GermEval 2015: LexSub – A Shared Task for German-language Lexical Substitution

Tristan Miller
Ubiquitous Knowledge Processing Lab
Department of Computer Science
Technische Universität Darmstadt

Darina Benikova
Research Training Group AIPHES
Department of Computer Science
Technische Universität Darmstadt

Sallam Abualhaija
Institute of Computer Technology
Technische Universität Hamburg-Harburg

Abstract

Lexical substitution is a task in which participants are given a word in a short context and asked to provide a list of synonyms appropriate for that context. This paper describes GermEval 2015: LexSub, the first shared task for automated lexical substitution on German-language text. We describe the motivation for this task, the evaluation methods, and the manually annotated data set used to train and test the participating systems. Finally, we present an overview and discussion of the participating systems' methodologies, resources, and results.

1 Introduction

Word sense disambiguation, or WSD (Agirre and Edmonds, 2007)—the task of determining which of a word's senses is the one intended in a particular context—has been a core research problem in computational linguistics since the very inception of the field. Approaches to WSD system evaluation can be categorized as *intrinsic* (or *in vitro*) or *extrinsic* (*in vivo*) (Ide and Véronis, 1998). In the former, the assessment is performed independently of any particular natural language processing application. Rather, evaluators directly compare the automatically produced sense assignments with a manually annotated gold standard (Palmer et al., 2007). In extrinsic evaluation, however, systems are scored according to their contribution to a dedicated NLP task, such as machine translation (Carpuat and Wu, 2005a,b; Chan et al., 2007; Carpuat and Wu, 2007) or information retrieval (Clough and Stevenson, 2004; Schütze and Pedersen, 1995; Sanderson, 1994; Zhong and Ng, 2012).

Most published WSD evaluations to date, such as those in the Senseval and SemEval workshop series, have been of the intrinsic variety. However, it is widely agreed that extrinsic evaluations are preferable, since the usual point of computational WSD is to support real-world NLP applications. The idea of using lexical substitution for *in vivo* WSD evaluation was proposed as far back as 2002 (McCarthy, 2002) and has led to a number of English, Italian, and crosslingual evaluation competitions since then (McCarthy and Navigli, 2007; Toral, 2009; Mihalcea et al., 2010). Until now, however, no one has conducted a rigorous evaluation of lexical substitution systems on German-language text. In this paper, we describe and present the results of GermEval 2015: LexSub, the scientific community's first shared task for German-language lexical substitution.

The remainder of this paper is structured as follows: §2 reviews the task of lexical substitution and the methodologies used to evaluate the performance of lexical substitution systems, §3 describes the data set used to train and test the systems participating in our task, and §4 describes the lexical-semantic resources made available to the participants and employed by some of the systems and baselines. In §§5 and 6 we briefly describe these systems and baselines, respectively, and in §7 we present and discuss their results on the test data set. Finally, we wrap things up in §8 with some general observations.

2 Task definition

Lexical substitution is the task of identifying appropriate substitutes for a target word in a given context. For example, consider the following two German-language contexts (abridged from Cholakov et al. (2014)) containing the word *Erleichterung*:

Proceedings of GermEval 2015: LexSub, pages 1–9, September 2015.

(1) *In der Legislaturperiode 1998–2002 wurden einige Reformen des Staatsbürgerschaftsrechts bezüglich der **Erleichterung** von Einwanderung verabschiedet.*

(In the legislative period of 1998–2002 a few reforms on citizenship law concering the **easing** of immigration were passed.)

(2) *Vor allem auf dem Lande war die Umstellung aber schwer durchsetzbar und die **Erleichterung** groß, als 1802 der Sonntagsrhythmus und 1805 der vorrevolutionäre Kalender insgesamt wieder eingeführt wurden.*

(The change was particularly difficult to enforce in the countryside, and there was great **relief** when in 1802 the Sunday routine and in 1805 the pre-revolutionary calendar were reintroduced.)

The word *Förderung* (meaning "facilitation") would be an appropriate substitute for *Erleichterung* (meaning "easing") in the first context, whereas the word *Freude* (meaning "delight") would not be. Conversely, *Freude* would indeed be a valid substitute for *Erleichterung* (meaning "relief") in the second context, whereas *Förderung* would not be.

Lexical substitution is a relatively easy task for humans, but potentially very challenging for machines because it relies—explicitly or implicitly—on word sense disambiguation, a longstanding core problem in computational linguistics. In fact, lexical substitution was originally conceived as a method for evaluating word sense disambiguation systems which is independent of any one sense inventory. However, it also has a number of uses in real-world NLP tasks, such as text summarization, question answering, paraphrase acquisition, text categorization, information extraction, text simplification, lexical acquisition, and text watermarking.

Evaluation of automated lexical substitution systems is effected by applying them on a large number of word–context combinations (*items* or *instances*) and then comparing the substitutions they propose to those made by human annotators. There are various scoring methodologies which have been used in past lexical substitution tasks. The following list briefly describes the ones employed in our task; for details of their derivation and precise computation the reader is referred to the cited papers.

Best (McCarthy and Navigli, 2009) allows a system to propose as many substitutes as it wishes for each item, but considers the first proposed substitute to be its "best guess". This methodology uses the following metrics:

- **Recall (R)** scores each item by finding the average human annotator response frequency of the system's substitutes and dividing by the number of system substitutes. The scores for all items are then summed and divided by the total number of items in the data set.
- **Precision (P)** is the same as recall, except that items for which the system declined to propose any substitutes are disregarded.
- **Mode recall (Mode R)** is the number of times the system's "best guess" corresponded to the one substitute most commonly chosen by the human annotators, divided by the number of items with such a human-annotated substitute.
- **Mode precision (Mode P)** is the same as mode recall, except that items for which the system declined to propose any substitutes are disregarded.

Out-of-ten (OOT) (McCarthy and Navigli, 2009) allows a system to propose up to ten substitutes for each item, though the order of these is not significant. The following scoring metrics are used:

- **Recall (R)** is the same as the *best* recall metric, except that the credit for each correct substitute is not divided by the number of proposed substitutes.
- **Precision (P)** is the same as the *best* precision metric, except that the credit for each correct substitute is not divided by the number of proposed substitutes.
- **Mode recall (Mode R)** is the number of times the one substitute most commonly chosen by the human annotators occurred among the system's substitutes, divided by the number of items for which there was a single most frequent human-annotated substitute.
- **Mode precision (Mode P)** is the same as mode recall, except that items for which the system declined to propose any substitutes are disregarded.

Generalized average precision (GAP) (Kishida, 2005) allows a system to propose a ranked list of substitutes and then assesses the quality of the entire ranked list. It is believed to be superior to *OOT* because of its sensitivity to the relative position of correct and incorrect candidates in the ranking.

3 Data set

For our training and test data, we use the German-language lexical substitution data set produced by Cholakov et al. (2014). The full data set consists of 2040 context sentences from the German edition of Wikipedia, each containing one target word. There are 153 unique target words, equally distributed across parts of speech (nouns, verbs, and adjectives) and three frequency groups according to the lemma frequency list of the German WaCky corpus (Baroni et al., 2009). There are ten context sentences for each noun and adjective target, and twenty for each verb. Two hundred of the sentences were annotated by four professional human annotators, and the remainder by one professional annotator and five additional annotators recruited via crowdsourcing. About half of this data (26 nouns, 26 verbs, and 26 adjectives in 1040 sentence contexts) forms the training set, which was made available to participants in full in advance of the task. The remainder forms the test set, which (excluding the list of substitutions) was given to the participants at the beginning of the task.

This German data set is similar in size and scope to past English and Italian data sets. The SemEval-2007 lexical substitution data set consists of 2010 sentences (ten sentences for each of 201 unique target words) and the EVALITA 2009 data contains 2310 sentences (also with ten sentences per word). In contrast to the English and Italian data sets, the Cholakov et al. (2014) data has a greater emphasis on verbs, and contains no adverbs since the distinction between adverbs and adjectives is less pronounced in German.

We have now published the entire data set, including the human-provided substitutions, under the Creative Commons Attribution-ShareAlike license.[1] This is, to our knowledge, the only published data set which makes possible the evaluation of WSD systems with an arbitrary sense inventory. (Existing collections of sense-annotated German text, such as WebCAGe (Henrich et al., 2012) and TüBa-D/Z (Henrich and Hinrichs, 2013), are all tied to GermaNet.)

The format of the files in the data set corresponds to that of lexical substitution tasks in other languages (McCarthy and Navigli, 2007; Toral, 2009). There are two types of files:

1. XML files containing single-sentence instances enclosed in `instance` and `context` elements. Within each instance, the target word is enclosed in a `head` element. Instances with the same target lemma are grouped together in a `lexelt` element. The `lexelt` elements are grouped together in a top-level `corpus` element. The entire format is illustrated in Figure 1.

2. Delimited *gold* files which are cross-referenced to the XML files and which contain the gold-standard substitutions. Each line has the format

   ```
   lexelt id :: subs
   ```

   where

   `lexelt` is the unique identifier for the target lemma, corresponding to the `item` attribute of the `lexelt` element in the XML file;

   `id` is the unique identifier for the instance, which matches the `id` attribute of the `instance` element; and

   `subs` is a semicolon-delimited list of lemmatized substitutes. Each substitute is followed by a space and its corresponding frequency count (indicating the number of annotators who provided that substitute).

   The *gold* file line corresponding to the instance shown in Figure 1 is shown in Figure 2.

4 Resources

We made available to all participants a number of language resources supporting the task of lexical substitution:

GermaNet (Hamp and Feldweg, 1997; Henrich and Hinrichs, 2010) is a lexical-semantic network that relates German-language nouns, verbs, and adjectives. It is the analogue of WordNet (Fellbaum, 1998) and ItalWordNet (Roventini et al., 2000) used in past English

[1] `https://www.ukp.tu-darmstadt.de/data`

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE corpus SYSTEM 'lexsub.dtd'>
<corpus lang="de">
  <lexelt item="Monarch.n">
    <instance id="Monarch_1">
      <context>
        Dies war die letzte britische Regierung, die ein <head>Monarch</head>
        ohne Mehrheit im Unterhaus ernannte, und scheiterte schon im April 1835.
      </context>
    </instance>
    ⋮
  </lexelt>
  ⋮
</corpus>
```

Figure 1: Format of the data set's XML files

```
Monarch.n Monarch_1 :: König 3; Herrscher 2; Adliger 1; Staatsoberhaupt 1;
```

Figure 2: Sample line from a *gold* file

Resource	Senses	Synsets
WordNet 2.1	207 016	117 597
WordNet 3.0	206 941	117 659
ItalWordNet	ca. 130 000	ca. 80 000
GermaNet 8.0	111 361	84 584
GermaNet 9.0	121 810	93 246
GermaNet 10.0	131 814	101 371

Table 1: Comparison of language resources used for lexical substitution

and Italian lexical substitution tasks, respectively. All three wordnets group word–sense pairs (*lexical units* or *senses*) expressing the same concept into structures called *synsets*.

The standalone version of GermaNet offered to GermEval 2015: LexSub participants was GermaNet 10.0, though they also had the choice of using GermaNet 9.0 as part of UBY (see below). The baselines described in §6 use GermaNet 8.0.

Table 1 shows the number of senses and synsets for the versions of WordNet, ItalWordNet, and GermaNet used in the current and past lexical substitution tasks.

UBY (Gurevych et al., 2012) is a large-scale lexical-semantic resource which links information from several expert- and collaboratively constructed resources for English and German. The linked resources include GermaNet 9.0, WordNet 3.0, and the English and German versions of Wikipedia and Wiktionary.

JoBimText (Biemann and Riedl, 2013) is an automatically induced resource for German by means of distributional semantics. Distributional thesauri, as well as distributional features of words, are provided as a RESTful API and as a database. These features were demonstrated to be beneficial for lexical substitution by Szarvas et al. (2013).

5 Participating systems

GermEval 2015: LexSub saw participation from two systems, from Hintz and Biemann (2015) and Jackov (2015), though as the former is connected with one of the task organizers, it was entered non-competitively.

Hintz and Biemann use a supervised delexicalized approach adapted from previous work on English-language lexical substitution by Szarvas et al. (2013). They made use of Wiktionary and GermaNet (via UBY) and the JoBimText distributional thesauri, as well as the online lexical resources Woxikon, Duden, and Leipzig Wortschatz. They employ a maximum entropy classifier, regarding the task as a binary classification problem on whether any given substitution fits or does not fit the context. In addition to the semantic resource features, they make use of frequency, co-occurrence, and embedding features.

Jackov applies a deep semantic and syntactic approach relying on machine translation techniques.

System	Best P	Best R	Best Mode P	Best Mode R	OOT P	OOT R	OOT Mode P	OOT Mode R	GAP
RandomSense	7.40	7.40	15.13	15.13	12.53	12.53	23.45	23.45	9.54
TopRankedSynonym	10.04	10.04	19.82	19.82	15.21	15.21	27.99	27.99	12.25
WeightedSense	7.50	7.50	13.46	13.46	**20.54**	**20.54**	**35.55**	**35.55**	14.28
Hintz and Biemann[a]	**11.20**	**11.10**	**24.28**	**24.21**	19.49	19.31	33.99	33.89	**15.96**
Jackov	6.73	6.45	13.36	12.86	20.14	19.32	33.18	31.92	11.26

[a] System co-authored by one of the task organizers

Table 2: Baseline and system results for the *best*, *OOT*, and *GAP* metrics

Apart from the English WordNet, the author employs a custom-built machine translation system and a dependency relation knowledge base. The approach first disambiguates the input text by tentatively mapping the lemmatized German words to concepts represented by WordNet synsets. Each parsing hypothesis is scored with reference to a knowledge base of dependency relations; the synonyms and hypernyms of the target concept in the highest-scoring parsing hypothesis are taken as the substitution candidates.

6 Baselines

In addition to the dedicated lexical substitution systems described in the previous section, we implemented three simple baselines, at least two of which have been used in previous lexical substitution tasks:

RandomSense selects a random sense of the target word from GermaNet and returns its synonyms, followed by its hypernyms, in the same order as retrieved from the GermaNet API.

TopRankedSynonyms (McCarthy and Navigli, 2009) builds a list of substitutes in the following order:

1. Synonyms from the first synset of the target word, ranked according to their frequency in a large corpus.
2. Synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) from the first synset, ranked according to their frequency in a large corpus.
3. Synonyms from all other synsets of the target word, ranked according to their frequency in a large corpus.
4. Synonyms from the hypernyms (verbs and nouns) or closely related classes (adjectives) of all other synsets of the target word, ranked according to their frequency in a large corpus.

WeightedSense (Toral, 2009) uses multiple lexical-semantic resources to build the list of candidates. In our case, we use GermaNet and Wiktionary to extract all synonyms and hypernyms of the target word. Synonyms are given a weight of 3, and hypernyms a weight of 1. If a substitute is extracted more than once (i.e., from different synsets or resources), the weights are summed. The list is then ordered by descending weight.

7 Results

Table 2 shows the baseline and participating systems' results for the various *best*, *OOT*, and *GAP* metrics, represented as percentages, on the test set. For each metric, the score for the best-performing system or baseline is set in boldface. Unsurprisingly, RandomSense is the worst-performing baseline. TopRankedSynonym performs best among the baselines by the *best* methodology and the WeightedSense baseline performs best according to both the *OOT* and the *GAP* methodologies.

With respect to the participants' systems, we observe that Hintz and Biemann's entry greatly outperforms Jackov's on the *best* and *GAP* metrics. In fact, the latter fails to beat even the baseline systems in *best*, pointing to the lack of an appropriate substitute ranking scheme. However, for *OOT*,

	System	Best P	Best R	Best Mode P	Best Mode R	OOT P	OOT R	OOT Mode P	OOT Mode R	GAP
adjectives	RandomSense	7.31	7.31	17.18	17.18	16.01	16.01	35.58	35.58	11.87
	TopRankedSynonym	9.63	9.63	23.31	23.31	16.01	16.01	35.58	35.58	13.85
	WeightedSense	6.10	6.10	11.66	11.66	20.73	**20.73**	**42.33**	**42.33**	15.06
	Hintz and Biemann[a]	**14.20**	**13.69**	**36.02**	**35.58**	**21.29**	20.53	42.24	41.72	**18.86**
	Jackov	4.58	4.07	10.20	9.20	16.94	15.04	29.93	26.99	7.35
nouns	RandomSense	8.42	8.42	14.02	14.02	16.79	16.79	23.78	23.78	12.46
	TopRankedSynonym	**12.73**	**12.73**	20.73	20.73	18.56	18.56	26.22	26.22	15.96
	WeightedSense	8.80	8.80	9.76	9.76	24.43	24.43	35.37	35.37	16.32
	Hintz and Biemann[a]	11.11	11.11	**21.95**	**21.95**	**26.38**	**26.38**	**39.63**	**39.63**	**19.61**
	Jackov	10.58	10.49	17.90	17.68	21.61	21.44	31.48	31.10	14.62
verbs	RandomSense	6.93	6.93	14.67	14.67	8.65	8.65	17.37	17.37	6.92
	TopRankedSynonym	8.89	8.89	17.66	17.66	13.14	13.14	25.15	25.15	9.59
	WeightedSense	7.55	7.55	16.17	16.17	18.50	18.50	32.34	32.34	**12.87**
	Hintz and Biemann[a]	**9.80**	**9.80**	**19.76**	**19.76**	15.17	15.17	27.25	27.25	12.69
	Jackov	5.75	5.62	12.54	12.28	**20.86**	**20.40**	**35.47**	**34.73**	11.53

[a] System co-authored by one of the task organizers

Table 3: Baseline and system results for the *best*, *OOT*, and *GAP* metrics, by part of speech

System	Best P	Best R	Best Mode P	Best Mode R	OOT P	OOT R	OOT Mode P	OOT Mode R
Yuret	**12.90**	**12.90**	20.65	20.65	46.15	46.15	61.30	61.30
Hassan et al.	12.77	12.77	**20.73**	**20.73**	49.19	49.19	**66.26**	**66.26**
Giuliano et al.	6.95	6.94	20.33	20.33	**69.03**	**68.90**	58.54	58.54
TopRankedSynonym	9.95	9.95	15.28	15.28	29.70	29.35	40.57	40.57

Table 4: Top-performing baseline and system results for SemEval-2007

System	Best P	Best R	Best Mode P	Best Mode R	OOT P	OOT R	OOT Mode P	OOT Mode R
Basile and Semeraro	8.16	7.18	10.58	10.58	**41.46**	**36.50**	**47.23**	**47.23**
WeightedSense[a]	**10.86**	**9.06**	**13.94**	**13.94**	23.00	19.20	26.97	26.97
WeightedSense[b]	9.71	8.19	13.16	13.16	27.52	23.23	37.24	32.39

[a] CLIPS only
[b] CLIPS and ItalWordNet

Table 5: Top-performing baseline and system results for EVALITA 2009

Jackov's performance is on par with, and occasionally exceeds, that of Hintz and Biemann. Neither system was able to beat the WeightedSense baseline for any of the metrics in *OOT*.

When broken down by part of speech (see Table 3), we observe that scores of the best-performing systems are generally higher for adjectives and nouns, but lower for verbs. It has long been known that verbs are the hardest category of words to process in traditional WSD (Agirre and Stevenson, 2007); it seems this holds for lexical substitution as well. The part-of-speech breakdown also allows us to see that some systems perform better, relative to the others, for different word categories. Of particular note is the TopRankedSynonym baseline's high precision and recall scores for nouns in the *best* methodology, and Jackov's outstanding performance on verbs across all *OOT* metrics. An optimal lexical substitution system may therefore benefit from adapting its strategy according to the target's part of speech.

We also performed an analysis of the relationship between system scores and target word frequency using the Pearson product-moment correlation coefficient. For each combination of system and scoring metric we observed only a negligible negative correlation ($-0.188 \leq r \leq -0.003$). The correlation between system scores and target word polysemy was also computed; this was weak at best ($-0.219 \leq r \leq -0.039$).

7.1 Comparison to SemEval and EVALITA

As previously mentioned, the English SemEval-2007 and Italian EVALITA 2009 shared tasks use similar data sets to our own, as well as some of the same baselines and evaluation methodologies. It is therefore interesting to compare the results of these baselines, and those of their top-performing systems, to our own.

Table 4 shows the results of the best-performing SemEval-2007 system for each of the *best* and *OOT* metrics (Yuret, 2007; Hassan et al., 2007; Giuliano et al., 2007). Also shown there are results for their TopRankedSynonym baseline, which uses WordNet 2.1. Again, for each column the best-performing system or baseline is set in boldface. We observe that the GermaNet-based TopRankedSynonyms baseline performs slightly better than its English counterpart for all the *best* metrics, but significantly worse for all the *OOT* metrics. As in GermEval 2015: LexSub, at least one participating system was able to beat the TopRankedSynonym baseline for any given metric. However, the relative improvement over the baseline was dramatically higher in the English-language task (29.6% to 132.4% in SemEval as compared to 10.2% to 37.6% in GermEval).

Table 5 shows a corresponding results table for the EVALITA 2009 shared task. Here we report scores for two implementations of the WeightedSense baseline; the first uses only the CLIPS lexical-semantic resource (Ruimy et al., 2002), whereas the second, like our own WeightedSense, uses two resources: CLIPS and ItalWordNet. The top-performing participating system here was one submitted by Basile and Semeraro (2009). As in GermEval, in EVALITA scores for the WeightedSense baseline frequently exceeded those of the participating systems. Interestingly, the circumstances under which this occurred were quite different: in GermEval, WeightedSense bested the participating systems for most of the *OOT* metrics, whereas in EVALITA, it was the *best* metrics in which the baseline excelled. German systems may be performing worse due to a lack of lexical coverage in GermaNet, or possibly, as Hintz and Biemann (2015) speculate, because its graph structure makes its lexical items harder to discover.

8 Concluding remarks

In this paper we have introduced GermEval 2015: LexSub, the first lexical substitution task using German text, and presented the results of three baselines and two participating systems. Due to the very low number of participating systems compared with previous lexical substitution tasks in other languages, it is difficult to draw any firm conclusions concerning the efficacy of the different approaches. On the one hand, one of the systems has shown that techniques proven to work well for English-language lexical substitution can work well for German too. But on the other hand, the second system, taking a completely novel approach, had comparable performance much of the time, and the rest of the time seemed to be held back only by its substitute ranking criteria.

Compared with previous lexical substitution tasks, our absolute scores in the *best* metrics were in about the same range, though relative to the baselines they were much lower than in SemEval and much higher than in EVALITA. Unlike in the English and Italian tasks, our participants' systems

had trouble beating the baselines for *OOT*, suggesting that the problem may be lack of lexical coverage in German language resources, or the systems' inability to exploit this coverage.

Acknowledgments

The GermEval 2015: LexSub shared task was supported by the DFG-funded project "Integrating Collaborative and Linguistic Resources for Word Sense Disambiguation and Semantic Role Labeling" (InCoRe, GU 798/9-1), the BMBF-funded Common Language Resources and Technologies Infrastructure (CLARIN, F-AG7), and DFG-funded research training group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES, GRK 1994/1).

References

Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Eneko Agirre and Mark Stevenson. Knowledge sources for WSD. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226, 2009. ISSN 1574-020X.

Pierpaolo Basile and Giovanni Semeraro, Baroni. UNIBA @ EVALITA 2009 lexical substitution task. In *Proceedings of EVALITA 2009*, 2009.

Chris Biemann and Martin Riedl. Text: Now in 2D! A framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95, 2013. ISSN 2299-856X.

Marine Carpuat and Dekai Wu. Evaluating the word sense disambiguation performance of statistical machine translation. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP 2005)*, 2005a.

Marine Carpuat and Dekai Wu. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43th Annual Meeting of the Association of Computational Linguistics (System Demonstrations) (ACL 2005)*, 2005b.

Marine Carpuat and Dekai Wu. Improving statistical machine translation using word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL 2007)*, June 2007.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL 2007)*, 2007.

Kostadin Cholakov, Chris Biemann, Judith Eckle-Kohler, and Iryna Gurevych. Lexical substitution dataset for German. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, pages 2524–2531, 2014.

Paul Clough and Mark Stevenson. Evaluating the contribution of EuroWordNet and word sense disambiguation to cross-language retrieval. In *Proceedings of the 2nd International Conference of the Global WordNet Association (GWC 2004)*, pages 97–105, 2004.

Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998. ISBN 978-0-262-06197-1.

Claudio Giuliano, Alfio Gliozzo, and Carlo Strapparava. FBK-irst: Lexical substitution task exploiting domain and syntagmatic coherence. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 145–148, 2007.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. UBY – A large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, April 2012.

Birgit Hamp and Helmut Feldweg. GermaNet – A lexical-semantic net for German. In *Proceedings of the ACL Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15, 1997.

Samer Hassan, Andras Csomai, Carmen Banea, Ravi Sinha, and Rada Mihalcea. UNT: SubFinder: Combining knowledge sources for automatic lexical substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 410–413, 2007.

Verena Henrich and Erhard Hinrichs. GernEdiT – The GermaNet editing tool. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 2228–2235, May 2010.

Verena Henrich and Erhard Hinrichs. Extending the TüBa-D/Z treebank with GermaNet sense annotation. In Iryna Gurevych, Chris Biemann, and Torsten

Zesch, editors, *Language Processing and Knowledge in the Web: 25th International Conference, GSCL 2013*, volume 8105 of *Lecture Notes in Artificial Intelligence*, pages 89–96. Springer, 2013.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. WebCAGe – A Web-harvested corpus annotated with GermaNet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 387–396, April 2012.

Gerold Hintz and Chris Biemann. Delexicalized supervised German lexical substitution. In *Proceedings of GermEval 2015: LexSub*, pages 11–16, September 2015.

Nancy Ide and Jean Véronis. Introduction to the special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40, 1998. ISSN 0891-2017.

Luchezar Jackov. Lexical substitution using deep syntactic and semantic analysis. In *Proceedings of GermEval 2015: LexSub*, pages 17–20, September 2015.

Kazuaki Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. Technical Report NII-2005-014E, National Institute of Informatics, Tokyo, Japan, October 2005.

Diana McCarthy. Lexical substitution as a task for WSD evaluation. In *Proceedings of the SIGLEX/ SENSEVAL Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pages 109–115, 2002.

Diana McCarthy and Roberto Navigli. SemEval-2007 Task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, 2007.

Diana McCarthy and Roberto Navigli. The English lexical substitution task. *Language Resources and Evaluation*, 43(2):139–159, 2009. ISSN 1574-020X.

Rada Mihalcea, Ravi Sinha, and Diana McCarthy. SemEval-2010 Task 2: Cross-lingual lexical substitution. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 9–14, July 2010.

Martha Palmer, Hwee Tou Ng, and Hoa Trang Dang. Evaluation of WSD systems. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*, volume 33 of *Text, Speech, and Language Technology*. Springer, 2007. ISBN 978-1-4020-6870-6.

Adriana Roventini, Antonietta Alonge, Nicoletta Calzolari, Bernardo Magnini, and Francesca Bertagna. ItalWordNet: A large semantic database for Italian. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, pages 783–790, 2000.

Nilda Ruimy, Monica Monachini, Raffaella Distante, Elisabetta Guazzini, Stefano Molino, Marisa Ulivieri, Nicoletta Calzolari, and Antonio Zampolli. Clips, a multi-level Italian computational lexicon: A glimpse to data. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, 2002.

Mark Sanderson. Word sense disambiguation and information retrieval. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1994)*, pages 142–151, 1994.

Hinrich Schütze and Jan O. Pedersen. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.

György Szarvas, Chris Biemann, and Iryna Gurevych. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 10th Conference of the North American Chapter of the Association for Computational Linguistics and the 18th Human Language Technologies Conference (NAACL–HLT 2013)*, pages 1131–1141, 2013.

Antonio Toral. The lexical substitution task at EVALITA 2009. In *Proceedings of EVALITA 2009*, 2009.

Deniz Yuret. Ku: Word sense disambiguation by substitution. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 207–214, 2007.

Zhi Zhong and Hwee Tou Ng. Word sense disambiguation improves information retrieval. In *Proceedings of the 50th Annual Meeting of the Association of Computational Linguistics (ACL 2012)*, pages 273–282, July 2012.