Rapid Bootstrapping of Word Sense Disambiguation Resources for German

Samuel Broscheit, Anette Frank, Dominic Jehle, Simone Paolo Ponzetto,
Danny Rehl, Anja Summa, Klaus Suttner and **Saskia Vola**
Department of Computational Linguistics
Heidelberg University
lastname@cl.uni-heidelberg.de

Abstract

This paper presents ongoing efforts on developing Word Sense Disambiguation (WSD) resources for the German language, using GermaNet as a basis. We bootstrap two WSD systems for German. (i) We enrich GermaNet with *predominant sense information*, following previous unsupervised methods to acquire predominant senses of words. The acquired predominant sense information is used as a type-based first sense heuristics for token-level WSD. (ii) As an alternative, we adapt a state-of-the-art *knowledge-based WSD system* to the GermaNet lexical resource. We finally investigate the hypothesis of whether the two systems are complementary by combining their output within a voting architecture. The results show that we are able to bootstrap two robust baseline systems for word sense annotation of German words.

1 Introduction

Word Sense Disambiguation (WSD), the task of computationally determining the meanings of words in context (Agirre & Edmonds, 2006; Navigli, 2009), is a well-studied Natural Language Processing (NLP) task. In comparison to other labeling tasks, WSD is highly challenging because of the large amount of different senses that words have in context, and the difficulty of discriminating them, given the fine-grained sense distinctions offered by existing lexical resources.

If one relies on a fixed sense inventory, two main approaches have been proposed in the literature: supervised and knowledge-based methods. Both methods crucially require language resources in the form of wide-coverage semantic lexica or annotated data. While the most successful approaches to WSD are based on supervised machine learning, these require large training sets of sense-tagged data, which are expensive to obtain. Knowledge-based methods minimize the amount of supervision by exploiting graph-based algorithms on structured lexical resources such as WordNet (Fellbaum, 1998). Following the model of WordNet, wordnets have been developed for a wide range of languages (Vossen, 1998; Lemnitzer & Kunze, 2002; Pianta et al., 2002; Atserias et al., 2004; Tufiş et al., 2004, *inter alia*). Moreover, research efforts recently focused on automatically acquiring wide-coverage multilingual lexical resources (de Melo & Weikum, 2009; Mausam et al., 2009; Navigli & Ponzetto, 2010, *inter alia*).

An alternative to supervised and knowledge-based approaches is provided by fully unsupervised methods (Schütze, 1992; 1998; Pedersen & Bruce, 1997, *inter alia*), also known as word sense induction approaches: these merely require large amounts of raw text, but do not deliver well-defined sense clusters, and therefore are more difficult to exploit.

For supervised WSD methods, corpora such as SemCor (Miller et al., 1993) and the ones developed for the SensEval (Mihalcea & Edmonds, 2004) and SemEval (Agirre et al., 2007) competitions represent widely-used training resources. However, in the case of German, the development of supervised WSD systems based on the sense inventory provided by GermaNet (Lemnitzer & Kunze, 2002) is severely hampered by the lack of annotated corpora.

This lack of a sense-annotated corpus implies in turn that no predominant sense information is available for GermaNet senses, in contrast to WordNet, which offers Most Frequent Sense (MFS) information computed from frequency counts over sense annotations in SemCor. As a result, no MFS baseline system can be produced for German data, and

no MFS heuristics, i.e. assigning the predominant sense in case no answer can be computed, is available. To overcome these limitations, we propose to leverage existing proposal for English and exploit them to bootstrap new WSD resources for German. Our contributions are the following:

1. We *enrich GermaNet with predominant sense information* acquired from large web-based corpora, based on previous work on unsupervised predominant sense acquisition for English words. This allows us to automatically label target words in context using the predominant sense as a type-level first-sense heuristics.

2. We adapt a *state-of-the-art knowledge-based WSD system* to tag words in context with GermaNet senses. This system performs an instance-based disambiguation based on contextual information, and allow us to move away from the type-based first-sense heuristics.

3. We explore the hypothesis of whether the word sense annotations generated by these two WSD approaches are complementary, and accordingly experiment with combining their outputs in a *voting architecture*.

The remainder of the paper is structured as follows. Section 2 presents how we adapted previous proposals for finding predominant senses and performing graph-based WSD in English for German. In Section 3 we present a gold standard we created for German WSD and report our experiments and evaluation results. Section 4 concludes the paper.

2 Rapid Bootstrapping of WSD Resources for German

Our approach to develop WSD resources for German is two-fold. We first apply state-of-the-art methods to find predominant senses for English (McCarthy et al., 2004; Lapata & Keller, 2007, Sections 2.1 and 2.2). Both methods are language-independent and require minimal supervision. However, they do not make use of the structured knowledge provided by the GermaNet taxonomy. Accordingly, in Section 2.3 we move on and adapt the state-of-the-art graph-based WSD system of Agirre & Soroa (2009) to use GermaNet as a lexical knowledge base. Finally, we propose to integrate the output of these methods in Section 2.4.

2.1 Using a Thesaurus-based Method

McCarthy et al. (2004) propose an unsupervised method for acquiring the predominant sense of a word from text corpora. Key to their approach is the observation that distributionally similar words of a given target word tend to be sense-related to the sense of the target word. Thus, for a set of distributionally similar words N_w of a target word w, they compute semantic similarity according to some WordNet similarity measure for each pair of senses of w and senses of n_j, for all $n_j \in N_w$. The WordNet-based semantic similarity scores (sss) are weighted by the distributional similarity scores (dss) of the respective neighbors:

$$prevalence\_score(w, s_i) = \sum_{n_j \in N_w} dss(w, n_j) \times \frac{sss(s_i, n_j)}{\sum_{s'_i \in \text{senses}(w)} sss(s'_i, n_j)} \quad (1)$$

$$\text{where } sss(s_i, n_j) = \max_{s_x \in \text{senses}(n_j)} sss(s_i, s_x)$$

Choosing the highest-scoring sense for w yields the predominant sense tailored to the domain of the underlying corpus on which distributional similarity is computed. McCarthy et al. (2004) make use of Lin's (1998) method of constructing a thesaurus of distributionally similar words. Such a thesaurus can be computed on the basis of grammatical relations or word proximity relations from parsed or raw text corpora, respectively. Syntactic relations were extracted with RASP (Briscoe et al., 2006) from 90M words of the BNC (Leech, 1992). WordNet-based semantic similarity was computed using the jcn (Jiang & Conrath, 1997) and $lesk$ (Banerjee & Pedersen, 2003) measures.

The acquired predominant senses were evaluated against SemCor for the different parts of speech, both at the type-level (measuring accuracy of predicting the predominant sense of words within SemCor) and for tokens (measuring the accuracy of using the predominant sense as a first sense heuristics in instance-based sense tagging). For both evaluations, the predominant senses calculated perform well over a random baseline. Compared to the most frequent sense computed from SemCor, the predominant senses score lower. For instance, for nouns using BNC and $lesk$ they report 24.7% random baseline, 48.7% predominant sense and 68.6% MFS accuracy. Verbs, adjectives and adverbs show the same pattern at lower performance levels.

Computing predominant senses for German. To acquire predominant sense information for German using McCarthy et al.'s (2004) method, we first need a large corpus for the computation of distributional similarity. We select the German part of the WaCky corpora (Baroni et al., 2009), deWAC henceforth, a very large corpus of 1.5G words obtained by web crawling, additionally cleaned and enriched with basic linguistic annotations (PoS and lemma). For parsing we selected a subcorpus of sentences (i) that are restricted to sentence length 12 (to ensure good parsing quality) and (ii) that contain target words from GermaNet version 5.1 (nouns, adjectives and verbs). From this subcorpus we randomly selected sample sets for each word for parsing. Parsing was performed using Malt parser (Hall & Nivre, 2008), trained on the German TüBa/DZ corpus. The parser output is post-processed for special constructions (e.g. prepositional phrases, auxiliary constructions), and filtered to reduce dependency triples to semantically relevant word pairs. The computation of distributional similarity follows Lin (1998), whereas semantic similarity is computed using Leacock & Chodorow's (1998) measure, built as an extension of the GermaNet API of Gurevych & Niederlich (2005).

Predominant sense scores are computed according to Equation 1. To determine optimal settings of system parameters, we made use of a held-out development set of 20 words. We obtained the best results for this set using subcorpora for words with 20-200 occurrences in deWAC, a selection of up to 200 sentences per word for dependency extraction, and restriction to 200 nearest neighbors from the set of distributionally similar words for prevalence score computation.

We developed two components providing predominant sense annotations using GermaNet senses: the computed predominant senses are included as an additional annotation layer in the deWAC corpus. Moreover, we extended the GermaNet API of Gurevych & Niederlich (2005) to return predominant senses, which implements a baseline system for predominant sense annotation.

2.2 Using an Information Retrieval Approach

Lapata & Keller (2007) present an information retrieval-based methodology to compute sense predominance which, in contrast to McCarthy et al. (2004), requires no parsed text. Key to their approach is to query an information retrieval system to estimate the degree of association between a word and its sense descriptions as defined by WordNet synsets. That is, predominant senses are automatically discovered by computing for each sense of a target word how often the word co-occurs with the synonyms of that sense. Let w be a target word and SD_{s_i} the sense description for s_i, namely the i-th sense of w. In practice, SD_{s_i} is a set of words $\{w_1 \dots w_n\}$ which are strongly semantically associated with s_i, e.g. its synonyms, and provide a context for sense ranking. The predominant sense is then obtained by selecting the sense description which has the highest co-occurrence score with w:

$$\hat{s} = \underset{s_i \in \text{senses}(w)}{\text{argmax}} \; df(\{w\} \cup SD_{s_i})$$

where df is a document frequency score, i.e. the number of documents that contain w and words from SD_{s_i} (which may or may not be adjacent), as returned from queries to a text search engine. The queries are compiled for all combinations of the target word with each of its synonyms, and the frequencies are combined using different strategies (i.e. sum, average or taking the maximum score).

Computing predominant senses for German. We start with a German polysemous noun, e.g. Grund and collect its senses from GermaNet:

```
nnatGegenstand.3    {Land}
nnatGegenstand.15   {Boden, Gewaesser}
nArtefakt.6305      {Boden, Gefaess}
nMotiv.2            {Motivation,
                     Beweggrund,
                     Veranlassung,
                     Anlass}.
```

We then compile the queries, e.g. in the case of the `nMotiv.2` sense the following queries are created

```
Grund AND Motivation
Grund AND Beweggrund
Grund AND Veranlassung
Grund AND Anlass
```

and submitted to a search engine. The returned document frequencies are then normalized by the document frequency obtained by querying the synonym alone. Finally the senses of a word are ranked according to their normalized frequency, and the one with the highest normalized frequency is taken as the predominant sense.

Lapata & Keller (2007) explore the additional expansion of the sense descriptors with the hypernyms of a given synset. While their results show that models that do not include hypernyms perform

better, we were interested in our work in testing whether this holds also for German, as well as exploring different kinds of contexts. Accordingly, we investigated a variety of extended contexts, where the sense descriptions include synonyms together with: (i) paraphrases, which characterize the meaning of the synset (PARA, e.g. Weg as a 'often not fully developed route, which serves for walking and driving'); (ii) hypernyms (HYPER, à la Lapata & Keller (2007)); (iii) hyponyms (HYPO) (iv) all hyponyms together in a disjunctive clause e.g. "Grund AND (Urgrund OR Boden OR Naturschutzgrund OR Meeresgrund OR Meeresboden)" (HYPOALL). The latter expansion technique is motivated by observing during prototyping that one hyponym alone tends to be too specific, thus introducing sparseness. In order to filter out senses which have sparse counts, we developed a set of heuristic filters:

- LOW FREQUENCY DIFFERENCE (FREQ) filters out *words* whose difference between the relative frequencies of their first two synsets falls below a confidence threshold, thus penalizing vague distinctions between senses.
- LOW DENOMINATOR COUNT (DENOM) removes *synsets* whose denominator count is too low, thus penalizing synsets whose information in the training corpus was too sparse.
- LOW SYNSET INFORMATION COUNT (SYN) filters out *synsets* whose number of synonyms falls under a confidence threshold.

In our implementation, we built the information retrieval system using Lucene[1]. Similarly to the setting from Section 2.1, the system was used to index the deWAC corpus (Baroni et al., 2009). Due the to productivity of German compounds, e.g. "Zigarettenanzündersteckdose" (cigarette lighter power socket), many words cannot be assigned word senses since no corresponding lexical unit can be found in GermaNet. Accordingly, given a compound, we perform a morphological analysis to index and retrieve its lexical head. We use Morphisto (Zielinski et al., 2009), a morphological analyzer for German, based on the SMOR-based SFST tools (Schmid et al., 2004).

2.3 GermaNet-based Personalized PageRank

While supervised methods have been extensively shown in the literature to be the best performing ones for monolingual WSD based on a fixed sense inventory, given the unavailability of sense-tagged data for German we need to resort to minimally supervised methods to acquire predominant senses from unlabeled text. Alternatively, we also experiment with extending an existing knowledge-based WSD system to disambiguate German target words in context.

We start by adapting the WSD system from Agirre & Soroa (2009, UKB)[2], which makes use of a graph-based algorithm, named Personalized PageRank (PPR). This method uses a lexical knowledge base (LKB), e.g. WordNet, in order to rank its vertices to perform disambiguation in context. First, a LKB is viewed as an undirected graph $G = \langle V, E \rangle$ where each vertex $v_i \in V$ represents a concept, e.g. a synset, and each semantic relation between edges, e.g. hypernymy or hyponymy, corresponds to an undirected edge $(v_i, v_j) \in E$. Given an input context $C = \{w_1 \dots w_n\}$, each content word (i.e. noun, verb, adjective or adverb) $w_i \in C$ is inserted in G as a vertex, and linked with directed edges to m associated concepts, i.e. the possible senses of w_i according to the sense inventory of the LKB. Next, the PageRank algorithm (Brin & Page, 1998) is run over the graph G to compute PR, the PageRank score of each concept in the graph given the input context as:

$$PR(v_i) = (1 - d) + d \sum_{j \in \text{deg}(v_i)} \frac{PR(v_j)}{|\text{deg}(v_j)|} \quad (2)$$

where $\text{deg}(v_i)$ is the set of neighbor vertices of vertex v_i, and d is the so-called damping factor (typically set to 0.85). The PageRank score is calculated by iteratively computing Equation 2 for each vertex in the graph, until convergence below a given threshold is achieved, or a fixed number of iterations, i.e. 30 in our case, is executed. While in the standard formulation of PageRank the PR scores are initialized with a uniform distribution, i.e. $\frac{1}{|V|}$, PPR concentrates all initial mass uniformly over the word vertices representing the context words in C, in order to compute the structural relevance of the concepts in the LKB given the input context. Finally, given the PageRank scores of the vertices in G and a target word w to be disambiguated, PPR chooses its associated concept (namely, the vertex in G corresponding to a sense of w) with the highest PageRank score.

[1] http://lucene.apache.org

[2] http://ixa2.si.ehu.es/ukb

In order to use UKB to find predominant senses of German words, we first extend it to use GermaNet as LKB resource. This is achieved by converting GermaNet into the LKB data format used by UKB. We then run PPR to disambiguate target words in context within a set of manually annotated test sentences, and select for each target word the sense which is chosen most frequently by PPR for the target word in these sentences.

2.4 System combination

All our methods for predominant sense induction are unsupervised in the sense that they do not require any sense-tagged sentences. However, they all rely on an external resource, namely GermaNet, to provide a minimal amount of supervision. McCarthy et al.'s (2004) method uses the lexical resource to compute the semantic similarity of words. Lapata & Keller (2007) rely instead on its taxonomy structure to expand the sense descriptors of candidate senses based on hypernyms and hyponyms. Finally, UKB uses the full graph of the lexical knowledge base to find structural similarities with the input context.

All these methods include a phase of weak supervision, while in different ways. We thus hypothesize that they are complementary: that is, by combining their sense rankings, we expect their different amounts of supervision to complement each other, thus yielding a better ranking. We accordingly experiment with a simple *majority voting* scheme which, for each target word, collects the predominant senses output by all three systems and chooses the sense candidate with the highest number of votes. In case of ties, we perform a random choice among the available candidates.

3 Experiments and Evaluation

We evaluate the performance of the above methods both on the detection of predominant senses and token-level WSD in context. For this purpose we created a gold standard of sense-annotated sentences following the model of the SensEval evaluation datasets (Section 3.1). The most frequent sense annotations are then used to provide gold standard predominant senses for German words as the most frequent ones found in the annotated data. Accordingly, we first evaluate our systems in an *intrinsic* evaluation quantifying how well the automatically generated sense rankings model the one from the gold standard sense annotations (Section 3.2). In addition, the gold standard of sense-annotated German words in context is used to *extrinsically* evaluate each method by performing type-based WSD, i.e. disambiguating all contextual occurrences of a target word by assigning them their predominant sense (Section 3.3). Finally, since UKB provides a system to perform token-based WSD, i.e. disambiguating each occurrence of a target word separately, we evaluate its output against the gold standard annotations and compare its performance against the type-based systems.

3.1 Creation of a gold standard

Given the lack of sense-annotated corpora for German in the SensEval and SemEval competitions, we created a gold standard for evaluation, taking the SensEval data for other languages as a model, to ensure comparability to standard evaluation datasets. The construction of our gold standard for predominant sense is built on the hypothesis that the *most frequent sense* encountered in a sample of sentences for a given target word can be taken as the *predominant sense*. While this is arguably an idealization, it follows the assumption that, given balanced data, the predominant sense will be encountered with the highest frequency. In addition, this reflects the standard definition of predominant sense found in WordNet.

We selected the 40 keys from the English SensEval-2 test set[3] and translated these into German. In case of alternative translations, the selection took into account part of speech, comparable ambiguity rate, and frequency of occurrence in deWAC (at least 20 sentences). We ensure that the data set reflects the distribution of GermaNet across PoS (the set contains 18 nouns, 16 verbs and 6 adjectives), and yields a range of ambiguity rates between 2 and 25 senses for all PoS. For each target word, we extracted 20 sentences for words with up to 4 senses, and an additional 5 sentences per word for each additional sense. This evaluation dataset was manually annotated with the contextually appropriate GermaNet senses.

3.2 Modeling human sense rankings from gold standard annotations

Experimental setting. The gold standard ranking of word sense we use is given by frequency of senses in the annotations, i.e. the most frequently annotated word sense represents the predominant

[3] http://www.d.umn.edu/~tpederse/data.html

one, and so on. For each method, we then evaluate in terms of standard measures of precision (P, the ratio of correct predominant senses to the total of senses output by the system), recall (R, the ratio of correct predominant senses to the total of senses in the gold standard) and F_1-measure ($\frac{2PR}{P+R}$). Since all methods provide a ranking of word senses, rather than a single answer, we also performed an additional evaluation using the ranking-sensitive metrics of *precision at rank* – $P@k$ i.e. the ratio of correct predominant senses found in the top-k senses to the total of senses output by the system – as well as *Mean Reciprocal Rank* – MRR, namely the average of the reciprocal ranks of the correct predominant senses given by the system.

Results and discussion. Tables 1 and 2 present results for the intrinsic evaluation of the German predominant senses. These are generated based on the methods of McCarthy et al. (2004, MCC), Lapata & Keller (2007, LK), the frequency of sense assignments of UKB to the sense-annotated test sentences, and the system combination (Merged, Section 2.4). In the case of LK, we show for the sake of brevity only results obtained with the best configuration (including counts from PARA, HYPER and HYPOALL with no filtering), as found by manually validating the system output on a held-out dataset of word senses. As a baseline, we use a random sense assignment to find the predominant sense of a word (Random), as well as a more informed method that selects as predominant sense of a word the one whose synset has the largest size (SynsetSize). Each system is evaluated on all PoS and the nouns-only subset of the gold standard.

All systems, except LK on the all-words dataset, perform above both baselines, indicating meaningful output. The drastic performance decrease of LK when moving from nouns only to all PoS is due to the fact that in many cases, i.e. typically for adverbs and adjectives but also for nouns, the GermaNet synsets contain none or few synonyms to construct the base sense descriptions with, as well as very few hyponyms and hypernyms to expand them (i.e. due to the paucity of connectivity in the taxonomy). Among the available methods, UKB achieves the best performance, since it indirectly makes use of the supervision provided by the words in context. System combination performs lower than the best system: this is because in many cases (i.e. 17 out of 40 words) we reach a tie, and the method randomly selects a sense out of the three available without considering their confidence scores. Following Lapata & Keller (2007), we computed the correlation between the sense frequencies in the gold standard and those estimated by our models by computing the Spearman rank correlation coefficient ρ. In the case of our best results, i.e. UKB, we found that the sense frequencies were significantly correlated with the gold standard ones, i.e. $\rho = 0.44$ and 0.49, $p << 0.01$, for nouns and all-words respectively.

3.3 Type and token-based WSD for German

Experimental setting. We next follow the evaluation framework established by McCarthy et al. (2004) and Lapata & Keller (2007) and evaluate the sense ranking for the *extrinsic* task of performing WSD on tokens in contexts. We use the sense rankings to tag all occurrences of the target words in the test sentences with their predominant senses. Since such a *type-based* WSD approach only provides a baseline system which performs disambiguation without looking at the actual context of the words, we compare it against the performance of a full-fledged *token-based* system, i.e. UKB.*inst*, which disambiguates each instance of a target word separately based on its actual context.

Results and discussion. The results on the WSD task are presented in Table 3. As in the other evaluation, we use the standard metrics of precision, recall and balanced F-measure, as well as the Random and SynsetSize baselines. We use SynsetSize as a back-off strategy in case no sense assignment is attempted by a system, i.e. similar to the use of the SemCor most frequent sense heuristic for standard English WSD systems. In addition, we compute the performance of the system using the most frequent sense from the test sentences themselves: this represents an oracle system, which uses the most frequent sense from the gold standard to provide an upper-bound for the performance of type-based WSD.

The WSD results corroborate our previous findings from the intrinsic evaluation, namely that: (i) all systems, except LK on the all-words dataset, achieve a performance above both baselines, indicating the feasibility of the task; (ii) the best results on type-based WSD are achieved by selecting the sense which is chosen most frequently by UKB for the target word; (iii) system combination based on a simple majority-voting scheme does not improve the results, due to the ties in the voting and the relative random choice among the three votes. As expected, performing token-based WSD performs bet-

	Nouns only			All words		
	P	R	F_1	P	R	F_1
Random	16.67	16.67	16.67	17.50	17.50	17.50
SynsetSize	33.33	33.33	33.33	32.50	32.50	32.50
MCC	44.44	44.44	44.44	35.90	35.00	35.44
LK	56.25	50.00	52.94	29.03	22.50	25.35
UKB	**66.67**	**66.67**	**66.67**	**50.00**	**50.00**	**50.00**
Merged	61.11	61.11	61.11	47.50	47.50	47.50

Table 1: Results against human sense rankings: precision/recall on exact, i.e. predominant-only, senses.

	Nouns only				All words			
	P@1	P@2	P@3	MRR	P@1	P@2	P@3	MRR
baseline	16.67	66.67	77.78	0.50	17.50	52.50	70.00	0.47
SynsetSize	33.33	72.22	88.89	0.61	32.50	55.00	67.50	0.54
MCC	44.44	88.89	88.89	0.69	35.90	66.67	79.49	0.58
LK	56.25	81.25	93.75	0.65	29.03	41.94	48.39	0.29
UKB	**66.67**	**88.89**	**100.00**	**0.81**	**50.00**	**77.50**	**87.50**	**0.68**
Merged	61.11	83.33	88.89	0.74	47.50	70.00	70.00	0.59

Table 2: Results against human sense rankings: precision @k and MRR on full sense rankings.

	Nouns only $P/R/F_1$	All words $P/R/F_1$
Random	22.49	15.42
SynsetSize	31.98	24.67
MCC	41.46	27.66
LK	42.28	21.31
UKB	**48.78**	**36.73**
Merged	44.72	33.55
UKB.*inst*	**55.49**	**38.90**
Test MFS	64.50	56.54

Table 3: Results for *type*- and *token*-based WSD on the gold standard.

ter than type-based: this is because, while labeling based on predominant senses represents a powerful option due to the skewness of sense distributions, target word contexts also provide crucial evidence to perform robust WSD.

4 Conclusions and Future Work

We presented a variety of methods to automatically induce resources to perform WSD in German, using GermaNet as a LKB. We applied methods for predominant sense induction, originally developed for English (McCarthy et al., 2004; Lapata & Keller, 2007), to German. We further adapted a graph-based WSD system (Agirre & Soroa, 2009) to label words in context using the GermaNet resource.

Our results show that we are able to robustly bootstrap baseline systems for the automatic annotation of word senses in German. The systems were evaluated against a carefully created gold standard corpus. Best results were obtained by the knowledge-based system, which profits from its limited supervision by the surrounding context. System integration based on majority voting could not improve over the best system, yielding an overall performance degradation. We leave the exploration of more refined ensemble methods, e.g. weighted voting, to future work.

While our evaluation is restricted to 40 words only, we computed predominant sense rankings for the entire sense inventory of GermaNet. We will make these available in the form of a sense-annotated version of deWAC, as well as an API to access this information in GermaNet.

The present work is *per-se* not extremely novel, but it extends and applies existing methods to create new resources for German. The annotations produced by the WSD systems can serve as a basis for rapid construction of a gold standard corpus by manually validating their output. A natural extension of our approach is to couple it with manual validation frameworks based on crowdsourcing, i.e. Amazon's Mechanical Turk (cf. Snow et al. (2008)). We leave such exploration to future work.

References

Agirre, Eneko & Philip Edmonds (Eds.) (2006). *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Agirre, Eneko, Lluís Màrquez & Richard Wicentowski (Eds.) (2007). *Proceedings of SemEval-2007*.

Agirre, Eneko & Aitor Soroa (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proc. of EACL-09*, pp. 33–41.

Atserias, Jordi, Luis Villarejo, German Rigau, Eneko Agirre, John Carroll, Bernardo Magnini & Piek Vossen (2004). The MEANING multilingual central repository. In *Proc. of GWC-04*, pp. 80–210.

Banerjee, Satanjeev & Ted Pedersen (2003). Extended gloss overlap as a measure of semantic relatedness. In *Proc. of IJCAI-03*, pp. 805–810.

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi & Eros Zanchetta (2009). The WaCky Wide Web: A collection of very large linguistically processed web-crawled corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Brin, Sergey & Lawrence Page (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Briscoe, Edward, John Carroll & Rebecca Watson (2006). The second release of the RASP system. In *Proc. of COLING-ACL-06 Interactive Presentation Sessions*, pp. 77–80.

de Melo, Gerard & Gerhard Weikum (2009). Towards a universal wordnet by learning from combined evidence. In *Proc. of CIKM-09*, pp. 513–522.

Fellbaum, Christiane (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Gurevych, Iryna & Hendrik Niederlich (2005). Accessing GermaNet data and computing semantic relatedness. In *Comp. Vol. to Proc. of ACL-05*, pp. 5–8.

Hall, J. & J. Nivre (2008). A dependency-driven parser for german dependency and constituency representations. In *Proceedings of the Parsing German Workshop at ACL-HLT 2008*, pp. 47–54.

Jiang, Jay J. & David W. Conrath (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th International Conference on Research in Computational Linguistics*.

Lapata, Miralla & Frank Keller (2007). An information retrieval approach to sense ranking. In *Proc. of NAACL-HLT-07*, pp. 348–355.

Leacock, Claudia & Martin Chodorow (1998). Combining local context and WordNet similarity for word sense identification. In C. Fellbaum (Ed.), *WordNet. An Electronic Lexical Database*, Chp. 11, pp. 265–283. Cambridge, Mass.: MIT Press.

Leech, Geoffrey (1992). 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Lemnitzer, Lothar & Claudia Kunze (2002). GermaNet – representation, visualization, application. In *Proc. of LREC '02*, pp. 1485–1491.

Lin, Dekang (1998). Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL-98*, pp. 768–774.

Mausam, Stephen Soderland, Oren Etzioni, Daniel Weld, Michael Skinner & Jeff Bilmes (2009). Compiling a massive, multilingual dictionary via probabilistic inference. In *Proc. of ACL-IJCNLP-09*, pp. 262–270.

McCarthy, Diana, Rob Koeling, Julie Weeds & John Carroll (2004). Finding predominant senses in untagged text. In *Proc. of ACL-04*, pp. 280–287.

Mihalcea, Rada & Phil Edmonds (Eds.) (2004). *Proceedings of SENSEVAL-3*.

Miller, George A., Claudia Leacock, Randee Tengi & Ross T. Bunker (1993). A semantic concordance. In *Proceedings of the 3rd DARPA Workshop on Human Language Technology*, pp. 303–308.

Navigli, Roberto (2009). Word Sense Disambiguation: A survey. *ACM Computing Surveys*, 41(2):1–69.

Navigli, Roberto & Simone Paolo Ponzetto (2010). BabelNet: Building a very large multilingual semantic network. In *Proc. of ACL-10*.

Pedersen, Ted & R. Bruce (1997). Distinguishing word senses in untagged text. In *Proc. EMNLP-97*, pp. 197–207.

Pianta, Emanuele, Luisa Bentivogli & Christian Girardi (2002). MultiWordNet: Developing an aligned multilingual database. In *Proc. of GWC-02*, pp. 21–25.

Schmid, Helmut, Arne Fitschen & Ulrich Heid (2004). SMOR: A German computational morphology covering derivation, composition and inflection. In *Proc. of LREC '04*.

Schütze, Hinrich (1992). Dimensions of meaning. In *Proceedings of Supercomputing '92,* Los Alamitos, Cal., 16–20 November 1992, pp. 787–796.

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

Snow, Rion, Brendan O'Connor, Daniel Jurafsky & Andrew Ng (2008). Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proc. of EMNLP-08*, pp. 254–263.

Tufiş, Dan, Dan Cristea & Sofia Stamou (2004). BalkaNet: Aims, methods, results and perspectives. a general overview. *Romanian Journal on Science and Technology of Information*, 7(1-2):9–43.

Vossen, Piek (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht, The Netherlands: Kluwer.

Zielinski, Andrea, Christian Simon & Tilman Wittl (2009). Morphisto: Service-oriented open source morphology for German. In Cerstin Mahlow & Michael Piotrowski (Eds.), *State of the Art in Computational Morphology: Proceedings of the Workshop on Systems and Frameworks for Computational Morphology*, Vol. 41, Communications in Computer and Information Science, pp. 64–75. Springer.