IJCLA VOL. 3, NO. 2, JUL-DEC 2012, PP. 47–62
RECEIVED 31/10/11 ACCEPTED 09/12/11 FINAL 28/09/12

An Automatic Method for Creating a Sense-Annotated Corpus Harvested from the Web

VERENA HENRICH, ERHARD HINRICHS, AND TATIANA VODOLAZOVA

University of Tübingen, Germany

ABSTRACT

This paper reports on an automatic and language-independent method for compiling a sense-annotated corpus of web data. To validate its language-independence, the method has been applied to English and German. The sense inventories are taken from the Princeton WordNet for English and from the German wordnet GermaNet. The web-harvesting utilizes existing mappings of WordNet and GermaNet to the English and German versions of the web-based dictionary Wiktionary, respectively. The data obtained by this method have resulted in the English WebCAP (short for: Web-Harvested Corpus Annotated with Princeton WordNet Senses) and the German WebCAGe (short for: Web-Harvested Corpus Annotated with GermaNet Senses) resources.

KEYWORDS: *Sense-annotated corpus, sense-tagged corpus, word sense disambiguation, WSD, Princeton WordNet, GermaNet, Wictionary*

1 INTRODUCTION

Sense-annotated corpora are an important resource for a variety of natural language processing tasks including word sense disambiguation, machine translation, and information retrieval. In past resource building, sense-annotated corpora have typically been constructed manually. This has

made the compilation of such resources costly and has put a natural limit on the size of such data sets. This in turn suggests that alternatives to manual annotation need to be explored and automatic, language-independent means of creating sense-annotated corpora need to be investigated. The purpose of this paper is therefore threefold:

1. To propose an automatic method for harvesting and sense-annotating data from the web.
2. To prove the viability and the language-independence of the proposed approach.
3. To make the resulting sense-annotated corpora freely available for other researchers.

The proposed method relies on the following resources as input: (i) a sense inventory and (ii) a mapping between the sense inventory in question and a web-based resource such as Wiktionary[1] or Wikipedia[2].

As a proof of concept and to validate its language-independence, this automatic method has been applied to two languages: To English, a language for which several sense-annotated corpora are already available, as well as to German, a language for which sense-annotated corpora are still in short supply. The sense inventories are taken from the Princeton WordNet for English [1] and from the German wordnet GermaNet [2, 3]. In order to be able to compare the resulting resources for the two languages, the web-harvesting for both languages relies on existing mappings of the wordnets in question with the English and German versions of the web-based dictionary Wiktionary described in [4] and [5], respectively. The resulting resources consist of the web-harvested corpora WebCAP (short for: *Web-Harvested Corpus Annotated with Princeton WordNet Senses*) and WebCAGe (short for: *Web-Harvested Corpus Annotated with GermaNet Senses*). These resources will be made freely available.[3]

The remainder of this paper is structured as follows: An overview of related work is given in Section 2. Section 3 introduces the three resources WordNet, GermaNet, and Wiktionary used in the present research. The algorithm for automatically harvesting and sense-annotating

[1] `http://www.wiktionary.org/`

[2] `http://www.wikipedia.org/`

[3] See `http://www.sfs.uni-tuebingen.de/en/general-and-computational-linguistics/resources/corpora/webcap` and `http://www.sfs.uni-tuebingen.de/en/general-and-computational-linguistics/resources/corpora/webcage`

textual materials from the web is described in Section 4. Section 5 evaluates the proposed approach applied to English and German, and compares the results for the two languages. Finally, the paper concludes with a summary of the results and with an outlook to future work in Section 6.

2 Related Work

With relatively few exceptions to be discussed shortly, the construction of sense-annotated corpora has focussed on purely manual methods. This is true for SemCor, the WordNet Gloss Corpus, and for the training sets constructed for English as part of the SensEval and SemEval shared task competitions [6–8]. Purely manual methods were also used for the German sense-annotated corpora constructed by Broscheit et al. [9] and Raileanu et al. [10] as well as for other languages including the Bulgarian and the Chinese sense-tagged corpora [11, 12]. The only previous attempts of harvesting corpus data for the purposes of constructing a sense-annotated corpus is the semi-supervised method developed by Yarowsky [13], the knowledge-based approach of Leacock et al. [14], later also used by Agirre and Lopez de Lacalle [15], and the automatic association of Web directories (from the Open Directory Project, ODP) to WordNet senses by Santamaría et al. [16].

The latter study [16] is closest in spirit to the approach presented here. It also relies on an automatic mapping between WordNet senses and a second web resource. While our approach is based on automatic mappings between WordNet/GermaNet and Wiktionary, their mapping algorithm maps WordNet senses to ODP subdirectories. Since these ODP subdirectories contain natural language descriptions of websites relevant to the subdirectory in question, this textual material can be used for harvesting sense-specific examples.

The approach of Yarowsky [13] first collects all example sentences that contain a polysemous word from a very large corpus. In a second step, a small number of examples that are representative for each of the senses of the polysemous target word is selected from the large corpus created in step 1. These representative examples are manually sense-annotated and then fed into a decision-list supervised WSD algorithm as a seed set for iteratively disambiguating the remaining examples collected in step 1. The selection and annotation of the representative examples in Yarowsky's approach is performed completely manually and is therefore limited to the amount of data that can reasonably be annotated by hand.

Leacock et al., Agirre and Lopez de Lacalle, and Mihalcea and Moldovan [14, 15, 17] propose a set of methods for automatic harvesting of web data for the purposes of creating sense-annotated corpora. By focusing on web-based data, their work resembles the research described in the present paper. However, the underlying harvesting methods differ. While our approach relies on a wordnet to Wiktionary mapping, their approaches all rely on the monosemous relative heuristic. Their heuristic works as follows: In order to harvest corpus examples for a polysemous word, the WordNet relations such as synonymy and hypernymy are inspected for the presence of unambiguous words, i.e., words that only appear in exactly one synset. The examples found for these monosemous relatives can then be sense-annotated with the particular sense of its ambiguous word relative. In order to increase coverage of the monosemous relatives approach, Mihalcea and Moldovan [17] have developed a gloss-based extension, which relies on word overlap of the gloss and the WordNet sense in question for all those cases where a monosemous relative is not contained in the WordNet dataset.

The approaches of Leacock et al., Agirre and Lopez de Lacalle, and Mihalcea and Moldovan as well as Yarowsky's approach provide interesting directions for further enhancing the WebCAP and WebCAGe resources (for some preliminary discussion on such an integration see Section 6 below).

In our own previous research, we have addressed the issue of automatically creating sense-annotated corpora for German. The creation of the resource WebCAGe described in the present paper relies on a mapping between GermaNet and the German Wiktionary [5] and is based on an earlier study [18]. With WikiCAGe, we have built a second sense-annotated corpus for German [19]. It consists of examples harvested from the German Wikipedia and was constructed by means of an automatic mapping between GermaNet and the German Wikipedia.

3 RESOURCES

3.1 *WordNet and GermaNet*

Both the Princeton WordNet for English [1] and the German wordnet GermaNet [2, 3] are lexical semantic networks that partition the lexical space into sets of concepts that are interlinked by semantic relations such as hypernymy, part-whole relations, entailment, causation, or antonymy. Wordnets are hierarchically structured in terms of the hypernymy relation. A semantic concept is modeled by a *synset*. A synset is a set of

words (called *lexical units*) where all the words are taken to have (almost) the same meaning. Thus a synset is a set-representation of the semantic relation of synonymy, which means that it consists of a list of lexical units.

The Princeton WordNet has served as inspiration and as best practice example for the construction of GermaNet as well as for the creation of other wordnets for a large number of typology diverse languages.[4]

The coverage of the Princeton WordNet includes the four word classes of adjectives, adverbs, nouns, and verbs. Its release 3.0 covers 206,941 word senses, which are grouped into 117,659 synsets. GermaNet covers the three word classes of adjectives, nouns, and verbs. GermaNet's version 6.0 (release of April 2011) covers 93,407 lexical units, which are grouped into 69,594 synsets.

3.2 *Wiktionary*

Wiktionary is a web-based dictionary that is available for many languages, including English and German. As is the case for its sister project Wikipedia, Wiktionary is constructed by contributions of a large number of volunteers and is freely available. The dictionary provides information such as part-of-speech, hyphenation, possible translations, inflection, etc. for each word. It covers, among others, the word categories of adjectives, adverbs, nouns, and verbs. Distinct word senses are distinguished by sense descriptions, accompanied with example sentences illustrating the usage of the sense in question. Further, Wiktionary provides relations to other words, e.g., in the form of synonyms, antonyms, hypernyms, hyponyms, holonyms, and meronyms. Different from WordNet and GermaNet, the relations are (mostly) not disambiguated.

Since Wiktionary is a dynamic resource, it is important to clearly identify the versions used for the present research. The construction of WebCAP is based on a dump of the English Wiktionary as of April 3, 2010, which consists of 335,748 English words comprising 421,847 word senses [4]. For WebCAGe, the German Wiktionary as of February 2, 2011 is utilized, consisting of 46,457 German words and 70,339 word senses [5]. The Wiktionary data is extracted by the freely available Java-based library JWKTL[5].

[4] See `http://www.globalwordnet.org/` for an informative overview.
[5] `http://www.ukp.tu-darmstadt.de/software/jwktl`

4 Creating a Sense-Annotated Corpus Harvested from the Web

The starting point for creating the English WebCAP (short for: *Web-Harvested Corpus Annotated with Princeton WordNet Senses*) and the German WebCAGe (short for: *Web-Harvested Corpus Annotated with GermaNet Senses*) resources are existing mappings of senses in WordNet and GermaNet with Wiktionary senses as described in [4] and [5], respectively. These mappings were created by automatic word sense alignment algorithms with high accuracy: 91.5% for English [4] and 93.8% for German [5]. For German, a manual post-correction step of the automatic alignment was performed that further improved the accuracy of the mapping.

4.1 *Web-Harvesting Sense-Annotated Materials*

Fig. 1 illustrates the existing WordNet-Wiktionary mapping using the example word *crutch*. The polysemous word *crutch* has two distinct senses in WordNet which directly correspond to two separate senses in the English Wiktionary[6]. Each Wiktionary sense entry contains a definition and one or more example sentences illustrating the sense in question. Since the target word in the example sentences for a particular Wiktionary sense (rendered in Fig. 1 in bold face) is linked to a WordNet sense via the sense mapping of WordNet to Wiktionary, the example sentences are automatically sense-annotated and can be included as part of WebCAP.

An example for the GermaNet-Wiktionary mapping using the example word *Option* is given in Fig. 2. As is the case for the English example *crutch*, the polysemous word *Option* has two distinct senses in GermaNet which directly correspond to two separate senses in the German Wiktionary. Again, each Wiktionary sense contains one or more example sentences, which can directly be mapped to a specific sense in GermaNet and thus be sense-annotated and included in WebCAGe. Furthermore, the examples in turn are linked to external references, including sentences contained in Wikipedia articles (see link in the second Wiktionary sense entry in Fig. 2) and in other web-based textual sources such as online newspaper materials and the German Gutenberg text archive[7] (see the topmost sense entry in Fig. 2).

[6] Note that there is one further sense in Wiktionary not displayed here for reasons of space.

[7] `http://gutenberg.spiegel.de/`

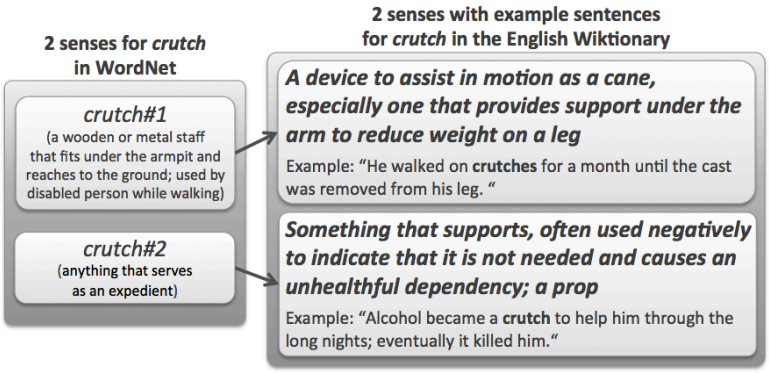

Fig. 1. Sense mapping of WordNet and Wiktionary using the example of *crutch*.

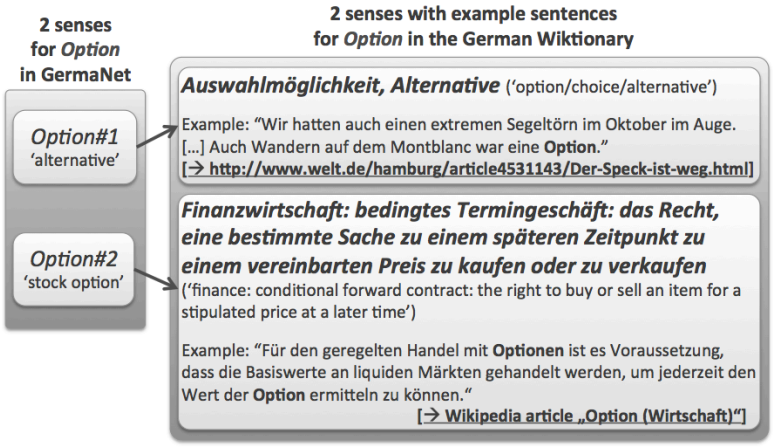

Fig. 2. Sense mapping of GermaNet and Wiktionary using the example of *Option*.

Additional data for WebCAGe is harvested by following the links to Wikipedia and other web-based resources referenced by Wiktionary. Since these links belong to particular Wiktionary sense entries that in turn are mapped to GermaNet senses, the target words contained in these materials are automatically sense-annotated.

Notice that the target word often occurs more than once in a given text. In keeping with the widely used heuristic of "one sense per discourse", multiple occurrences of a target word in a given text are all as-

signed to the same wordnet sense. An inspection of the annotated data shows that this heuristic proves to be highly reliable in practice.[8]

WebCAP and WebCAGe are developed primarily for the purpose of the word sense disambiguation task. Therefore, only those target words that are ambiguous are included in these resources. For the German WebCAGe, this means that each target word has at least two GermaNet senses, i.e., belongs to at least two distinct synsets in GermaNet. For the English WebCAP, each target word has at least two senses in WordNet regardless of word class; i.e., the target word belongs to at least two distinct synsets in WordNet which may belong to more than one word class. Taking into account polysemy across word classes is important for English. In contrast to German, this type of conversion involving the same orthography for different word classes with possibly distinct meanings is a frequent phenomenon in English.

Both the WordNet-Wiktionary and the GermaNet-Wiktionary mappings are not always one-to-one. For example, sometimes one WordNet/GermaNet sense is mapped to more than one sense in Wiktionary. In those cases, all example sentences from all mapped Wiktionary senses are assigned to the WordNet/GermaNet sense in question.

4.2 *Target Word Identification*

The next step for creating a sense-annotated corpus is the target word identification. For highly inflected languages such as German, target word identification is more complex compared to languages with a simplified inflectional morphology such as English and requires automatic lemmatization. Moreover, the target word in a text to be sense-annotated is not always a simplex word, but can also appear as subpart of a complex word such as a compound. Since the constituent parts of a compound are not separated by blank spaces or hyphens, German compounding poses a particular challenge for target word identification. Another challenging case for automatic target word detection in German concerns particle verbs such as *an-kündigen* 'announce'. Here, the difficulty arises when the verbal stem (e.g., *kündigen*) is separated from its particle (e.g., *an*) in German verb-initial and verb-second clause types.

[8] Henrich et al. [18] show that for German the heuristic works correctly in 99.96% of all target word occurrences in the Wiktionary example sentences, in 96.75% of all occurrences in the external webpages, and in 95.62% of the Wikipedia files.

```
Radioaktivität, radioaktiver <tag luids="l88831" lemma="Zerfall"
1 wcat="NN">Zerfall</tag> oder Kern<tag luids="l88831"
2 lemma="Zerfall" wcat="NN">zerfall</tag> ist die Eigenschaft
  instabiler Atomkerne, sich spontan unter Energieabgabe
  umzuwandeln. [...]

  Der Zeitpunkt eines radioaktiven <tag luids="l88831"
3 lemma="Zerfall" wcat="NN">Zerfalls</tag> ist im Voraus nicht
  bestimmbar. [...]

  Im Allgemeinen sind die <tag luids="l88831" lemma="Zerfall"
4 wcat="NN">Zerfall</tag>sprodukte nicht stabil. In den meisten
  Fällen sind die Tochterkerne ihrerseits wieder radioaktiv und
  zerfallen gemäß ihrer eigenen Halbwertszeiten. Auf diese Weise
  entsteht eine Abfolge von radioaktiven <tag luids="l88831"
5 lemma="Zerfall" wcat="NN">Zerfällen</tag>, bis schließlich ein
  stabiler Kern als Endprodukt übrig bleibt. Diese Aufeinander-
  folge radioaktiver <tag luids="l88831" lemma="Zerfall"
6 wcat="NN">Zerfälle</tag> heißt <tag luids="l88831"
7 lemma="Zerfall" wcat="NN">Zerfall</tag>sreihe.[...]

            Source: http://de.wikipedia.org/wiki/Radioaktivität
```

Fig. 3. Excerpt from Wikipedia article *Radioaktivität* 'radioactivity' tagged with the target word *Zerfall* 'radioactive decay'.

As a preprocessing step for target word identification, the web-harvested texts are split into individual sentences, tokenized, and lemmatized. For this purpose, the sentence detector and the tokenizer of the suite of Apache OpenNLP tools[9] and the TreeTagger [20] are used both for English and German. Further, for German, compounds are split by using BananaSplit[10]. Since the automatic lemmatization obtained by the tagger (and the compound splitter) are not a 100% accurate, target word identification also utilizes the full set of inflected forms for a target word whenever such information is available in Wiktionary.

Fig. 3 shows a German example of a sense-annotated text for the target word *Zerfall* in the sense of 'radioactive decay'. The text is an excerpt from the Wikipedia article *Radioaktivität* 'radioactivity' and contains many occurrences of the target word (rendered in bold face). Only the first occurrence shown in Fig. 3 (marked with a `1` on the left margin) exactly matches the word *Zerfall* as is. All other occurrences are either the genitive form *Zerfalls* (occurrence `3`), the genitive plural *Zerfälle* (occurrence `6`), the dative plural *Zerfällen* (occurrence `5`), or part of a compound such as *Kernzerfall*, *Zerfallsprodukte*, or *Zerfallsreihe* (occurrences `2`, `4`, and `7`).

[9] `http://incubator.apache.org/opennlp/`
[10] `http://niels.drni.de/s9y/pages/bananasplit.html`

4.3 *Data Encoding*

For expository purposes, the data format shown in Fig. 3 has been simplified compared to the actual XML data encoding used for both WebCAP and WebCAGe. This data encoding is inspired by the best practise format for sense-annotated corpora established by the sense-annotated corpora used in the SensEval and SemEval shared task competitions [6–8].

Fig. 3 illustrates the information provided for each sense-annotated target word in WebCAGe: (i) a sense ID referring to a lexical unit in GermaNet, (ii) the lemma of the target word, and (iii) the word class of the target word. The target word information in WebCAP following exactly the same data format. However, in the case of WebCAP, the sense information of each target word points to a WordNet synset rather than a WordNet lexical unit. The reason for this difference in encoding stems from the WordNet/GermaNet-Wiktionary mappings: The WordNet-Wiktionary mapping links synset IDs in WordNet to Wiktionary senses, whereas the GermaNet-Wiktionary mapping links lexical unit IDs in GermaNet to Wiktionary senses.

5 Evaluation and Discussion of the Results

In order to assess the effectiveness of the approach, we examine and compare the overall sizes of WebCAP and WebCAGe (see Table 1) and present a precision and recall based evaluation for the algorithm that is used for automatically identifying the target words in the harvested texts (see Table 2).

The target words in WebCAP belong to 3628 distinct polysemous words contained in WordNet, among which there are 934 adjectives, 174 adverbs, 1480 nouns, and 1040 verbs. These words have on average 3.7 senses in WordNet (1.9 for adjectives, 2.6 for adverbs, 4.1 for nouns, and 5.0 for verbs). The target words in WebCAGe belong to 2607 distinct polysemous words contained in GermaNet (211 adjectives, 1499 nouns, and 897 verbs) which have on average 2.9 senses in GermaNet (2.4 for adjectives, 2.6 for nouns, and 3.6 for verbs).

Table 1 shows the overall sizes of WebCAP and WebCAGe: The numbers of tagged word tokens (i.e., the target word occurrences), the number of sentences containing those tags, and the number of overall sentences (i.e., all sentences in the corpora including those where no target word has been tagged) separately for the four word classes of adjectives, adverbs, nouns, and verbs. The numbers for WebCAP describe the Wiktionary

example sentences only, whereas the numbers for WebCAGe are given separately for the Wiktionary example sentences (in order to be comparable with WebCAP), for the external materials, and overall (the sum of the Wiktionary example sentences and the external materials). WebCAGe contains a total of 10750 tagged word tokens whereas WebCAP only contains 6526 word tokens. Even if we compare the numbers of the Wiktionary example sentences in WebCAP (6526 tagged word tokens) with those in WebCAGe (7644 tagged word tokens), i.e., excluding the external materials from WebCAGe, the German resource is larger than the English one. This is especially astonishing considering that the English input resources constitute a multiple of their German counterparts: The Princeton WordNet contains 1.7 times as many word senses as GermaNet and the English Wiktionary contains 6 times as many word senses as the German Wiktionary (see Section 3). The explanation for the German Wiktionary examples outnumbering those for English has to do with the online instructions given to Wiktionary contributors for English. For the English Wiktionary, contributors are asked to accompany each word sense definition by a quotation that illustrates the definition in question and to compose example sentences on their own only if no suitable quotation sentence can be found.[11] Accordingly, the English Wiktionary contains fewer example sentences compared to German.

According to the guidelines for the English Wiktionary, a *quotation* is an attested example taken from a literary work or from some other published textual material. Such quotations are accompanied by the appropriate reference to their textual source. The version of the API that was used to extract the Wiktionary data does not support the harvesting of the quotations themselves and the textual sources from which those quotations are taken. We anticipate that the size of WebCAP would increase significantly if the harvesting functionality is extended to the set of quotations that contributors are encouraged to provide for each sense definition. For the German Wiktionary, the situation is different in that example sentences are a mixture of made-up materials and attested examples that are often cross-referenced with their online sources and can thus be harvested automatically by the API.

It is also noticeable that the relative numbers of the different word classes are rather equally distributed in WebCAP, whereas there are con-

[11] See `http://en.wiktionary.org/wiki/Wiktionary:Entry_layout_explained` for the relevant instructions.

Table 1. Current sizes of WebCAP and WebCAGe.

		WebCAP	WebCAGe		
		Wiktionary examples	**Wiktionary examples**	**External materials**	**All texts**
Number of tagged word tokens	adjectives	1522	575	138	713
	adverbs	311	0	0	0
	nouns	2596	4103	2744	6847
	verbs	2097	2966	224	3190
	all word classes	6526	7644	3106	10750
Number of tagged sentences	adjectives	1488	565	133	698
	adverbs	302	0	0	0
	nouns	2526	3965	2448	6413
	verbs	2056	2945	224	3169
	all word classes	6372	7475	2805	10280
Total number of sentences	adjectives	1578	623	66757	67380
	adverbs	317	0	0	0
	nouns	2726	4184	392640	396824
	verbs	2181	3087	152303	155390
	all word classes	6802	7894	611700	619594

siderably more texts in WebCAGe contributed by nouns than by adjectives and verbs (see Table 1).[12]

Apart from the size of the resources in question, the usefulness of the compiled data sets depends crucially on the quality of the annotated data. WebCAP and WebCAGe are the results of an automatic harvesting method. Such an automatic method will only constitute a viable alternative to the labor-intensive manual method of creating sense-annotated corpora if the results are of sufficient quality so that the harvested data set can be used as is or can be further improved with a minimal amount of manual post-editing. For the purposes of the present evaluation, a precision and recall based analysis was conducted, and the tagged target words are manually verified. For WebCAGe, all textual materials have been manually checked, while for WebCAP, only the first 1,000 Wiktionary example sentences for nouns and the first 500 sentences for adjectives, adverbs, and verbs could be manually verified. Table 2 shows that precision and recall for all word classes are above 97% in WebCAP and above 93% in WebCAGe. The only deviations are the results for verbs

[12] The reason why there are no tagged adverbs in WebCAGe is due to the GermaNet resource which covers adjectives, nouns, and verbs, but no adverbs.

Table 2. Evaluation of the algorithm of identifying the target words.

			WebCAGe		
		WebCAP	**Wiktionary examples**	**External materials**	**All texts**
Precision	adjectives	97.98%	97.70%	98.39%	98.21%
	adverbs	98.68%	–	–	–
	nouns	97.62%	98.17%	95.52%	96.18%
	verbs	97.88%	97.38%	87.37%	89.80%
	all word classes	97.90%	97.32%	93.29%	94.30%
Recall	adjectives	99.19%	97.70%	97.48%	97.54%
	adverbs	99.01%	–	–	–
	nouns	99.27%	98.30%	95.37%	96.10%
	verbs	98.99%	97.51%	96.26%	96.58%
	all word classes	99.16%	97.94%	96.36%	96.01%

that occur in WebCAGe, which are slightly lower than the results for the other word classes. Apart from this one exception, the results in Table 2 prove the viability of the proposed method for automatic harvesting of sense-annotated data. The average precision for all three word classes is of sufficient quality to be used as is if approximately 2-5% noise in the annotated data is acceptable. In order to eliminate such noise, manual post-editing would be required.

6 Conclusion and Future Work

This paper has described an automatic method for harvesting and sense-annotating data from the web. In order to validate the language-independence of the approach, the proposed method has been applied to both English and German. The publication of this paper will be accompanied by making the two sense-annotated corpora WebCAP and WebCAGe freely available. In the case of WebCAGe, the automatic sense-annotation of all target word has been manually verified.

In order to further enlarge the WebCAP and WebCAGe resources, it would be interesting and worthwhile to use the automatically harvested sense-annotated examples as the seed set for Yarowsky's iterative method for creating a large sense-annotated corpus. Another fruitful direction for further automatic expansion of WebCAP and WebCAGe consists of using the heuristic of monosemous relatives used by Leacock et al., by Agirre

and Lopez de Lacalle, and by Mihalcea and Moldovan. However, we have to leave both of these matters for future research.[13]

Finally, we plan to apply our method to further languages. A precondition for such an experiment are existing mappings between the sense inventories in question and web-based resources such as Wiktionary or Wikipedia. With BabelNet, Navigli and Ponzetto [21] have created a multilingual resource that allows the testing of our approach with languages other than English and German.

ACKNOWLEDGEMENTS We are very grateful to Emanuel Dima, Yana Panchenko, Klaus Suttner, and Yannick Versley for their support in obtaining the external web-based materials. We would like to thank Reinhild Barkey, Sarah Schulz, and Johannes Wahle for their help with the evaluation. Special thanks go to Christian M. Meyer, who has provided both the English Wiktionary and the JWKTL API in the same versions that were used for the WordNet-Wiktionary mapping, and to Tristan Miller, who provided helpful input to the final data format of WebCAGe. This work was supported by the CLARIN-D grant of the BMBF and the SFB 833 grant of the DFG.

REFERENCES

1. Fellbaum, C., ed.: WordNet – An Electronic Lexical Database. The MIT Press (1998)
2. Henrich, V., Hinrichs, E.: GernEdiT – the GermaNet editing tool. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta. (2010) 2228–2235
3. Kunze, C., Lemnitzer, L.: GermaNet – representation, visualization, application. In: Proceedings of the 3rd International Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands. (2002) 1485–1491
4. Meyer, C.M., Gurevych, I.: What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In: Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, Thailand. (2011) 883–892
5. Henrich, V., Hinrichs, E., Vodolazova, T.: Semi-automatic extension of GermaNet with sense definitions from Wiktionary. In: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC'11), Poznan, Poland. (2011) 126–130

[13] For a description of these approaches, see Section 2.

6. Agirre, E., Marquez, L., Wicentowski, R.: Proceedings of the 4th International Workshop on Semantic Evaluations. Assoc. for Computational Linguistics, Stroudsburg, PA, USA (2007)
7. Erk, K., Strapparava, C.: Proceedings of the 5th International Workshop on Semantic Evaluation. Assoc. for Computational Linguistics, Stroudsburg, PA, USA (2010)
8. Mihalcea, R., Chklovski, T., Kilgarriff, A.: Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain. Association for Computational Linguistics (2004)
9. Broscheit, S., Frank, A., Jehle, D., Ponzetto, S.P., Rehl, D., Summa, A., Suttner, K., Vola, S.: Rapid bootstrapping of word sense disambiguation resources for German. In: Proceedings of the 10th Konferenz zur Verarbeitung Natürlicher Sprache, Saarbrücken, Germany. (2010) 19–27
10. Raileanu, D., Buitelaar, P., Vintar, S., Bay, J.: Evaluation corpora for sense disambiguation in the medical domain. In: Proceedings of the 3rd International Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands. (2002) 609–612
11. Koeva, S., Leseva, S., Todorova, M.: Bulgarian sense tagged corpus. In: Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, Italy. (2006) 79–87
12. Wu, Y., Jin, P., Zhang, Y., Yu, S.: A Chinese corpus with word sense annotation. In: Proceedings of 21st International Conference on Computer Processing of Oriental Languages (ICCPOL'06), Singapore. (2006) 414–421
13. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics (ACL'95), Stroudsburg, PA, USA, Association for Computational Linguistics (1995) 189–196
14. Leacock, C., Chodorow, M., Miller, G.A.: Using corpus statistics and wordnet relations for sense identification. Computational Linguistics **24**(1) (1998) 147–165
15. Agirre, E., Lopez de Lacalle, O.: Publicly available topic signatures for all WordNet nominal senses. In: Proceedings of the 4th International Conference on Languages Resources and Evaluations (LREC'04), Lisbon, Portugal. (2004) 1123–1126
16. Santamaría, C., Gonzalo, J., Verdejo, F.: Automatic association of web directories to word senses. Computational Linguistics **29**(3) (2003)
17. Mihalcea, R., Moldovan, D.: An automatic method for generating sense tagged corpora. In: Proceedings of the American Association for Artificial Intelligence (AAAI'99), Orlando, Florida. (1999) 461–466
18. Henrich, V., Hinrichs, E., Vodolazova, T.: 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), Avignon, France. (2012) 387–396

19. Henrich, V., Hinrichs, E., Suttner, K.: Automatically linking GermaNet to Wikipedia for harvesting corpus examples for GermaNet senses. Journal for Language Technology and Computational Linguistics (JLCL) **27**(1) (2012) 1–19
20. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing, Manchester, UK. (1994)
21. Navigli, R., Ponzetto, S.P.: BabelNet: Building a very large multilingual semantic network. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10), Uppsala, Sweden. (2010) 216–225

VERENA HENRICH
DEPARTMENT OF LINGUISTICS,
UNIVERSITY OF TÜBINGEN,
WILHELMSTR. 19, 72074 TÜBINGEN, GERMANY
E-MAIL: <VERENA.HENRICH@UNI-TUEBINGEN.DE>

ERHARD HINRICHS
DEPARTMENT OF LINGUISTICS,
UNIVERSITY OF TÜBINGEN,
WILHELMSTR. 19, 72074 TÜBINGEN, GERMANY
E-MAIL: <ERHARD.HINRICHS@UNI-TUEBINGEN.DE>

TATIANA VODOLAZOVA
DEPARTMENT OF LINGUISTICS,
UNIVERSITY OF TÜBINGEN,
WILHELMSTR. 19, 72074 TÜBINGEN, GERMANY
E-MAIL: <TATIANA.VODOLAZOVA@UNI-TUEBINGEN.DE>