Using Nominal Compounds for Word Sense Discrimination

Yannick Versley
University of Tübingen
Department of Linguistics
versley@sfs.uni-tuebingen.de

Verena Henrich
University of Tübingen
Department of Linguistics
verena.henrich@uni-tuebingen.de

Abstract

In many morphologically rich languages, conceptually independent morphemes are glued together to form a new word (a compound) with a meaning that is often at least in part predictable from the meanings of the contributing morphemes. Assuming that most compounds express a subconcept of exactly one sense of its nominal head, we use compounds as a higher-quality alternative to simply using general second-order collocate terms in the task of word sense discrimination. We evaluate our approach using lexical entries from the German wordnet GermaNet (Henrich and Hinrichs, 2010).

1 Introduction

In several morphologically rich languages such as German and Dutch, compounds are usually written as one word: In a process where nouns, verbs and other prefixes combine with a head noun (called the *simplex* when it occurs on its own), a novel word can be formed which is typically interpretable by considering its parts and the means of combination. The process of compounding is both highly productive and subject to lexicalization (i.e., the creation of non-transparent compounds that can only be interpreted as a whole rather than as a combination of parts). The analysis of compounds have been subject to interest in machine translation as well as in the semantic processing of morphologically rich languages. The analysis of compounds is generally challenging for many reasons. In particular, compounds leave us with the dilemma of either modeling them as complete units, yielding a more accurate picture for lexicalized compounds but creating a more severe sparse data problem in general, or trying to separate out their parts and ending up with problems of wrongly split lexicalized compounds, or of incurring mis-splits where spurious ambiguities occur.

The purpose of this paper is to address the question of whether semantic information of compound occurrences can be used to learn more about the sense distribution of the simplex head, with respect to a text collection. Specifically, this paper focuses on the task of *word sense discrimination*, where the goal is to find different senses of a word without assuming a hand-crafted lexical resource as training material (in contrast to word sense *disambiguation*, where the exact sense inventory to be tagged is known at training and inference time, and where making effective use of a resource such as WordNet (Miller and Fellbaum, 1991) or GermaNet (Henrich and Hinrichs, 2010) is an important part of the problem to be solved).

While the present paper focuses on nominal compounds in German, the method as such can also be applied to other languages where compounds are written as one word.

2 Related Work

Automatic word sense discrimination (WSD) is a task that consists of the automatic discovery of a sense inventory for a word and of associated examples for each sense.

To evaluate systems performing word sense discrimination, earlier research such as Schütze (1998)

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 36–41,
Jeju, Republic of Korea, 12 July 2012.

uses either *pseudowords* – two words that have been artificially conflated to yield an ambiguous concept such as *wide range/consulting firm* – or use (expensive) manually annotated data. Subsequently, the contexts of these occurrences are clustered into groups that correspond to training examples for each postulated sense.

A different approach to the idea of word sense discrimination can be found in the work of Pantel and Lin (2002): they retrieve a set of most-similar items to the target word, and then cluster these similar items according to distributional semantic properties. In Pantel and Lin's approach, the output of the word sense induction algorithm is not a group of contexts with the target word that will be used to represent a sense, but instead one or more words that are (hopefully) related to one particular sense. The contexts in which the related words occur could then be used as positive examples for that particular sense of the target word.

Pantel and Lin aim at a principled approach to compare the soft-clustering approaches they propose, in conjunction with a fixed set of related words. While the main interest of this paper lies in comparing different methods for generating the candidate set of related words, the exact clustering method is only of marginal interest. In this paper, a simpler hard clustering method is used and only the assignment for the tight center of a cluster is considered since the non-central items can be different or even incomparable for the different methods.

3 Our Approach

Our approach to word sense discrimination is based on the idea that different compounds that have the same simplex word as their head (e.g. *Blütenblatt* 'petal', and *Revolverblatt* 'tabloid rag') are less ambiguous than the simplex (*Blatt* 'leaf', 'newspaper') itself. This assumption is along the lines with what the "one sense per collocation" heuristic of Yarowsky (1993) would predict.

Yarowsky noted that in a corpus of homographs/homophones/near-homographs, translation distinctions, and pseudo-words, a single collocation (such as "foreign" or "presidential") is often enough to disambiguate the occurrence of a near-homograph such as aid/aide. While Yarowsky claims that most of the problems of such an approach would be due to absent or unseen collocates, it is easily imaginable that collocates such as *old* or *big* can occur with multiple senses of a word.

In German, noun compounds usually involve at least a minimum degree of lexicalization: In English, 'red flag' and 'red beet' are lexicalized (i.e., denote something more specific than the compositional interpretation would suggest), but 'red rag' or 'red box' are purely compositional. In German, *Rotwein* 'red wine' is a compound, but the more compositional *roter Apfel/*Rotapfel* 'red apple' is not a compound and points to the fact that 'red apple' only has a compositional interpretation. Because of this minimal required degree of lexicalization, we would expect that German nominal compounds (as well as any compounds in a language that has a similar distinction between affixating and non-affixating compounds) are, for the largest portion, compositional enough to be interpretable, but lexicalized enough that a compound is always specific to only one sense of its head simplex.

3.1 Finding Committees

The method of finding committees that form sense clusters is illustrated in Figure 1 using the target word example *Blatt*. To generate a candidate list of related terms, our method first retrieves all words (compounds) that have the target word as a suffix (step 3 in Figure 1). This candidate set is then sorted according to distributional similarity with the target word and cut off after N items (step 2 in Figure 1) to reduce the influence of spurious matches and non-taxonomic compounds and to avoid too much noise in the candidate list.

In order to evaluate the method of selecting compounds as candidate words, we first cluster the set of candidate words into as many clusters as there are target word senses represented in the candidate words (step 3 in Figure 1, again using the distributional similarity vectors of the words described in the following subsection 3.2).

To avoid biasing our method towards any particular method of choosing the candidate words, we simply assume that it is possible to produce a 'reasonable' number of clusters. In the next step, the most central items of each cluster (the 'committee') are determined, purely by closeness to the cluster's

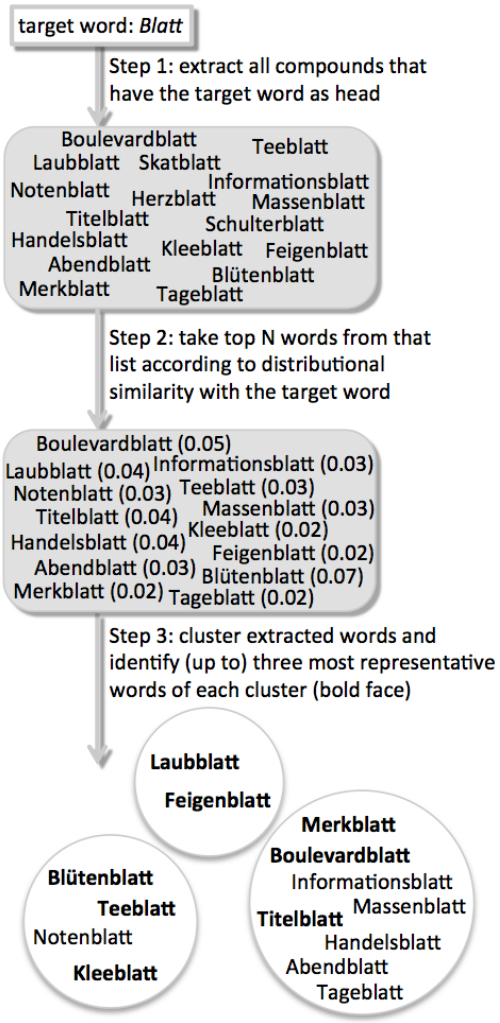

Figure 1: Steps in the clustering method

centroid and disregarding similarity with the target word. The committee words are rendered in bold face in the circles in Figure 1. The quality of the approach is then evaluated according to whether the committees form a suitable representation for the set of senses that the target word possesses.

An advantage of only including compounds in the candidate list of related terms, instead of all words, is that the related words generated by such an approach are conceptually considerably closer to the target word than those using all words as candidates: Using all words, the top candidates include the coordinate terms *Frucht* 'fruit' and *Blüte* 'flower', as well as more faraway terms such as *Tuch* 'cloth' or *Haar* 'hair'; using only compounds of the simplex, the candidate list contains mostly hyponyms such as *Laubblatt* 'leaf', *Titelblatt* 'title page' or *Notenblatt* 'sheet of music'.

3.2 Distributional Similarity and Clustering

Both for the initial selection of candidate words (where the list is cut off after the top-N similar terms) and for the subsequent clustering step, frequency profiles from a large corpus are used to create a semantic vector from the target word and each (potential or actual) candidate word.

To construct these frequency profiles, the *webnews* corpus of Versley and Panchenko (2012) is used, which contains 1.7 billion words of text from various German online newspapers. The text is parsed using MALTParser (Hall et al., 2006) and the frequency of collocates with the `ATTR` (premodifying adjective) and `OBJA` (accusative object) relations is recorded. Vectors are weighted using the conservative pointwise mutual information estimate from Pantel and Lin (2002). For selecting most-similar words in candidate selection, we use a kernel based on the Jensen-Shannon Divergence across both grammatical relations, similar to the method proposed by Ó Séaghdha and Copestake (2008).

The resulting vector representations of words are then clustered using average-link hierarchical agglomerative clustering using the CLUTO toolkit (Zhao and Karypis, 2005), which uses cosine similarity to assess the similarity of two vectors. In the study of Pantel and Lin (2002), agglomerative clustering was among the best-performing off-the-shelf clustering methods.

As we initially found that many features that were used in clustering were less relevant to the different senses of the head word that were targeted, we also introduce a method to enforce a focus on target word compatible aspects of those words. In the basic approach (**raw**), the normal vector representation of each word is used. In the modified approach (**intersect**), only the features that are relevant for the target word are selected, by using for each feature the minimum value of (i) that feature's value in the candidate word's vector and (ii) that feature's value in the target word's vector.

3.3 Competing and Upper Baselines

To see how well our method performs in relation to other approaches for finding related terms describ-

ing each sense of a synset, two lower baselines and one upper baseline have been implemented.

One lower baseline uses general distributionally similar items. This is an intelligent (but realistic) **general** baseline method – close in spirit to Pantel and Lin (2002). It simply consists in retrieving the distributionally most similar words for the clustering task. Effectively, this resembles our own method, but without the compound filtering step.

The second lower baseline assumes that it should always be possible to find one word that is related to one of the senses (yielding poor coverage but trivializing the clustering problem). This trivial baseline is called **one-cluster**.

The upper baseline (called **profile**) assumes that it knows which senses of the word should be modeled and that errors can only be introduced by the clustering step not reproducing the original sense. This baseline retrieves the synsets corresponding to each sense of the word from GermaNet, and, among the terms in the neighbouring synsets (synonyms, hypernyms, hyponyms as well as sibling synsets), select those that are both unambiguous (i.e., do not have other synsets corresponding to that word) and are distributionally most similar to the (ambiguous) target word.

4 Evaluation Framework

Our evaluation framework is based on retrieving a set of words related to the target item (the candidate set), and then using collocate vectors extracted from a corpus to cluster the candidate set into multiple subsets.

Once we have a clustering of the generated terms, we want a **quantitative evaluation** of the clustering. The underlying idea for this is that we would like to have, for each sense of the target word, a cluster that has one or several words describing it. (We should not assume that it is always possible to find many related words for a particular sense).

4.1 Evaluation Data

As target items, we used a list of simplexes that are most productive in terms of compounding, using a set of gold-standard compound splits that were created by Henrich and Hinrichs (2011); candidate words (both compounds and general neighbours)

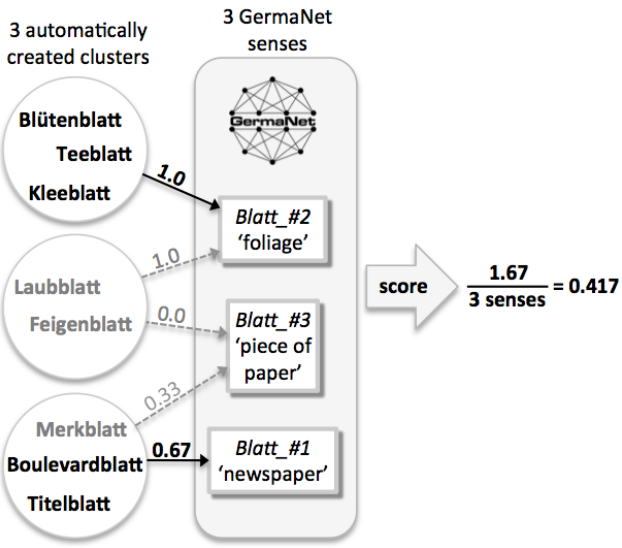

Figure 2: Evaluation procedure for the committees of related words

were selected using a frequency list extracted from the TüPP-D/Z corpus (Müller, 2004). For the experiments themselves, no information about correct splits of the compounds was assumed and potential compounds were simply retrieved as lemma forms that have the target word as a suffix.

The subsets from clustering the candidate set are then evaluated according to whether the most-central related words in that cluster are related to the same sense of the target word, and how many senses of the target word are covered by the clusters.

4.2 Evaluation Metric

Given the committee lists that are output by the candidate selection and output, we calculate an evaluation score by creating a mapping between senses of the target word and the committees that are the output of the clustering algorithm, choosing that mapping according to a quality measure that describes how well the committee members match that synset (the *precision* of that possible pairing between a committee and a sense of the target word), as shown in figure 2. Each candidate word is assigned a sense of the target word, either because it is a hyponym of that sense (for the compound-based method) or because its path distance in GermaNet's taxonomy is less than four (for the general terms method). If a candidate word is not near any of the target word's sense synsets, it is assigned no sense (and always

candidates	num	vectors	score	quality	coverage
compound	5	intersect	0.399	0.882	0.468
compound	30	intersect	0.489	0.721	0.702
compound	100	intersect	0.419	0.586	0.769
general	5	intersect	0.433	0.882	0.510
general	30	intersect	0.528	0.696	0.784
general	100	intersect	**0.573**	0.650	**0.896**
compound	5	raw	0.406	0.898	0.468
compound	30	raw	0.479	0.712	0.702
compound	100	raw	0.422	0.591	0.769
general	5	raw	0.441	**0.902**	0.510
general	30	raw	0.526	0.694	0.784
general	100	raw	0.551	0.630	0.896
profile	10n	intersect	0.737	0.781	0.945
profile	10n	raw	0.753	**0.801**	**0.946**
one-cluster	1	—	0.325	1.000	0.325

Table 1: Evaluation scores for the different methods and baselines

counted wrong).[1]

Given a committee C of these (at most) k most-central candidate words in a cluster, we can calculate a measure $P(C, s) = \frac{|w \in C : \text{sense}(w) = s|}{|C|}$ that describes how well this cluster corresponds to a given sense. (Ideally, the committee would contain words only related to one sense).

Using the Kuhn-Munkres algorithm (Kuhn, 1955), we compute a mapping between each represented synset s and a cluster C_s such that $\sum_s P(C_s, s)$ is maximized. The final score for one target word is this sum divided by the total number of synsets for the target word – this means that a method that yields a less representative set of candidate words will normally not get a better score, unless the clusters are of higher enough quality, than one that has candidate terms for each cluster.

In addition to the *score* metric, we calculated a *quality* metric that divides the raw sum by the number of senses covered in the candidate set, and a *coverage* metric that corresponds to the fraction of senses covered by the candidate set in the first place (see Table 1).

5 Results and Discussion

Table 1 contains quantitative results for the different methods and also evaluation statistics for some lower and upper baselines: Selecting exactly one related word as a candidate (and putting it in a cluster of its own) would yield a quality of 1.0, since that cluster is related to exactly one synset, but a very poor coverage of 0.325. For the *profile* upper baseline, which takes related terms from GermaNet and uses imperfect information only in clustering, we see that our clustering approach is able to reconstruct committees of sense-identical terms out of the candidate list fairly well: given related terms for each sense, distributional similarity yields fairly good quality (0.801) and, unsurprisingly, near-perfect coverage for all senses (0.946).

For the actual methods using compounds of a word (*compound* rows in Table 1) or distributionally similar words (*general* rows), we find that the compound-based candidate selection only reaches very limited coverage numbers and furthermore gives the best results with a smaller number of candidate words (30 for compounds versus 100 for general). Whether this effect is due to minority senses being less productive in compounding or whether compounds of the minority senses are not represented in GermaNet is left to be investigated in future work.

6 Conclusion

We used compounds in the selection of candidate words for representing a target word's senses in a word sense discrimination approach. Because compounds are less-frequent overall than the similar-frequency coordinate terms that are retrieved in the general baseline approach, the proposed approach does less well in covering all senses encoded in the gold standard and gets lower results in our evaluation metric. In contrast to previous work by Pantel and Lin, our evaluation approach allows a principled comparison between approaches to generate candidate lemmas in such a task and would be applicable also to other alternative methods to do so.

Acknowledgements We would like to thank Anne Brock, as well as several anonymous reviewers, for helpful comments of an earlier version of this paper. The research in this paper was partially funded by the Deutsche Forschungsgemeinschaft as part of Collaborative Research Centre (SFB) 833.

[1] If a candidate word is not represented in GermaNet at all, it is discarded before the committee-building step, so that all committee words are in fact GermaNet-represented terms.

References

Agirre, E., Aldezabal, I., and Pociello, E. (2006). Lexicalization and multiword expression in the Basque WordNet. In *Proceedings of the first International WordNet Conference*.

Bentivogli, L. and Pianta, E. (2004). Extending WordNet with syntagmatic information. In *Proceedings of the Second Global WordNet Conference (GWC 2004)*.

Hall, J., Nivre, J., and Nilsson, J. (2006). Discriminative classifiers for deterministic dependency parsing. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*.

Henrich, V. and Hinrichs, E. (2010). GernEdiT - the GermaNet editing tool. In *LREC 2010*.

Henrich, V. and Hinrichs, E. (2011). Determining immediate constituents of compounds in GermaNet. In *Proc. International Conference Recent Advances in Language Processing (RANLP 2011)*.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Miller, G. A. and Fellbaum, C. (1991). Semantic networks of English. *Cognition*, 41:197–229.

Müller, F. H. (2004). Stylebook for the Tübingen partially parsed corpus of written German (TüPP-D/Z). Technischer Bericht, Seminar für Sprachwissenschaft, Universität Tübingen.

Ó Séaghdha, D. and Copestake, A. (2008). Semantic classification with distributional kernels. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING 2008)*.

Pantel, P. and Lin, D. (2002). Discovering word senses from text. In *Proceedings of ACM Conference on Knowledge Discovery and Data Mining (KDD-02)*.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Versley, Y. and Panchenko, Y. (2012). Not just bigger: Towards better-quality Web corpora. In *Proceedings of the 7th Web as Corpus Workshop (WAC-7)*.

Yarowsky, D. (1993). One sense per collocation. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop Held at Plainsboro*.

Zhao, Y. and Karypis, G. (2005). Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168.