An Unsupervised Semantic Tagger Applied to German

Paul Buitelaar, Jan Alexandersson, Tilman Jaeger, Stephan Lesch, Norbert Pfleger, Diana Raileanu

DFKI GmbH
Stuhlsatzenhausweg 3
D-66123 Saarbruecken, Germany
{paulb,janal,jaeger,lesch,pfleger,raileanu}@dfki.de

Tanja von den Berg, Kerstin Klöckner, Holger Neis, Hubert Schlarb

Department of Computational Linguistics,
Universität des Saarlandes
Postfach 151150
66041 Saarbrücken, Germany
{kekl,hone,husc}@coli.uni-sb.de

Abstract

We describe an unsupervised semantic tagger, applied to German, but which could be used with any language for which a corresponding "XNet" (WordNet, GermaNet, etc.), POS tagger and morphological analyzer are available. Disambiguation is performed by comparing co-occurrence weights on pairs of semantic classes (synsets from GermaNet). Precision is around 67% at a recall of around 65% (for all *ambiguous* words -- 81% for all words at a recall of 80%). Our results show the influence of context size and of semantic class frequency in the training corpus.

1 Introduction

Natural language applications such as (concept-based) information retrieval, information extraction and machine translation requires a certain level of semantic analysis. An important part of this process is *semantic tagging*: the annotation of each content word with a semantic category. This allows for semantic normalization of different lexical realizations (words) for the same concept (semantic class). However, as words often correspond to more than one concept, or *sense*, the semantic tagger needs to disambiguate between these different senses.

Word sense disambiguation has a rich tradition in natural language processing, originating with the use of hand crafted knowledge bases in the 70`s and early 80`s, (e.g. Small, 1980; Hirst, 1988), followed by the advent of machine readable dictionaries in the late 80`s, (e.g. Lesk, 1986), and down to the current, much more robust methods that use WordNet (Miller, 1995) and similar resources in combination with corpora (Yarowsky, 1992; Agirre and Rigau 1996; Ng and Lee, 1996; Resnik, 1997). Although a lot of work on corpus based word sense disambiguation has been reported in recent years (for an overview, see: Ide and Veronis, 1998; Kilgarriff and Palmer, 2000), most of these approaches use supervised training over manually annotated, English corpora like SEMCOR (Fellbaum, 1997) and DSO (Ng and Lee 1996). Supervised semantic taggers therefore are dependent on manually annotated corpora for every new application. Obviously this is not feasible given the high cost of manual annotation. Also, porting the system to a different language depends on the availability of semantically annotated corpora for this language. Given these restrictions, we chose to implement an unsupervised semantic tagger that builds on the use of co-occurrence information between words and/or semantic classes (as used e.g. by Yarowsky 1992; Agrirre and Rigau 1996; Resnik, 1997 in word sense disambiguation and by Seligman et al., 1999 in speech understanding) both in training and disambiguation. The basic idea is to collect co-occurrences using a thesaurus. By smoothing (using more abstract semantic classes higher up in the hierarchy), a co-occurrence probability can be computed although no data has been collected between the particular words in question. By generalizing this over co-occurrences between semantic classes within a text window, we have developed a robust semantic tagging system that does neither depend on a semantically nor a syntactically annotated corpus, but only needs raw text of any kind as a training corpus and semantic classes from a lexical semantic resource like WordNet or GermaNet to derive a statistical model of class co-occurrence. In disambiguation, this statistical model is then used to decide which sense (semantic class) is most likely, given its co-occurrence with other semantic classes in its context. In this paper we discuss only experiments with GermaNet on German text, but our system can be trained also on English in combination with WordNet, a POS tagger and morphological analyzer for English. In fact, our system can be trained for any language as long as a corresponding "Xnet," POS tagger and morphological analyzer are available.

The rest of the paper consists of a description of the system in section 2, a detailed account of the results obtained in evaluation in section 3, and an outlook on further research in section 4.

2 System Description

The system has a training and a disambiguation part. In training, raw text (Training Text) is processed to obtain co-occurrence statistics (Co-Oc Statistics) that are used in disambiguation to annotate a text document with semantic tags. There are three exchangeable components: A semantic resource ("XNet"), a Part-of-Speech tagger and a morphological analyzer.

2.1 Preprocessing

In order to acquire co-occurrence information over semantic classes in the training corpus, the system first annotates all words with part-of-speech (to find all content words: noun, verb, adjective) and morphological information (to look up the lemma in GermaNet). For the experiments reported on in this paper we have used the TnT-tagger for German, trained with the Stuttgart-Tübinger POS-tag set (Brants 2000). For morphological analysis, an implementation of the Mmorph algorithm is used, which has been developed within the context of the Multext project (Petitpierre and Russell 1995).

As in Resnik`s approach, our system treats all synsets and their hypernyms as semantic classes to which a word may belong. Each sense plus its hypernyms then form a so called *class path*. Obviously, not all semantic classes will be equally informative. To measure this effect, we counted all occurrences of each semantic class and cut off either very frequent, or infrequent classes for further processing. At the same time this has the added positive effect that taking less classes into account reduces combinatorial complexity in training.

2.2 Training

During the training phase, weights are computed for each pair of semantic classes that co-occur within a certain text window. For instance, consider the following text window:

die **Arbeit** wird nie **gemacht**
(the work never gets done)

For this window, co-occurrence weights will be computed between the 20 semantic classes (covering 4 senses) that `Arbeit` belongs to and the 10 semantic classes (covering 6 senses) that the verb `machen (to make)` belongs to.

The training corpus is processed by moving a window over a sequence of segments into which the corpus is divided. Each segment consists of *n* relevant words (content words for which a sense definition in GermaNet exists) and the words in-between. For instance, consider the following fragment from the training corpus of newspaper text from the *Frankfurter Rundschau* that we used in our experiments:

Landesbank schlägt Verträge zwischen Stadt und privaten Investoren vor Überall wird gebuddelt und gemauert. Hamburg erlebt den größten Geschäftsbau-Boom. Jährlich kommen rund 300.000 Quadratmeter an Büroräumen hinzu.

(Landesbank proposes contracts between city and private investors Everywhere there is digging and building. Hamburg experiences the biggest office building boom. Every year 300.000 square meters of office space are added.)

If $n=3$, the segment includes 3 relevant words and all words in between:

- Landesbank **schlägt** Verträge zwischen **Stadt** und **privaten**
- Investoren vor Überall wird **gebuddelt** und **gemauert**. Hamburg **erlebt**
- den **größten** Geschäftsbau-Boom. Jährlich **hinzukommen**[1] rund 300.000 **Quadratmeter**

If $n=0$, we define that the segment includes all words between two sentence boundaries:

- Landesbank **schlägt** Verträge zwischen **Stadt** und **privaten** Investoren vor Überall wird **gebuddelt** und **gemauert**.
- Hamburg **erlebt** den **größten** Geschäftsbau-Boom.
- Jährlich **hinzukommen** rund 300.000 **Quadratmeter** an Büroräumen.

Unfortunately, newspaper text includes headlines that are not closed off by punctuation markers. Therefore, segmentation in sentences is not always successful as the first segment shows. Instead of one sentence, this segment concatenates the headline with the first sentence of the article, which obviously will influence training results. At the same time, this example is an indication of the kind of problems to expect when dealing with raw text of any possible kind as we advocate in our approach.

In training, two weights are computed for each co-occurring pair of semantic classes in the training corpus (weights are based on those used in Resnik, 1997):

- a *conditional probability* P($c|c'$) on the occurrence of c, given the co-occurrence of c' in the context of c. Context is defined by a segmentation in windows as discussed above.

$$P(c \mid c') = \frac{P(c,c')}{P(c')}$$

- a *mutual information* score between c and c' based on the *conditional probability*

$$MI(c,c') = \log_2 \frac{P(c \mid c')}{P(c)}$$

2.3 Disambiguation

The semantic tagger proceeds by moving a window of segments over the text that is to be semantically annotated. Each time, the middle segment is annotated, moving from one relevant[2] word to the next. At the end of the segment, the window is moved one segment ahead in the text.

In the case of an ambiguous word w in context C, the most likely sense s_{max} is determined by computing for each sense s of word w its sense weight $sw(s)$ on basis of the co-occurrence weights that were computed in training. The sense s in S with the highest average score (s_{max}) is taken as most likely in the particular context. If no co-occurrences for the (semantic classes of the) word were computed in training, none of the senses is selected and the word does not receive a semantic tag.

$$S \quad : \; set\ of\ all\ senses\ of\ word\ w$$

$$s_{max} = \underset{S}{argmax}\ sw(s)$$

$sw(s)$ is the average over the sum of all class weights $cw(c)$ for each semantic class (c) on the class path of sense s:

$$sw(s) = |CP_s|^{-1} \sum_{c \in CPs} cw(c)$$

$cw(c)$ is the average over the sum of multiplied co-occurrence weights -- $MI(c,c')\ P(c|c')$ -- of pairs (c,c') with c in CPs and c' in K:

$$CP_s : \; set\ of\ all\ classes\ c\ on\ the\ class\ path\ of\ sense\ s$$

$$K \quad : \; set\ of\ all\ classes\ c\ in\ context\ C$$

$$cw(c) = |K|^{-1} \sum_{c' \in K} MI(c,c')\, P(c|c')$$

3 Evaluation

Disambiguation results are compared to a manually annotated evaluation corpus in order to determine precision and recall performance of the approach.

[1] Verbs with verb particles are concatenated in a preprocessing step to be lemmatized correctly.

[2] Non-content words and those content words that are not in GermaNet are not semantically tagged.

3.1 Evaluation Corpus: NEGRA-LexSem

The evaluation corpus consists of 604 sentences from the *Frankfurter Rundschau* corpus (as collected for the NEGRA project: Brants and Skut, 1998). All content words in this corpus (NEGRA-LexSem) have been manually annotated with GermaNet synsets by two annotators. Differences in annotation were solved by arbitration.

Annotators were given the option of choosing more than one sense if they were not able to distinguish between them. The reason for this is twofold. First, all semantic resources (WordNet, GermaNet, as well as most dictionaries) have sense distinctions that are too fine grained for practical use. Although lexicographers can distinguish one sense from the other, an average language user, let alone an automatic system, cannot (see Kilgarriff, 1997 for a critical overview of this topic). Secondly and in connection to this, some senses may be systematically related and should therefore not be "separated." Instead, the different senses of such *systematic polysemous* words should be left *underspecified* (Buitelaar 1998). For example, take the 6 GermaNet senses of the noun `Geschichte:`

Sense1 Geschichte, Vergangenheit (past, past times, yesteryear)
Sense2 Geschichte, Erzählung (report, account)
Sense3 Geschichte, Story (narration, story, tale, yarn)
Sense4 Geschichte, Geisteswissenschaft (history)
Sense5 Geschichte, Angelegenheit (personal business, affairs)
Sense6 Entwicklungsgeschichte (history)

In the following sentence from the NEGRA-LexSem corpus the annotators were not able to decide between sense 1 and 6:

> ... rechnet Alfredo Joskowics in "Playa Azul" mit der jüngsten Geschichte ab ...
>
> (... deals Alfredo Joskowics in "Playa Azul" with the *recent past* [sense 1] /
> recent history [sense 6]...)

The evaluation corpus covers 8,897 words, of which 1,872 (nouns, verbs and adjectives that are covered by GermaNet) have been manually annotated. Average ambiguity of all annotated words is 3.1 and of all ambiguous words 4.6 1,095 words are ambiguous between two or more senses, of which 303 were annotated with more than one sense.

3.2 Results

We evaluated the semantic tagger in a series of experiments to determine its performance against a theoretical baseline, but also to see how different parameters (window size, class frequency) influence disambiguation accuracy. The training corpus we used for the experiments that are described here consists of 10.000 newspaper sentences (around 1.000.000 tokens) from the *Frankfurter Rundschau* (as collected for the NEGRA and TIGER projects: Brants and Skut, 1998).

In word sense disambiguation for English (using WordNet) often the so-called "most-frequent" baseline is used to compare more sophisticated methods with. This baseline uses sense frequency as obtained from SEMCOR (Miller et al., 1994). Unfortunately, for GermaNet no such large manually annotated corpus is available and therefore also no information on sense frequency is available. We therefore compare our results with a theoretical baseline that is computed in the following way. For each word in the evaluation corpus we compute the probability for assigning (one of) the right sense(s) by chance, by dividing the number of annotated senses with the number of senses in GermaNet. We obtain the average precision by summing the probabilities and dividing this by the number of words in the evaluation corpus.

Then the baseline probability for a random assignment of senses to all annotated words in the evaluation corpus is:

$$P_{rand} = \left|EC\right|^{-1} \sum_{w \in EC} \frac{|AS_w|}{|GS_w|}$$

$$EC\ :\ The\ Evaluation\ Corpus$$

$$AS_w\ :\ The\ Annotated\ Sense\ of\ word\ w\ in\ EC$$

$$GS_w\ :\ The\ GermaNet\ Sense\ of\ word\ w$$

For our evaluation corpus (NEGRA-LexSem) this baseline comes to: $P_{rand} = 45.8\%$ If we ignore the influence of underspecified semantic tags (annotation with more than one sense) in the evaluation corpus by instead assuming single tags (annotation with one sense) then $P_{rand} = 32.6\%$

The influence of window size is shown by the following results[3] of experiments with *s=3* (number of segments) and a varying size of *n* (number of relevant words per segment):

n	**0 (Sentence)**	**2**	**7**	**15**
Rec. (amb)	64.08	53.42	65.45	65.63
Prec. (amb)	66.95	52.84	67.29	66.30
Rec. (all)	78.95	72.70	79.75	79.86
Prec. (all)	80.99	72.23	81.05	80.33

Table 1: Results with varying window size

These results show that a small context window (*n=2*) negatively influences both precision and recall. A larger context window (*n=7*) improves results, but by making it even larger (*n=15*) precision drops although recall slightly increases. Using sentence boundaries (*n=0*) for segmentation gives results that are comparable to those obtained with a (somewhat) larger context window (*n=7*).

Recall is determined by the number of words for which co-occurrence information could be computed during training. As explained in section 2.3, if no co-occurrences are computed for a word then the system cannot disambiguate between senses.

In order to measure the influence of class frequency we conducted the following experiments with *s=3, n=7* and varying lower (*cf_l*) and upper (*cf_u*) thresholds on class frequency. In this way we were able to compare the influence of using only high, low or middle frequency classes relative to using all classes (see also results in table 1).

cf_l-cf_u	**all**	**0-1.000**	**1.000-50.000**	**10.000-50.000**
Rec. (amb.)	65.45	60.07	21.51	1.82
Prec. (amb.)	67.29	68.50	69.82	76.92
Rec. (all)	79.75	76.60	54.01	42.47
Prec. (all)	81.05	82.56	90.84	99.25

Table 2: Results with varying range of classes

The experiments show that overall (precision and recall) the best results are obtained when taking into account all classes. However, they also show that taking into account only the more frequent classes (*cf_l-cf_u=1.000-50.000, cf_l-cf_u=10.000-50.000*) does significantly improve precision.

3.3 Discussion

Since there exist, to our knowledge, no other broad-coverage semantic taggers for German, we had to compare our results with taggers for other languages. Our results are thus best compared to those of the unsupervised systems that competed in the SENSEVAL exercise (Kilgarriff and Rosenzweig 2000). The best systems (University of Sunderland and CL Research) reached around 55% to 65% precision at a recall of 50% to 60%. However, apparently both systems were not strictly unsupervised (Kilgarriff and Rosenzweig 2000) which makes the comparison questionable. Therefore a more adequate comparison would be with the XRCE-CELI system which reaches a precision of about 46% at a recall of roughly 37%.

An additional complicating factor when comparing our results with other systems is our use of underspecified semantic tags in the evaluation corpus. This somewhat simplifies the disambiguation task although only in a (theoretical) range of 10% to 15% as shown by our baseline computation. This means, however, that even considering this aspect our best results are still comparable to those of CL Research and outperform those of XRCE-CELI.

Finally, a meaningful comparison of our results can only be made with other systems that work with German and use GermaNet. Alternatively, our system has to be evaluated under the same conditions as the systems previously mentioned as discussed in future work.

4 Conclusions and Future Work

We presented a generic unsupervised semantic tagger, which uses only raw text as training material. The exchangeable components ("Xnet", POS tagger and morphological analyzer) allow us to use the system for any language for which such modules are available.

Future work will be in the following directions:

[3] Results in this table are based on using *all* classes.

- Continue to test our system with other parameter settings
- Apply our system to other languages, especially English and using WordNet in the context of SENSEVAL-II in order to make a direct comparison possible with other systems
- Investigate the effects of using linguistically motivated segmentation methods
- Investigate using the system for prediction purposes

5 Acknowledgements

Thanks to Thorsten Brants and Oliver Plaehn for their contributions in producing the NEGRA-LexSem evaluation corpus and to Sebastian Thull, Olaf Holz and Stephanie Becker for their annotation work. Also, thanks to Detlef Prescher for help in defining some of the formulas, and Tilman Becker for comments on earlier versions. This research has in part been supported by EC/NSF grant *IST-1999-11438* for the MUCHMORE project.

References

Agirre E., and Rigau G. 1996. *Word sense disambiguation using conceptual density*. In Proceedings of COLING'96, pages 16--22, Copenhagen, Denmark.

Brants, T. 2000. *TnT - A Statistical Part-of-Speech Tagger*. In: Proceedings of the 6th Applied Natural Language Processing Conference, Seattle, WA.

Brants, T., and W. Skut. 1998. *Automation of Treebank Annotation*. In: Proceedings of the Conference on New Methods in Language Processing (NeMLaP-3), Australia.

Buitelaar, P. 1998. *CoreLex: Systematic Polysemy and Underspecification.* PhD Dissertation , Brandeis University.

Fellbaum Chr. 1997. *Analysis of a hand-tagging task*. Proceedings of ANLP-97 Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington D.C., USA.

Hirst, G. 1988. *Semantic Interpretation and the Resolution of Ambiguity*. Cambridge University Press.

Ide, N., & Véronis, J. (Eds.). 1998. *Word Sense Disambiguation*. Special issue of Computational Linguistics, 24(1).

Kilgarriff, A. 1997. *I don't believe in word senses*. Computers and the Humanities 31 (2), pp 91--113.

Kilgarriff, A., and M. Palmer. 2000. *Introduction to the special issue on SENSEVAL*. Computers and the Humanities 34(1/2):1-13.

Kilgarriff, A. and Rosenzweig J. 2000. *English SENSEVAL: Report and Results*. In: Proceedings of LREC2000, Athens, Greece.

Lesk, M.E. 1986. *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cone*. In Proceedings of the SIGDOC Conference.

Miller G., Chodorow M., Landes S., Leacock C., Thomas R. 1994 *Using a Semantic Concordance for Sense Identification*. In: ARPA Workshop on Human Language Technology, Plainsboro NJ.

Miller, G.A. 1995. *WordNet: A Lexical Database for English*. Communications of the ACM 11.

Ng, H.T., and H.B. Lee. 1996. *Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach*. In Proceedings of ACL96.

Dominique Petitpierre and Graham Russell, 1995. *MMORPH - The Multext Morphology Program*. Multext deliverable report for the task 2.3.1, ISSCO, University of Geneva.

Resnik, P. 1997. *Selectional preference and sense disambiguation*. In Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How? Washington, D.C., USA.

Seligman, M., Alexandersson J. and Jokinen K. 1999. *Tracking Morphological and Semantic Co-occurrences in Spontaneous Dialogues* In: Proceedings of the IJCAI Workshop Knowledge and Reasoning in Practical Dialogue Systems, Stockholm, Sweden.

Small, S.L. 1980. *Word Expert Parsing: A Theory of Distributed Word-based Natural Language Understanding*. Ph.D. thesis, The University of Maryland, Baltimore, MD.

Yarowsky, D. 1992. *Word-sense disambiguation using statistical models of Roget's categories*. In Proceedings of COLING-92, Nantes, France.