan, is near Tohoku area where the 2011 Great East Japan Earthquake occurred which also affected Tokyo residences. In order to select only useful informations for emotion analysis, we need to pre-processed the tweets. From the original tweets, we parsed them and changed their encoding to UTF-8. From this parsed tweets, we selected only the tweets on the days of the big earthquakes or aftershocks according to `http://en.wikipedia.org/wiki/ List_of_foreshocks_ and_aftershocks_of_the_2011_Tohoku_earthquake`. We chose the 6 earthquakes and aftershocks from this site which shown in Table 1. All of the tweets were daily selected by UTC time.

As the result, the corpus consists of 4 files for 4 days of 6 earthquakes. This corpus has totally 110,715 tweets. For the purpose of processing only Japanese tweets, we selected only Japanese tweets by a language detection program. We then removed the spam tweets from these Japanese tweets. The spam tweets in our research context are advertising

[2] http://dev.twitter.com/

Table 1. Significant earthquakes in Tohoku area from March to July, 2011

Japan Time	Magnitude	Intensity (shindo)
2011-03-11 14:46	Mw 9.0	7
2011-03-11 15:15	Mw 7.9	upper 6
2011-03-11 15:25	Mw 7.7	4
2011-04-07 23:32	Mw 7.1	upper 6
2011-04-11 17:16	Mw 6.6	lower 6
2011-07-10 09:57	Mw 7.0	4

tweets, automatic tweets generated by applications, and location check-in information because they do not show human's emotions. Totally, the number of the spam-filtered Japanese tweets is 70,725. Table 2 shows the concrete spam-filtered Japanese tweet numbers of 4 days.

Table 2. Spam-filtered Japanese tweet numbers in 4 days

UTC Date	Tweet number
2011-03-11	19,420
2011-04-07	15,893
2011-04-11	16,700
2011-07-10	18,712
Total tweets	70,725

4.2 *Data annotation*

A part of the spam-filtered-Japanese tweets was annotated with 6 emotions and earthquake not-related tweets following our emotion definitions and annotation guideline. The tweets annotated with 6 emotions are considered as Concerned tweets while earthquake not-related tweets are Unconcerned tweets. The Concerned data includes messages directly convey the emotions with obvious emotion words and messages with inferred emotions. The tweets were annotated to 6 emotions by two annotators. The annotator chose only one emotion class or Unconcerned class for a tweet. The inter-annotator agreement was calculated using Fleiss' Kappa statistics [18]. The measured Kappa coefficients for Concerned and 6 emotions are 0.96 and 0.684, respectively. Only the tweets annotated with the same class were examined as the training tweets.

Table 3 shows the distribution of the annotated data in which the sum of 6 emotion tweets equals the number of Concerned tweets and Unconcerned tweets for the sake of balancing the training set.

Table 3. Distribution of annotated data

Unconcerned	Concerned	Calm	Unpleasantness	Sadness	Anxiety	Fear	Relief
1905	1905	155	310	4	580	635	221

From observing the annotated data, we found that Sadness has only 4 tweets over 1905 tweets. There are two reasons of this low appearance of Sadness tweets in the annotated data: a) The unbalanced selection of data; b) Not many people in Tokyo feel sad because they were not effected severely by the earthquake and tsunami like the people of the Tohoku area. Therefore, we remove Sadness from our training data set for emotion classification. The training data now has 5 emotion classes: Calm, Unpleasantness, Anxiety, Fear, and Relief. Accordingly, these 5 emotions are our targeted emotions for analysis.

5 Experiments and evaluation

5.1 *Experiment settings*

We implemented the proposed methods with following specifications:

Features

Emoticons Although we do not use the emoticon detection methods mentioned in [11], we could identify emoticons effectively with our regular expression for our data. As the Japanese emoticons often start with non-word characters and different kinds of brackets, the beginning of regular expression are the representation of non alphabetical and Japanese word character with a set of brackets. The central part of emoticons are three or more than three word characters. The ending parts of these emoticon regular expressions are similar to their starting parts.

After detecting emoticons in tweets, we remove them for testing non-emoticon feature or assign the emotions for these emoticons using the two emoticon lists of CAO: the list of full characters of each emoticon and the list of only three main characters of each emoticon. Firstly, we

used the full character emoticon list for identifying emoticons that appear in this list. If the emoticons did not appear in this list, we then used the list of three main characters to replace the emoticons with special emotion names.

Bag of words Mecab[3], a Japanese dependency structure analyzer was used for word segmentation. We did not change the segmented words to the dictionary form. We only changed the wrongly segmented words to the correct ones by using our manual normalization dictionary. We selected the features with frequencies equal or greater than 5 for the experimental purpose and reducing the numbers of features.

N-grams We applied frequency pruning for bi-gram, tri-gram, the combination of uni-gram and bi-gram, and the combination of unigram, bi-gram and tri-gram. We selected the features with frequencies equal or greater than 5.

CLASSIFIER MODELS We use Weka [19], the collection of machine learning algorithms, for classification tasks. Options of all algorithms were set as default values for 10-fold cross-validation classification.

5.2 *Earthquake Concerned/Unconcerned tweet classification*

Table 4 shows a part of the classification results in Precision - Recall - F-measure order. List of models are in the left column while some features are in the first row. The first listed feature is bi-gram with removing emoticons out of tweets. The second feature is bi-gram with emoticons grouped into 10 emotions. The third feature is the combination of unigram, bi-gram and tri-gram with the emoticons grouped into 10 emoticons. All of these features were selected based on their appearance frequencies in the feature sets.

The best result (F-measure = 87.8) comes to the combination of unigram, bi-gram, and tri-gram with Multinomial Naive Bayes model.

5.3 *Emotion classification*

The combination of uni-gram, bi-gram, and tri-gram with MNB model again bring the best results. We classified 5 emotions with this feature

[3] http://code.google.com/p/mecab

Table 4. A part of earthquake Concerned/Unconcerned tweet classification results (Precision - Recall - F-measure)

Models	No emoticon - Bi	Grouped emoticons - Bi	Grouped emoticons - Uni-bi-tri
Naive Bayes	79.9 - 79.9 - 79.9	80.2 - 80.1 - 80.1	82.2 - 82.1 - 82.1
SVM	25.6 - 50.6 - 34	25.6 - 50.6 - 34	55.1 - 61.7 - 55.3
MaxEnt	69.6 - 69.4 - 69.3	67 - 66.7 - 66.7	55.1 - 61.7 - 55.3
J48	81.2 - 81.2 - 81.2	81.2 - 81.1 - 81.2	84.6 - 82.1 - 83.1
MNB	87.2 - 87.1 - 87	87.2 - 87.2 - 87.2	**87.9 - 87.8 - 87.8**

and obtained **64.7** and 62.2 F-measure for **MNB** and SMO, respectively. However, Calm emotion is classified with 22.4 F-measure by MNB and 22.3 by SMO. This springs from the fuzzy characteristics of this emotion class. This Calm class includes the tweets about the earthquake without distinctive emotions towards earthquake problems.

5.4 *Emotion analysis in earthquake dates*

We used the MNB model of uni-gram, bi-gram, and tri-gram feature for both classifying earthquake Concerned/Unconcerned tweet and emotions of tweets in earthquake dates. Because of the purpose of tracking emotions during earthquake situations, we only consider the emotions from the time when each earthquake occurred until the end of that day by UTC time. More clearly, we plot the time in Japan time from the period of each earthquake until 9 AM of the next day because Japan time is GMT+9. Emotions are tracked in 30 minute time unit. Figure. 1, figure. 2, and figure. 3 show the emotions plotted in March 11th, April 07th, April 11th, and July 10th, 2012.

Table 5. Pearson correlation coefficient of Fear and Anxiety

Date	Pearson coefficient
2011-03-11	0.85
2011-04-07	0.96
2011-04-11	0.96
2011-07-10	0.89

A clear notice of these earthquakes is that Fear emotion is always dominant when earthquakes started and rapidly decreases to nearly cor-

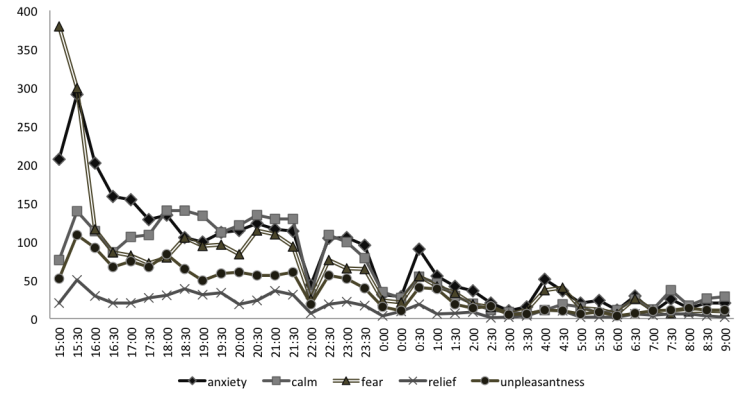


Fig. 1. Twitter emotions in March 11th, 2012

related with Anxiety emotion. Table 5 shows the Pearson correlation coefficient of Fear and Anxiety of 4 earthquake dates. The peaks of Fear and Anxiety mostly associate with the real earthquakes. For example, earthquakes in March 11th, 2011 that showed in Table 1 and a small earthquake around 8 AM of April 12th, Japan time. The earthquakes with higher intensity (March 11th, April 07th, and April 11th) result more tweets than the earthquake with lower intensity (July 10th). Because of the small amount of tweets in July 10th, the small peaks of Fear emotions (around 13:00, 16:00, and 21:00) are not really correlated with the real earthquakes. Therefore, we need to improve the Concerned tweet classification with other features such as the tweet amount in a specific time scale.

Except for March 11th when all the emotions significantly variate, other earthquakes show the low level of Calm, Relief, and Unpleasantness emotion because March 11th earthquake was the biggest earthquake brought various issues including the unpleasantness of transportations, phone communications, and the relief of surviving from this big earthquake. March 11th earthquake also has the Anxiety emotion has higher volumes than Fear emotion because of the intensity of this earthquake and other worries for the Fukushima Power Plant and tsunami. Calm emotion changes from lower than Anxiety and Fear from the earthquakes happened at nearly 15:00 to higher than these emotions at 18:00 that shows Twitter users became calmer after the first three hours of Anxiety and Fear.

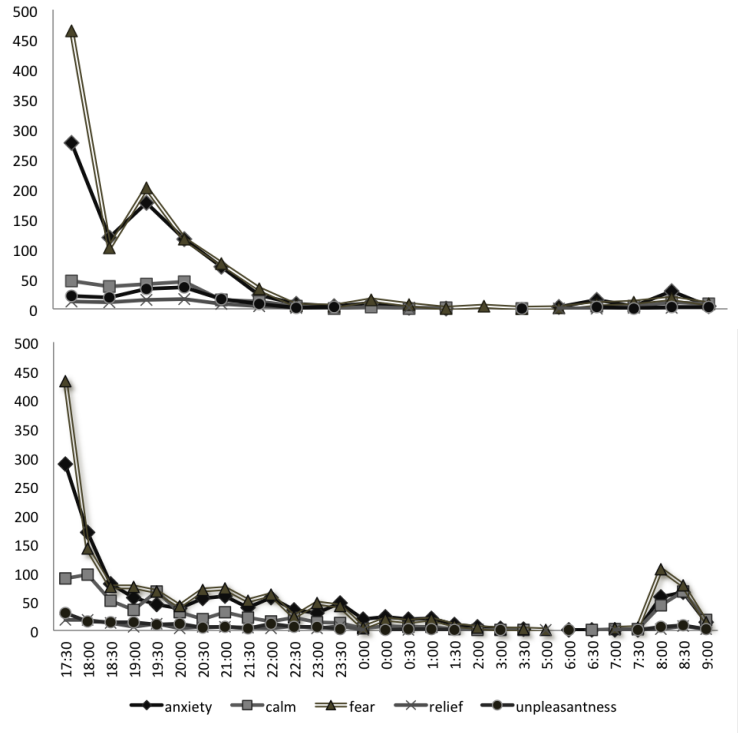


Fig. 2. Twitter emotions in April 07th (left) and April 11th (right), 2012

6 CONCLUSION

In this paper, we presented a novel application of Twitter sentiment analysis: tracking emotions in earthquakes for better managing the situations. To accomplish this purpose, we are the pioneer to propose the appropriate emotions for earthquake situations. We also propose the earthquake Concerned/Unconcerned tweet classification and emotion tweet classification which are completely different from the available works related to earthquake. Simple n-gram features are the best choice for classifying the agglutinative Japanese language and noisy Twitter language. Emotions in the time interval of earthquake dates reveal the insights of Twitter users during such hard time. Fear and Anxiety emotion always correlate with the occurrences of big earthquakes. Calm emotion will dominant after

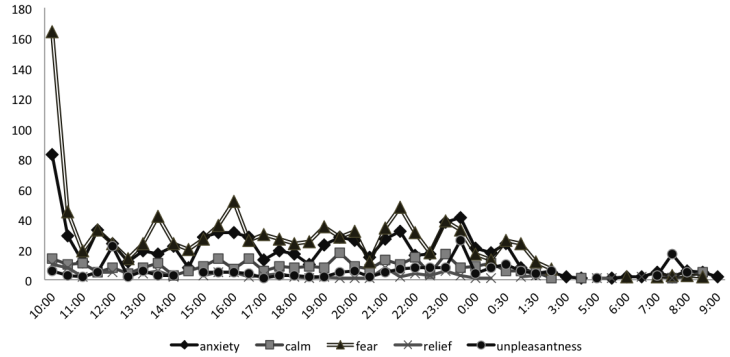


Fig. 3. Twitter emotions in July 10th, 2012

the first hours of Fear and Anxiety because of the earthquakes. Although Relief and Unpleasantness do not present the significant tweet amounts in earthquakes, they are important for the management purposes.

REFERENCES

1. Strapparava, C., Mihalcea, R.: Learning to identify emotions in text. In: 2008 ACM symposium on Applied computing, ACM (March 2008) 1556–1560
2. Scherer, K.R.: Emotion as a multicomponent process: A model and some cross-cultural data. Review of Personality and Social Psych **5** (1984) 37–63
3. Mandel, B., Culotta, A., Boulahanis, J., Stark, D., Lewis, B.: A demographic analysis of online sentiment during hurricane Irene. In: NAACL-HLT Workshop on Language in Social Media. (2012)
4. Bollen, J., Mao, H., Zeng, X.J.: Twitter mood predicts the stock market. Journal of Computational Science **2**(1) (2010) 1–8
5. Kramer, A.D.: An unobtrusive behavioral model of "gross national happiness". In: 28th International Conference on Human Factors in Computing Systems, April 10-15. (2010)
6. Neviarouskaya, A., Prendinger, H., Ishizuka, M.: Affect analysis model: Novel rule-based approach to affect sensing from text. International Journal of Natural Language Engineering **17**(1) (2011) 95–135
7. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1–2) (January 2008) 1–135
8. Alm, C.O., Roth, D., Sproat, R.: Emotions from text: machine learning for text-based emotion prediction. In: Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing. (2005) 579–586

9. Barbosa, L., Feng, J.: Robust sentiment detection on Twitter from biased and noisy data. In: 23rd International Conference on Computational Linguistics, COLING. (2010) 36–44
10. Nagy, A., Stamberger, J.: Crowd sentiment detection during disasters and crises. In: 9th International Conference on Information Systems for Crisis Response and Management, ISCRAM. (2012)
11. Ptaszynski, M., Dybala, P., Rzepka, R., Araki, K., Momouchi, Y.: YACIS: A five-billion-word corpus of Japanese blogs fully annotated with syntactic and affective information. In: 2nd Symposium on Linguistic and Cognitive Approaches To Dialog Agents, LaCATODA. (2012)
12. Tokuhisa, R., Inui, K., Matsumoto, Y.: Emotion classification using massive examples extracted from the Web. In: 22nd International Conference on Computational Linguistics, COLING. Volume 1. (2008)
13. Torii, Y., Das, D., Bandyopadhyay, S., Okumura, M.: Developing Japanese WordNet Affect for analyzing emotions. In: Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, WASSA 2011, 49th Annual Meeting of the Association for Computational Linguistics (ACL). (2011) 80–86
14. Sakaki, T., Okazaki, M., Matsuo, Y.: Earthquake shakes Twitter users: Real-time event detection by social sensors. In: 19th International Conference on World Wide Web, WWW 2010, April 26-30. (2010)
15. Ekman, P.: Facial expression and emotion. American Psychologist **48**(4) (1993) 384–392
16. Nakamura, A.: Kanjo hyogen jiten (in Japanese). Tokyodo Publishing (1993)
17. Teramura, H.: Japanese Syntax and Meaning (in Japanese). Kurosio Publishers (1982)
18. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin **76**(5) (1971) 378–382
19. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, H.I.: The weka data mining software: An update. SIGKDD Explorations **11**(1) (2009)

BAO-KHANH H. VO
NATIONAL INSTITUTE OF INFORMATICS,
2-1-2 HITOTSUBASHI, CHIYODA-KU,
TOKYO 101-8430, JAPAN
E-MAIL: <KHANHVO@NII.AC.JP>

NIGEL COLLIER
NATIONAL INSTITUTE OF INFORMATICS,
2-1-2 HITOTSUBASHI, CHIYODA-KU,
TOKYO 101-8430, JAPAN
E-MAIL: <COLLIER@NII.AC.JP>

IJCLA VOL. 4, NO. 1, JAN-JUN 2013, PP. 175–194
RECEIVED 07/12/12 ACCEPTED 11/01/13 FINAL 12/03/13

Facet-Driven Blog Feed Retrieval

LIFENG JIA,[1] CLEMENT YU,[1] WEIYI MENG,[2] AND LEI ZHANG[1]

[1] *University of Illinois at Chicago, USA*
[2] *Binghamton University, USA*

ABSTRACT

The faceted blog distillation task retrieves blogs (i.e. RSS feeds) that are not only relevant to a query but also satisfy an interested facet. The facets under consideration are opinionated vs. factual, personal vs. official and in-depth vs. shallow. For the opinionated/factual facets, we propose a classifier that uses syntactic and semantic features to determine whether an opinion in blog documents is relevant to a given query. For the personal/official facets, we propose three classifiers that are learned based on different assumptions to categorize a blog document into either the personal or the official class. For the in-depth/shallow facets, we propose to calculate the depth of the coverage of a blog document on a given query by the occurrences of the concepts related to the query. Dependencies among different facets are also discussed. Experimental results on TREC Blogs06 and Blogs08 collections show that our techniques are not only effective in finding faceted blogs but also significantly outperform the best known results over both collections.

1 INTRODUCTION

Faceted blog distillation task is simply defined as: "*find me a quality blog with a principal, recurring interest in X*" [1]. Three pairs of quality aspects (called facets) of blogs are defined: 1)*Opinionated vs. Factual*: Some blogs convey opinionated comments on the topics of interest while others report factual information; 2)*Personal vs. Official*: Some blogs are written by individuals to depict their personal experiences while others

are written by companies to deliver their commercial influences; 3) *In-depth vs. Shallow*: Some blogs express in-depth thoughts and analysis on the reported issues while others simply provide quick bites on these topics without analyzing the implications of the provided information.

In this paper our aim is to achieve high retrieval effectiveness for faceted blogs, such as opinionated blogs. A blog (i.e. a RSS feed) consists of a set of blog documents (or called blog posts). We use the term *document* to represent a blog document (post) and the term *feed* to represent a blog. Faceted blog distillation can be seen as a two-phase task. Given a query Q and one of three pairs of facets mentioned above, 1) feeds (or documents) are ranked by only addressing their topical relevance to Q; 2) the feeds (or documents) from Phrase 1 as the baseline are re-ranked based on the pair of facets. Since TREC provided three baselines (i.e. the ranking of feeds in Phase 1), we only present the facet-finding techniques. There are three challenges discussed below for faceted blog distillation.

The retrieval of the opinionated blog documents is exactly the opinion retrieval problem [2]. It aims at retrieving the documents that convey the opinions relevant to a query. Since a blog document may contain opinions about multiple topics, the first challenge is how to capture opinions in a document that are related to the query. The state-of-the-art techniques are proximity-based [3, 4, 5, 6, 7]. If an opinion is close to the terms of a query within an blog document, it is likely to be relevant to the query. But the proximity-based determination is not sufficiently accurate, so we propose to use both the syntax and the semantics of a sentence to determine the opinion relevance. In addition, the query-referencing pronouns are identified by co-reference resolution and the key concepts (to be defined in Section 3) related to the query are extracted from knowledge bases. In this way, opinions not directly applicable to a query but applicable to those query-referencing pronouns or the key related concepts can be determined to be relevant to the query. Determining whether a blog document delivers the personal experiences or official information with commercial interests is the second challenge. To address this challenge, we study a research issue: should the personal or official facet of a blog document be independent of the query i.e. should a blog document be considered as a personal or official one irrespective of the query? This issue which has not been examined by other researchers has a direct impact on effectiveness. Moreover, we observe that people often express some opinions when describing their personal experiences. Thus we propose to examine whether the personal or official facet of a blog document is dependent on its opinionated or factual facet. The third challenge is

how to identify the in-depth or shallow blog documents. In-depth documents should provide in-depth thoughts and analysis about the queries. Since "thoughts" may be expressed as "opinions" and "analysis" means the depth of the coverage of blog documents, we explore these two characteristics in our solution.

This paper has the following contributions. (1) We propose a classifier to determine whether the opinions in blog documents are relevant to a given query. (2) We propose a set of classifiers to classify blog documents into personal or official classes. (3) We propose an approach to measure the in-depth or shallow facet of documents. Experiments show that the proposed techniques are effective.

2 Related Work

Besides the opinion retrieval studies, there is extensive research on opinion mining. Most opinion mining studies ignore the determination of the relevance of opinion and assume the opinions in their corpora (mainly product reviews) are always related to the object (product). They focus on how to relate an opinion to the different aspects of the object or to the opinion holder (who expresses the opinions). Instead of the opinion relevance to the opinion holder [8, 9, 10], our work studies the relevance of an opinion toward the object (mentioned in the query). The aspects in opinion mining roughly correspond to the key-related concepts in our work. The key differences between their works [11, 12, 13, 14, 15, 16, 17] and our work are: 1) the objects in their works are mainly products in reviews, while the objects mentioned in TREC queries come from diversified domains. Their techniques of mining the aspects of products are applied to product reviews and may not be applicable to the key related concepts of TREC queries over blog corpora. Therefore, we develop techniques to extract the key related concepts of query concepts from knowledge bases. 2) The utilization of key related concepts aims at recognizing the relevant opinions. Some relevant opinions in blog documents are not directed toward the objects (mentioned in the queries) but applicable to those key related concepts.

For finding the personal or official documents, some studies [18, 19, 20, 21] simply assume that the personal or official documents are the opinionated or factual documents respectively. Other studies calculate the personal or official facet scores based on dictionaries [22], heuristics [23, 24] and classifiers [25, 26]. No previous work studied our first research

issue: whether a document being personal or official is independent of the query.

To identify in-depth or shallow documents, the cross entropy between a blog document *d* and the whole collection is calculated as the in-depth score of *d* [20, 21]. Various heuristics [23, 22, 24] are proposed to measure the in-depth and shallow facets of documents. For example, an in-depth document is likely to be longer in terms of the number of terms than a shallow document. We propose to measure the depth of the coverage of a document *d* on a query topic by the occurrences of concepts in d which are closely related to the query.

3 OPINIONATED VS. FACTUAL

In this section, we introduce how to measure the extents of blog documents being opinionated and factual. Given a query Q and a blog document *d*, we first utilize an classifier [7] to classify the sentences in d into opinionated or factual ones. This classifier assigns each sentence an opinion or a factual score. Then, we determine whether the opinionated or factual sentences are relevant to Q. Finally, we calculate the opinionated (or factual) facet score of d is the sum of the opinion (or factual) scores of the relevant opinionated (or factual) sentences.

The key is how to recognize the opinionated/factual sentences relevant to Q. For each opinionated sentence s, we decide s is relevant to Q if the following two conditions are satisfied. The first condition is that s and Q co-occur within a window of five sentences consisting of s, two proceeding ones and two succeeding ones [7]. But this proximity-based condition alone is not sufficient to accurately determine the relevance of s to Q. Therefore, we stipulate a second condition to further determine whether s is indeed relevant to Q. Specifically, we first identify the occurrences of Q in s, then resolve the query-referencing pronouns in s and finally identify the hypernyms of Q or the key related concepts of Q in s, if present. We denote the occurrences of Q, the query-referencing pronouns and the hypernyms and the key related concepts of Q as *target terms*. We also identify the opinion terms in s by two opinion lexicons [27, 28]. The second condition is whether s has an opinion term related to one of the target terms. If *s* contains no target terms, the opinion in *s* is irrelevant to Q, in spite of its close proximity to Q. The hypernyms of Q and the key-related concepts of Q are essential as illustrated. For example, given Q = "*Brokeback Mountain*", the opinion terms that are related to a hypernym

Table 1. A Sample of *Syntactic* (italics) and **Semantic** (bold) Features.

Feature Name	Feature Description (*O* = Opinion Term, *T* = a Target Term)
TSub	Valued *TRUE* when *T* is the subject (*Sub*) of the opinionated sentence;
OPred	Valued *TRUE* when *O* is the predicate (*Pred*) of the opinionated sentence;
OModNNT	Valued *TRUE* when *O* modifies a noun *N*; *T* and *N* satisfy the following condition: *T* is a non-person concept but *N* is the hyponym of person or vice versa;
SpecialPhrase	Valued *TRUE* when *O* forms some special phrases without opinions, such as "as **well** as";

of Q, "*movie*" or a key related concept of Q, "*Health Ledger*" can represent the relevant opinions to Q. Factual sentences do not have "factual terms" to signify their factualness as there is no "factual lexicon". So we determine a factual sentence s to be relevant to Q, if s and Q co-occur within a window of five sentences.

Query-Referencing Pronouns, Hypernyms and Key Related Concepts. To determine whether an opinionated sentence s is relevant to a query Q, at least one opinion term in s is related with Q. Some opinion terms that are not directly related with Q but related with the pronouns referencing Q can convey the opinions relevant to Q. Specifically, Illinois Coreference toolkit [29] is used on the paragraph containing s to resolve the pronouns referencing Q. Besides the pronouns, the opinion terms related with the hypernyms or the key related concepts of queries are relevant to the queries. Specifically, key concepts are related to Q by the "part-of" and "equivalence" relationships. There are other possible relationships, such as the "associative" relationship between two concepts. However, in our opinion, they are unsuitable for determining the opinion relevance toward the query. We use three knowledge bases: YAGO [30], DBPedia[3] and Freebase[4], to extract the hypernyms and the key related concepts of queries. The hypernyms of a query Q can be automatically identified by their associations with Q by the relationships 'IsA" in YAGO, "type" in DBPedia or "category" in Freebase. But the key related concepts cannot be directly extracted from the knowledge bases, because relationships in these knowledge bases are defined in free-text and determining which free-text relationships correspond to the "part-of" and the "equivalence" relationships is difficult. We manually examine the free-text relationships in these knowledge bases to determine whether they can simulate either the "part-of" or the "equivalence" relationship. For example, given a relationship, "starring", two concepts, "*Leonardo DiCaprio*" and "*Titanic*" are associated by "starring" in the knowledge bases. "*Leonardo*

[3] http://wiki.dbpedia.org/Ontology

[4] http://www.freebase.com/

DiCaprio" can be considered as a part of movie "*Titanic*", so "starring" is determined to be qualified for simulating the "part-of" relationship. Following the selection criteria above, a list of 313 relationships (10 from YAGO, 104 from DBPedia and 199 from Freebase) is established. Note that the manual examination of relationships is carried out only once before the query processing. No query is involved in the examination. Given such a list of relationships, the key-related concepts of any query can be retrieved from these knowledge bases automatically.

Syntactic and Semantic Features. Given an opinionated sentence s, after all the target terms are identified, if present, we determine whether an opinion term O is related to one of the target terms T in terms of s's syntax and semantics. We treat the relevance of O to T within s as a classification problem. We propose a set of features based on syntax and semantics. Table 1 presents a sample of the proposed features and those features described below are excluded. The syntax of a sentence can be expressed by typed dependencies and a parse tree, both of which are obtained by Stanford parser [31]. We propose typed dependency (TD) features and (parse) tree node (TN) features. *Typed Dependency*: given the TDs of an opinionated sentence, an undirected TD graph is built where the vertices are the terms and the edges are TDs between terms. A TD path between term *A* and term *B* is a sequence of TDs between vertex *A* and vertex *B*. Given the shortest TD path SP between the opinion term and a target term, for each TD t_d in SP, we prefix t_d's name with SP's length and suffix t_d's name with its sequential position in SP. It is a TD feature. *Tree Node*: given a parse tree of an opinionated sentence, we ignore the directions of tree edges. We then find the shortest path SP from a leaf node A representing an opinionated term to a leaf node B representing a target term. We represent SP by a sequence of intermediate tree nodes by excluding A and B. For each tree node t_n in SP, we prefix t_n's name with SP's length and suffix t_n's name with its sequential position in SP. It is a TN feature. A short distance between an opinion term and a target term in a TD graph or in a parse tree indicates relevance of opinion.

Moreover, the Boolean features in Table 1 can indicate the relevance of the opinion terms to the target terms within an opinionated sentence. For example, a syntactic feature named *OTDiffC* is valued true when an opinion term and the target terms occur in different clauses, which indicates that they are unlikely to be related. We propose some semantic features too. For example, a semantic feature named *Comparison* is valued true when the opinionated sentence is a comparative or superlative one. The intuition is that an opinion in such a sentence is always directed

toward all entities involving the comparison and thus the opinion term is likely to be related to the target terms.

We sample 1108 training examples from TREC Blogs06 collection w.r.t. 50 TREC 2006 queries. Each example is a triple consisting of an opinion term, a query and an opinionated sentence containing them. The opinion term is manually labeled to be either relevant or irrelevant to the query. The query-referencing pronouns, hypernyms and key related concepts if present are identified. An opinion relevance classifier is trained by using the training data and the features.

4 PERSONAL VS. OFFICIAL

In this section, we present three classifiers. Each of them classifies the blog documents into either the personal or the official class. These classifiers examines the following two research issues. First, is the class of a document (personal vs. official) independent of the query i.e. should a document be considered as personal or official irrespective of the query? Second, is the class of a document dependent on whether the document is opinionated or factual? To build classifiers, a set of features and the training data are essential. TREC relevance judgements are used as the training data but they only provide facet judgments on feeds, instead of documents. Table 2 shows a sample of proposed features. The proposed features can be generally categorized into query independent ones (QID and QIF groups) and query dependent ones (QDD and QDF groups). The answers to these two issues influence the feature selection and the usage of the training data. Our proposed features can be partitioned into two classes: query-independent and query-dependent. Each class can be further partitioned into two subclasses: document level or feed level. These 4 subclasses are sketched below.

1) *Query Independent Document Level Features* (QID). A document can show some clues of its personal or official facet. For example, people are more interested in commenting the personal documents than the official ones. Thus the number of comments in a document is a good indicator of its personal or official facet. The more comments a document has, the more likely it is personal. In TREC Blogs08 collection, the average number of comments per document in personal feeds is 4.9 while that of official feeds is 1.1. We propose 22 QID features.

2) *Query Dependent Document Level Features* (QDD). An example feature is the number of sentences that are classified to be opinionated relevant ones to a given query. We propose 8 QDD features.

Table 2. A Sample of Features for Personal or Official Classification.

Group	ID	Feature Description (d = document, f = the feed containing d)	#
QID	D_1	No. of images in d;	1
QID	D_2	No. of sentences in d classified to be opinionated and the sum of their opinion scores;	2
QDD	D_3	No. of query terms in the title of d;	1
QDD	D_4	Similar to D_2, except the classified opinionated and relevant sentences to a given query are utilized;	2
QIF	F_1	The mean and the standard deviation of the feature D_1 of documents in f;	2
QDF	F_2	The mean and the standard deviation of the feature D_3 of documents in f;	2

3) *Query Independent Feed Level Features* (QIF). An example feature is the percentage of documents in a feed that have no first person pronouns. A higher percentage more likely indicates an official feed. We propose 51 QIF features.

4) *Query Dependent Feed Level Features* (QDF). An example feature is the percentage of documents in a feed whose titles contain at least one query term. We propose 3 QDF features

Three personal/official (PS/OF) classifiers are built based on different assumptions about those two research issues. Accordingly, three PS/OF modules are constructed. Each module uses a classifier and ranks the documents as below.

1) *Query Independent with Opinionated and Factual Features* (QIOPFT): By assuming that a document being personal or official is independent of queries but depends on its opinionated or factual facet, the first classifier QIOPFT is built as follows. Given a labeled feed *f*, all documents in *f* are used as the training data and they are assigned the same facet label as that of *f*. All query-independent features (QID and QIF groups) are utilized. After QIOPFT is learned over the training data by those features, a module using QIOPFT is established. In this module, a document *d* is first classified into either the personal or the official class. Then *d* is assigned by QIOPFT a classification score $PS(d)$ (or $OF(d)$), if it is classified into the personal (or official) class. Let $F_{OP}(d)$ and $F_{FT}(d)$ be the opinionated and factual facet scores of *d* respectively. Since we assume that the class of a document depends on its being opinionated or factual, the module using QIOPFT assigns the personal facet score, $F_{PS}(d)$, and the official facet score, $F_{OF}(d)$ of the document *d* as follows. Here, λ is empirically learned:

$$\begin{aligned} F_{PS}(d) &= \lambda F_{OP}(d) + (1-\lambda)PS(d), \\ F_{OF}(d) &= \lambda F_{FT}(d) + (1-\lambda)OF(d). \end{aligned} \tag{1}$$

2) *Query Dependent with Opinionated and Factual Features* (QDOPFT). By assuming that a document being personal or official is not only dependent on queries but also dependent on its opinionated or factual facet, the second classifier QDOPFT is built as follow. Given a labeled feed f, only the subset of documents that contains at least a query concept is considered to inherit the label of f. These query-dependent documents are utilized as the training data. QDOPFT is trained over this subset of training data but involves all features including both query-independent and query-dependent features (QID, QIF, QDD and QDF groups). Due to the assumption that the class of a document d depends on its being opinionated or factual, the module using QDOPFT assigns d a facet score by Equation (1) too.

3) *Query Independent without Opinionated and Factual Features* (QIwoOPFT). To train the third classifier QIwoOPFT, we make the assumption that a document being personal or official is independent of not only the query but also its opinionated or factual facet. QIwoOPFT is constructed using the same training data as the first classifier. However, the features used by OIwoOPFT are those query independent features (QID and QIF groups) with the exclusion of those features which are calculated based on the opinionated or factual sentences of documents, such as D_2 in Table 2. After QIwoOPFT is constructed, a document d is first categorized into the personal or the official class and is then assigned a classification score by QIwoOPFT accordingly. Due to the independent assumption between the personal/official facets and the opinionated/factual facets, we just use the classification score of d as its corresponding facet score.

We build QIOPFT and QDOPFT by the different assumptions about whether the class (personal or official) of documents is independent of queries, so we can answer the first issue by comparing their effectiveness. Experimental results in Section 7 show that QIOPFT yields better effectiveness than QDOPFT and we conclude that the class of documents is independent of queries. Acknowledging such a conclusion, we build QIOPFT and QIwoOPFT by the different assumptions about whether the class of documents depends on its opinionated or factual facet. We can answer the second research issue by comparing their effectiveness.

5 IN-DEPTH VS. SHALLOW

In this section, we present our techniques for in-depth and shallow facets. Intuitively, an in-depth analysis about a query Q should not only contain

Q, but also contain the related concepts of Q. So we propose an approach that identifies the related concepts of Q. The method is described in two steps: 1) Given the Wikipedia entry of each concept of Q, collect the anchor texts and the noun phrases in the subtitles as the candidates of the related concepts. 2) Calculate the association between a candidate *e* and Q by Pointwise Mutual Information [32]; $P(e, Q)$ is the co-occurrence probability of e and Q. $P(e)$ (or $P(Q)$) is the occurrence probability of e (or Q). They are estimated by Google.

$$PMI(e, Q) = \log\left(\frac{P(e, Q)}{P(e)P(Q)}\right) \tag{2}$$

We propose two methods to measure the in-depth (or shallow) facet score of a document *d*, $F_{ID}(d)$(or $F_{SW}(d)$). The first method computes $F_{ID}(d)$ or $F_{SW}(d)$ without considering whether d is opinionated or factual. It assumes that d is a in-depth document if it provides deep analysis; otherwise, d is a shallow document. Let $RC(Q)$ be the top k (k = 30 in this paper) related concepts of Q and $CNT(e, d)$ be the count of *e* in *d*. $F_{ID}(d)$ and $F_{SW}(d)$ are calculated as below.

$$\begin{aligned} F_{ID} = Dep(d) &= \sum_{e \in RC(Q)} CNT(e, d) \cdot PMI(e, Q), \\ F_{SW}(d) &= 1 - Dep(d). \end{aligned} \tag{3}$$

The second method assumes that an in-depth document is likely to be opinionated and provides deep analysis; a shallow document is likely to be factual and provides no deep analysis. Let $F_{OP}(d)$ and $F_{FT}(d)$ be the opinionated and factual facet scores of d. After $Dep(d)$ score is normalized between 0 and 1, $F_{ID}(d)$ and $F_{SW}(d)$ can be alternatively calculated as below. λ is empirically learned.

$$\begin{aligned} F_{ID}(d) &= \lambda F_{OP}(d) + (1 - \lambda) Dep(d), \\ F_{SW}(d) &= \lambda F_{FT}(d) + (1 - \lambda)(1 - Dep(d)). \end{aligned} \tag{4}$$

6 AGGREGATION MODULE

In this section, we propose an aggregation method to calculate the facet score of each feed by the facet scores of its documents. Let Q be a query topic and D_Q be the set of documents retrieved by a topical retrieval system w.r.t. Q; given a feed *f*, D_f is the set of the documents in *f*; $IR(d)$

and $F_t(d)$ are the ad-hoc score of a document *d* from the topical retrieval system and the facet score of *d* for the facet *t* respectively. t is one of six facets discussed. In this paper we use TREC baselines to obtain $IR(d)$ and D_Q. For any feed *f*, its ad-hoc score, $IR(f)$ and its facet score, $F_t(f)$, are calculated as below:

$$IR(f) = \frac{|D_Q \cap D_f|}{|D_f|} \cdot \sum_{d \in D_Q \cap D_f} IR(d), \tag{5}$$

$$F_t(f) = \frac{|D_Q \cap D_f|}{|D_f|} \cdot \sum_{d \in D_Q \cap D_f} F_t(d)$$

An aggregated score $AS_t(f)$ is computed as below. All feeds are ranked in descending order of their aggregated scores. Here, α is empirically learned:

$$AS_t(f) = \alpha IR(f) + (1 - \alpha) F_t(f) \tag{6}$$

7 EXPERIMENTS

Experimental Setup. Since opinion retrieval plays a central role in faceted blog distillation, we first evaluate our opinion-finding techniques using 100 TREC 2007-2008 queries over five TREC baselines (of documents) from TREC Blogs06 collection. The performance metrics are Mean Average Precision (MAP), R-Precision (R-Prec), binary Preference (bPref) and Precision at top 10 documents (P@10). MAP is the most important metric. Another set of experiments is designed to evaluate the proposed facet-finding techniques using 70 TREC 2009-2010 queries over three TREC baselines (of feeds) from TREC Blogs08 collection. These 70 queries consist of 20 queries with opinionated/factual facets, 18 queries with personal/official facets and 32 queries with in-depth/shallow facets. TREC Blogs08 collection is the only official blog collection for faceted blog distillation. MAP as the most important metric in TREC 2009-2010 is used here.

Opinion Retrieval Evaluation. Our opinion-finding technique is characterized by three sub-techniques: 1) the syntax and semantics features, 2) the hypernyms and the key related concepts from knowledge bases and 3) co-reference resolution for identifying query-referencing pronouns. We evaluate their impacts individually as follows.

We first use the opinion retrieval system [7] as baseline. It determines the opinion relevance to a query only based on the proximity condition

of five sentences. Denote it by System I. Then, we configure a second system (denoted by System II) that in addition to the proximity condition, employs the proposed classifier to further determine the relevance of opinionated sentences. It uses the syntactic and semantic features but the hypernyms and the key related concepts of the query concepts and the query-referencing pronouns are not identified. The target terms are only the query concepts. The third system (denoted by System III) uses not only the classifier but also the hypernyms and the key related concepts as target terms. Co-reference resolution is not used. The fourth system employs all three sub-techniques and performs co-reference resolution by Illinois Co-reference toolkit [29]. Denote it by System IV.

All systems are given the same ad-hoc baseline obtained by the topical retrieval system [33] as input and re-rank the documents by addressing the opinionated facet. Since 50 TREC 2006 queries are used for training the opinion relevance classifier, all systems are evaluated by 100 TREC 2007-2008 queries. Table 3 shows their performance. System II achieves statistically significant improvements over System I in all measures. It indicates the classifier that is based on syntax and semantics is effective in determining the opinion relevance, even though only query concepts are used as target terms. The utilization of the hypernyms and the key related concepts in System III contributes to consistent improvements over System II in all measures. Specially, the improvements in MAP and bPref are statistically significant. These improvements indicate that the utility of the hypernyms and the key related concepts are beneficial for determining the opinion relevance. In comparing System IV with System III, the resolution of pronouns contributes to the marginal improvements in all measures. We employed a different co-reference resolution toolkit OpenCalais[5] without observing significant performance difference.

Overall, System IV achieves statistically significant improvements over System I in all measures. It indicates the proposed techniques together are very effective. In addition, we compare System VI with the state-of-the-art opinion retrieval method called laplaceInt [3]. It determines the relevance of opinions to queries by their proximities, achieving the best performance over five TREC baselines from TREC Blogs06 collection by using 50 TREC 2008 queries. We evaluate System IV over those five baselines by the same set of queries and compare its performance with that of laplaceInt. Table 4 shows that System IV consistently and significantly outperforms laplaceInt over these five baselines in MAP,

[5] http://www.opencalais.com

Table 3. Comparison of System K with System K-1 (K = 2, 3, 4); $\triangle$ denotes statistically significant improvements over System K-1 by System K at $p < 0.05$; ▲ denotes statistically significant improvements over System I by System IV at $p < 0.05$.

	MAP	R-Prec	bPref	P@10
System I	0.4304	0.4497	0.4790	0.6560
System II	$0.4771^{\triangle}$	$0.4875^{\triangle}$	$0.5091^{\triangle}$	$0.7060^{\triangle}$
System III	$0.4835^{\triangle}$	0.4900	$0.5188^{\triangle}$	0.7100
System IV	$\mathbf{0.4843}^{\blacktriangle}$	$\mathbf{0.4904}^{\blacktriangle}$	$\mathbf{0.5192}^{\blacktriangle}$	$\mathbf{0.7120}^{\blacktriangle}$

R-Prec and bPref. For P@10, System IV outperforms laplaceInt by 5.0% averagely.

Faceted Blog Distillation Evaluation. We now evaluate all proposed faceted-finding techniques over the three TREC baselines from TREC Blogs08 collection. In addition, we compare the performance of our techniques with the best performance in TREC 2010 [34]. They are the "hitFeeds" runs [26] and the "LexMIRuns" runs [20]. Note that the parameters λ and α (from Equations (1), (4) and (6)) are learned as follows. We try all possible values for λ and α from 0.1 to 1.0 with interval of 0.1 respectively. The values of λ and α that perform best for TREC 2009 queries are used to evaluate TREC 2010 queries and vice versa. Moreover, the opinion relevance classifier used in this set of experiments is trained by using TREC 2006 queries while tested by 70 TREC 2009-2010 queries.

Opinionated and Factual Effectiveness. Tables 5 and 6 show the evaluation of our opinionated (OP) and factual (FT) blog distillation method (denoted by OPFT) over three baselines by using 20 TREC queries with OP and FT facets. OPFT consistently achieves significant improvements in both facet performance over all three baselines. We also compare OPFT with the state-of-the-art methods.

Tables 5 and 6 show that OPFT consistently and significantly outperforms the best performance in both facet performance. Xu et al. [35] only studied opinionated blog distillation by using those 20 TREC queries over the same baselines. Our performance outperforms theirs by 4.0% in mean MAP score. We show the average improvement without showing their results due to space limit.

Personal and Official Effectiveness. We evaluate three proposed personal (PS) and official (OF) blog distillation methods by 18 TREC queries with PS and OF facets over the three baselines. The three methods use

Table 4. Comparison of System IV with laplaceInt; ▲ denotes statistically significant improvements over baselines by System IV at $p < 0.05$.

	MAP	R-Prec	bPref	P@10
baseline1	0.3239	0.3682	0.3514	0.5800
laplaceInt	0.4020	0.4412	0.4326	0.6920
System IV	**0.4294**▲	**0.4610**▲	**0.4631**▲	**0.7260**▲
baseline2	0.2639	0.3145	0.2902	0.5500
laplaceInt	0.2886	0.3411	0.3166	0.5860
System IV	**0.3526**▲	**0.4108**▲	**0.3862**▲	**0.6400**▲
baseline3	0.3564	0.3887	0.3677	0.5540
laplaceInt	0.4043	0.4389	0.4247	**0.6660**
System IV	**0.4192**▲	**0.4447**▲	**0.4374**▲	**0.6660**▲
baseline4	0.3822	0.4284	0.4112	0.6160
laplaceInt	0.4292	0.4578	0.4485	**0.7140**
System IV	**0.4540**▲	**0.4836**▲	**0.4811**▲	0.7040▲
baseline5	0.2988	0.3524	0.3395	0.5300
laplaceInt	0.3223	0.3785	0.3715	0.6120
System IV	**0.3535**▲	**0.4015**▲	**0.3944**▲	**0.6860**▲

three proposed PS/OF classifiers and are named as QDOPFT, QIwoOPFT and QIOPFT respectively. Note that TREC 2009 query topics are tested over the PS/OF classifiers that are trained over the relevance judgments of TREC 2010 query topics and vice versa. The comparison between QDOPFT and QIOPFT can answer our first research issue: "Is a document being personal or official independent of the query?" Tables 5 and 6 show that QIOPFT outperforms QDOPFT in terms of the mean MAP score of PS and OF performance over the three baselines. So we believe that a document being personal or official is independent of the query. To address the second research issue: "Is a document being personal or official dependent on whether it is opinionated or factual?", we conduct the comparison between QIwoOPFT and QIOPFT.

Tables 5 and 6 show that the QIOPFT consistently outperforms the QIwoOPFT over three baselines in terms of PS and OF facet performance. So we conclude that a document being personal or official is dependent on its opinionated or factual nature. Since a document being personal or official is independent of the query, a possible concern is that there may be feeds that are judged to be personal (or official) for some TREC 2009 queries and they have the same judgments for some TREC 2010 queries. This may cause our classifiers to be overfitting, because the

Table 5. Performance of faceted blog distillation modules, part 1. ▲ (▼) and △ (▽) denote statistically significant improvements (deteriorations) at $p < 0.05$ and $p < 0.1$.

	Opinionated Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2426	0.1318	0.1001
OPFT	**0.2742**$^{(13.0\%)}$	**0.1721**$^{(30.6\%)}$	**0.1533**$^{(53.1\blacktriangle)}$
hitFeeds	0.2436	0.1319	0.1015
LexMIRuns	0.2518	0.1428	0.1199
OPFT	**0.2742**$^{(12.6\%,8.9\%\triangle)}$	**0.1721**$^{(30.5\%\triangle,20.5\%)}$	**0.1533**$^{(51.0\%\blacktriangle,27.9\%)}$

	Personal Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2097	0.1527	0.0895
QDOPFT	0.2444$^{(16.5\%)}$	0.1667$^{(9.2\%)}$	0.1677$^{(87.3\%\triangle)}$
QIwoOPFT	0.1751$^{(-16.5\%,N/A)}$	0.1167$^{(-23.6\%,N/A)}$	0.0872$^{(-2.6\%)}$
QIOPFT	0.2440$^{(16.4\%,-0.2\%,39.3\%)}$	**0.1966**$^{(28.7\%,17.2\%\triangle,68.5\%\blacktriangle)}$	**0.1683**$^{(88.0\%\triangle,0.4\%,93.0\%\blacktriangle)}$
hitFeeds	0.2126	0.1533	0.0911
LexMIRuns	**0.2727**	0.1607	0.0875
QIOPFT	0.2440$^{(14.8\%,-10.5\%)}$	**0.1966**$^{(28.2\%,22.3\%)}$	**0.1683**$^{(84.7\%\triangle,92.3\%\blacktriangle)}$

	In-depth Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.3298	0.2185	0.1616
IDSW	0.3283$^{(-0.5\%)}$	0.2357$^{(7.8\%)}$	0.2097$^{(29.8\%\triangle)}$
IDSWOPFT	**0.3339**$^{(1.2\%,1.7\%)}$	**0.2421**$^{(10.8\%,2.7\%\triangle)}$	**0.2141**$^{(32.5\%\blacktriangle,2.1\%)}$
hitFeeds	0.3300	0.2184	0.1643
LexMIRuns	0.3311	0.2185	0.1616
IDSWOPFT	**0.3339**$^{(1.2\%,0.8\%)}$	**0.2421**$^{(10.9\%,10.8\%)}$	**0.2141**$^{(30.3\%\blacktriangle,32.5\%\blacktriangle)}$

PS/OF classifiers are trained over the facet-judged feeds for TREC 2010 queries and then tested by TREC 2009 queries and vice versa. However, after examining the facet-judged feeds of 18 TREC PS/OF queries, the set of 181 facet-judged feeds for 8 TREC 2009 PS/OF queries and the set of 205 facet-judged feeds for 10 TREC 2010 PS/OF queries are disjoint.

Table 6. Performance of faceted blog distillation modules, part 2. ▲ (▼) and △ (▽) denote statistically significant improvements (deteriorations) at $p < 0.05$ and $p < 0.1$.

	Factual Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2520	0.1911	0.1419
OPFT	**0.2802**$^{(11.2\%)}$	**0.2032**$^{(6.3\%)}$	**0.1639**$^{(15.5\%\blacktriangle)}$
hitFeeds	0.2529	0.1913	0.1417
LexMIRuns	0.2477	0.1942	0.1622
OPFT	**0.2802**$^{(10.8\%,13.1\%)}$	**0.2032**$^{(6.2\%,4.6\%)}$	**0.1639**$^{(15.6\%\blacktriangle,1.0\%)}$

	Official Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.2673	0.1957	0.2016
QDOPFT	0.2570$^{(-3.9\%)}$	0.2046$^{(4.5\%)}$	0.2102$^{(4.3\%)}$
QIwoOPFT	0.2658$^{(-0.6\%,N/A)}$	0.1674$^{(-14.5\%,N/A)}$	0.2085$^{(3.4\%,N/A)}$
QIOPFT	0.2690$^{(0.6\%,4.7\%,0.6\%)}$	**0.2449**$^{(25.1\%,19.7\%,46.3\%\blacktriangle)}$	**0.2366**$^{(17.4\%,12.3\%,13.5\%)}$
hitFeeds	**0.2700**	0.1957	0.1985
LexMIRuns	0.2662	0.1882	0.2016
QIOPFT	0.2690$^{(-0.4\%,1.0\%)}$	0.2449$^{(25.1\%,30.1\%)}$	**0.2366**$^{(19.2\%,17.4\%)}$

	Shallow Facet Effectiveness		
	BASELINE 1	BASELINE 2	BASELINE 3
baseline	0.1370	0.1125	0.0921
IDSW	**0.1450**$^{(5.8\%\triangle)}$	**0.1293**$^{(14.9\%)}$	0.1043$^{(13.2\%)}$
IDSWOPFT	0.1421$^{(3.1\%,-0.2\%)}$	0.1289$^{(14.6\%,-0.3\%)}$	**0.1144**$^{(24.2\%,9.7\%)}$
hitFeeds	0.1378	0.1123	0.0913
LexMIRuns	0.1279	0.1046	0.0910
IDSWOPFT	0.1421$^{(3.1\%,11.1\%)}$	0.1289$^{(14.8\%,23.2\%)}$	**0.1144**$^{(25.3\%,25.7\%)}$

QIOPFT is the most effective and robust one among all three methods. So we compare its performance with the best known performance. QIOPFT significantly outperforms the "hitFeeds" runs and the "LexMIRuns" runs in both faceted performance. Gerani et al. [18] only studied personal blog distillation. We perform experiments using their queries and outperform their results by 18.1% in MAP. We show the average improvement without showing their results due to space limit.

In-depth and Shallow Effectiveness. We evaluate our in-depth (ID) or shallow (SW) methods by using 32 TREC queries with ID and SW facets. We first configure a method where the facet scores are calculated by Equation (2). The depth of documents are measured by the extent of the occurrences of related concepts to queries. Let IDSW denote this method. We then configure another method where the facet scores are calculated by Equation (4). It considers the depth or shallowness of a document not only by the related concepts but also by the OP or FT facet scores. Let IDSWOPFT denote this method. Tables 5 and 6 show that IDSW significantly improve the baselines in the ID and SW performance, which indicates the effectiveness of the usage of related concepts of queries to measure the depth of blog documents. IDSWOPFT is more robust and more effective than IDSW, because it not only outperforms IDSW in terms of the mean MAP scores for ID and SW performance but also consistently and significantly improves all three baselines in ID and SW performance. Thus, we believe that an in-depth document is likely to contain opinionated contents and a shallow document is likely to be factual. We observe that IDSWOPFT consistently and significantly outperforms those best performance in both faceted performance.

8 CONCLUSION

In this paper, we proposed techniques to classify and rank facet-oriented feeds. Moreover, we carefully studied a number of research issues in the construction of the classifiers. Some of these issues have not been addressed by earlier researchers. We set up different experiments to answer these research issues. Experiments demonstrated that our facet-finding techniques not only consistently outperform the three TREC baselines but also outperform the best results.

REFERENCES

[1] Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2009 blog track. In: TREC. (2009)
[2] Ounis, I., de Rijke, M., Macdonald, C., Mishne, G., Soboroff, I.: Overview of the trec-2006 blog track. In: TREC'06. (2006)
[3] Gerani, S., Carman, M.J., Crestani, F.: Proximity-based opinion retrieval. In: SIGIR '10. (2010)

[4] Santos, R.L.T., He, B., Macdonald, C., Ounis, I.: Integrating proximity to subjective sentences for blog opinion retrieval. In: ECIR '09. (2009)
[5] Vechtomova, O.: Facet-based opinion retrieval from blogs. Inf. Process. Manage. **46**(1) (January 2010) 71–88
[6] Zhang, M., Ye, X.: A generation model to unify topic relevance and lexicon-based sentiment for opinion retrieval. In: SIGIR '08. (2008)
[7] Zhang, W., Yu, C., Meng, W.: Opinion retrieval from blogs. In: CIKM '07. (2007)
[8] Breck, E., Choi, Y., Cardie, C.: Identifying expressions of opinion in context. In: IJCAI'07. (2007)
[9] Choi, Y., Breck, E., Cardie, C.: Joint extraction of entities and relations for opinion recognition. In: EMNLP '06. (2006)
[10] Johansson, R., Moschitti, A.: Reranking models in fine-grained opinion analysis. In: COLING'10. (2010)
[11] Ding, X., Liu, B.: Resolving object and attribute coreference in opinion mining. In: COLING '10. (2010)
[12] Ding, X., Liu, B., Yu, P.S.: A holistic lexicon-based approach to opinion mining. In: WSDM '08. (2008)
[13] Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD '04. (2004)
[14] Joshi, M., Penstein-Rosé, C.: Generalizing dependency features for opinion mining. In: ACL-IJCNLP '09. (2009)
[15] Kobayashi, N., Inui, K., Matsumoto, Y.: Extracting aspect-evaluation and aspect-of relations in opinion mining. In: EMNLP-CoNLL'07. (2007)
[16] Popescu, A.M., Etzioni, O.: Extracting product features and opinions from reviews. In: HLT-EMNLP '05. (2005)
[17] Wu, Y., Zhang, Q., Huang, X., Wu, L.: Phrase dependency parsing for opinion mining. In: EMNLP '09. (2009)
[18] Gerani, S., Keikha, M., Carman, M., Crestani, F.: Personal blog retrieval using opinion features. In: ECIR'11. (2011)
[19] Jia, L., Yu, C.T.: Uic at trec 2010 faceted blog distillation. In: TREC. (2010)
[20] Keikha, M., Mahdabi, P., Gerani, S., Inches, G., Parapar, J., Carman, M.J., Crestani, F.: University of lugano at trec 2010. In: TREC. (2010)
[21] Zhou, Z., Zhang, X., Vines, P.: Rmit at trec 2010 blog track: Faceted blog distillation task. In: TREC. (2010)

[22] Li, S., Li, Y., Zhang, J., Guan, J., Sun, X., Xu, W., Chen, G., Guo, J.: Pris at trec 2010 blog track: Faceted blog distillaton. In: TREC. (2010)
[23] Guo, L., Zhai, F., Shao, Y., Wan, X.: Pkutm at trec 2010 blog track. In: TREC. (2010)
[24] Mejova, Y., Ha-Thuc, V., Foster, S., Harris, C., Arens, R.J., Srinivasan, P.: Trec blog and trec chem: A view from the corn fields. In: TREC. (2009)
[25] Santos, R.L.T., McCreadie, R.M.C., Macdonald, C., Ounis, I.: University of glasgow at trec 2010: Experiments with terrier in blog and web tracks. In: TREC. (2010)
[26] Yang, J., Dong, X., Guan, Y., Huang, C., Wang, S.: Hit_ltrc at trec 2010 blog track: Faceted blog distillation. In: TREC. (2010)
[27] Taboada, M., Grieve, J.: Analyzing appraisal automatically. In: Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications. (2004)
[28] Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phraselevel sentiment analysis. In: HLT-EMNLP'05. (2005)
[29] Bengtson, E., Roth, D.: Understanding the value of features for coreference resolution. In: EMNLP '08. (2008)
[30] Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW '07. (2007)
[31] Marneffe, M.c.D., Maccartney, B., Manning, C.D.: Generating typed dependency parses from phrase structure parses. In: LREC'06. (2006)
[32] Fano, R.: Transmission of Information: A Statistical Theory of Communications. The MIT Press, Cambridge, MA (1961)
[33] Liu, S., Liu, F., Yu, C., Meng, W.: An effective approach to document retrieval via utilizing wordnet and recognizing phrases. In: SIGIR '04. (2004)
[34] Macdonald, C., Ounis, I., Soboroff, I.: Overview of the trec 2010 blog track. In: TREC. (2010)
[35] Xu, X., Tan, S., Liu, Y., Cheng, X., Lin, Z., Guo, J.: Find me opinion sources in blogosphere: a unified framework for opinionated blog feed retrieval. In: WSDM '12. (2012)

LIFENG JIA
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <LJIA2@UIC.EDU>

CLEMENT YU
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <CYU@UIC.EDU>

WEIYI MENG
BINGHAMTON UNIVERSITY,
BINGHAMTON, NY 13902, USA
E-MAIL: <MENG@CS.BINGHAMTON.EDU>

LEI ZHANG
UNIVERSITY OF ILLINOIS AT CHICAGO,
CHICAGO, IL 60607, USA
E-MAIL: <LZHANG32@GMAIL.COM>

Author Index

ADDITIONAL REFEREES FOR THIS VOLUME

Alexandra Balahur
Somnath Banerjee
Liliana Barrio-Alvers
Adrián Blanco
Francis Bond
Dave Carter
Chen Chen
Jae-Woong Choe
Simon Clematide
Geert Coorman
Victor Darriba
Dipankar Das
Orphee De Clercq
Ariani Di Felippo
Maud Ehrmann
Daniel Eisinger
Ismail El Maarouf
Tilia Ellendorff
Milagros Fernández Gavilanes
Santiago Fernández Lanza
Daniel Fernández-González
Karën Fort
Sofia N. Galicia-Haro
Koldo Gojenola
Gintare Grigonyte
Francisco Javier Guzman
Masato Hagiwara
Kazi Saidul Hasan
Eva Hasler
Stefan Hoefler
Chris Hokamp
Stefan Höfler
Adrian Iftene
Iustina Ilisei
Leonid Iomdin
Pistol Ionut Cristian
Milos Jakubicek
Nattiya Kanhabua
Kurt Keena
Natalia Konstantinova
Vojtech Kovar
Kow Kuroda
Gorka Labaka
Shibamouli Lahiri
Egoitz Laparra
Els Lefever
Lucelene Lopes
John Lowe
Oier López de La Calle
Shamima Mithun
Tapabrata Mondal
Silvia Moraes
Mihai Alex Moruz
Koji Murakami
Vasek Nemcik
Zuzana Neverilova
Anthony Nguyen
Inna Novalija
Neil O'Hare
John Osborne
Santanu Pal
Feng Pan
Thiago Pardo
Veronica Perez Rosas
Michael Piotrowski
Soujanya Poria
Luz Rello
Francisco Ribadas-Pena
Tobias Roth
Jan Rupnik
Upendra Sapkota
Gerold Schneider
Djamé Seddah
Keiji Shinzato
João Silva
Sara Silveira
Sen Soori
Sanja Stajner
Tadej Štajner
Zofia Stankiewicz

Hristo Tanev
Irina Temnikova
Mitja Trampus
Diana Trandabat
Yasushi Tsubota
Srinivas Vadrevu
Josh Weissbock
Clarissa Xavier
Victoria Yaneva
Manuela Yapomo
Hikaru Yokono
Taras Zagibalov
Vanni Zavarella