Delexicalized Supervised German Lexical Substitution

Gerold Hintz and **Chris Biemann**
Research Training Group AIPHES / FG Language Technology
Department of Computer Science, Technische Universität Darmstadt
{hintz,biem}@lt.informatik.tu-darmstadt.de

Abstract

We address the German lexical substitution task, which requires retrieving a ranked list of meaning-preserving substitutes for a given target word within an utterance. With *GermEval-2015: LexSub*[1], this challenge is posed for the first time using German language data. In this work we build upon the existing state of the art for English lexical substitution, employing a delexicalized supervised system. In adapting the existing approach, we consider in particular the available lexical resources for German and evaluate their suitability to the task at hand. We report first results on German lexical substitution and observe a similar performance as English systems evaluated on the SemEval dataset.

1 Introduction

Lexical substitution is a special form of contextual paraphrasing which aims to predict substitutes for a target word instance within a sentence. This implicitly addresses the problem of resolving the ambiguity of polysemous terms. In contrast to Word Sense Disambiguation (WSD) this is achieved without requiring a predefined inventory of senses. A vector of substitute words for a given target can be regarded as an alternative contextualized meaning representation that can be used in similar downstream tasks such as Information Retrieval or Question Answering. In contrast to WSD, lexical substitution systems are not limited by the coverage or granularity of the underlying sense inventory, and is still applicable to languages in which no such resource is available at all. As a result, lexical substitution systems have become very popular for evaluating context-sensitive lexical inference since the introduction of the first *SemEval-2007 lexical substitution* task (McCarthy and Navigli, 2007). Whereas this and earlier variants of this task were posed without any training data and a relatively small evaluation set of a few thousand instances, later datasets were scaled up by the use of crowdsourcing, containing nearly 24k sentences with substitutes for a lexical sample of 1012 frequent nouns (Biemann, 2013). With *GermEval 2015*, German lexical substitution data (Cholakov et al., 2014) is provided for the first time. The dataset contains 153 unique target words, with 10 (nouns and adjectives) or 20 (verbs) sample sentences being selected from the German Wikipedia for annotation. About half of this data (1040 sentences) is released as training data and is available at the time of writing.

In this work, we apply the current state of the art for English lexical substitution to this German dataset. In Section 2 we briefly cover the related work in lexical substitution. Section 3 discusses German lexical resources for obtaining substitution candidates and evaluates their suitability to the task at hand. In Section 4 we describe the final system and report on the results in Section 5.

2 Related Work

2.1 Unsupervised systems

Unsupervised approaches to the lexical substitution task typically use a contextualized word instance representation and rank substitute candidates according to their similarity to this representation. Early methods employed syntactic vector space models (Erk and Padó, 2008; Thater et al., 2011) or a clustering of instance representations (Erk and Padó, 2010). Later approaches have explored various other models, including probabilistic graphical models (Moon and Erk, 2013), LDA topic models (O Séaghdha and Korhonen, 2014), graph centrality (Sinha and Mihalcea, 2011), and distributional models (Melamud et al., 2015a).

[1] *GermEval-2015: LexSub*: https://sites.google.com/site/germeval2015/

Proceedings of GermEval 2015: LexSub, pages 11–16, September 2015.

A recent line of research takes advantage of word embeddings, which are low-dimensional continuous vector representations popularized by the skip-gram model (Mikolov et al., 2013). A simple but effective embedding-based model for lexical substitution is proposed by Melamud et al. (2015b): They decompose the semantic similarity between a target and a substitute word into a second-order target-to-target similarity based on their similarity in the embedding space, and a first-order target-to-context similarity. For this, they consider the learned context embeddings (which are usually discarded after training a Skip-gram model) and compute a substitute-to-context similarity. They achieve state-of-the-art results by just considering a (balanced) geometric mean of these two components.

2.2 Supervised systems

Supervised systems can be divided into *per-word* systems, which are trained on target instances per lexeme, and *all-words* systems, which aim to generalize over all lexical items. It could be shown that per-word supervised systems perform very well (with a precision > 0.8 on SemEval-2007 data) given a sufficient amount of training data for the target lexemes (Biemann, 2013). The downside of this approach is the inability to scale to unseen targets. A successful remedy to this is proposed by Szarvas et al. (2013) by the use of *delexicalized features*. The features extracted from the training data is generalized in such a way that it can generalize across lexical items beyond the training set. In this work, we build upon this framework and apply delexicalized features to German lexical substitution.

3 Candidate set evaluation

The lexical substitution task generally relies on lexical semantic resources to obtain a set of substitution candidates for a given lexeme. Most prevalently, WordNet (Fellbaum, 1998) is chosen as a standard resource for the English version of this task. Given multiple resources, a supervised combination of all resources was found to lead to the best results (Sinha and Mihalcea, 2009).

GermaNet (Hamp and Feldweg, 1997) can be considered an out-of-the-box replacement for *WordNet*. It groups lexical units into *synsets* and denotes semantic relations between these synsets. To obtain a candidate set from *GermaNet*, clearly synonyms of the substitute target should be considered (all lexemes sharing a common synset). It is further reasonable to consider both hyponyms and hypernyms of the target, as well as the transitive hull (*Transporter* → *Automobil* → *Fahrzeug* → ..) of these relations. Although higher level nodes of the *GermaNet* taxonomy include highly abstract terminology (.. → *Artefakt* → *Objekt* → *Entität*), no effort was done to exclude these terms from the candidate set. For this candidate extraction stage, no sense disambiguation of target words is performed and all senses of a given target lemma are aggregated into the candidate list.

candidate set	*R*	*P*
GermaNet syn	0.05	0.15
GermaNet syn + hy	0.14	0.15
GermaNet syn + hy + ho	0.17	0.09
GermaNet all (transitive)	0.20	0.04
Wiktionary	0.17	0.14
Woxikon	0.44	0.08
Duden	0.34	0.15
Wortschatz	0.40	0.07
all lexical resources	0.61	0.04
DT (top 200 similar)	0.46	0.01
DT + lexical resources	0.71	0.02

Table 1: Candidate set evaluation on GermEval training data. The abbreviations *syn*, *hy*, and *ho* specify synonyms, direct hypernyms and direct hyponyms respectively, whereas *all* refers to pairs with an arbitrary semantic relation between them

We use UBY (Gurevych et al., 2012) to access *GermaNet* (version 9.0) and *Wiktionary*[2]. Additionally we crawl lexical resources available on the web: *Woxikon*[3], *Duden*[4] and *Leipzig Wortschatz*[5]. From these websites we scrape all listed synonyms, and in case of *Leipzig Wortschatz* all their semantic relations such as *referenced-by*, *compared-to*, and *Dornseiff-Bedeutungsgruppen* (Dornseiff, 1959).

In order to evaluate the suitability of each of these resources to the GermEval task, we construct a binary test set: each substitute pair which is present at least once in the gold data is considered a "good" expansion, whereas substitute pairs not present in the gold data are considered "bad". For each resource, we consider the recall and precision of "good" expansion pairs, as shown in Table 1. As we perform ranking on the given candidate sets,

[2] https://www.wiktionary.org/
[3] http://www.woxikon.com/
[4] http://www.duden.de/
[5] http://wortschatz.uni-leipzig.de/

we are mostly interested in the recall, as it constitutes an upper bound for the final system. We also perform a preliminary error analysis of available substitution candidates: while all target words, and 85% of their substitutes were found in *GermaNet*, only for 20% of the GermEval pairs a semantic relation existed between these pairs. This indicates that the main problem with obtaining substitution candidates from a semantic resource is not necessarily its lexical coverage, but missing semantic relations between substitution pairs.

As an alternative to using a lexical semantic resource, fully knowledge-free approaches to lexical substitution have been proposed by the use of a distributional thesaurus (DT) (Biemann and Riedl, 2013). Although we do not follow this direction in-depth in the scope of this work, we observe that candidates obtained from a DT already yielded a better coverage than any lexical resource ($R = 0.4$) when pruned to the 200 most similar words. In line with the findings in Biemann and Riedl (2013) these candidates do not yield competitive performance within our system when compared to knowledge-based substitutes and we leave this direction open as future work.

4 System setup

Our system is roughly equivalent to *LexSub*[6] (Szarvas et al., 2013), although a reimplementation was used to obtain the experimental results. We follow their approach of ranking a given set of candidates based on a small set of training examples using delexicalized features. The ranking problem is cast into a binary classification task by labeling all lexical substitutions with their presence in the gold data. Hence, all substitutes which occur at least once as a gold item for a given instance are used as positive examples, whereas the remaining substitutes based on the candidate set are negative examples. We use a Maximum Entropy classifier[7] and obtain a ranking score based on the posterior probability of the positive label.

As a pre-processing step we only apply tokenization and part-of-speech tagging. We obtain the lemmatized target words directly from the gold data and have no further need to lemmatize all lexical items within the sentence, nor for syntactic parsing.

4.1 Features

We use most features from *LexSub*, and therefore do not cover in detail here those which can be easily adapted.

Frequency features A language model is used to obtain *frequency ratio* features, where an n-gram sliding window around a target t is used to generate a set of features $\frac{freq(c_l,s,c_r)}{freq(c_l,t,c_r)}$, where c_l and c_r is the left and right context of t. We also include the different normalization variants of this feature as described in Szarvas et al. (2013), and the conjunctive phrase ratio based on the conjunctions {"*und*","*oder*",","}. For obtaining frequency counts, we evaluated 5-gram counts from *web1t* (Brants and Franz, 2009) and *German Web Counts* (Biemann et al., 2013), which both yielded nearly equivalent results.

DT features We create a DT from a German news corpus of 70 million sentences (Biemann et al., 2007) and obtain first-order context-features, as well as a second-order word-to-word similarity measure as described in Biemann and Riedl (2013): We prune the data, keeping only the 1000 most salient features according to a log-likelihood test (Dunning, 1993) and obtain a ranked list of 200 similar terms for each word in the corpus, based on the overlap in these context features. In particular we use as context features tuples of left and right neighbors (*de_70M_trigram*) as well as dependency features obtained using the Mate-tools[8] parser (*de_70M_mate*) to construct two distinct DTs[9].

We define delexicalized features based on the overlap in the top k shared similar words ($k = 1$, 5, 10, 20, 50, 100, 200) and top k shared salient features ($k = 1$, 5, 10, 20, 50, 100, 1000) and directly use the similarity measure between target and substitute as a feature. Lastly, we define a feature based on the accumulated LL significance measures of DT context features occurring in the sentential context. Their computation is equivalent to *coocurence* features which are explained next.

Cooccurence features We obtained word co-occurrence counts as described in Quasthoff et al. (2006) and define the following features: For a given sentence regarded as a

[6] Original *LexSub* system: `https://sourceforge.net/projects/lexsub/`

[7] We use the *MaxEnt* implementation of Mallet: `http://mallet.cs.umass.edu/`

[8] `https://code.google.com/p/mate-tools/`

[9] The DTs are available at `https://sourceforge.net/projects/jobimtext/files/data/models/`

bag-of-words S, target word t and candidate set C, we consider the set of context words $W = S \setminus \{t\}$. For each substitute $s \in C$ we then compute the feature

$$\frac{\sum_{w \in W} LL(s,w)}{\sum_{s' \in C, w \in W} LL(s',w)}$$

where LL is the log-likelihood measure of coocurence. We also compute a simple overlap version $|Co_s \cap W|/|W|$, where Co_s denotes the set of words co-occurring with the substitute s.

Embedding features We roughly follow Melamud et al. (2015b) to define features in a word embedding space. To obtain German word embeddings we run the *word2vec*[10] toolkit to obtain a *CBOW* model with default parameters (200 dimensions, window-size of 8) on our German news corpus. Based on this embedding, we define two features: A second-order similarity measure between target and substitute based on cosine distance in the embedding space, as well as a very simple contextualized first-order target-to-context similarity measure. In contrast to Melamud et al. (2015b), we do not use the internal context embeddings to compute a similarity to the syntactic dependents of a target, and our embeddings are not syntax-based (Levy and Goldberg, 2014). Instead, we directly compute the similarities between a target word and a given set of context words in the embedding space, based on an n-gram sliding sliding window around the target. This is analogous to the delexicalized n-gram frequency features: For a given n-gram window around a target word t, with the context words $c_1 \dots c_k, t, c_{k+2} \dots c_n$, we compute for each substitute s the difference in similarity to the context words with respect to the target t:

$$\sum_{i \leq n} |\, cos\,(v_s, v_{c_i}) - cos\,(v_t, v_{c_i})\,|$$

where v_x denotes the embedding of x. This is motivated by the assumption that a substitute word should behave in the same way to each context word, as the original target t.

Semantic resource features As illustrated in Section 3 we make use of various semantic relation labels from multiple semantic resources. For each lexical resource, we obtain a set of labels for a given pair of lexemes and prefix it with the name of the resource. For *GermaNet* relations, we additionally encode the length of the transitive chain, denoting an n^{th}-level hyponymy/hypernymy relation. For instance, the semantic relation labels for the pair (*wünschen.v*, *postulieren.v*) are {`gn_hypernym_2`, `Wortschatz_synonym`}.

[10] `https://code.google.com/p/word2vec/`

dataset	*mean* (1 − dice coefficient)			
	noun	verb	adj	all
SemEval-2007	0.750	0.830	0.755	**0.760**
GermEval-2015	0.594	0.667	0.604	**0.645**

Table 2: Degree of variation within lexical substitution gold answers

Some features were discarded from the original *LexSub* system, as they could not directly be ported to German resources, or they did not prove useful. This includes the number of senses of target and substitute within *GermaNet*, the path between target and substitute within *GermaNet*, and binary features for their respective synset IDs.

5 Experimental results

As a preface to our evaluation, we comment briefly on the GermEval data. Upon inspection we noted that very few target lexemes in fact exhibit an ambiguous behavior. Most training instances refer to the same (or a close) meaning of a given target word, resulting in a low variance in gold answers between multiple instances of the same lexeme. We quantify this statement by calculating the mean dice coefficient between all pairwise sets of gold answers for a given lexeme. In Table 2 we compare these results to the SemEval-2007 data and observe a much lower degree of variation. A consequence of this is that a lexical substitution system based on GermEval data is less reliant on sentential context, and is primarily influenced by good prior expansions for a given word. In fact, we report a high performance on the ranking-only task (GAP=84.16% with candidate oracles), which is in line with our expectations.

System evaluation For evaluating the final system we perform a 10-fold cross-validation (splitting is based on target lexeme level) on the training data and report on the measures P_{best}, P_{oot}, GAP as provided by the official *GermEval* scoring tool. We disregard any multiword expressions in the gold data, as none of our candidate sets included any viable multiword expression present in the training set, and their inclusion negatively impacted results. We considered various lexical resources as potential candidate sets filtered to only single-word

candidate set	P_{best}	GAP	P@1
GermaNet	**15.04**	**19.12**	**55.77**
Wortschatz	12.26	14.84	19.39
Duden	6.41	12.25	24.74
Woxikon	4.09	10.25	22.44
Wiktionary	3.22	7.50	22.53
candidate oracle	28.06	84.16	(100)

Table 3: Evaluation of the final system using different lexical resources as substitution candidates

	GN candidates		
	P_{best}	P_{oot}	GAP
w/o frequency feat.	13.43	24.44	16.80
w/o DT feat.	14.77	**24.67**	17.59
w/o sem. relation feat.	12.26	23.22	14.84
w/o embedding feat.	14.26	24.64	17.73
w/o POS feat.	13.18	24.60	16.95
full system (train-cv)	**15.04**	24.35	**19.12**
full system (testset)	11.20	19.49	15.96

Table 4: Final system results and feature ablation using 10-fold cross-validation on the training set and final results

expressions. Table 3 shows the output of the full system, restricted to candidates of each resource. Despite their promising coverage of gold items in the training data (see Table 1), all lexical resources perform notably worse than *GermaNet*. This may be due to the nature of these resources: Whereas the candidate set from *GermaNet* is very accurate in enforcing the denoted semantic relationship. e.g. in case of synonymy, the other resources contain a much broader spectrum of terms that are considered "synonymous". Furthermore, the false positives in the *GermaNet* candidate set contain very obscure terms from upper levels in the ontology (*Artefakt*, *Objekt*, ..) which are easily downranked - the ranking of e.g. *Duden* candidates appears to be more difficult, as they contain mostly words which are in fact suitable in the given context. We also compare the performance to a *candidate oracle*, which serves as an upper bound for candidate sets as well as a general evaluation for the ranking-only task. Despite the bad performance as candidate sets, we find that extracting the *semantic relations* from all of these lexical resources as a feature could still notably improve the final system performance.

We further perform feature ablation test for the full system using *GermaNet* candidates as shown in Table 4. Although some features seem to exhibit redundancy (e.g. DT features and semantic relation features) all features yield a significant relative gain. It can be seen that the addition of semantic relation features yielded a relative improvement of nearly 23% for P_{best}, indicating that this is a strong feature for German lexical substitution. Final performance on the testset (see Table 4) is significantly worse ($P_{best} = 11.20$ compared to $P_{best} = 15.04$ on the training set with cross-validation). The reason for this is partly that candidates obtained from GermaNet have less coverage of the test data, and the test data containing more (non-covered) multiword expressions. However, when exchanging the datasets, a reasonable performance is obtained ($P_{best} = 14.68$) indicating that the issue is not related to a discrepancy between the datasets. Instead, the testset may contain generally harder instances.

6 Conclusion and Future Work

In this work we have successfully applied state of the art methods to German lexical substitution. We find that approaches applicable to the English version of this task can be readily adapted to German. We experimented with various lexical resources which can be used in place of their conventional English counterparts, and observe that *GermaNet* is a high quality resource which has however slight shortcomings in terms of coverage. We observe that in particular in the case of *GermaNet*, obtaining lexical substitution candidates based on the semantic relations *synonymy*, *hyponymy* and *hypernymy* is not sufficient for matching the substitutes provided by human annotators. Extracting semantic relations from other lexical resources notably improved system performance. While this is a delexicalized feature that is sufficient to generalize across all German lexical items, it is very language-dependent. In future work, we plan to overcome this dependency by generalizing features even more and experiment with delexicalized features in a multilingual setting. Additionally, we aim for a completely knowledge-free approach, obtaining substitution candidates from large background corpora.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1."

References

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The Leipzig Corpora Collection: monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Chris Biemann, Felix Bildhauer, Stefan Evert, Dirk Goldhahn, Uwe Quasthoff, Roland Schäfer, Johannes Simon, Leonard Swiezinski, and Torsten Zesch. 2013. Scalable construction of high-quality web corpora. *Journal for Language Technology and Computational Linguistics*, 28(2):23–60.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Thorsten Brants and Alex Franz. 2009. Web 1T 5-gram, 10 European languages version 1. *Linguistic Data Consortium*.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Lexical Substitution Dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

Franz Dornseiff. 1959. *Der deutsche Wortschatz nach Sachgruppen*. Walter de Gruyter.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Waikiki, Honolulu.

Katrin Erk and Sebastian Padó. 2010. Exemplar-based models for word meaning in context. In *Proceedings of the ACL 2010 conference short papers*, pages 92–97, Uppsala, Sweden.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a Lexical-Semantic Net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308, Baltimore, USA.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53, Prague, Czech Rep.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015a. Modeling Word Meaning in Context with Substitute Vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, Denver, USA.

Oren Melamud, Omer Levy, Ido Dagan, and Israel Ramat-Gan. 2015b. A Simple Word Embedding Model for Lexical Substitution. *VSM Workshop*. Denver, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, Harrah's Lake Tahoe, USA.

Taesun Moon and Katrin Erk. 2013. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):42.

Diarmuid O Séaghdha and Anna Korhonen. 2014. Probabilistic Distributional Semantics with Latent Variable Models. *Computational Linguistics*, 40(3):587–631.

Uwe Quasthoff, Matthias Richter, and Christian Biemann. 2006. Corpus portal for search in monolingual corpora. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, Genova, Italy.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the Conference in Recent Advances in Natural Language Processing*, pages 404–410, Borovets, Bulgaria.

Ravi Som Sinha and Rada Flavia Mihalcea. 2011. Using Centrality Algorithms on Directed Graphs for Synonym Expansion. In *FLAIRS Conference*, Palm Beach, USA.

György Szarvas, Chris Biemann, Iryna Gurevych, et al. 2013. Supervised All-Words Lexical Substitution using Delexicalized Features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1131–1141, Atlanta, USA.

Stefan Thater, Hagen Fürstenau, and Manfred Pinkal. 2011. Word Meaning in Context: A Simple and Effective Vector Model. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1134–1143, Stroudsburg, PA, USA.