NAACL-HLT 2007

Doctoral Consortium
Proceedings of the Workshop

22 April 2007
University of Rochester
Rochester, New York, USA

Production and Manufacturing by
Omnipress Inc.
Post Office Box 7214
Madison, WI 53707-7214

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
75 Paterson Street, Suite 9
New Brunswick, NJ 08901
USA
Tel: +1-732-342-9100
Fax: +1-732-342-9339
acl@aclweb.org

Introduction

These are the proceedings of the NAACL-HLT 2007 Doctoral Consortium. Ten participants were selected from a total of eighteen applicants based on submission quality, advisor recommendation, and expected completion date for the PhD degree. The goals of this event are to provide these senior Ph.D. students with the opportunity to discuss and explore their research and career objectives with a panel of established researchers in the fields of natural language processing, speech technology, and information retrieval, and to develop the skills necessary to effectively communicate their research in preparation for future job talks.

Organizers:

Jackson Liscombe, Columbia University
Phillip Michalak, University of Rochester

Faculty Advisor:

Julia Hirschberg, Columbia University

Panelists:

James Allen
Chris Brew
Ciprian Chelba
Mona Diab
Graeme Hirst
Ed Hovy
Kevin Knight
Roland Kuhn
Gina Levow
Mitch Marcus
Bob Moore
Ani Nenkova
Mari Ostendorf
Michael Riley
Brian Roark
Stephen Robertson
Candace Sidner
Mark Steedman
Ellen Voorhees
Bonnie Webber
Ralph Weischedel

Table of Contents

Doctoral Consortium Program

Sunday, April 22, 2007

8:30–9:00 Breakfast

9:00–9:40 *Query Expansion Using Domain Information in Compounds*
Karin Friberg

9:40–10:20 *Learning Structured Classifiers for Statistical Dependency Parsing*
Qin Iris Wang

10:20–11:00 *Creating a Knowledge Base from a Collaboratively Generated Encyclopedia*
Simone Paolo Ponzetto

11:00–11:40 *Knowledge-Based Labeling of Semantic Relationships in English*
Alicia Tribble

11:40–12:20 *Analysis of Summarization Evaluation Experiments*
Marie-Josée Goulet

12:20–13:20 Lunch

13:20–14:00 *Exploiting Event Semantics to Parse the Rhetorical Structure of Natural Language Text*
Rajen Subba

14:00–14:40 *Dynamic Use of Ontologies in Dialogue Systems*
Joana Paulo Pardal

14:40–15:20 *Semantic Frames in Romanian Natural Language Processing Systems*
Diana Marie Trandabăţ

15:20–16:00 *Combining Evidence for Improved Speech Retrieval*
J. Scott Olsson

16:00–16:40 *Unsupervised Natural Language Processing Using Graph Models*
Chris Biemann

16:40–17:00 Coffee Break

17:00–18:00 Panel Discussion

Query Expansion Using Domain Information in Compounds

Karin Friberg
Department of Swedish Language
Göteborg University
Göteborg, Sweden
karin.friberg@svenska.gu.se

Abstract

This paper describes a query expansion strategy for domain specific information retrieval. Components of compounds are used selectively. Only parts belonging to the same domain as the compound itself will be used in expanded queries.

1 Introduction

Compounds are semantic units containing at least two content-bearing morphemes. They function as one word, and are, in many languages, written as one word. In Swedish newspapers around 10% of the words have been found to be compounds (Hedlund, 2002). Since a compound has at least two content-bearing morphemes, a great part of the information is contained in the compounds, information which can be essential in retrieving relevant documents.

I will study medical compounds, examining possible ways to expand queries in information retrieval using domain information. This information will guide the decision of when to include compound parts in search queries. The hypothesis is that components from the same domain as the compound itself, in this case the medical domain, will increase the effectiveness of the search, while components from other domains or standard language will not.

2 Information Retrieval

Information retrieval is about storing and organizing documents so that they can be found and retrieved when relevant to an information need (Baeza-Yates, and Ribiero-Neto, 1999). The words of the documents are stored in indexes. The user poses a query to the system containing words describing the information need. Words in the queries are matched against the indexed words. A ranking function finally ranks the documents in order of calculated relevance. The better the match, the higher a document is ranked.

The goal of information retrieval is to retrieve as many documents relevant to an information need as possible, **high recall**, and to have as low proportion of irrelevant documents in the output as possible, **high precision**.

2.1 Query expansion

Query expansion is modification of a query to improve retrieval effectiveness. This can be done by changing or increasing the term content of a query.

In my work the strategy of expanding queries containing compounds, with selected compound components, is discussed. The strategy should result in higher recall, since more documents are likely to be retrieved. There is, however, a risk of lower precision, since irrelevant documents with certainty also will be retrieved. To minimize the decrease of precision, only components from the same domain as the compound itself will be used. Here, dealing with medical compounds, the objective is to decide if the components are from the medical domain.

3 Compounds

A compound is, as mentioned above, a semantic unit with more than one content-bearing morpheme. In Swedish, compounding is a very productive mor-

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 1–4,
Rochester, April 2007.

phological process. There is an infinite number of possible compounds, so it is impossible to list them all. They are also written as one word without the boundary between the parts marked in any way.

3.1 Compositional/non-compositional compounds

Occasional compounds, not lexicalized but constructed when needed, usually have a transparent meaning, where the meaning can be derived from the meaning of the parts. These are called **compositional compounds**. Other compounds, with a meaning that has strayed from the combined meaning of the components, are called **non-compositional compounds** (Hedlund, 2002). Non-compositional compounds are often lexicalized with a fixed meaning. An example of a lexicalized non-compositional compound is *trädgård* 'tree yard', Swedish for 'garden', not necessarily a garden containing trees.

In information retrieval, compositional and non-compositional compounds are best treated in different ways. Non-compositional compounds are often found in dictionaries and can be processed as they are. Using the components in queries would not benefit the result. If a query contains a compositional compound, the compound components might very well be used to expand the query, since they build up the meaning of the whole.

3.2 Decomposition not always beneficial

When expanding queries with compound components, to increase recall, it is important to be aware that this could result in lower precision. This might be the case if the compound is non-compositional or if the parts are too general or used in other domains. In Ahlgren (2004) the author gives examples of when decomposition of compounds is useful and when it is not. For a compound such as *fotboll* 'foot ball' (soccer), expanding a query with *fot* and *boll* would probably result in lower precision. On the other hand, expanding a query containing the compound *narkotikapolitik* 'drug politics', with *narkotika* and *politik* would probably be more useful. Documents containing phrases like *politik mot narkotika* 'politics against drugs' could be retrieved. Documents containing *narkotika* or *politik* alone would also be found. Here one can speculate that documents containing *narkotika* have a good chance of being relevant, while the concept *politik* is broader and could cause retrieval of many irrelevant documents.

My idea is to expand queries containing medical compounds by selecting components that also belong to the medical domain. Take the compound *korsband* 'cross band/tape' (cruciate ligament). Both parts belong to standard language. Including them would do more harm than good. In the case of *åderbråcksstrumpa* 'varicose-veins stocking' the component *åderbråck* seems to be a good candidate for query expansion, unlike *strumpa*, which belongs to standard language.

4 The MeSH thesaurus

One way to determine which compound parts belong to the medical domain is to use a medical thesaurus, a controlled vocabulary with words organized according to conceptual relations. I have used the Swedish MeSH (Medical Subject Headings) (Svensk MeSH, www), which is based on a translation of the original American MeSH (MeSH, www).

4.1 The MeSH tagger

A Swedish MeSH tagger (Kokkinakis, 2006) is being developed at Språkdata, Department of Swedish Language, Göteborg University. The tagger tags maximal length strings from six subdomains of the Swedish MeSH: **A**: Anatomy, **B**: Organisms, **C**: Diseases, **D**: Chemicals and Drugs, **E**: Analytical, Diagnostic, and Therapeutic Techniques and Equipment, and **F**: Psychiatry and Psychology. If a string is tagged, the tagger will not mark a substring of this string unless it is from another subdomain. The tagger does not tag any substrings shorter than five letters.

In the Swedish MeSH the compound *kransartär* 'wreath artery' (coronary artery) is not listed, thus it is not tagged. On the other hand *artär* is found and tagged accordingly. The word *krans* is not a medical term. It is not included in MeSH and consequently not tagged:

```
krans<mesh:A07>artär</mesh>
```

4.2 Expansion using MeSH

As mentioned above, one expansion strategy for queries containing medical compounds is to add do-

main specific parts of the compounds to the query. This should work with compositional compounds. An example is *patellaluxation* 'patella dislocation' (dislocation of the knee cap). Chances are that documents containing any or both of the simplex words *patella* and *luxation* will be relevant to the needs of a user including *patellaluxation* in a query.

Baseline query:
`#sum(...patellaluxation...)`

Expanded query:
```
#sum (...#syn(patellaluxation
patella luxation)...)
```

Expanding queries with components not from the domain, especially those common in standard language, will probably result in lower precision. In the example *kransartär* the strategy would be to keep the original compound, add *artär* which is found by the MeSH tagger, but not *krans* which is not tagged.

Baseline query:
`#sum(...kransartär...)`

Expanded query:
`#sum(...#syn(kransartär artär)...)`

5 Experiments

To test the MeSH tagger, a run was made with 5205 compounds extracted from the on-line medical lexicon Medlex (Kokkinakis, 2004), created at Språkdata, Department of Swedish Language, Göteborg University. Medlex was created by adding medical vocabulary to a learner's dictionary, thus a great part of the compounds in Medlex are from the medical domain.

895 of the 5205 compounds were tagged. Among compounds not tagged, around 10% were medical. This figure should improve with a more comprehensive tagger.

233 compounds which were not tagged as a whole, had one or two components correctly tagged. This is where the strategy described should be most beneficial, suggesting that an expanded query contain the compound itself and the tagged substring(s).

Examples of tagging which may improve effectiveness in query expansion, are shown below:

`<mesh:D22/D27>cellgift</mesh>sbehandling`
'cell-poison treatment' (chemotherapy treatment)
`dotter<mesh:C04>tumör</mesh>`
'daughter tumor'
`fot<mesh:C17/C04/C02>vårta</mesh>`
'foot wart'
`<mesh:D06/D12>insulin</mesh>chock`
'insulin chock'

63 compounds had tagged components not used in medical senses. Those strings were homonymic, polysemic, or had several facets. **Homonymy** is when a string represents different words that by chance are alike. **Polysemy** is when one word has several meanings. For example, the 'leg' of a person and the 'leg' of a table. **Facets** are different aspects of one concept. A 'person' has a body aspect as well as a personality aspect (Croft and Cruse, 2004).

It is tagging of words that are homonymic, polysemic, or with medical and non-medical facets that I predict will cause difficulties. An example is *hästansikte* 'horse face'. Although *ansikte* is a medical term, it is not used in a medical sense here. If you say that a person has a *hästansikte* it is a comment about looks, not health. The word *ansikte* has a medical facet, but also a personal appearance facet.

Other examples of compounds with problematic components are listed below:

`död<mesh:A02>skalle</mesh>`
'death skull' (skull referred to in a pirate or scary sense)
`femdygns<mesh:E01>prognos</mesh>`
'five-days prognosis' (weather domain)
`<mesh:A01>finger</mesh>borg`
'finger castle' (thimble)

Only four compounds had spurious substrings tagged. An example is *röntgen+apparat* 'x-ray device', tagged as below, *nappar* meaning 'pacifiers':

`röntge<mesh:E07>nappar</mesh>at`

5.1 A pre-decomposed run

As mentioned above, the MeSH tagger tags only maximal length (sub)strings from each subdomain of MeSH. The tagger also does not tag short strings unless they are separate words. This entails that

short components will not be tagged unless decomposition of the compound is done first.

In order to see how these features affect the outcome of the tagger, I ran the Medlex list through the tagger after decomposing the compounds.

This time, 1095 compounds had one or both components tagged. 819 of these were used in the medical sense. This is a number which should be compared with 233 in the previous run.

276 compounds had components that were tagged although not used in a medical sense. Only one compound had a spurious substring tagged.

5.2 Standard language versus medical language

One problem in decomposing compounds and using the medically tagged components to expand queries, is that many words that are medical in some meaning or facet are common in standard language. Even if we know that such a component is used in the medical sense in a query, expanding the query with that component would bring on irrelevant documents. Examples of words with such properties are *hand* 'hand' and *hjärta* 'heart'. Even though these words are used in medical senses they are also common in standard language, for example in lexicalized compounds or in phrases.

In the tagger run with the decomposed list, most of the 276 words that were tagged as medical, though not used in the medical sense, had as a component one of only 16 basic words. Below are a few such compounds that have as a component a word from that list, *hand*:

<mesh:A01>hand</mesh> bok
'hand book'
<mesh:A01>hand</mesh> broms
'hand brake'
<mesh:A01>hand</mesh> duk
'hand cloth' (towel)

6 Future work

I have presented a strategy of how to use domain information to decide when parts of a compound should be used in query expansion. Unfortunately I have not been able to test the effectiveness of this strategy. To get a true evaluation of the strategy, a Swedish medical test collection is needed. At present there is no such collection. However, my research at the moment is concentrated on creating a Swedish medical test collection, by the name of MedEval.

The first step in my research is thus to create a Swedish medical test collection, the second to test query expansion strategies based on domain information, such as the one described here. The strategy described could be carried through not only in queries, but also in indexes. That is, if a document from the medical domain contains a medical compound, the index could contain not only the compound, but also its medical components. Still, a big challenge will be to work out how to deal with polysemic and homonymic words and words with medical and non-medical facets.

References

Ahlgren, Per. 2004. *The effects of indexing strategy-query term combination on retrieval effectiveness in a Swedish full text database*. Publications from Valfrid, nr 28. University College of Borås/Göteborg University.

Baeza-Yates Ricardo and Berthier Ribeiro-Neto. 1999. *Modern information retrieval*. ACM-press, New York, NY.

Croft, William and D. Alan Cruse. 2004. *Cognitive Lingustics*. Cambridge University Press. Cambridge.

Hedlund, Turid. 2002. Compounds in dictionary-based cross-language information retrieval. *Information Research*, volume 7 No.2. 2002. Department of Information Studies. University of Tampere, Finland.

Kokkinakis, Dimitrios. 2004. *MEDLEX: Technical Report*. Department of Swedish Language, Språkdata, Göteborg University. [www]. <http://demo.spraakdata.gu.se/svedk/pbl/MEDLEX_work2004.pdf> Retrieved January 9, 2007.

Kokkinakis, Dimitrios. 2006. Developing Resources for Swedish Bio-Medical Text Mining. *Proceedings of the 2nd International Symposium on Semantic Mining in Biomedicine (SMBM)*. Jena, Germany.

MeSH. *Medical Subject Headings*. U.S. National Library of Medicine, Bethesda, MD. [www]. <http://www.nlm.nih.gov/mesh/>. Retrieved January 9, 2007.

Svensk MeSH. *MeSH-resurser vid KIB*. Karolinska Institutet Universitetsbiblioteket, Stockholm. [www]. <http://mesh.kib.ki.se/>. Retrieved January 9, 2007.

Learning Structured Classifiers for Statistical Dependency Parsing

Qin Iris Wang
Department of Computing Science
University of Alberta
Edmonton, Canada T6G 2E8
wqin@cs.ualberta.ca

Abstract

My research is focused on developing machine learning algorithms for inferring dependency parsers from language data. By investigating several approaches I have developed a unifying perspective that allows me to share advances between both probabilistic and non-probabilistic methods. First, I describe a generative technique that uses a strictly lexicalised parsing model, where all the parameters are based on words and do not use any part-of-speech (POS) tags nor grammatical categories. Then, I incorporate two ideas from probabilistic parsing—word similarity smoothing and local estimation—to improve the large margin approach. Finally, I present a simpler and more efficient approach to training dependency parsers by applying a boosting-like procedure to standard training methods.

1 Introduction

Over the past decade, there has been tremendous progress on learning parsing models from treebank data (Magerman, 1995; Collins, 1999; Charniak, 1997; Ratnaparkhi, 1999; Charniak, 2000; Wang et al., 2005; McDonald et al., 2005). Most of the early work in this area was based on postulating *generative* probability models of language that included parse structures (Magerman, 1995; Collins, 1997; Charniak, 1997). Learning in this context consisted of estimating the parameters of the model with simple likelihood based techniques, but incorporating various smoothing and back-off estimation tricks to cope with the sparse data problems (Collins, 1997; Bikel, 2004). Subsequent research began to focus more on *conditional* models of parse structure given the input sentence, which allowed discriminative training techniques such as maximum conditional likelihood (i.e. "maximum entropy") to be applied (Ratnaparkhi, 1999; Charniak, 2000). Currently, the work on conditional parsing models appears to have culminated in large margin training approaches (Taskar et al., 2004; McDonald et al., 2005), which demonstrates the state of the art performance in English dependency parsing.

Despite the realization that maximum margin training is closely related to maximum conditional likelihood for conditional models (McDonald et al., 2005), a sufficiently unified view has not yet been achieved that permits the easy exchange of improvements between the probabilistic and non-probabilistic approaches. For example, smoothing methods have played a central role in probabilistic approaches (Collins, 1997; Wang et al., 2005), and yet they are not being used in current large margin training algorithms. Another unexploited connection is that probabilistic approaches pay closer attention to the individual errors made by each component of a parse, whereas the training error minimized in the large margin approach—the "structured margin loss" (McDonald et al., 2005)—is a coarse measure that only assesses the total error of an entire parse rather than focusing on the error of any particular component. I have addressed both of these issues, as well as others in my work.

2 Dependency Parsing Model

Given a sentence $W = (w_1, ..., w_n)$, I consider the problem of computing an accurate directed depen-

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 5–8,
Rochester, April 2007.

dency tree, T, over W. Note that T consists of ordered pairs of words $(w_i \rightarrow w_j)$ in W such that each word appears in at least one pair and each word has in-degree at most one. Dependency trees are usually assumed to be projective (no crossing arcs), which means that if there is an arc $(w_i \rightarrow w_j)$, then w_i is an ancestor of all the words between w_i and w_j. Let $\Phi(W)$ denote the set of all the directed, projective trees that span W.

From an input sentence W, one would like to be able to compute the best parse; that is, a projective tree, $T \in \Phi(W)$, that obtains the highest "score". In particular, I follow Eisner (1996) and McDonald et al. (2005) and assume that the score of a complete spanning tree T for a given sentence, whether probabilistically motivated or not, can be decomposed as a sum of local scores for each link (a word pair). In which case, the parsing problem reduces to

$$T^* = \arg \max_{T \in \Phi(W)} \sum_{(w_i \rightarrow w_j) \in T} s(w_i \rightarrow w_j) \quad (1)$$

where the score $s(w_i \rightarrow w_j)$ can depend on any measurable property of w_i and w_j within the tree T. This formulation is sufficiently general to capture most dependency parsing models, including probabilistic dependency models (Wang et al., 2005; Eisner, 1996) as well as non-probabilistic models (McDonald et al., 2005; Wang et al., 2006).

For the purpose of learning, the score of each link can be expressed as a weighted linear combination of features

$$s(w_i \rightarrow w_j) = \boldsymbol{\theta}^\top \mathbf{f}(w_i \rightarrow w_j) \quad (2)$$

where $\boldsymbol{\theta}$ are the weight parameters to be estimated during training.

3 Lexicalised Dependency Parsing

To learn an accurate dependency parser from data, the first approach I investigated is based on a strictly lexical parsing model where all the parameters are based on words (Wang et al., 2005). The advantage of this approach is that it does not rely on part-of-speech tags nor grammatical categories. Furthermore, I based training on maximizing the *conditional* probability of a parse tree given a sentence, unlike most previous generative models (Magerman, 1995; Collins, 1997; Charniak, 1997), which focus on maximizing the joint probability of the parse tree and the sentence.

An efficient training algorithm can be achieved by maximizing the conditional probability of each parsing decision, hence minimizing a loss based on each local link decision independently. Importantly, inter-dependence between links can still be accommodated by exploiting *dynamic features* in training—features that take into account the *labels* of (some) of the surrounding components when predicting the label of a target component. To cope with the sparse data problem, I use distributional word similarity (Pereira et al., 1993; Grefenstette, 1994; Lin, 1998) to generalize the observed frequency counts in the training corpus. The experimental results on the Chinese Treebank 4.0 show that the accuracy of the conditional model is 13.6% higher than corresponding joint models, while similarity smoothing also allows the strictly lexicalised approach to outperform corresponding models based on part-of-speech tags.

4 Extensions to Large Margin Parsing

The approach presented above has a limitation: it uses a local scoring function instead of a global scoring function to compute the score for a candidate tree. The structured large margin approach, on the other hand, uses a global scoring function by minimizing a training loss—the "structured margin loss" (McDonald et al., 2005)—which is directly coordinated with the global tree. However, the training error minimized in the large margin approach is a coarse measure that only assesses the total error of an entire parse rather than focusing on the error of any particular component. Also, smoothing methods, which have been widely used in probabilistic approaches, are not currently being used in large margin training algorithms. In the second approach, I improve structured large margin training for parsing in two ways (Wang et al., 2006). First, I incorporate local constraints that enforce the correctness of each individual link, rather than just scoring the global parse tree. Second, to cope with sparse data and generalize to unseen words, I smooth the lexical parameters according to their underlying word similarities. To smooth parameters in the large margin framework, I introduce the technique of Laplacian

regularization in large margin parsing. Finally, to demonstrate the benefits of my approach, I reconsider the problem of parsing Chinese treebank data using only lexical features, as in Section 3. My results improve current large margin approaches and show that similarity smoothing combined with local constraint enforcement leads to state of the art performance, while only requiring word-based features that do not rely on part-of-speech tags nor grammatical categories in any way.

5 Training via Structured Boosting

Finally, I have recently demonstrated the somewhat surprising result that state of the art dependency parsing performance can be achieved through the use of conventional, local classification methods. In particular, I show how a simple form of structured boosting can be used to improve the training of standard local classification methods, in the context of structured predictions, without modifying the underlying training method (Wang et al., 2007). The advantage of this approach is that one can use off-the-shelf classification techniques, such as support vector machines or logistic regression, to achieve competitive parsing results with little additional effort.

The idea behind structured boosting is very simple. To produce an accurate parsing model, one combines the local predictions of multiple weak predictors to obtain a score for each link, which a parser can then use to compute the maximum score tree for a given sentence. Structured boosting proceeds in rounds. On each round a local "link predictor" is trained merely to predict the existence and orientation of a link between two words given input features encoding context—without worrying about coordinating the predictions in a coherent global parse. Once a weak predictor is learned, it is added to the ensemble of weak hypotheses, the training corpus is re-parsed using the new predictor, and the local training contexts are re-weighted based on errors made by the *parser*'s output. Thus, a wrapper approach is used to successively modify the training data so that the training algorithm is encouraged to facilitate improved global parsing accuracy.

Table 1: Comparison with State of the Art (Dependency Accuracy)

Model	Chinese	English
Yamada&Matsumoto 03	-	90.3
Nivre&Scholz 04	-	87.3
Wang et al. 05 (Sec. 3)	79.9*	-
McDonald et al. 05	-	90.9
McDonald&Pereira 06	82.5*	91.5
Corston-Oliver et al. 06	73.3†	90.8
Structured Boosting (Sec. 5)	86.6* 77.6†	89.3

* Obtained with Chinese Treebank 4.0 using the data split reported in Wang et al. (2005).
† Obtained with Chinese Treebank 5.0 using the data split reported in Corston-Olivr et al. (2006).

6 Current Results

Table 1 compares my results[1] with those obtained by other researchers, on both English and Chinese data.[2] The English results are obtained using the same standard training and test set splits from English Penn Treebank 3.0. The results on Chinese are obtained on two different data sets, Chinese Treebank 4.0 and Chinese Treebank 5.0 as noted.[3]

Table 1 shows that the results I am able to achieve on English are competitive with the state of the art, but are still behind the best results of (McDonald and Pereira, 2006). However, perhaps surprisingly, Table 1 also shows that the structured boosting approach actually surpasses state of the art accuracy on Chinese parsing for both treebank collections.

7 Future Work

Although the three pieces of my work above look very different superficially, they are actually closely related by the "scoring" formulation and, more

[1] I did not include the results of the technique described in Section 4, because we were only able to conveniently train on sentences with less than or equal to 15 words.

[2] McDonald et al. (2005) have tried MIRA on Chinese Treebank 4.0 with the same data split reported here, obtaining a dependency accuracy score of 82.5 (Ryan McDonald, personal communication).

[3] The results on Chinese Treebank 5.0 are generally worse than on Chinese Treebank 4.0, since the former is a superset of the latter, and moreover the additional sentences come entirely from a Taiwanese Chinese source that is more difficult to parse than the rest of the data.

specifically, by the equations introduced in Section 2. In other words, they all compute a linear classifier.[4] The only differences among them are: (1) What features are used? (2) How are the parameters $\boldsymbol{\theta}$ estimated?

A general perspective I bring to my investigation is the desire to delineate the effects of domain engineering (choosing good features for representing and learning parsing models) from the general machine learning principles (training criteria, regularization and smoothing techniques) that permit good results. In fact, combined features have been proved to be useful in dependency parsing with support vector machines (Yamada and Matsumoto, 2003), and I have already obtained some preliminary results on generating useful feature combinations via boosting. Therefore, I will consider combining all the projects I presented above. That is, I plan to incorporate all the useful features, the morphological features and the combined features as discussed above, into the training algorithms presented in Section 4 or Section 5, to train a dependency parser globally. Then I am going to augment the training with the existing smoothing and regularization techniques (as described in Section 4), or new developed ones. I expect the resulting parser to have better performance than those I have presented above.

There are a lot of other ideas which can be explored in my future work. First and most important, I plan to investigate new advanced machine learning methods (e.g., structured boosting or unsupervised / semi-supervised algorithms (Xu et al., 2006)) and apply them to the dependency parsing problem generally, since the goal of my research is to learn natural language parsers in an elegant and principled manner. Next, I am going to apply my approaches to parse other languages, such as Czech, German, Spanish and French, and analyze the performance of my parsers on these different languages. Furthermore, I plan to apply my parsers in other domains (e.g., biomedical data) (Blitzer et al., 2006) besides treebank data, to investigate the effectiveness and generality of my approaches.

[4] In general, for any probabilistic model, the product of probabilities can be converted to sums of scores in the log space, which makes the search identical to a score based discriminative model.

References

D. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4).

J. Blitzer, R. McDonald, and F. Pereira. 2006. Doman adaptation with structural correspondence learning. In *Proc. of EMNLP*.

E. Charniak. 1997. Statistical parsing with a context-free grammar and word statistics. In *Proc. of AAAI*, pages 598–603.

E. Charniak. 2000. A maximum entropy inspired parser. In *Proc. of North American ACL*, pages 132–139.

M. Collins. 1997. Three generative, lexicalized models for statistical parsing. In *Proc. of ACL*, pages 16–23.

M. Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

S. Corston-Oliver, A. Aue, K. Duh, and E. Ringger. 2006. Multilingual dependency parsing using Bayes' point machines. In *Proc. of HLT/NAACL*.

J. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *Proc. of COLING*.

G. Grefenstette. 1994. Corpus-derived first, second and third-order word affinities. In *Proc. of Euralex*.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. of COLING/ACL*, pages 768–774.

D. Magerman. 1995. Statistical decision-tree model for parsing. In *Proc. of ACL*, pages 276–283.

R. McDonald and F. Pereira. 2006. Online learning of approximate dependency parsing algorithms. In *Proc. of EACL*.

R. McDonald, K. Crammer, and F. Pereira. 2005. Online large-margin training of dependency parsers. In *Proc. of ACL*.

F. Pereira, N. Tishby, and L. Lee. 1993. Distributional clustering of English words. In *Proc. of ACL*, pages 183–190.

A. Ratnaparkhi. 1999. Learning to parse natural language with maximum entropy models. *Mach. Learn.*, 34(1-3):151–175.

B. Taskar, D. Klein, M. Collins, D. Koller, and C. Manning. 2004. Max-margin parsing. In *Proc. of EMNLP*.

Q. Wang, D. Schuurmans, and D. Lin. 2005. Strictly lexical dependency parsing. In *Proc. of IWPT*, pages 152–159.

Q. Wang, C. Cherry, D. Lizotte, and D. Schuurmans. 2006. Improved large margin dependency parsing via local constraints and Laplacian regularization. In *Proc. of CoNLL*.

Q. Wang, D. Lin, and D. Schuurmans. 2007. Simple training of dependency parsers via structured boosting. In *Proc. of IJCAI*, pages 1756–1762.

L. Xu, D. Wilkinson, F. Southey, and D. Schuurmans. 2006. Discriminative unsupervised learning of structured predictors. In *Proc. of ICML*.

H. Yamada and Y. Matsumoto. 2003. Statistical dependency analysis with support vector machines. In *Proc. of IWPT*.

Creating a Knowledge Base From a Collaboratively Generated Encyclopedia

Simone Paolo Ponzetto
EML Research gGmbH
Schloss-Wolfsbrunnenweg 33
69118 Heidelberg, Germany
http://www.eml-research.de/~ponzetto

Abstract

We present our work on using Wikipedia as a knowledge source for Natural Language Processing. We first describe our previous work on computing semantic relatedness from Wikipedia, and its application to a machine learning based coreference resolution system. Our results suggest that Wikipedia represents a semantic resource to be treasured for NLP applications, and accordingly present the work directions to be explored in the future.

1 Introduction

The last decade has seen statistical techniques for Natural Language Processing (NLP) gaining the status of standard approaches to most NLP tasks. While advances towards robust statistical inference methods (cf. e.g. Domingos et al. (2006) and Punyakanok et al. (2006)) will certainly improve the computational modelling of natural language, we believe that crucial advances will also come from rediscovering the use of symbolic knowledge, i.e. the deployment of large scale knowledge bases.

Arguments for the necessity of symbolically encoded knowledge for AI and NLP date back at least to McCarthy (1959). Symbolic approaches using knowledge bases, however, are expensive and time-consuming to maintain. They also have a limited and arbitrary coverage. In our work we try to overcome such problems by relying on a wide coverage on-line encyclopedia developed by a large amount of users, namely Wikipedia. That is, we are interested in whether and how Wikipedia can be integrated into NLP applications as a knowledge base. The motivation comes from the necessity to overcome the brittleness and knowledge acquisition bottlenecks that NLP applications suffer.

2 Previous Work: WikiRelate! and Semantic Knowledge Sources for Coreference Resolution

Ponzetto & Strube (2006) and Strube & Ponzetto (2006) aimed at showing that 'the encyclopedia that anyone can edit' can be indeed used as a semantic resource for research in NLP. In particular, we assumed its category tree to represent a semantic network modelling relations between concepts, and we computed measures of semantic relatedness from it. We did not show only that Wikipedia-based measures of semantic relatedness are competitive with the ones computed from a widely used standard resource such as WordNet (Fellbaum, 1998), but also that including semantic knowledge mined from Wikipedia into an NLP system dealing with coreference resolution is in fact beneficial.

2.1 WikiRelate! Computing Semantic Relatedness Using Wikipedia

Semantic relatedness measures have been proven to be useful in many NLP applications such as word sense disambiguation (Kohomban & Lee, 2005; Patwardhan et al., 2005), information retrieval (Finkelstein et al., 2002), information extraction pattern induction (Stevenson & Greenwood, 2005), interpretation of noun compounds (Kim & Baldwin, 2005), paraphrase detection (Mihalcea et al., 2006) and spelling correction (Budanitsky & Hirst, 2006). Approaches to measuring semantic relatedness that

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 9–12,
Rochester, April 2007.

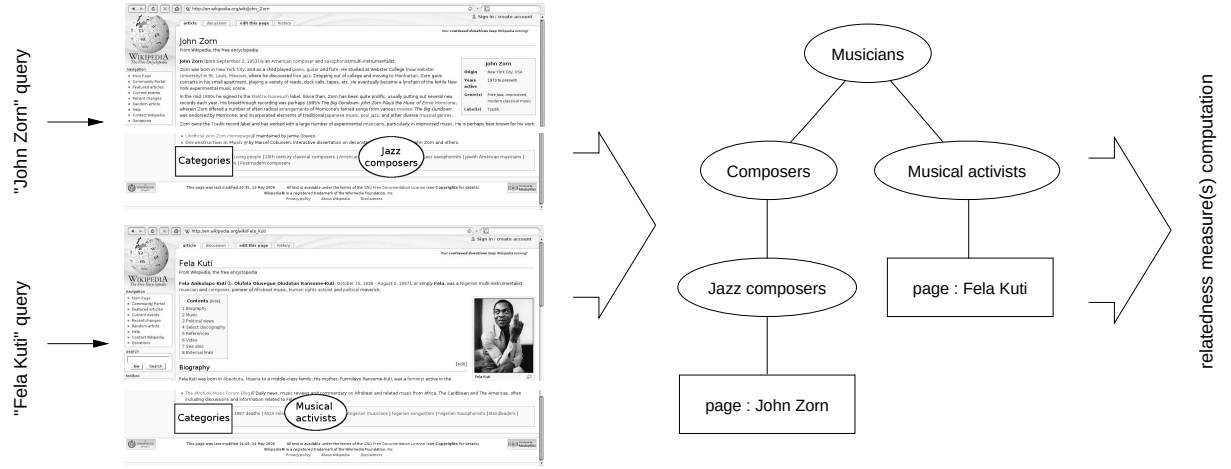


Figure 1: Wikipedia-based semantic relatedness computation. First, target pages for the given queries are retrieved, possibly via disambiguation. Next, categories are extracted to provide an entry point to the category network. Connecting paths are then searched along the category network using a depth-limited search. The paths found are scored and the ones satisfying the measure definitions (i.e. the shortest one for path-length measures, and the most informative one for information-content measures) are returned.

use lexical resources transform that resource into a network or graph and compute relatedness using paths in it[1]. For instance, Rada et al. (1989) traverse MeSH, a term hierarchy for indexing articles in Medline, and compute semantic relatedness as the edge distance between terms in the hierarchy. Jarmasz & Szpakowicz (2003) use the same approach with *Roget's Thesaurus* while Hirst & St-Onge (1998) apply a similar strategy to WordNet.

The novel idea presented in Strube & Ponzetto (2006) was to induce a semantic network from the Wikipedia categorization graph to compute measures of semantic relatedness. Wikipedia, a multilingual Web-based free-content encyclopedia, allows for structured access by means of *categories*: the encyclopedia articles can be assigned one or more categories, which are further categorized to provide a so-called "category tree". Though not designed as a strict hierarchy or tree, the categories form a graph which can be used as a taxonomy to compute semantic relatedness. We showed (1) how to retrieve Wikipedia articles from textual queries and resolve ambiguous queries based on the articles' link structure; (2) compute semantic relatedness as a function of the articles found and the paths between them along the categorization graph (Figure 1). We evaluated the Wikipedia-based measures against the ones computed from WordNet on benchmarking datasets from the literature (e.g. Miller and Charles' (1991) list of 30 noun pairs) and found Wikipedia to be competitive with WordNet.

2.2 Semantic Knowledge Sources for Coreference Resolution

Evaluating measures of semantic relatedness on word pair datasets poses non-trivial problems, i.e. all available datasets are small in size, and it is not always clear which linguistic notion (i.e. similarity vs. relatedness) underlies them. Accordingly, in Ponzetto & Strube (2006) we used a machine learning based coreference resolution system to provide an *extrinsic* evaluation of the utility of WordNet and Wikipedia relatedness measures for NLP applications. We started with the machine learning based

[1] An overview of lexical resource-based approaches to measuring semantic relatedness is presented in Budanitsky & Hirst (2006). Note that here we do not distinguish between *semantic similarity* (computed using hyponymy/hyperonymy, i.e. *is-a*, relations only) and *semantic relatedness* (using all relations in the taxonomy, including antonymic, meronymic, functional relations such as *is-made-of*, etc.), since the relations between categories in Wikipedia are neither semantically typed nor show a uniform semantics (see Section 3).

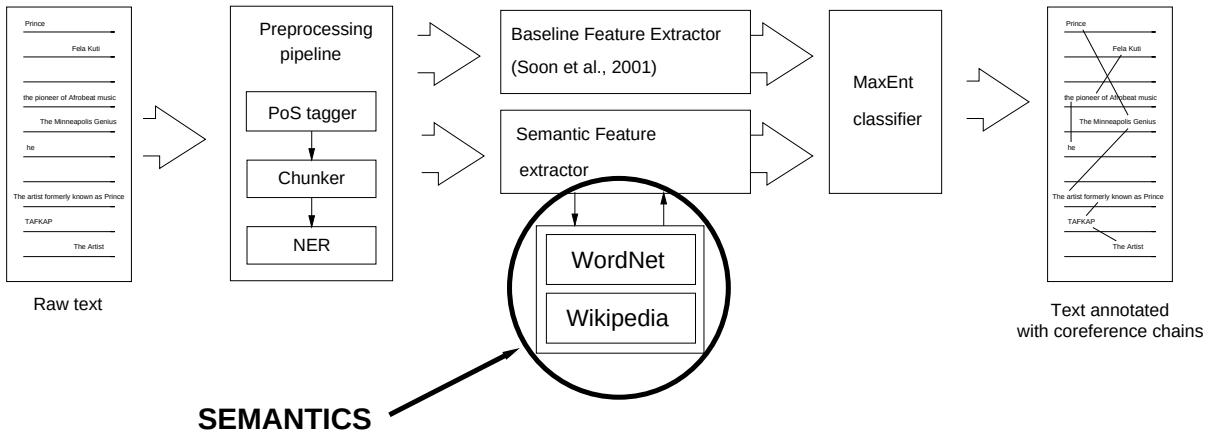


Figure 2: Overview of the coreference system for extrinsic evaluation of WordNet and Wikipedia relatedness measures. We start with a baseline system from Soon et al. (2001). We then include at different times features from WordNet and Wikipedia and register performance variations.

baseline system from Soon et al. (2001), and analyzed the performance variations given by including the relatedness measures in the feature set (Figure 2). The results showed that coreference resolution benefits from information mined from semantic knowledge sources and also, that using features induced from Wikipedia gives a performance only slightly worse than when using WordNet.

3 Future Work: Inducing an Ontology from a Collaboratively Generated Encyclopedia

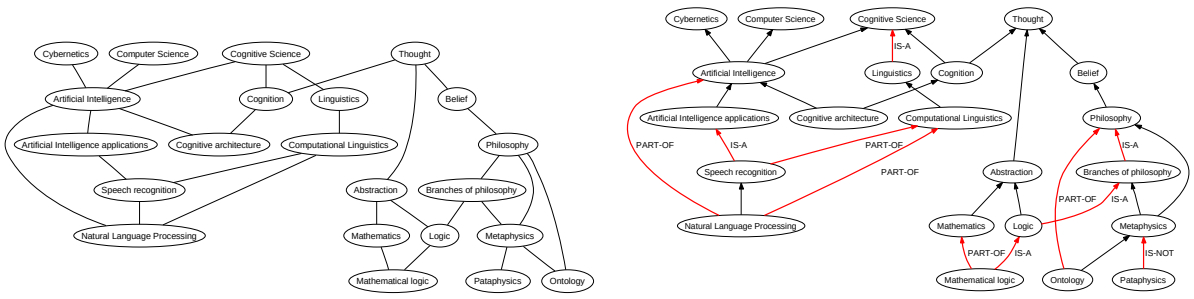
Our results so far suggest that Wikipedia can be considered a semantic resource in its own right. Unfortunately, the Wikipedia categorization still suffers from some limitations: it cannot be considered an ontology, as the relations between categories are not semantically-typed, i.e. the links between categories do not have an explicit semantics such as *is-a*, *part-of*, etc. Work in the near future will accordingly concentrate on automatically inducing the semantics of the relations between Wikipedia categories. This aims at transforming the unlabeled graph in Figure 3(a) into the semantic network in Figure 3(b), where the links between categories are augmented with a clearly defined semantics.

The availability of explicit semantic relations would allow to compute *semantic similarity* rather than *semantic relatedness* (Budanitsky & Hirst, 2006), which is more suitable for coreference resolution. That is, we assume that the availability of hyponymic/hyperonymic relations will allow us to compute lexical semantic measures which will further increase the performance of our coreference resolution system, as well as further bringing forward Wikipedia as a direct competitor of manually-designed resources such as WordNet.

In order to make the task feasible, we are currently concentrating on inducing *is-a* vs. *not-is-a* semantic relations. This simplifies the task, but still allows us to compute measures of semantic similarity. As we made limited use of the large amount of text in Wikipedia, we are now trying to integrate text and categorization. This includes extracting semantic relations expressed in the encyclopedic definitions by means of *Hearst patterns* (Hearst, 1992), detection of *semantic variations* (Morin & Jacquemin, 1999) between category labels, as well as using the categorized pages as bag-of-words to compute scores of *idf-based semantic overlap* (Monz & de Rijke, 2001) between categories. Further work will then concentrate on making this information available to our coreference resolution system, e.g. via semantic similarity computation.

Finally, since Wikipedia is available in many languages, we believe it is worth performing experiments in a multilingual setting. Accordingly, we are currently testing a website[2] that will allow us to collect word relatedness judgements from native speak-

[2] Available at `http://www.eml-research.de/nlp/353-TC`.



(a) current category graph

(b) category graph augmented with semantic relations

Figure 3: Inducing explicit semantic relations between categories in Wikipedia

ers of German, French and Italian, in order to translate the semantic relatedness dataset from Finkelstein et al. (2002) and test our methodology with languages other than English.

4 Conclusions

In this paper we presented our previous efforts on using Wikipedia as a semantic knowledge source. We aim in the future to induce an ontology from its collaboratively generated categorization graph. We believe that our work opens up exciting new challenges for the AI and NLP research community, e.g. how to handle the noise included in such knowledge bases and how to fully structure the information given in the form of only partially structured text and relations between knowledge base entries.

Acknowledgements: This work has been funded by the Klaus Tschira Foundation, Heidelberg, Germany. The author has been supported by a KTF grant (09.003.2004).

References

Budanitsky, A. & G. Hirst (2006). Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1).

Domingos, P., S. Kok, H. Poon, M. Richardson & P. Singla (2006). Unifying logical and statistical AI. In *Proc. of AAAI-06*, pp. 2–7.

Fellbaum, C. (Ed.) (1998). *WordNet: An Electronic Lexical Database*. Cambridge, Mass.: MIT Press.

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman & E. Ruppin (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Hearst, M. A. (1992). Automatic acquisition of hyponyms from large text corpora. In *Proc. of COLING-92*, pp. 539–545.

Hirst, G. & D. St-Onge (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, pp. 305–332. Cambridge, Mass.: MIT Press.

Jarmasz, M. & S. Szpakowicz (2003). Roget's Thesaurus and semantic similarity. In *Proc. of RANLP-03*, pp. 212–219.

Kim, S. N. & T. Baldwin (2005). Automatic interpretation of noun compounds using WordNet similarity. In *Proc. of IJCNLP-05*, pp. 945–956.

Kohomban, U. S. & W. S. Lee (2005). Learning semantic classes for word sense disambiguation. In *Proc. of ACL-05*, pp. 34–41.

McCarthy, J. (1959). Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, pp. 75–91.

Mihalcea, R., C. Corley & C. Strapparava (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Proc. of AAAI-06*, pp. 775–780.

Miller, G. A. & W. G. Charles (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Monz, C. & M. de Rijke (2001). Light-weight entailment checking for computational semantics. In *Proc. of ICoS-3*, pp. 59–72.

Morin, E. & C. Jacquemin (1999). Projecting corpus-based semantic links on a thesaurus. In *Proc. of ACL-99*, pp. 389–396.

Patwardhan, S., S. Banerjee & T. Pedersen (2005). SenseRelate::TargetWord – A generalized framework for word sense disambiguation. In *Proc. of AAAI-05*.

Ponzetto, S. P. & M. Strube (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proc. of HLT-NAACL-06*, pp. 192–199.

Punyakanok, V., D. Roth, W. Yih & D. Zimak (2006). Learning and inference over constrained output. In *Proc. of IJCAI-05*, pp. 1117–1123.

Rada, R., H. Mili, E. Bicknell & M. Blettner (1989). Development and application of a metric to semantic nets. *IEEE Transactions on Systems, Man and Cybernetics*, 19(1):17–30.

Soon, W. M., H. T. Ng & D. C. Y. Lim (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4):521–544.

Stevenson, M. & M. Greenwood (2005). A semantic approach to IE pattern induction. In *Proc. of ACL-05*, pp. 379–386.

Strube, M. & S. P. Ponzetto (2006). WikiRelate! Computing semantic relatedness using Wikipedia. In *Proc. of AAAI-06*, pp. 1419–1424.

Knowledge-Based Labeling of Semantic Relationships in English

Alicia Tribble
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, USA
atribble@cs.cmu.edu

Abstract

An increasing number of NLP tasks require semantic labels to be assigned, not only to entities that appear in textual elements, but to the relationships between those entities. Interest is growing in shallow semantic role labeling as well as in deep semantic distance metrics grounded in ontologies, as each of these contributes to better understanding and organization of text. In this work I apply knowledge-based techniques to identify and explore deep semantic relationships in several styles of English text: nominal compounds, full sentences in the domain of knowledge acquisition, and phrase-level labels for images in a collection. I also present work on a graphical tool for exploring the relationship between domain text and deep domain knowledge.

1 Introduction

As our command of NLP techniques has grown over the decades, the tasks which we can accomplish have become more useful and complex: we can (to an increasing extent) answer questions, create summaries, and even create new knowledge by extracting and merging facts from large text corpora. To make our systems reach their potential on these tasks, we need to extend our analysis of text into deep semantics, often grounded in world knowledge.

In this work, I explore the semantic relationships in several styles of English text using knowledge-driven NLP techniques as well as a novel graphical tool for the navigation of knowledge bases (KBs, or ontologies).

I begin by describing a system based on augmented LFG-style grammar rules, appropriate for the domain-limited sentences that are required for knowledge entry by knowledge base engineers. In a subsequent system for interpreting nominal compounds, I rely more heavily on the knowledge already stored in the knowledge base to guide a heuristic search for meaning (Tribble and Fahlman, 2006).

These systems demonstrate how a knowledge base can contribute to NLP performance. During development of the systems, knowledge acquisition and organization became important sub-topics. In response I began work on a graphical tool, SconeEdit (Tribble, Lambert, and Fahlman, 2006). SconeEdit allows users to navigate the semantic concepts and relations in a text corpus, guided by the rich, grounded features of these concepts in a knowledge base.

With this interface as a scaffold, future work entails improving the analysis systems for noun compounds and full sentences, and incorporating these systems in a comparative evaluation of the graphical and NLP-based methods for exploring semantic relationships in domain-restricted text. In addition, I will use this framework to evaluate a knowledge-

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 13–16,
Rochester, April 2007.

based approach for the task of retrieving labeled images from a collection.

2 Semantic Analysis for Knowledge Engineering

One of the motivating goals of this work is to leverage the power of NLP tools to ease the burden of knowledge engineers who develop ontological resources. By converting English sentences into a semantic representation automatically, a system provides an intuitive input method for adding new knowledge.

2.1 Knowledge Engineering in Scone

The context for this work is the Scone Knowledge Representation (KR) Project (Fahlman, 2005). The Scone KR System encompasses an inference engine along with a set of upper-level domain ontologies. As with other large KR systems along the lines of CYC (Lenat, 1986), knowledge engineers create much of the upper-level KB content by hand.

To develop a system that would address the needs of these engineers, I collected a corpus of English sentences covering the six core structure-building tasks in Scone:

- Defining a type
- Adding an instance of a type
- Defining a relation between types
- Adding an instance of a relation
- Defining a new role (HAS-A) relation
- Instantiating a role-filling relation

2.2 A Grammar-Based System

The resulting corpus displayed a high degree of semantic cohesion, as expected, but with a wide degree of syntactic variation. To transform these sentences automatically into the Scone KR, I developed a set of semantic interpretation functions and added them as callouts in an existing LFG-style syntactic grammar. The resulting augmented English grammar is applied to new sentences using the LCFlex parser of Rosé and Lavie (2000). In this way, every parse constituent can be conditioned on queries to the knowledge base, allowing not only flat semantic features (e.g. "is the noun animate?") but rich structural knowledge ("does this person own a pet?") to be applied during the parse.

The new grammar rules produce output in the Scone KR formalism. As a result, the output can be read as the knowledge-grounded meaning of an input sentence, and it can also become additional input to the Scone inference engine, adding to the store of background knowledge or making a new query. However, the appeal of this design is limited by the fact that, as in many grammar-based systems, the rules themselves are costly to write and maintain.

2.3 Adding Generalization

For this reason, I modified the approach and examined the effectiveness of a few general "preference" rules, based on syntax. In contrast with the grammar system, the search for interpretations can now be driven, rather than pruned, by domain knowledge. I tested this approach on the interpretation of noun compounds, where the lack of syntactic cues requires heavy reliance on semantic interpretation (Tribble and Fahlman, 2006). I found that a majority of compounds, even in a new textual domain, could be analyzed correctly using the new set of rules along with an appropriate domain-specific KB.

3 A Graphical Tool for Exploring Semantic Relationships

While the cost of grammar writing can be reduced with updated algorithms, developing and maintaining large knowledge repositories is one of the key challenges in knowledge-based NLP: the knowledge acquisition "bottleneck". My hypothesis is that a natural-language (NL) interface is an important tool for easily modifying and adding knowledge in a complex KR system like Scone; language is an intuitive way for users to express what they want from the knowledge base.

In the course of developing NL tools for the Scone Project, I also recognized the need to view domain text, domain knowledge, and the semantic relationships that they share in a "snapshot". Integrating textual and graphical exploration gives users a comfortable handle on the knowledge base, even when they don't know exactly what they want.

I designed the SconeEdit knowledge- and text-browsing tool (Tribble, et al. 2006) in response to this need. The tool provides an annotated view of text chosen by the user, allowing him to see what concepts and vocabulary from the text are currently in the KB. Alongside this Text View, SconeEdit provides a navigable snapshot of the knowledge base (KB View), centered on concepts that appear in the text. This unified browser establishes a principled, coverage-driven way to "surf" the KB and add new knowledge. A screenshot of SconeEdit, recently updated to view images as well as text, is shown in Figure 1.

The SconeEdit tool has already been used by groups outside the Scone Project, for the purpose of qualitatively evaluating knowledge bases for use in new subdomains. My goal for the conclusion of this work is to synergize the lines of research described so far, building our English analysis tools into the SconeEdit interface. With the resulting tool I can run a detailed evaluation of my English analyzers, as well as shed light on the usability of text-based versus graphical knowledge entry.

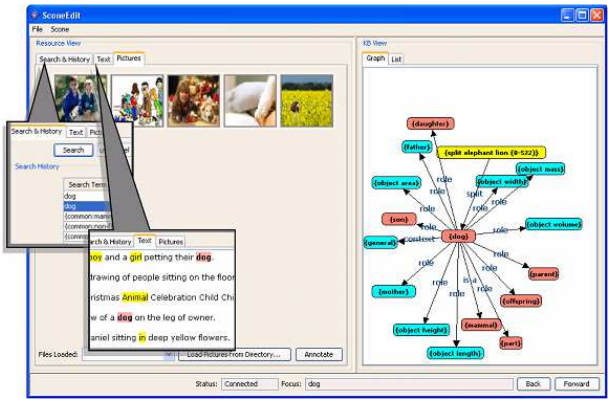


Figure 1. Screenshot of SconeEdit, updated to display images as well as text.

4 Task-Based Evaluation: Retrieving Labeled Images

To bring this work to bear in a task-based evaluation, I have also started developing a system for labeled image retrieval. To retrieve images of interest from large collections, traditional systems rely on matching between a high-level query and low-level image features, or on matching the query with an unordered bag-of-words that has been attached to each image. In current work I am investigating sentence fragments, which retain some syntactic structure, as a useful style of image annotation that is complementary to the current bag-of-words style. Analysis of 2,776 image titles downloaded from the web establishes that fragment-style labels are intuitive, discriminative, and useful.

These labels can be used to retrieve images from a collection in the following way: first, a typed query is given to the system (e.g. "people petting their dogs"). An English analyzer, using improvements to the techniques described in Section 2, produces the Scone semantic representation of this query (a semantic graph). Next, the Scone inference engine is used to match the query against pre-computed semantic representations of the image labels. The system retrieves the image whose label matches best. Figure 2 is an example retrieved for this query by Google Image Search.


Figure 2. Image retrieved by Google Image Search for "people petting their dogs".

4.1 Development Data

In order to train the functions that measure a "match" in the knowledge base, as well as to improve the English-to-Scone analysis, I need training data in the form of images, their fragment-style labels, and one or more query that matches each image and its label.

I collected one corpus of images with their fragment-style labels from the publicly available collection on Flickr (http://www.flickr.com). A second corpus of fragment-labeled images has been provided by one the authors of von Ahn and Dabbish (2004). In many cases, a single image has

multiple fragment-style labels. To convert this data into the format I need, I can use the redundant labels as substitute "queries", under the assumption (which should be validated experimentally) that image-retrieval queries often take the form of sentence fragments, as well.

An evaluation that uses these labels for image retrieval will proceed as follows: A subset of the labeled images which were not seen or used in previous work will be reserved as test data. Remaining images with their labels and queries will be used to improve the English-to-Scone analysis system and the semantic similarity functions within Scone. Finally, the queries for the test set will be submitted to the retrieval system, and system results will be compared to the "correct" images given by the test set. Precision and recall can be calculated under a variety of conditions, including one-image-per-query and several-images-per-query. Comparison to shallow techniques for label matching, as used with bag-of-words style labels, will also be a feature of this evaluation.

5 Conclusion

In summary, I have presented a body of work on exploring and labeling the deep semantic relationships in English text. A grammar-based system for sentences and a heuristic search system for noun compounds explore the role of domain knowledge in tools for syntactic and deep semantic analysis. In addition, I designed and demonstrated graphical tool for exploring rich semantic features in text, grounded in a knowledge base or ontology. The tool has been used by our own knowledge engineers as well by other research teams at CMU.

I will build on this work in the coming months as I prepare for two evaluations: a study on the usability of natural language and graphical tools for navigating a knowledge base, and a task-based evaluation on labeled image retrieval. These evaluations should bring closure to the work as a contribution in the field of semantic analysis of text.

References

Scott E. Fahlman. 2005. The Scone User's Manual. *http://www.cs.cmu.edu/~sef/scone.*

Alicia Tribble, Benjamin Lambert and Scott E. Fahlman. 2006. SconeEdit: A Text-Guided Domain Knowledge Editor. In *Demonstrations of HLT-NAACL 2006*. New York.

Alicia Tribble and Scott E. Fahlman. 2006. Resolving Noun Compounds with Multi-Use Domain Knowledge. In *Proceedings of FLAIRS-2006*. Melbourne Beach, Florida.

Alicia Tribble and Carolyn P. Rosé. 2006. Usable Browsers for Knowledge Acquisition. In *Proceedings of CHI-2006*. Montreal, Quebec.

Carolyn P. Rosé and Alon Lavie. 2001. Balancing Robustness and Efficiency in Unification-Augmented Context-Free Parsers for Large Practical Applications. In J.C. Junqua and G. Van Noord, eds. *Robustness in Language and Speech Technology.* Kluwer Academic Press.

D. B. Lenat, M. Prakash and M. Shepherd. 1986. Cyc: using common sense knowledge to overcome brittleness and knowledge acquisition bottlenecks.. In *AI Magazine.* 6:4.

Luis von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of ACM CHI* (pp 319—326).

Analysis of Summarization Evaluation Experiments

Marie-Josée Goulet
CIRAL, Department of Linguistics
Laval University, Quebec City
G1K 7P4, Canada
marie-josee.goulet.1@ulaval.ca

Abstract

The goals of my dissertation are: 1) to propose a French terminology for the presentation of evaluation results of automatic summaries, 2) to identify and describe experimental variables in evaluations of automatic summaries, 3) to highlight the most common tendencies, inconsistencies and methodological problems in summarization evaluation experiments, and 4) to make recommendations for the presentation of evaluation results of automatic summaries. In this paper, I focus on the second objective, i.e. identifying and describing variables in summarization evaluation experiments.

1 Introduction

The general subject of my dissertation is summarization evaluation. As stated in my thesis proposal, my work aims at four goals: 1) proposing a French terminology for the presentation of evaluation results of automatic summaries, 2) identifying and describing experimental variables in evaluations of automatic summaries, 3) highlighting the most common tendencies, inconsistencies and methodological problems in summarization evaluations, and 4) making recommendations for the presentation of evaluation results of automatic summaries. In this paper, I will focus on the second objective.

My ultimate goal is to provide the francophone scientific community with guidelines for the evaluation of automatic summaries of French texts. Evaluation campaigns for NLP applications already exist in France, the EVALDA project[1]. However, no campaign has yet been launched for French automatic summaries, like Document Understanding Conferences for English texts or Text Summarization Challenge for Japanese texts. I hope that such a campaign will begin in the near future and that my thesis work may then serve as a guide for its design.

2 Completed Work

I collected 22 scientific papers about summarization evaluation, published between 1961 and 2005. Each paper has been the subject of an in-depth analysis, where every detail regarding the evaluation has been carefully noted, yielding a quasi-monstrous amount of experimental variables. These variables have been classified into four categories: 1) information about source texts, 2) information about automatic summaries being evaluated, 3) information about other summaries used in the evaluation process, and 4) information about evaluation methods and criteria. At the current stage of my research work, the first three types of variables have been analyzed and will be presented here.

2.1 Variables about source texts

Four types of information about source texts emerged from the analysis: 1) the number of source texts, 2) the length, 3) the type of text, and 4) the language. First, the number of source texts is an indicator of the significance of the evaluation. In my study,

[1] http://www.elda.org/rubrique25.html

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 17–20,
Rochester, April 2007.

all the evaluations used less than 100 source texts, except for Mani and Bloedorn (1999) (300 source texts), Brandow et al. (1995) (250 source texts), Kupiec et al. (1995) (188 source texts) and Teufel and Moens (1999) (123 source texts).

Secondly, regarding source text length, it is expressed in different ways from one evaluation to another. For example, Edmundson (1969) gives the number of words, Klavans et al. (1998) give the number of sentences and Minel et al. (1997) give the number of pages. In some papers, the length of the shortest and of the longest text is provided (Marcu, 1999) while in others it is the average number of words, sentences or pages that is given (Teufel and Moens, 1999). Obviously, it would be wise to standardize the way source texts length is given in evaluation experiments.

In my corpora, there are three main types of source texts: 1) scientific papers, 2) technical reports, and 3) newspapers. Also, Minel et al. (1997) used book extracts and memos, and Farzindar and Lapalme (2005) used judgments of the Canadian federal court. All evaluations used only one type of source texts, except for Kupiec et al. (1995) and for Minel et al. (1997).

Finally, the majority of the evaluations used English texts. Some authors used French texts (Minel et al., 1997; Châar et al., 2004), Korean texts (Myaeng and Jang, 1999) or Japanese texts (Nanba and Okumura, 2000).

2.2 Variables about automatic summaries being evaluated

In this section, I describe variables about automatic summaries being evaluated. The variables have been classified into six categories: 1) the total number of automatic summaries evaluated, 2) the number of automatic summaries produced per source text, 3) if they are multiple document summaries, 4) the length, 5) if they are extracts or abstracts, and 6) their purpose.

First, concerning the total number of automatic summaries, Brandow et al. (1995), Mani and Bloedorn (1999), Kupiec et al. (1995), Salton et al. (1997) and Teufel and Moens (1999) evaluated respectively 750, 300, 188, 150 and 123 automatic summaries. All the other studies for which this information is given evaluated less than 100 automatic summaries. It may appear redundant to give the number of source texts and the number of automatic summaries in an evaluation, but sometimes more than one automatic summary per source text may have been produced. This is the case in Brandow et al. (1995) and Barzilay and Elhadad (1999) where automatic summaries of different lengths have been evaluated.

Automatic summaries can either be produced from one text or more than one text. In my corpora, only Mani and Bloedorn (1999) and Châar et al. (2004) evaluated multiple document summaries.

As for source texts, automatic summary length is expressed in different ways from one evaluation to another. Moreover, it is not always expressed in the same way than source text length, which is inconsistent.

On a different note, most experiments evaluated extracts, except for Maybury (1999) and Saggion and Lapalme (2002) who evaluated abstracts, reflecting the predominance of systems producing extracts in the domain of summarization. Extracts are summaries produced by extracting the most important segments from texts while abstracts are the result of a comprehension process and text generation. Most extracts evaluated are composed of sentences, except for Salton et al. (1997) and Châar et al. (2004) where they are respectively composed of paragraphs and passages. The type of automatic summaries is crucial information because it normally influences the choice of the evaluation method and criteria. Indeed, we do not evaluate extracts and abstracts in the same way since they are not produced in the same way. Also, their purposes generally differ, which can also influence the choice of the evaluation method and criteria.

Last, some papers contain the specific purpose of automatic summaries, not only if they are indicative or informative, which is interesting because it can sometimes explain the choice of the evaluation method. Only 9 experiments out of 22 give this information in my corpora.

2.3 Variables about other summaries used in the evaluation process

One of the most common evaluation methods consists of comparing automatic summaries with other summaries. During my analysis, I identified seven

types of information about these other summaries: 1) the total number of other summaries, 2) the type of summaries, 3) the length, 4) the total number of human summarizers, 5) the number of human summarizers per source text, 6) the instructions given to the human summarizers, and 7) the human summarizers' profile.

The number of other summaries does not necessarily correspond to the number of automatic summaries evaluated, depending on many factors: the use of other summaries of different types or different lengths, the number of persons producing the other summaries, the number of other systems producing the other summaries, and so on.

There are two general types of summaries used for comparison with the automatic summaries being evaluated. First, *gold standard summaries* (or *target summaries*) can be author summaries, professional summaries or summaries produced specifically for the evaluation. Second, *baseline summaries* are generally produced by extracting random sentences from source texts or produced by another system.

In my corpora, gold standard summaries are often produced specifically for the evaluation. In most cases, they are produced by manually extracting the most important passages, sentences or paragraphs, allowing automatic comparison between automatic summaries and gold standard summaries.

On the other hand, many evaluations used baseline summaries. For example, Barzilay and Elhadad (1999) used summaries produced by *Word AutoSummarize*, Hovy and Lin (1999) used summaries produced by automatically extracting random sentences from source texts. In Brandow et al. (1995), Kupiec et al. (1995) and Teufel and Moens (1999), baseline summaries were produced by automatically extracting sentences at the beginning of the texts, and in Myaeng and Jang (1999) by extracting the first five sentences of the conclusion.

Logically, the length of the summaries used for the comparison should be equivalent to the length of the automatic summaries being evaluated. If automatic summaries of different lengths are evaluated, there should be corresponding baselines and/or gold standard summaries for each length, unless the goal of the evaluation is to determine if the length plays a role in the quality of automatic summaries.

Many of the evaluations analyzed do not indicate the number of human summarizers participating in the production of gold standard summaries. A few of them specify the total number of persons involved, but not the number for each source text. This is an important variable because summarizing, either by extracting or abstracting, is a subjective task. The more people involved in the summarization of one text, the more we can consider the final summary to be reliable. From the pieces of information I was able to gather, the number of summarizers per source text ranges from 1 to 13 in my corpora.

In analyzing the evaluations of my corpora, I realized that some authors gave clear instructions to the human summarizers, for example Edmundson (1969). In other cases, authors asked the summarizers to extract the most "important" sentences. The term "important" includes other terms like representative, informative, relevant, and eligible. It is rarely mentioned however if those words were explained to the summarizers.

I also noticed that some evaluations used people coming from different backgrounds, for example in Salton et al. (1997), while others used more homogeneous groups, for example in Barzilay and Elhadad (1999) and Kupiec et al. (1995).

3 Future Directions

In the next couple of months, I plan to analyze evaluation methods identified in my corpora, for example comparing automatic summaries with gold standard or baseline summaries, and asking judges to give their opinion on the quality of automatic summaries. I will also describe evaluation criteria used to assess the quality of the automatic summaries, for example informativeness and readability. Next, I will make recommendations for the presentation of summarization evaluation results, based on the knowledge acquired from my analysis of 22 scientific papers, and from previous evaluation campaigns.

4 Conclusion

In this paper, I described variables about source texts, about automatic summaries being evaluated and about other summaries used in summarization evaluation experiments. These variables provide important information for the understanding of the

evaluation results presented in a scientific paper. My analysis is based on 22 scientific papers on summarization evaluation, which is to my knowledge the largest study on the variables found in evaluation experiments. This constitutes a notable contribution in the domain of summarization. In another paper (in French) to appear, I propose a French terminology for the presentation of evaluation results in the domain of summarization, which is also a major contribution.

To conclude, the analysis presented in this paper gave an overview of summarization evaluation habits since 1961. Also, it showed that there is no common agreement as to how evaluation results should be presented in a scientific paper about automatic summaries.

Acknowledgements

I would like to thank the SSHRC and the FQRSC for granting me doctoral scholarships. I would also like to thank Joël Bourgeoys, Neil Cruickshank, Lorraine Couture and the anonymous reviewer for their useful comments.

References

R. Barzilay and M. Elhadad. 1999. Using lexical chains for text summarization. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 111–121, Cambridge, Massachusetts. MIT Press.

R. Brandow, K. Mitze, and L. Rau. 1995. Automatic condensation of electronic publications by sentence selection. *Information Processing Management*, 31(5):675–685.

S. L. Châar, O. Ferret, and C. Fluhr. 2004. Filtrage pour la construction de résumés multidocuments guidée par un profil. *Traitement automatique des langues*, 45(1):65–93.

H. P. Edmundson. 1969. New methods in automatic abstracting. *Journal of the Association for Computing Machinery*, 16(2):264–285.

A. Farzindar and G. Lapalme. 2005. Production automatique de résumé de textes juridiques : évaluation de qualité et d'acceptabilité. In *TALN*, pages 183–192, Dourdan.

E. Hovy and C.-Y. Lin. 1999. Automated text summarization in SUMMARIST. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94, Cambridge, Massachusetts. MIT Press.

J. L. Klavans, K. R. McKeown, M.-Y. Kan, and S. Lee. 1998. Resources for the evaluation of summarization techniques. In Antonio Zampolli, editor, *LREC*, pages 899–902, Granada, Spain.

J. Kupiec, J. Pedersen, and F. Chen. 1995. A trainable document summarizer. In *SIGIR*, pages 68–73, Seattle.

I. Mani and E. Bloedorn. 1999. Summarizing similarities and differences among related documents. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 357–379, Cambridge, Massachusetts. MIT Press.

D. Marcu. 1999. Discourse trees are good indicators of importance in text. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 123–136, Cambridge, Massachusetts. MIT Press.

M. Maybury. 1999. Generating summaries from event data. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 265–281, Cambridge, Massachusetts. MIT Press.

J.-L. Minel, S. Nugier, and G. Piat. 1997. How to appreciate the quality of automatic text summarization? Examples of FAN and MLUCE protocols and their results on SERAPHIN. In *EACL*, pages 25–31, Madrid.

S. H. Myaeng and D.-H. Jang. 1999. Development and evaluation of a statistically-based document summarization system. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 61–70, Cambridge, Massachusetts. MIT Press.

H. Nanba and M. Okumura. 2000. Producing more readable extracts by revising them. In *18th International Conference on Computational Linguistics*, pages 1071–1075, Saarbrucker.

H. Saggion and G. Lapalme. 2002. Generating indicative-informative summaries with SumUM. *Computational Linguistics*, 28(4):497–526.

G. Salton, A. Singhal, M. Mitra, and C. Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–207.

S. Teufel and M. Moens. 1999. Argumentative classification of extracted sentences as a first step towards flexible abstracting. In I. Mani and M. T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 155–171, Cambridge, Massachusetts. MIT Press.

Exploiting Event Semantics to Parse the Rhetorical Structure of Natural Language Text

Rajen Subba
Department of Computer Science
University of Illinois at Chicago
Chicago, IL 60613
rsubba@cs.uic.edu

Abstract

Previous work on discourse parsing has mostly relied on surface syntactic and lexical features; the use of semantics is limited to shallow semantics. The goal of this thesis is to exploit event semantics in order to build discourse parse trees (DPT) based on informational rhetorical relations. Our work employs an Inductive Logic Programming (ILP) based rhetorical relation classifier, a Neural Network based discourse segmenter, a bottom-up sentence level discourse parser and a shift-reduce document level discourse parser.

1 Introduction

Discourse is a structurally organized set of *coherent* text segments. The minimal unit of discourse is called an elementary discourse unit (EDU). An EDU or a span of EDUs constitute a segment. When we read text, we automatically assign *rhetorical* (coherence) relations to segments of text that we deem to be related. Consider the segmented text below:

(Example 1) [Clean the walls thoroughly$_{(1a)}$] [and allow them to dry.$_{(1b)}$] [If the walls are a dark color,$_{(2a)}$] [apply primer.$_{(2b)}$] [Put a small amount of paste in the paint tray;$_{(3a)}$] [add enough water$_{(4a)}$] [to thin the paste to about the consistency of cream soup.$_{(4b)}$]

It is plausible to state that the rhetorical relation between $(1a)$ and $(1b)$ is *preparation:act*. We can also posit that the relation *act:goal* holds between $(4a)$ and $(4b)$. Figure 1 shows the complete annotation of the full text. Now, if we were to reorder these segments as $[(1b), (4a), (2a), (4b), (3a), (2b), (1a)]$, the text would not make much sense. Therefore, it is imperative that the contiguous spans of discourse be coherent for comprehension. Rhetorical relations help make the text coherent.

Rhetorical relations based on the subject matter of the segments are called informational relations. A common understanding in discourse study is that informational relations are based on the underlying content of the text segments. However, previous work (Marcu, 2000; Polanyi et al., 2004; Soricut and Marcu, 2005; Sporleder and Lascarides, 2005) in discourse parsing has relied on syntactic and lexical information, and *shallow* semantics only.

The goal of this thesis is to build a computational model for parsing the informational structure of instructional text that exploits "deeper semantics", namely event semantics. Such discourse structures can be useful for applications such as information extraction, question answering and intelligent tutoring systems. Our approach makes use of a neural network discourse segmenter, a rhetorical relation classifier based on ILP and a discourse parsing model that builds sentence level DPTs bottom-up and document level DPTs using a shift-reduce parser.

In section 2, we describe how we collected our data. In section 3, we present our automatic discourse segmenter. Section 4 details our discourse parsing model based on event semantics followed by the conclusion in section 5.

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 21–24,
Rochester, April 2007.

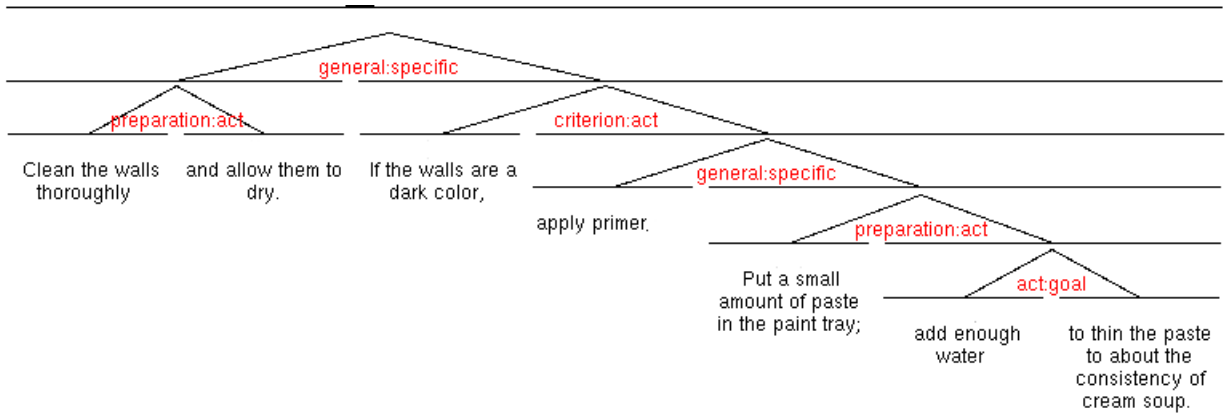


Figure 1: Discourse Annotation for Example 1

2 Data Collection

Our work calls for the use of a supervised machine learning approach. Therefore, we have manually annotated a corpus of instructional text with rhetorical relations and event semantic information. We used an existing corpus on home repair manuals (5Mb).[1]

2.1 Manual Discourse Annotation

In order to carry out the manual discourse annotation, a coding scheme was developed based on Marcu (1999) and RDA (Moser et al., 1996). The annotated data consists of 5744 EDUs and 5131 relations with a kappa value of 0.66 on about 26% of the corpus. We analyzed a total of 1217 examples to determine whether a cue phrase was present or not. Only 523 examples (43%) were judged to be signalled. Furthermore, discourse cues can be ambiguous with regard to which relation they signal. In order to account for cases where discourse cues are not present and to resolve such ambiguities, we intend to exploit event semantics.

2.2 Semi-Automatic Event Semantic Annotation

Informational relations describe how the content of two text segments are related. Therefore, it makes intuitive sense that verb semantics can be useful in determining these relations.[2] In Subba et al. (2006), we integrated LCFLEX (Rose and Lavie, 2000) with VerbNet (Kipper et al., 2000) and CoreLex (Buitelaar, 1998) to compositionally build verb based event semantic representations of our EDUs.

VerbNet groups together verbs that undergo the same syntactic alternations and share similar semantics. It accounts for about 4962 distinct verbs classified into 237 main classes. The semantic information is described in terms of an event that is decomposed into four stages, namely *start*, *during*, *end* and *result*. Semantic predicates like *motion* and *together* describe the participants of an event at various stages. CoreLex provides meaning representations for about 40,000 nouns that are compatible with VerbNet.

The parser was used to semi-automatically annotate both our training and test data. Since the output of the parser can be ambiguous with respect to the verb sense, we manually pick the correct sense.[3]

3 Automatic Discourse Segmentation

The task of the discourse segmenter is to segment sentences into EDUs. In the past, the problem of sentence level discourse segmentation has been tackled using both symbolic methods (Polanyi et al., 2004; Huong et al., 2004) as well as statistical models (Soricut and Marcu, 2003; Marcu, 2000) that have exploited syntactic and lexical features.

We have implemented a Neural Network model

[1]The corpus was collected opportunistically off the internet and from other sources, and originally assembled at the Information Technology Research Institute, University of Brighton.

[2]Especially in instructional manuals where the meaning of most sentences is centered on verbs.

[3]In addition, the parser generates semantic representations for fragments of the sentence to handle ungrammatical sentences, etc.

for sentence level discourse segmentation that uses syntactic features and discourse cues. Our model was trained and tested on RST-DT (2002) and achieves a performance of up to 86.12% F-Score, which is comparable to Soricut and Marcu (2003). We plan to use this model on our corpus as well.

4 Discourse Parsing

Once the EDUs have been identified by the discourse segmenter, the entire discourse structure of text needs to be constructed. This concerns determining which text segments are related and what relation to assign to those segments. Our discourse parsing model consists of a rhetorical relation classifier, a sentence level discourse parser and a document level discourse parser.

4.1 Rhetorical Relation Classifier

In a preliminary investigation (Subba et al., 2006), we modeled the problem of identifying rhetorical relations as a classification problem using rich verb semantics only.

Most of the work in NLP that involves learning has used more traditional machine learning paradigms like decision-tree algorithms and SVMs. However, we did not find them suitable for our data which is represented in first order logic (FOL). We found Progol (Muggleton, 1995), an ILP system, appropriate for our needs. The general problem specification for Progol (ILP) is given by the following posterior sufficiency property:

$$B \wedge H \models E$$

Given the background knowledge B and the examples E, Progol finds the simplest consistent hypothesis H, such that B and H entails E. The rich verb semantic representation of pairs of EDUs form the background knowledge and the manually annotated rhetorical relations between the pairs of EDUs serve as the positive examples.[4] An A*-like search is used to search for the most probable hypothesis. Given our model, we are able to learn rules such as the ones given in Figure 2. Due to the lack of space we only explain RULE1 here. RULE1 states that

```
RULE1:
relation(EDU1,EDU2,'before:after') :- motion(EDU1,event0,during,C),
        location(EDU2,event0,start,C,D).
RULE2:
relation(EDU1,EDU2,'act:goal') :- cause(EDU1,C,event0),
        together(EDU1,event0,end,physical,F,G),cause(EDU2,C,event0).
```

Figure 2: Examples of Rules learned by Progol

there is a theme (C) in motion during the event in EDU1 (the first EDU) and that C is located in location D at the start of the event in EDU2 (the second EDU).

We trained our classifier on 423 examples and tested it on 85 examples.[5] A majority function baseline performs at a 51.7 F-Score. Our model outperforms this baseline with an F-Score of 60.24.

Relation	Precision	Recall	F-Score
goal:act	31.57	26.08	28.57
step1:step2	**75**	**75**	**75**
before:after	54.5	54.5	54.5
criterion:act	71.4	71.4	71.4
Total	**61.7**	**58.8**	**60.24**

Table 1: Rhetorical Relation Classifier Result

This study has shown that it is possible to learn rules from FOL semantic representations using Inductive Logic Programming to classify rhetorical relations. However, it is not yet clear how useful event semantics is for discourse parsing. In the future, we intend to extend our model to incorporate syntactic and lexical information as well. Such an extension will allow us to assess the contribution of event semantics.

4.2 Building Discourse Parse Trees

In addition to extending the rhetorical relation classifier, our future work will involve building the discourse parse tree at the sentence level and at the document level. At the document level, the input will be the sentence level discourse parse trees and the output will be the discourse structure of the entire

[4] The output from the parser was further processed into definite clauses. Positive examples are represented as ground unit clauses.

[5] For this preliminary experiment, we decided to use only those relation sets that had more than 50 examples and those that were classified as *goal:act*, *step1:step2*, *criterion:act* or *before:after*

document.

When combining two text segments, promotion sets that approximate the most important EDUs of the text segments will be used. As a starting point, we propose to build sentence level DPTs bottom-up. EDUs that are subsumed by the same syntactic constituent (usually an S, S-Bar, VP) will be combined together into a larger text segment recursively until the the DPT at the root level has been constructed. At the document level, the DPT will be built using a shift-reduce parser as in Marcu (2000). However, unlike Marcu (2000), there will only be one shift and one reduce operation. The reduce operation will be determined by the rhetorical relation classifier and an additional module that will determine all the possible attachment points for an incoming sentence level DPT. An incoming sentence level DPT may be attached to any node on the right frontier of the left DPT. Lexical cohesion will be used to rank the possible attachment points. For both sentence level discourse parsing and document level discourse parsing, the rhetorical relation classifier will be used to determine the informational relation between the text segments.

5 Conclusion

In conclusion, this thesis will provide a computational model for parsing the discourse structure of text based on informational relations. Our approach exploits event semantic information of the EDUs. Hence, it will provide a measurement of how helpful event semantics can be in uncovering the discourse structure of text. As a consequence, it will also shed some light on the coverage of the lexical resources we are using. Other contributions of our work include a parser that builds event semantic representations of sentences based on rich verb semantics and noun semantics and a data driven automatic discourse segmenter that determines the minimal units of discourse.

References

Buitelaar, P.: CoreLex: Systematic Polysemy and Underspecification. Ph.D. thesis, Computer Science, Brandeis University, February 1998.

Huong Le Thanh, G. A. and Huyck., C.: Automated discourse segmentation by syntactic information and cue phrases. International Conference on Artificial Intelligence and Applications, 2004.

Kipper, K., H. T. D. and Palmer., M.: Class-based construction of a verb lexicon. AAAI-2000, Proceedings of the Seventeenth National Conference on Artificial Intelligence, 2000.

Livia Polanyi, Christopher Culy, M. H. v. d. B. G. L. T. and Ahn., D.: Sentential structure and discourse parsing. ACL 2004, Workshop on Discourse Annotation, 2004.

Marcu, D.: Instructions for Manually Annotating the Discourse Structures of Texts. Technical Report, University of Southern California, 1999.

Marcu, D.: The theory and practice of discourse parsing and summarization. Cambridge, Massachusetts, London, England, MIT Press, 2000.

Moser, M. G., Moore, J. D., and Glendening, E.: Instructions for Coding Explanations: Identifying Segments, Relations and Minimal Units. University of Pittsburgh, Department of Computer Science, 1996.

Muggleton., S. H.: Inverse entailment and progol. In New Generation Computing Journal, 13:245–286, 1995.

Rosé, C. P. and Lavie., A.: Balancing robustness and efficiency in unification-augmented context-free parsers for large practical applications. In Jean-Clause Junqua and Gertjan van Noord, editors, Robustness in Language and Speech Technology, 2000.

RST-DT.: Rst discourse treebank. Linguistic Data Consortium., 2002.

Sporleder, C. and Lascarides., A.: Exploiting linguistic cues to classify rhetorical relations. Recent Advances in Natural Language Processing, 2005.

Soricut, R. and Marcu., D.: Sentence level discourse parsing using syntactic and lexical information. Proceedings of the HLT and NAACL Conference, 2003.

Subba, R., Di Eugenio, B., E. T.: Building lexical resources for princpar, a large coverage parser that generates principled semantic representations. LREC 2006, 2006.

Subba, R., Di Eugenio, B., S. N. K.: Learning FOL rules based on rich verb semantic representations to automatically label rhetorical relations. EACL 2006, Workshop on Learning Structured Information in Natural Language Applications, 2006.

Wellner, B., Pustejovsky, J., C. H. R. S. and Rumshisky., A.: Classification of discourse coherence relations: An exploratory study using multiple knowledge sources. SIGDIAL Workshop on Discourse and Dialogue, 2006.

Dynamic Use of Ontologies in Dialogue Systems

Joana Paulo Pardal
Department of Information Systems and Computer Engineering
Instituto Superior Técnico, Technical University of Lisbon
Lisbon, Portugal
joana@l2f.inesc-id.pt

Abstract

Most dialogue systems are built with a single task in mind. This makes the extension of an existing system one of the major problems in the field as large parts of the system have to be modified. Some recent work has shown that ontologies have a role on the domain knowledge representation as the knowledge collected in an ontology can be used in all the modules. This work aims to follow the footsteps of the use of ontologies in dialogue systems and take it further as the current state of the art only uses taxonomical knowledge.

1 Introduction

At the present time, the Spoken Language Systems Lab (L^2F) integrates a project in the "House of the Future" at the Portuguese Communications Foundation. The house has a spoken dialogue system (Mourão et al., 2004) based on TRIPS architecture (Allen et al., 2005) where a virtual butler named "Ambrósio" helps the user in daily tasks that deal with devices and services, through speech commands. Whenever clarification is needed, further dialogue is entailed. To act in response to the user, the system needs to know which devices are connected, which services are available and what actions can be performed. Currently, this information is stored for each service or device: the available operations, the needed parameters and the possible values for each one. This kind of architecture is very common in the field. Nevertheless it's still hard to extend an existing system because it's always necessary to adapt lots of features in the system.

Recent work from Filipe (2006) has enhanced the access to the services and abstracted the database view in order to create an Application Programming Interface (API). The main contribution of that work is a Domain Knowledge Manager (DKM) advisor service, which suggests the best task-device pairs to satisfy a request. Additionally, a DKM recognizer service to identify the domain concepts from a natural language request is proposed. A hybrid approach is used to design ubiquitous domain models to allow the dialogue system to recognize the available devices and tasks they provide on-the-fly.

But more work is still needed to ease the dynamic configuration of dialogue systems and to deal with a set of arbitrary plug-and-play devices. The main goal of this work is to pursue the work done by Filipe.

2 State of the art

This work encompasses knowledge and techniques from two different areas: dialogue systems and ontologies. This work has to deal with the challenges from all these areas.

2.1 Dialogue Systems

Since the 1980s, the Natural Language Processing community has used spoken dialogue systems as a case study (Colea et al., 1997). This option is explained by the simplicity that comes from the treatment of restricted domains. The multidisciplinarity involved is one of the richnesses of this field as

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 25–28,
Rochester, April 2007.

it brings together people from several communities like signal processing – for speech recognition (Jurafsky and Martin, 2000) and synthesis (Huang et al., 2001); artificial intelligence – for interpretation of the spoken utterances (Allen, 1987); and software engineering – for more efficient architectures (McTear, 2002). But the complexity of these systems makes them expensive to develop (Allen et al., 2000) and difficult to adapt to new types of users, services, languages and scenarios (Turunen and Hakulinen, 2003).

With the proliferation of databases, some work has been done to take advantage of the knowledge structure and organization to dynamically extend existing systems to new domains, devices and services.

2.2 Ontologies

Ontologies aim at capturing static domain knowledge in a generic way and providing a commonly agreed understanding of a given domain. The main purpose is to share and reuse that knowledge across applications. The field of Ontologies appeared in the 1990s (Gruber, 1993), but only lately has been perceived as more valuable, as some effective results are being achieved with their use, reuse and sharing. Being so, an ontology is a formalized shared specification of a conceptualization. Mainly, a domain ontology collects the relevant concepts of a domain and the relations between them. An ontology usually also represents some formal restrictions verified in the domain. Therefore, ontologies usually have three types of entities: classes, relations, and axioms.

Currently the main challenges in this area include the definition of a clear building process (Pinto and Martins, 2004), automatic learning of ontologies (Maedche and Staab, 2004), transparent access to information (Gil et al., 2005) and efficient inference based on the available knowledge (Baader et al., 2003). Some work has been done where databases and other legacy knowledge sources are replaced by ontologies in different types of domains with success (Grau et al., 2005).

2.3 Use of Ontologies in Dialogue Systems

Separating the domain knowledge from the language features of the spoken dialogue systems has proven to reduce the complexity of a dialogue system's components. Moreover, if the domain knowledge is already available, reusing it is crucial to reduce the effort needed to build a new dialogue system or to extend an existing one into a new subject. Some recent work has shown the advantages of the use of Ontologies for these tasks.

Milward and Beveridge (2003) maintain that the ontology-based dialogue system for home information and control provides a dynamically reconfigurable system were new devices can be added and users can subscribe to new ones; asynchronous device input is allowed; unnatural scripted dialogues are avoided; and a flexible multimodal interaction for all users including the elderly and the disabled is provided. Also, the recognition, interpretation, generation and dialogue management are more flexible as the knowledge coded on the ontology can be used dynamically.

Flycht-Eriksson (2004) argues that the separation of the dialogue management from the domain knowledge management is crucial to reduce the complexity of the systems and enhance further extensions.

Both these works focus on the IS-A and PART-OF relations to solve under/over specification. This is helpful in medical-related dialogue systems that need taxonomical knowledge of the domain. Using more relations is still a challenge as the complexity increases.

3 Main goals

The main goal of this project is to enhance spoken dialogue systems to make them more general and domain-independent. This means that knowledge should be introduced in the system more easily and transparently. To do this, the dialog management should be separated from the domain knowledge management. This should be done not only by assigning a system module to it (the service manager) that has to be adapted to each domain, but, additionaly, by defining the kind of domain knowledge needed and creating an abstraction to represent it. For example, the dialogue system needs to know the possible words in the next expected response from the user and that depends mainly on the domain. This separation eases the creation of mechanisms to treat the common dialogue phenomena. A library

for these phenomena should be reused in dialogue systems across all domains.

Contributions from the ontologies field will be explored in regard to knowledge manipulation in a generic spoken dialogue system. As said before, some work has been done in the field but, at least for now, most of the work is reduced to the hierarchical knowledge (classes and IS-A relations) and under/over specification (PART-OF relations) that usually are represented on the ontologies. The extra-taxonomical knowledge is still being ignored but should be considered as that is the main richness of ontologies.

The most interesting topic is whether ontologies can enrich a spoken dialogue system and be used by it in such a way that the system can abstract the knowledge source thus allowing the system to focus only on dialogue phenomena and rather than the architecture adaptation that has to be done in order to include new domains.

The definition of the dialogue system as the instantiation of a spoken dialogue system will be explored after the existing dialogue systems and ontologies have been studied and categorized according to the tasks they perform and the used knowledge sources.

4 Completed Work

An ontology on the cooking domain has been built (Ribeiro et al., 2006; Batista et al., 2006). This ontology still hasn't been used but it will be included in our dialogue systems to provide help during the execution of a recipe. Currently an undergraduate student is enriching this ontology with a collection of recipes automatically extracted from text.

Also, a first prototype version of a cooking butler has been implemented. It lets the user choose from a list of recipes one to be dictated to him. Forward and rewind commands are available. This work is still preliminary as it doesn't use any ontology. It was done by two undergraduate students as a proof of concept that our current system can be extended to a dictating task.

5 Future directions

Since the PhD is still on going, lots of work is yet to be done. The next step to achieve the main goal of this work is to study the existing dialogue systems with emphasis on the performed tasks and the used knowledge sources. Beyond the simple enumeration of all the published systems, the aim is to create a categorization of dialogue systems according to the tasks they allow and to the type of knowledge they use independent of the used knowledge representation primitives (classes, relations and axioms).

5.1 Tasks to be performed

- A survey on the existing ontologies according to the coded information: classes, relations and axioms.
- Exploratory work on how to manage the domain knowledge transparently, focusing on the integration of ontologies in dialogue systems.
- Arrange the current architecture to consider not only the TRIPS architectural proposal, but the contributions coming from the ontological field. The separation of the dialogue manager in two modules should be considered here: one module for the dialogue features independent from the domain and other for the domain knowledge management.
- Adapt the existing L^2F's spoken dialogue system to the identified requirements in order to use domain knowledge from an ontology.
- Use the proposed methodology to include a cooking ontology on the L^2F's dialogue system to extend it to new domains.
- Include ontologies from different domains. An entertainment (Theatre, Movies, etc) domain ontology is being build.

5.2 Intelectual Contributions

- Classification of the existing dialogue systems according to the type of information they need and use;
- Classification of the used ontologies in dialogue systems according to the information coded and the used classes, relations and axioms;
- Propose an architecture where the contribution of each module is clearer and where the information flows both forward and backward;

- Propose a methodology for the integration of ontologies into general dialogue systems according to their classification;
- Integration of a cooking ontology into the existing dialogue system;
- Integration of another ontology into another dialogue system (from UoR).

References

James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering*, 6(3).

James Allen, George Ferguson, Mary Swift, Amanda Stent, Scott Stoness, Lucian Galescu, Nathanael Chambers, Ellen Campana, and Gregory Aist. 2005. Two diverse systems built using generic components for spoken dialogue (recent progress on TRIPS). In Ann Arbor, editor, *Proc. of the Interactive Poster and Demonstration Sessions at the 43rd Annual Meeting of ACL*, pages 85–88, Michigan, USA.

James F. Allen. 1987. *Natural Language Understanding*. Benjamin Cummings, 2nd edition.

Franz Baader, Diego Calvanese, Deborah McGuinness, Daniele Nardi, and Peter Patel-Schneider, editors. 2003. *The Description Logic Handbook: Theory, Implementation, and Applications*. Cambridge University Press.

Fernando Batista, Joana Paulo Pardal, Paula Vaz Nuno Mamede, and Ricardo Ribeiro. 2006. Ontology construction: cooking domain. Technical report, INESC-ID, Lisboa, Portugal.

Ron Colea, Joseph Mariani, Hans Uszkoreit, Giovanni Batista Varile, Annie Zaenen, Antonio Zampolli, and Victor Zue (editors), editors. 1997. *Survey of the State of the Art in Human Language Technology*. CSLU, CMU, Pittsburgh, PA.

Porfírio Pena Filipe and Nuno J. Mamede. 2006. A domain knowledge advisor for dialogue systems. In *International Joint Conference IBERAMIA/SBIA/SBRN 2006 – 4th Workshop in Information and Human Language Technology*.

Annika Flycht-Eriksson. 2004. *Design and Use of Ontologies in Information-providing Dialogue Systems*. Ph.D. thesis, School of Engineering at Linköping University.

Yolanda Gil, Enrico Motta, Richard Benjamins, and Mark Musen, editors. 2005. *The Semantic Web – 4th ISWC*, volume 3729 of *LNCS*. Springer, Ireland.

Bernardo Cuenca Grau, Ian Horrocks, Bijan Parsia, and Peter Patel-Schneider, editors. 2005. *What Have Ontologies Ever Done For Us: Potential Applications at a National Mapping Agency*, volume 188.

Thomas R. Gruber. 1993. A translation approach to portable ontology specifications. *Knowledge Acquisition*, 5(2):199–220.

Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon. 2001. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall.

Daniel Jurafsky and James H. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Speech Recognition, and Computational Linguistics*. Prentice-Hall.

Alexander Maedche and Steffen Staab, 2004. *Handbook on Ontologies*, chapter Ontology learning. International Handbooks on Information Systems. Springer.

Michael McTear. 2002. Spoken dialogue technology: enabling the conversational interface. *ACM Computing Surveys*, 34(1):90–169.

David Milward and Martin Beveridge. 2003. Ontology-based dialogue systems. In *3rd Workshop on Knowledge and Reasoning in Practical Dialogue Systems – 18th IJCAI03*.

Márcio Mourão, Renato Cassaca, and Nuno Mamede. 2004. An independent domain dialog system through a service manager. In *Proc. of 4th Intl. Conf. EsTAL*, pages 161–171. Springer-Verlag.

H. Sofia Pinto and João Pavão Martins. 2004. Ontologies: How can they be built? *Knowledge Information System*, 6(4):441–464.

Ricardo D. Ribeiro, Fernando Batista, Nuno J. Mamede Joana Paulo Pardal, and H. Sofia Pinto. 2006. Cooking an ontology. In *12th Intl. Conf. on AI: Methodology, Systems, Applications*, volume 4183, pages 213–221, Berlin.

Markku Turunen and Jaakko Hakulinen. 2003. Jaspis^2 - an architecture for supporting distributed spoken dialogues. In *Proc. of Eurospeech*, pages 1913–1916.

Semantic Frames in Romanian Natural Language Processing Systems

Diana Maria Trandabăț
Faculty of Computer Science, "Al.I.Cuza" University of Iaşi &
Institute for Computer Science, Romanian Academy, Iaşi Branch
16, Gen. Berthelot, 700483-Iaşi, Romania
dtrandabat@info.uaic.ro

Abstract

Interests to realize semantic frames databases as a stable starting point in developing semantic knowledge based systems exists in countries such as Germany (the Salsa project), England (the PropBank project), United States (the FrameNet project), Spain, Japan, etc. I thus propose to create a semantic frame database for Romanian, similar to the FrameNet database. Since creating language resources demands many temporal, financial and human resources, a possible solution could be the import of standardized annotation of a resource developed for a specific language to other languages. This paper presents such a method for the importing of the FrameNet annotation from English to Romanian.

1 Introduction

The realization of human-computer interaction in natural language represents a major challenge in the context of aligning Romania to existing technologies.

The proposed project aims to introduce the semantic frames and contexts, which define a concept's sense according to its facultative or mandatory valences (Baker and Fillmore, 1998), to Romanian NLP systems. The behavior of the Romanian clauses – mainly the verbal group, around which all the other sentence complements gravitates in a (more or less) specific order – has been closely debated in the last years (Irimia, 1997; Dobrovie-Sorin, 1994; Monachesi, 1998; Barbu, 1999), creating a proper frame for the introduction of semantic roles.

This paper presents the steps considered for the achievement of this project. Thus, Section 2 gives a very brief description of the frame semantics, and Section 3 presents the realization of a semantic structures database for the Romanian language, similar to those existing for English, German, or Spanish, containing detailed information about the relations between the semantic meaning and the syntax of the words. In the last section, some possible applications of the detection of semantic roles to written and spoken texts are mentioned (question answering systems, summarization systems, prosody prediction systems), before drawing some final conclusions.

2 Frame Semantics

The FrameNet (FN) lexical-semantic resource is based on the principles of Frame Semantics (FS). From FS point of view, the semantic/syntactic features of "*predicational words*"[1] (Curteanu, 2003-2004) are defined in a particular semantic frame. The sentences are schematic representations of different situations, including different participants, objects or other conceptual roles. Being a linguistically transposed experience, a sentence represents an event scenario that is structured around a semantic head. The meaning of this head

[1] Words, mostly verbs, but also several nouns and adjectives, bearing a *predicational feature*, viz. demanding a specific semantic argument structure in order to complete their meaning.

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 29–32,
Rochester, April 2007.

can be understood only by expressing the core frame elements and can, optionally, be enriched with other semantic features, by expressing some non-core frame elements.

Fillmore (1968) divides the language representation into two structures: Surface Structure (the syntactic knowledge) and Deep Structure (the semantic knowledge). The language process begins at the Deep Structure level with a non-verbal representation (an idea or a thought) and ends in the Surface Structure, as we express ourselves.

The Case Notions are representations at a semantic level of the lexical arguments. This inventory of cases comprises universal concepts, possible innate, sufficient for the classification of the verbs of a language and reusable in all languages. The list of Fillmore Cases, which will be considered for the project, includes: Agent, Instrument, Dative, Experiencer, Locative, Object, etc.

3 A Parallel Romanian/English FrameNet Using Annotation Import

The first step in the realization of the Romanian corpus of annotated semantic frames was the manual translation of 110 randomly selected sentences from the English FN. In order to align the Romanian version with the English one, a larger corpus was needed, so the translation continued with the *Event* frame, summing up to 1094 sentences. This frame was selected due to its rich frame to frame relations (Inheritance – *Change_of_consistency, Process_start,* etc., Subframe - *Change_of_state_scenario* and Using - *Process_end*). After the selection of the clauses and their translation, the Romanian sentences were aligned with the English ones using the aligner developed by the Institute of Research in Artificial Intelligence (Tufiş et al., 2005). The next step was the automatic import of the English annotation, followed by a manual verification, a detection of the mismatching cases and an optimization process which, based on inference rules, corrects the automatic annotation.

3.1 Automatic annotation import

The intuition behind the importing program (Trandabăț et al., 2005) is that most of the frames defined in the English FN are likely to be valid cross-linguistically, because semantic frames express conceptual structures, language independent at the deep structure level. The surface realization is realized according to each language syntactic constraints.

The automatic importing program is based on the correlation of the semantic roles expressed in English with the translation equivalents in Romanian of the words that realize a specific role. The automatic import is manually checked in order to establish the method efficiency.

3.2 The algorithm

The starting point for the German, Japanese and Spanish FN creation was the manual annotation at FE level of existing corpora for each language. For Romanian, I propose creating a corpus of semantic roles starting from the translation of (a part of) the English corpus of annotated sentences (see Figure 1).

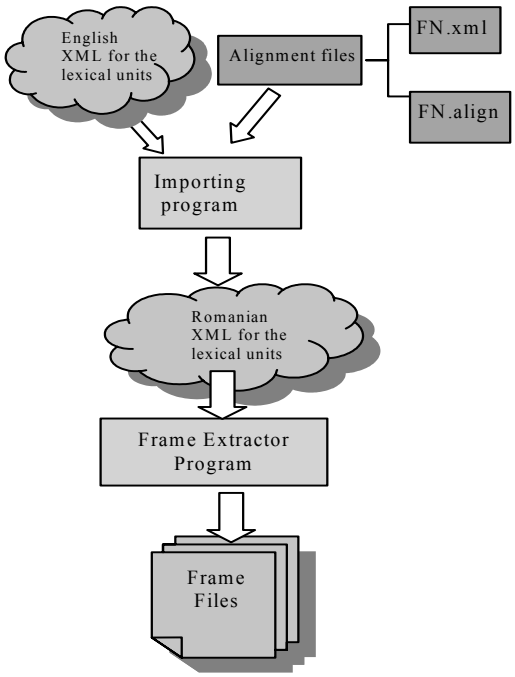


Figure 1. The architecture of the importing program

Using the XML files of the annotated English sentences, and the alignment files where each English word is linked to its corresponding Romanian translation, I automatically created a set of XML files containing a corpus of FE annotated sentences for the Romanian language. An example of imported annotation for the English lexical unit „occur" is presented in figure 2.

```
<annotationSet ID="1" status="AUTOMATIC">
 <layers>
  <layer ID="6375447" name="FE">
   <labels>
    <label name="Event" ID="19909459"
cDate="june 2006" start="0" end="9" />
    <label name="Time" ID="19909462"
cDate="june 2006" start="20" end="59" />
    <label name="Place" ID="19909465"
cDate="june 2006" start="61" end="101" />
   </labels>
  </layer>
..........
  <layer ID="6375452" name="Target">
   <labels>
    <label name="Target" ID="19905041"
cDate="june 2006" start="11" end="18" />
   </labels>
  </layer>
  <layer ID="6375453" name="Verb" />
 </layers>
 <sentence ID="671" aPos="103724676">
  <text>Incidentul a apărut după o dispută
între individ şi personal la o filială a
Băncii Irlandeze din Cahir .
  </text>
 </sentence>
</annotationSet>
```

Figure 2. Example of annotation set for English

The <*annotationSet*> tag indicates that a new sentence is annotated. Inside this tag, the <layers> tag sets the annotation layer (FE - Frame Element, GF - Grammatical Function or PT - Phrase Type) and the <*sentence*> tag encloses the text. The labels are applied to the words in <*text*>, indexed by their character. For example, the tag:

```
<label name="Event" ID="19909459"
cDate="june 2006" start="0" end="9"
/>
```

indicates that the *Event* frame element is starting with the first character of the sentence an stops at the 9th, meaning that the *Event* FE corresponds to „*Incidentul*" (en. *The incident*).

The general algorithm of the automatic importing program focuses on:

- reading of the input XML files;
- labeling of each English word with the corresponding semantic role (FE)
- converting the character indexes into a word level annotation;
- mapping the English words with the aligned Romanian correspondences, hence with the respective semantic role;
- writing an output XML file containing the Romanian annotated corpus.

For example, the lexical unit "occur.v" will appear in English and Romanian annotated as:

[Incidental]$_{Event}$ A APĂRUT [după o dispută între individ şi persona]$_{time/cause}$ [la o filială a Băncii Irlandeze din Cahir]$_{Place}$.
The incident]$_{Event}$ OCCURRED [after a dispute between the man and staff]$_{time/cause}$ [at a branch of the Bank of Ireland in Cahir]$_{Place}$

3.3 Optimization

My initial experiment has involved the translation of approx. 1000 sentences from English FN. The translations have been realized by professional translators, so the errors propagated in the corpus should be minimal. The reported problems during the translation relate mainly to the lack of the context of English sentences, which generate different translation variants. However, if the English semantic frame is considered, this problem is surmountable.

The alignment process was performed with the aligner developed by the Institute of Research in Artificial Intelligence (Tufiş, 2005), which is considered to have a precision of 87.17% and a recall of 70.25%. However, the aligner results were manually validated before entering the annotation import program.

The assessment of the correctitude of the obtained Romanian corpus is preformed manually. The first results of the annotation import show an overall accuracy of approx. 80%. The validation focuses on detecting the cases where the import has failed, trying to discover if the problems are due to the translation or to the semantic or syntactic specificities of Romanian. Only few translation errors were found, and even then, the meaning has been kept and the semantic roles were correctly assigned. However, there were cases where the FEs are expressed in English, but are implicit in the Romanian translation, as in:

[Blood]$_{Undergoer}$ had CONGEALED [thickly]$_{Manner}$ [on the end of the smashed fibula]$_{Place}$.

[Sângele]$_{Undergoer}$ se ÎNGROŞĂ [spre capătul fibulei zdrobite]$_{Place}$.

or not-expressed in English, but expressed in Romanian, as the *Protagonist* role in :

QUIT [smoking]$_{Process}$.
LĂSAŢI-[vă]$_{Protagonist}$ [de fumat]$_{Process}$.

The frame generation program based on the generated Romanian corpus is currently under development.

4 Conclusions

In this paper, I have presented a fast method for the realization of a Romanian corpus annotated with semantic frame relations. The main purpose of creating a quick semantic annotated database is using it as training corpus for automatic labeled semantic frames detection. Nowadays, expensive linguistic resources demanding a lot of time, money and human resources are created for different languages. After their utility is proved, those resources begin to be imported to other languages (see for instance the MultiSemCor project[2]). In this context, the realization of a Romanian FN is a challenging project in the frame of Romance FN.

The import method was preferred to the 'classical' creation by hand of a manually annotated corpus because of its possible automation. I investigate currently the possibility of using a translation engine for the most time consuming task, namely the translation of the English sentences. The project will be further developed by adding to the automatic import program rules discovered though the analysis of the mismatching cases.

The lack of semantic information was very obvious while working on the QA@CLEF competition[3] (Question Answering task within the Cross Language Evaluation Forum Competition) last year (Puşcaşu et al., 2006); having the semantic frames database (thus a semi-automatic role labeling system) can improve the precision of selecting an appropriate snippet for the desired answer, not to mention also the benefits for answer generation. Another application of the semantic frames I am interested in is prosody prediction. Within the Institute of Computer Science, I have begin to work at a syntax-prosody interface for Romanian based on FDG trees of sentences and other syntactical information to discover the phonological entities underlying the written text and the topic/focus articulation. The algorithm for finding sentence focus uses the semantic roles as a main component.

References

Baker, C., Fillmore, Ch., Lowe, J., *The Berkeley FrameNet project*, in Proceedings of the COLING-ACL, Montreal, Canada, 1998

Barbu, A-M, *The Verbal Complex*. Linguistic Studies and Enquires, L, no.1, Bucharest, p. 39-84 (In Romanian). 1999

Curteanu, N.: *Contrastive Meanings of the Terms "Predicative" and "Predicational" in Various Linguistic Theories (I, II)*. Computer Science Journal of Moldova (R. Moldova), Vol. 11, No. 4, 2003 (I); Vol. 12, No. 1, 2004 (II)

Curteanu, N., Trandabăţ, D., Moruz, M.: *Substructures of the (Romanian) Predicate and Predication Using FX-bar Projection Functions on the Syntactic Interface,* in Proc. of the 4th European Conference on Intelligent Systems and Technologies - ECIT2006, Iaşi, Romania, 2006.

Dobrovie-Sorin, C, *The syntax of Romanian. Comparative Studies.* Berlin: Mouton de Gruyter, 1994

Fillmore, Ch., *The case for case*; in Bach and Harms (Eds.), Universals in Linguistic Theory, Ed. Holt, Rinehart, and Winston, New York, 1968

Husarciuc M, Trandabăţ D., Lupu M., *Inferring Rules in Importing Semantic Frames from English FrameNet onto Romanian FrameNet,* 1st ROMANCE FrameNet Workshop, EUROLAN, Cluj, Romania 2005

Irimia, D. *The Morphosyntax of the Romanian* Verb. Ed. of the "Al. I. Cuza" Iaşi University (in Romanian). 1997

Monachesi, P., *The Morphosyntax of Romanian Cliticization*. in: P. A. Coppen et al. (Eds.), Proceedings of Computational Linguistics in The Netherlands, pp. 99-118, Amsterdam-Atlanta: Rodopi. 1998

Puşcaşu, G., Iftene, A., Pistol, I., Trandabăţ, D., Tufiş, D., Ceauşu, A., Ştefănescu, D., Ion, R., Orăşan, C., Dornescu, I., Moruz, A., Cristea, D., *Developing a Question Answering System for the Romanian-English Track at CLEF 2006*, CLEF 2006 Workshop, Alicante, Spain, to be published in LNCS

Trandabăţ, D., Husarciuc, M., Lupu, M., *Towards an automatic import of English FrameNet frames into the Romanian language*, 1st ROMANCE FrameNet Workshop, EUROLAN, Cluj, Romania, 2005

Tufiş, D., Ion R., Ceauşu, Al., Stefănescu, D., *Combined Aligners* in Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan, June, 2005

[2] http://multisemcor.itc.it/
[3] http://clef-qa.itc.it/2006bis/CLEF-2006.html

Combining Evidence for Improved Speech Retrieval

J. Scott Olsson
Department of Mathematics
University of Maryland
College Park, MD 20742
olsson@math.umd.edu

Abstract

The goal of my dissertation research is to investigate the combination of new evidence sources for improving information retrieval on speech collections. The utility of these evidence sources is expected to vary depending on how well they are matched to a collection's domain. I outline several new evidence sources for speech retrieval, situate them in the context of this domain dependency, and detail several methods for their combination with speech recognition output. Secondly, I highlight completed and proposed work for the production of this evidence.

1 Introduction and Goal

Early research in spoken document retrieval (SDR) was spurred by a new way to overcome the high cost of producing metadata (e.g., human assigned topic labels) or manual transcripts for spoken documents: large vocabulary continuous speech recognition. In this sense, SDR research has always been about making do with the available evidence. With the advent of automatic speech recognition (ASR), this available evidence simply grew from being only expensive human annotations to comparatively low-cost machine producible transcripts.

But today even more evidence is available for retrieving speech: (1) Using ASR text as input features, text classification can be applied to spoken document collections to automatically produce topic labels; (2) vocabulary independent spoken term detection (STD) systems have been developed which can search for query words falling outside of an ASR system's fixed vocabulary. These evidence sources can be thought of as two bookends to the spectrum of domain dependence and independence. On one end, topic labels can significantly improve retrieval performance but require the creation of a (presumably domain-dependent) topic thesaurus and training data. Furthermore, classification accuracy will be poor if the ASR system's vocabulary is badly matched to the collection's speech (e.g., we shouldn't expect a classifier to sensibly hypothesize automotive topics if the ASR system can not output words about *cars* or *driving*). On the other end, STD systems offer the most promise precisely when the ASR system's vocabulary is poorly matched to the domain. If the ASR system's vocabulary already includes every word in the domain, after all, STD can hardly be expected to help.

The primary goal of this dissertation is (1) to explore the combination of these new evidence sources with the features available in ASR transcripts or word lattices for SDR and (2) to determine their suitability in various domain-matching conditions. Secondarily, I'll explore improving the production of these new resources themselves (e.g., by classifying with temporal domain knowledge or more robust term detection methods).

Research in SDR has been inhibited by the absence of suitable test collections. The recently available MALACH collection of oral history data will, in large part, make this dissertation research possible (Oard et al., 2004). The MALACH test collection

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 33–36,
Rochester, April 2007.

contains about 1,000 hours of conversational speech from 400 interviews with survivors of the Holocaust[1]. The interviews are segmented into 8,104 documents with topic labels manually assigned from a thesaurus of roughly 40,000 descriptors. The collection includes relevance assessments for more than 100 topics and has been used for several years in CLEF's cross-language speech retrieval (CLSR) track (Oard et al., 2006).

Participants in the CLEF CLSR evaluations have already begun investigating evidence combination for SDR, through the use of automatic topic labels—although label texts are presently only used as an additional field for indexing. In monolingual English trials, this topic classification represents a significant effort both in time and money (i.e., to produce training data), so that these evidence combination studies have so far been rather domain dependent. Participants have also been using what are probably unnaturally good ASR transcripts. The speech is emotional, disfluent, heavily accented, and focused on a somewhat rare topic, such that the ASR system required extensive tuning and adaptation to produce the current word error rate of approximately 25%. In this setting, we'd expect STD output and topic labels to have low and high utility, respectively. To investigate the domain mismatch case, I will apply an off-the-shelf ASR system to produce new, comparatively poor, transcripts of the collection. In this setting, we'd expect STD output and topic labels to instead have high and low utility, respectively.

2 Proposed Combination Solutions

I will investigate improving SDR performance in both the poorly and well matched domain conditions through: (1) multiple approaches for utilizing automatically produced topic labels and (2) the utilization of STD output.

Throughout this paper, completed work will be denoted with a '$\star$', while proposed (non-complete, future) work will be denoted with a '†'.

[1]This is only a small subset of the entire MALACH collection, which contains roughly 116,000 hours of speech from 52,000 interviews in 32 languages. This additional data also provides training examples for classifi cation.

2.1 Speech Classification for SDR

I outline three methods of incorporating evidence from automatic classification for speech retrieval.

Creating Additional Indexable Text*

The simplest way to combine classification and speech retrieval is to use the topic labels associated with the classes as indexable text. As a participant on the MALACH project, I produced these automatic topic labels ("keywords") for the collection's speech segments. These keywords were used in this way in both years of the CLEF CLSR track. For a top system in the track, using solely automatically produced data (e.g., ASR transcripts and keyword text), indexing keyword text gave a relative improvement in mean average precision of 40.6% over an identical run without keywords (Alzghool and Inkpen, 2007).

Runtime Query Classification for SDR†

Simply using keyword text as an indexing field is probably suboptimal because information seekers don't necessarily speak the same language as the thesaurus constructors. An alternative is to classify the queries themselves at search time and to use these label assignments to rank the documents. We might expect this to be superior, insofar as information seekers use language more like interviewees (from which classification features are drawn) than like thesaurus builders.

Class Guided Document Expansion†

A third option for using classification output is as seed text for document expansion. The intuition here is that ASR text may be a strong predictor for a particular class label even if the ASR contains few terms which a user might consider for a query. In this sense, the class label text may represent a more semantically dense representation of the segment's topical content. This denser representation may then be a superior starting source for document centered term expansion.

2.2 Unconstrained Term Detection for SDR†

It is not yet clear how best to combine a STD and topical relevance IR system. One difficulty is that IR systems count words (or putative occurrences of words from an ASR system), while STD systems

report a score proportional to the confidence that a word occurs in the audio. As a solution, I propose normalizing the STD system's score for OOV query terms by a function of the STD system's score on putative occurrences of in-vocabulary terms. The intuition here is that the ASR transcript is roughly a ground truth representation of in-vocabulary term occurrences and the score on OOV query terms ought to reflect the STD system's confidence in prediction (which can be modeled from the STD system's score on "ground truth" in-vocabulary term occurrences). In this way, the presence or absence of in-vocabulary terms and their associated STD confidence scores can be used to learn a normalizer for the STD system's scores.

3 Producing the Evidence

In this section, I highlight both completed and proposed work to improve the production of evidence for combination.

3.1 Classifying with Temporal Evidence*

In spoken document collections, features beyond merely the automatically transcribed words may exist. Consider, for example, the oral history data contained in the MALACH collection. Each interview in this collection can be thought of as a time ordered set of spoken documents, produced by the guided interview process. These documents naturally arise in this context, and this temporal information can be used to improve classification accuracy.

This work has so far focused on MALACH data, although we expect the methods to be generally applicable to speech collections. For example, the topical content of a television episode may often be a good predictor of the subsequent episode's topic. Likewise, topics in radio, television, and podcasts may tend to be seasonally dependent (based on Holidays, recurring political or sporting events, etc.).

Time-shifted classification* One source of temporal information in the MALACH data is the features associated with temporally adjacent segments. Terms may be class-predictive for not only their own segment, but for the subsequent segments as well. This intuition may be easily captured by a *time shifted classification* (TSC) scheme. In TSC, each training segment is labeled with the *subsequent* segment's labels. During classification, each test segment is used to assign labels to its subsequent segment.

Temporal label weighting* We can also benefit from non-local temporal information about a segment. For example, because interviewees were instructed to relate their story in chronological order, we are more likely to find a discussion of childhood at an interview's beginning than at its end. We can estimate the joint probability of labels and segment times on held-out data and use this to bias new label assignments. We call this approach *temporal label weighting* (TLW).

In Olsson and Oard (2007), we showed that a combined TSC and TLW approach on MALACH data yields significant improvements on two separate label assignment tasks: conceptual and geographic thesaurus terms, with relative improvements in mean average precision of 8.0% and 14.2% respectively.

3.2 Classifying across languages*

In multilingual collections, training data for metadata creation may not be available for a particular language—a good example of domain mismatch. If however, training examples are available in a second language, the metadata may still be produced through *cross-language* text classification. In Olsson (2005), we used a probabilistic Czech-English dictionary to transform Czech document vectors into an English vector space before classifying them with k-Nearest Neighbors and English training examples. In this study, the cross-language performance achieved 73% of the monolingual English baseline on conceptual topic assignment.

3.3 Vocabulary Independent Spoken Utterance Retrieval*

In Olsson (2007), we examined a low resource approach to utterance retrieval using the expected posterior count of n-grams in phonetic lattices as indexing units. A query's phone subsequences are then extracted and matched against the index to produce a ranking on the lattices. Against a 1-best phone sequence baseline, the approach was shown to significantly improve the mean average precision of retrieved utterances on five human languages.

3.4 Improving Spoken Term Detection†

Phonetic lattices improve spoken term detection performance by more accurately encoding the recognizer's uncertainty in prediction. Even so, a correct lattice may not always contain a path with the query's entire phone sequence. This is so not only because of practical constraints on the size (i.e., depth) of the lattice, but also because speakers don't always pronounce words with dictionary precision. We'd like to allow approximate matching of a query's phone sequence with the phonetic lattices, and to do this as quickly as possible. This time requirement will prevent us from linearly scanning through lattices for near matches. I am currently investigating two solutions to this problem: phonetic query degradation and query expansion.

Phonetic query degradation† The idea in phonetic query degradation is to build an error model for the phone recognition system and to then degrade the query phone sequence such that it, hopefully, will more closely resemble recognized sequences. This approach incurs only a very slight cost in time and is query independent (in the sense that any term can be pushed through the degradation model—not, for example, only terms for which we can find recognized examples).

Phonetic query expansion† The idea of phonetic query expansion is, again, to transform the clean phone sequence of the query into the degraded form hypothesized by a recognizer. Instead of using a degradation model however, we simply run a first pass at STD with the non-degraded query term and use the putative occurrences to learn new, alternative, degraded forms for a second search pass. This can be thought of as blind relevance feedback or query by (putative) example.

The advantage of this approach is that we are not required to explicitly model the degradation process. Disadvantages are that we (1) require examples which may not be available and (2) assume that the degradation process is well represented by only a few examples.

4 Contributions

This dissertation will significantly contribute to speech retrieval research in several ways.

Can we improve SDR by evidence combination? By exploring evidence combination, this dissertation will advance the state of the art in speech retrieval systems and their applicability to diverse domains. I will investigate multiple methods for combining the evidence presented by both STD and classification systems with conventional ASR output (transcripts or word lattices). This work will develop upon previous research which studied, in depth, the use of only one evidence source, e.g., (Ng, 2000).

Can evidence combination decrease domain dependency? I will investigate how combining evidence sources can increase their applicability to new content domains. This will include, for example, understanding how (vocabulary independent) STD systems can be paired with fixed vocabulary ASR.

How can these evidence sources be improved? Lastly, I will explore how these new evidence sources may themselves be improved. This will include utilizing temporal domain knowledge for classification and improving the robustness of phone-based STD systems.

References

M. Alzghool and D. Inkpen. Experiments for the Cross Language Spoken Retrieval Task at CLEF 2006. In *Not yet published*.

K. Ng. 2000. Subword-based approaches for spoken document retrieval. MIT dissertation.

D.W. Oard, et al. Building an Information Retrieval Test Collection for Spontaneous Conversational Speech. In *Proceedings of SIGIR'04*.

D.W. Oard, et al. 2006. Evaluation of Multilingual and Multi-modal Information Retrieval. In *Seventh Workshop of the Cross-Language Evaluation Forum*, Alicante, Spain. Selected Papers Series: Lecture Notes in Computer Science.

J.S. Olsson and D.W. Oard. 2007. Improving Text Classifi cation for Oral History Archives with Temporal Domain Knowledge. In *Not yet published*.

J.S. Olsson, et al. Cross-language text classifi cation. In *Proceedings of SIGIR'05*.

J.S. Olsson, et al. Fast Unconstrained Audio Search in Numerous Human Languages. In *ICASSP'07*.

Unsupervised Natural Language Processing using Graph Models

Chris Biemann

NLP Dept., University of Leipzig

Johannisgasse 26

04103 Leipzig, Germany

biem@informatik.uni-leipzig.de

Abstract

In the past, NLP has always been based on the explicit or implicit use of linguistic knowledge. In classical computer linguistic applications *explicit* rule based approaches prevail, while machine learning algorithms use *implicit* knowledge for generating linguistic knowledge. The question behind this work is: how far can we go in NLP without assuming explicit or implicit linguistic knowledge? How much efforts in annotation and resource building are needed for what level of sophistication in text processing? This work tries to answer the question by experimenting with algorithms that do *not presume any* linguistic knowledge in the system. The claim is that the knowledge needed can largely be acquired by knowledge-free and unsupervised methods. Here, graph models are employed for representing language data. A new graph clustering method finds related lexical units, which form word sets on various levels of homogeneity. This is exemplified and evaluated on language separation and unsupervised part-of-speech tagging, further applications are discussed.

1 Introduction

1.1 Unsupervised and Knowledge-Free

A frequent remark on work dealing with unsupervised methods in NLP is the question: "Why not take linguistic knowledge into account?" While for English, annotated corpora, classification examples, sets of rules and lexical semantic word nets of high coverage do exist, this does not reflect the situation for most of even the major world languages. Further, as e.g. Lin (1997) notes, handmade and generic resources often do not fit the application domain, whereas resources created from and for the target data will not suffer from these discrepancies.

Shifting the workload from creating resources manually to developing generic methods, a one-size-fits-all solution needing only minimal adaptation to new domains and other languages comes into reach.

1.2 Graph Models

The interest in incorporating graph models into NLP arose quite recently, and there is still a high potential exploiting this combination (cf. Widdows, 2005). An important parallelism between human language and network models is the small world structure of lexical networks both built manually and automatically (Steyvers and Tenenbaum, 2005), providing explanation for power-law distributions like Zipf's law and others, see Biemann (2007). For many problems in NLP, a graph representation is an intuitive, natural and direct way to represent the data.

The pure vector space model (cf. Schütze, 1993) is not suited to highly skewed distributions omni-present in natural language. Computationally expensive, sometimes lossy transformations have to be applied for effectiveness and efficiency in processing. Graph models are a veritable alternative, as the equivalent of zero-entries in the vector representation are neither represented nor have to

Proceedings of the NAACL-HLT 2007 Doctoral Consortium, pages 37–40,
Rochester, April 2007.

be processed, rendering dimensionality reduction techniques unnecessary while still retaining the exact information.

1.3 Roadmap

For the entirety of this research, nothing more is required as input data than plain, tokenized text, separated into sentences. This is surely quite a bit of knowledge that is provided to the system, but unsupervised word boundary and sentence boundary detection is left for future work. Three steps are undertaken to identify similar words on different levels of homogeneity: same language, same part-of-speech, or same distributional properties. Figure 1 shows a coarse overview of the processing steps discussed in this work.

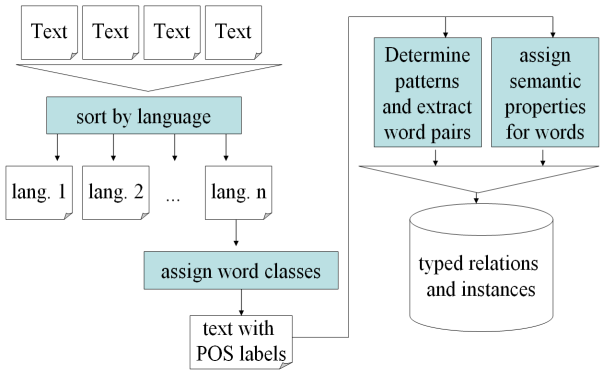


Figure 1: Coarse overview: From multilingual input to typed relations and instances

2 Methods in Unsupervised Processing

Having at hand neither explicit nor implicit knowledge, but in turn the goal of identifying structure of equivalent function, the only possibility that is left in unsupervised and knowledge-free processing is statistics and clustering.

2.1 Co-occurrence Statistics

As a building block, co-occurrence statistics are used in several components of the system described here. A significance measure for co-occurrence is a means to distinguish between observations that are there by chance and effects that take place due to an underlying structure. Throughout, the likelihood ratio (Dunning, 1993) is used as significance measure because of its stable performance in various evaluations, yet many more measures are possible. Dependent on the context range in co-occurrence calculation, they will be called sentence-based or neighbor-based co-occurrences in the remainder of this paper. The entirety of all co-occurrences of a corpus is called its co-occurrence graph. Edges are weighted by co-occurrence significance; often a threshold on edge weight is applied.

2.2 Graph Clustering

For clustering graphs, a plethora of algorithms exist that are motivated from a graph-theoretic viewpoint, but often optimize NP-complete measures (cf. Šíma and Schaeffer, 2005), making them non-applicable to lexical data that is naturally represented in graphs with millions of vertices. In Biemann and Teresniak (2005) and more detailed in Biemann (2006a), the Chinese Whispers (CW) Graph Clustering algorithm is described, which is a randomized algorithm with edge-linear run-time. The core idea is that vertices retain class labels which are inherited along the edges: In an update step, a vertex gets assigned the predominant label in its neighborhood. For initialization, all vertices get different labels, and after a handful of update steps per vertex, almost no changes in the labeling are observed – especially small world graphs converge fast. CW can be viewed as a more efficient modification and simplification of Markov Chain Clustering (van Dongen, 2000), which requires full matrix multiplications.

CW is parameter-free, non-deterministic and finds the number of clusters automatically – a feature that is welcome in NLP, where the number of desired clusters (e.g. in word sense induction) is often unknown.

3 Results

3.1 Language Separation

Clustering the sentence-based co-occurrence graph of a multilingual corpus with CW, a language separator with almost perfect performance is implemented in the following way: The clusters represent languages; a sentence gets assigned the label of the cluster with the highest lexical overlap between sentence and cluster. The method is evaluated in (Biemann and Teresniak, 2005) by sorting monolingual material that has been artificially mixed together. Dependent on similarities of languages, the method works almost error-free from about 100-1,000 sentences per language on. For

languages with different encoding, it is possible to un-mix corpora of size factors up to 10,000 for the monolingual parts.

In a nutshell, comparable scores to supervised language identifiers are reached without training. Notice that the number of languages in a multilingual chunk of text is unknown. This prohibits any clustering method that needs the number of clusters to be specified be-forehand.

3.2 Unsupervised POS Tagging

Unlike in standard POS tagging, there is neither a set of predefined categories, nor annotation in a text. As POS tagging is not a system for its own sake, but serves as a preprocessing step for systems building upon it, the names and the number of categories are very often not important.

The system presented in Biemann (2006b) uses CW clustering on graphs constructed by distributional similarity to induce a lexicon of supposedly non-ambiguous words w.r.t. POS by selecting only safe bets and excluding questionable cases from the lexicon. In this implementation, two clusterings are combined, one for high and medium frequency words, the other collecting medium and low frequency words. High and medium frequency words are clustered by similarity of their stop word context feature vectors: a graph is built, including only words that are involved in highly similar pairs. Clustering this graph of typically 5,000 vertices results in several hundred clusters, which are further used as POS categories. To extend the lexicon, words of medium and low frequency are clustered using a graph that encodes similarity of neighbor-based co-occurrences. Both clusterings are mapped by overlapping elements into a lexicon that provides POS information for some 50,000 words. For obtaining a clustering on datasets of this size, an effective algorithm like CW is crucial. Using this lexicon, a trigram tagger with a morphological extension is trained, which assigns a tag to every token in the corpus.

The tagsets obtained with this method are usually more fine-grained than standard tagsets and reflect syntactic as well as semantic similarity. Figure 2 demonstrates the domain-dependence on the tagset for MEDLINE: distinguishing e.g. illnesses and error probabilities already in the tagset might be a valuable feature for relation extraction tasks.

Size	Sample words
1613	colds, apnea, aspergilloma, ACS, breathlessness, lesions, perforations, ...
1383	proven, supplied, engineered, distinguished, constrained, omitted, …
589	dually, circumferentially, chronically, rarely, spectrally, satisfactorily, ...
124	1-min, two-week, 4-min, 2-day, …
6	P<0.001, P<0.01, p<0.001, p<0.01, ...

Figure 2: Some examples for MEDLINE tagset: Number of lex. entries per tag and sample words.

In Biemann (2006b), the tagger output was directly compared to supervised taggers for English, German and Finnish via information-theoretic measures. While it is possible to compare the contribution of different components of a system relatively along this scale, it only gives a poor impression on the utility of the unsupervised tagger's output. Therefore, the tagger was evaluated indirectly in machine learning tasks, where POS tags are used as features. Biemann et al. (2007) report that for standard Named Entity Recognition, Word Sense Disambiguation and Chunking tasks, using unsupervised POS tags as features helps about as much as supervised tagging: Overall, almost no significant differences between results could be observed, supporting the initial claim.

3.3 Word Sense Induction (WSI)

Co-occurrences are a widely used data source for WSI. The methodology of Dorow and Widdows (2003) was adopted: for the focus word, obtain its graph neighborhood (all vertices that are connected via edges to the focus word vertex and edges between these). Clustering this graph with CW and regarding clusters as senses, this method yields comparable results to Bordag (2006), tested using the unsupervised evaluation framework presented there. More detailed results are reported in Biemann (2006a).

4 Further Work

4.1 Word Sense Disambiguation (WSD)

The encouraging results in WSI enable support in automatic WSD systems. As described by Agirre et al. (2006), better performance can be expected if the WSI component distinguishes between a large number of so-called micro-senses. This illustrates a

principle of unsupervised NLP: It is not important to reproduce word senses found by introspection; rather, it is important that different usages of a word can be reliably distinguished, even if the corresponding WordNet sense is split into several subsenses.

4.2 Distributional Thesaurus with Relations

It is well understood that distributional similarity reflects semantic similarity and can be used to automatically construct a distributional thesaurus for frequent words (Lin, 1997; inter al). Until now, most works aiming at semantic similarity rely on a parser that extracts dependency relations. The claim here again is that similarity on parser output might be replaced by similarity on a pattern basis, (cf. Davidov and Rappoport 2006). For class-based generalization in these patterns, the system described in section 3.2 might prove useful. Preliminary experiments revealed that similarity on significantly co-occurring patterns is able to produce very promising similarity rankings. A clustering of these with CW leads to thesaurus entries comparable to thesauri like Roget's.

Clustering not only words based on similarity of patterns, but also patterns based on similarity of words enables us to identify clusters of patterns with different relations they manifest.

5 Conclusion

The claim of this work is that unsupervised NLP can support and/or replace preprocessing steps in NLP that have previously been achieved by a large amount of manual work, i.e. annotation, rule construction or resource building. This is proven empirically on the tasks of language identification and part-of-speech tagging, exemplified on WSD and discussed for thesaurus construction and relation extraction. The main contributions of the dissertation that is summarized here are:

- A framework for unsupervised NLP
- An efficient graph clustering algorithm
- An unsupervised language separator
- An unsupervised POS tagger

The main advantage of unsupervised NLP, namely language independence, will enable the immediate processing of all languages and domains for which a large amount of text is electronically available.

References

E. Agirre, D. Martínez, O. López de Lacalle and A. Soroa. 2006. *Evaluating and optimizing the parameters of an unsupervised graph-based WSD algorithm*. Proceedings of Textgraphs-06, New York, USA

C. Biemann and S. Teresniak. 2005. *Disentangling from Babylonian Confusion – Unsupervised Language Identification*. Proc. CICLing-2005, Mexico City

C. Biemann. 2006a. *Chinese Whispers - an Efficient Graph Clustering Algorithm and its Application to Natural Language Processing Problems*. Proceedings of Textgraphs-06, New York, USA

C. Biemann. 2006b. *Unsupervised Part-of-Speech Tagging Employing Efficient Graph Clustering*. Proceedings of COLING/ACL-06 SRW, Sydney, Australia

C. Biemann. 2007. *A Random Text Model for the Generation of Statistical Language Invariants*. Proceedings of HLT-NAACL-07, Rochester, USA

C. Biemann, C. Giuliano and A. Gliozzo. 2007. *Unsupervised POS tagging supporting supervised methods*. Manuscript to appear

S. Bordag. 2006. *Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation*. Proceedings of EACL-06. Trento, Italy

D. Davidov, A. Rappoport. 2006. *Efficient Unsupervised Discovery of Word Categories Using Symmetric Patterns and High Frequency Words*. Proceedings of COLING/ACL-06, Sydney, Australia

S. van Dongen. 2000. *A cluster algorithm for graphs*. Technical Report INS-R0010, CWI, Amsterdam.

B. Dorow and D. Widdows. 2003. *Discovering Corpus-Specific Word Senses*. In EACL-2003 Conference Companion, Budapest, Hungary

T. Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, 19(1), pp. 61-74

D. Lin. 1997. *Automatic Retrieval and Clustering of Similar Words*. Proc. COLING-97, Montreal, Canada

H. Schütze. 1993. *Word Space*. Proceedings of NIPS-5, Morgan Kaufmann, San Francisco, CA, USA

J. Šíma and S.E. Schaeffer. 2005. *On the NP-completeness of some graph cluster measures*. Technical Report cs.CC/0506100, http://arxiv.org/.

M. Steyvers, J. B. Tenenbaum. 2005. The large-scale structure of semantic networks. *Cog. Science*, 29(1)

D. Widdows. 2005. *Geometry and Meaning*. CSLI Lecture notes #172, Stanford, USA

Author Index