Verena Henrich, Erhard Hinrichs, Klaus Suttner

Automatically Linking GermaNet to Wikipedia for Harvesting Corpus Examples for GermaNet Senses

The comprehension of a word sense is much easier when its usages are illustrated by example sentences in linguistic contexts. Hence, examples are crucially important to better understand the sense of a word in a dictionary. The goal of this research is the semi-automatic enrichment of senses from the German wordnet GermaNet with corpus examples from the online encyclopedia Wikipedia. The paper describes the automatic mapping of GermaNet senses to Wikipedia articles, using proven, state-of-the-art word sense disambiguation methods, in particular different versions of word overlap algorithms and PageRank as well as classifiers that combine these methods. This mapping is optimized for precision and then used to automatically harvest corpus examples from Wikipedia for GermaNet senses. The paper presents details about the optimization of the model for the GermaNet-Wikipedia mapping and concludes with a detailed evaluation of the quantity and quality of the harvested examples. Apart from enriching the GermaNet resource, the harvested corpus examples can also be used to construct a corpus of German nouns that are annotated with GermaNet senses. This sense-annotated corpus can be used for a wide range of NLP applications.

1 Introduction

Different senses of a word are often hard to distinguish – not only for second language learners. This is especially the case when a dictionary makes fine-grained sense distinctions for polysemous words (Palmer et al., 2007). Although the usefulness of meaningful sense descriptions for identifying word senses is self-evident, descriptions alone are often not sufficient to discriminate senses. Kilgarriff et al. (2008) point out that humans grasp the sense of a word in a dictionary much easier when example sentences illustrating the usage of a word in context are available. Consequently, corpus examples are crucially important for comprehensive understanding of senses in dictionaries and other lexical resources such as wordnets.

The purpose of this paper is to describe an automatic method for adding corpus examples to the word senses of a wordnet. While the method described is language-independent, the present paper will focus on the German wordnet GermaNet. Using German as a test case is particularly appropriate since – with the exception of its verb entries – GermaNet's word senses are still lacking illustrative example sentences. This gap in coverage is particularly evident in the case of nouns, which have a total of 77 925 word senses in GermaNet (release 6.0) and which are – with few exceptions

– not accompanied by any example sentences. Due to the large number of missing examples, the task of adding them by purely manual, lexicographic work would be at best an arduous task and require considerable effort and person power. Therefore, the possibility of employing automatic or semi-automatic methods for adding corpus examples would be extremely valuable.

Such an automatic method has to rely on an electronically available resource that should ideally satisfy the following criteria: (i) it should be of sufficient size in order to provide the necessary lexical coverage, (ii) since nouns are the focus of the present paper, the resource should have a comprehensive coverage of nominal word senses and a significant overlap in coverage with GermaNet, and (iii) it should be freely available so that the corpus examples harvested from the resource in question can be freely shared. The requirement that word senses are to be mapped to example sentences by automatic means imposes a further restriction on the type of textual material to be used. Such a mapping needs to perform automatic word sense disambiguation so as to ensure that the candidate word senses from GermaNet are mapped to the appropriate example sentences. The precision of this word-sense-to-example mapping should be extremely high so as to be usable with minimal amount of manual post-correction. Such high precision can be realized only if automatic word sense disambiguation can be performed with high reliability. This is, in turn, the case if the texts from which the examples are harvested exhibit a high degree of thematic coherence so as to provide sufficient cues for contextual disambiguation.

If one takes the requirements just mentioned into account, the web-based encyclopedia Wikipedia[1] becomes a natural choice. Its thematic coverage focuses on articles that typically describe nominal concepts and thus provides the type of lexical coverage needed for the present purpose. It is freely available, of sufficient size, and thematically diverse and comprehensive. Moreover, there is a 76.7% overlap in coverage between Wikipedia and the 4 358 polysemous nouns in GermaNet. In addition, the articles attempt to illustrate a particular target concept and are thus thematically highly coherent. This in turn facilitates automatic word sense disambiguation.

In short, the task at hand consists of an automatic mapping of word senses in GermaNet to articles in Wikipedia and the actual harvesting of corpus examples from the linked Wikipedia articles. The nature of the task of harvesting corpus examples for word senses is closely related to the task of creating a sense-annotated corpus. Both tasks focus on harvesting textual materials whose words will be assigned the corresponding word senses of the sense inventory (i.e., wordnet) in question. Because of this close similarity between the two tasks, it is appropriate to combine all harvested corpus examples into a sense-annotated corpus.

In recent years, the use of Wikipedia has gained considerable popularity in empirically oriented research in theoretical and computational linguistics. The present paper wants to contribute to this growing body of research which thus far has mostly focused on English. To the best of our knowledge the present study is the first of its kind for German

[1] http://www.wikipedia.org/

that links word senses in GermaNet to the corresponding articles in Wikipedia. There has been a considerable body of research for English that investigates the alignment of the Princeton WordNet with Wikipedia (see Section 3). However, we are not aware of any other previous research that tries to align the German Wikipedia to GermaNet.

The semi-automatic enrichtment of GermaNet with examples taken from Wikipedia is valuable not only for users of GermaNet, but also for lexicographers involved in the further construction of GermaNet. The Wikipedia examples offer authentic language materials and thereby free lexicographers from having to construct made-up examples that are not validated by actual language corpora.

The remainder of this paper is structured as follows: After a short description of the resources GermaNet and Wikipedia in Section 2, Section 3 provides an overview of related work. Sections 4 and 5 introduce the mapping of GermaNet to Wikipedia and describe how this mapping can be used to automatically harvest corpus examples for GermaNet senses, respectively. The approach is evaluated in Section 6. Finally, there are concluding remarks and an outlook to future work in Section 7.

2 Resources

2.1 GermaNet

GermaNet (Henrich and Hinrichs, 2010; Kunze and Lemnitzer, 2002) is a lexical semantic network that is modeled after the Princeton WordNet for English (Fellbaum, 1998). It represents word meanings by *lexical units* and groups lexical units that express the same semantic concept into *synsets* (*synonymy sets*). Thus, a synset is a set-representation of the semantic relation of synonymy.

Synsets and lexical units are interlinked by two types of semantic relations: by *conceptual* and by *lexical* relations. Conceptual relations hold between two semantic concepts, i.e., synsets. They include relations such as hypernymy, part-whole relations, entailment, or causation. Lexical relations hold between two individual lexical units. Antonymy, a pair of opposites, is an example of a lexical relation.

GermaNet covers the three word categories of adjectives, nouns, and verbs, each of which is hierarchically structured in terms of the hypernymy relation of synsets. The development of GermaNet started in 1997, and is still in progress. GermaNet's version 6.0 (release of April 2011) contains 93 407 lexical units, which are grouped into 69 594 synsets. At present, GermaNet provides comprehensive example sentences only for its verbs senses.

2.2 Wikipedia

Wikipedia is a web-based encyclopedia that is available for many languages, including German. It is written collaboratively by volunteers and is freely available[2]. The general

[2] Wikipedia is available under the Creative Commons Attribution/Share-Alike license `http://creativecommons.org/licenses/by-sa/3.0/deed.en`

structure of a Wikipedia article starts with a paragraph that briefly defines the presented concept. The rest of the article consists of a detailed description optionally containing references that proof the source of the text, hyperlinks to other Wikipedia articles as well as pictures illustrating the described context. Further, the encyclopedia divides its articles into thematic categories. For those words that have multiple articles, Wikipedia provides disambiguation pages with a short description of each article.

For the present project, a dump of the German Wikipedia as of June 21, 2011 is utilized, consisting of 2.27 mio. pages. The Wikipedia data was extracted by the freely available Java-based library JWPL (Zesch et al., 2008).

3 Related Work

As mentioned in the Introduction, the purpose of this paper is to describe an automatic method for adding corpus examples to the word senses of GermaNet. This task is twofold: (i) it involves the automatic mapping of word senses in GermaNet to articles in Wikipedia and (ii) on the basis of this mapping, it harvests corpus examples for GermaNet's senses. Related work for both these tasks is discussed in the following two subsections.

3.1 Mapping Wikipedia to a Wordnet

Several authors have investigated ways of aligning the Princeton WordNet with the English Wikipedia, with some studies focusing on an alignment of Wikipedia categories to WordNet synsets and others investigating the alignment between Wikipedia articles and WordNet. Toral et al. (2009) utilize several text similarity measures to match Wikipedia categories to WordNet synsets. For the same task, Ponzetto and Navigli (2009) apply a knowledge-rich method which maximizes the structural overlap between the WordNet taxonomy and the category graph extracted from Wikipedia.

Other approaches align articles in Wikipedia – instead of categories – with WordNet synsets. In the study of Wolf and Gurevych (2010), the actual alignment between Wikipedia articles and WordNet synsets has been performed manually on the basis of an automatically extracted set of potential sense alignments. A vector-based similarity measure is applied by Ruiz-Casado et al. (2005) to map articles of the Simple English Wikipedia to their most similar WordNet synset. Suchanek et al. (2007) ignore ambiguity while aligning Wikipedia and WordNet and solve ambiguous mappings manually. Ponzetto and Navigli (2010) calculate conditional probabilities relying on a normalized word overlap measure of the textual sense representation. A threshold-based Personalized PageRank to automatically align articles in Wikipedia with synsets in WordNet is utilized by Niemann and Gurevych (2011). The most recent study we are aware of is the one by Fernando and Stevenson (2012), who first compute similarity between WordNet synsets and Wikipedia articles to perform the alignment and then apply heuristics based on the link structure of Wikipedia to refine their resulting mappings.

All these accounts differ in certain aspects from our approach. Most follow the idea of extending the coverage of an ontology, whereas we focus on the systematic enrichment of an existing resource, i.e., GermaNet, by corpus examples. This is the reason why we perform the mapping on word senses (i.e., lexical units) in GermaNet and not on synsets, as the above-mentioned studies do. Moreover, these studies all focus on English, while our work concerns German. Our approach allows the alignment of multiple Wikipedia articles to a sense in GermaNet, whereas some of the other algorithms assign only the most likely WordNet synset to an article in Wikipedia.

3.2 Harvesting Corpus Examples

The nature of the task of harvesting corpus examples for word senses is closely related to the task of creating a sense-annotated corpus. Both tasks focus on harvesting textual materials whose words will be assigned the corresponding word senses of the wordnet in question. Because of this close similarity between the two tasks, it is appropriate and relevant to review and to characterize the state of the art in creating sense-annotated corpora.

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 (Agirre et al., 2007; Erk and Strapparava, 2010; Agirre et al., 2004). Purely manual methods were also used for the German sense-annotated corpora constructed by Broscheit et al. (2010) and Raileanu et al. (2002) as well as for other languages including the Bulgarian and the Chinese sense-tagged corpora (Koeva et al., 2006; Wu et al., 2006).

Few previous attempts of (semi-)automatically harvesting corpus data for the purpose of constructing a sense-annotated corpus exist. Yarowsky (1995), for example, developed a semi-supervised method based on a decision-list supervised WSD algorithm that iteratively disambiguates examples starting with a manually created seed set of annotated sentences. The knowledge-based approach of Leacock et al. (1998) – later also used by Agirre and Lacalle (2004) and Mihalcea and Moldovan (1999) – relies on the monosemous relative heuristic for the automatic harvesting of web data for the purposes of creating sense-annotated corpora. By focussing on web-based data, their work resembles the research described in the present paper. However, the underlying harvesting methods differ.

The three studies that are closest in spirit to the approach presented here are those of Santamaría et al. (2003), Henrich et al. (2012), and Henrich et al. (to appear). These studies also rely on automatic mappings between wordnet senses and a second web resource: an automatic association of Web directories (from the Open Directory Project, ODP) to WordNet senses for English (in the case of Santamaría et al. (2003)), a mapping between the German version of the web-based dictionary Wiktionary and GermaNet created by Henrich et al. (2011) (in the case of Henrich et al., 2012), and a mapping between the English Wiktionary and the Princeton WordNet created by Meyer

and Gurevych (2011) (in the case of Henrich et al., to appear). Henrich et al.'s (2012) work has produced the German WebCAGe corpus (short for: *Web-Harvested Corpus Annotated with GermaNet Senses*). WebCAGe has been constructed by harvesting sense-specific example sentences from Wiktionary itself and by harvesting additional textual materials from other web-based textual sources such as Wikipedia, online newspaper materials, and the German Gutenberg text archive[3]. These additional materials were harvested by following the links that accompany example sentences in Wiktionary. The work by Henrich et al. (to appear) applies Henrich et al.'s (2012) approach to English and has led to a sense-annotated corpus for English which they call WebCAP (short for: *Web-Harvested Corpus Annotated with Princeton WordNet Senses*). For both these corpora (Henrich et al., 2012; Henrich et al., to appear) it has to be kept in mind that the example sentences contained in Wiktionary are often artificially constructed by the authors of a Wiktionary entry and are, thus, not authentic materials taken from actual text corpora. Harvesting example sentences from Wikipedia articles – the goal of the present research – results in authentic corpus examples and, thus, provides a significant extension of Henrich et al.'s work.

4 Mapping GermaNet to Wikipedia

As mentioned above, harvesting of corpus examples from Wikipedia presupposes the existence of a mapping from GermaNet to Wikipedia in order to be able to link each target word in question to the appropriate GermaNet sense. Since the words contained in GermaNet and Wikipedia are often ambiguous, this mapping involves lexical disambiguation. The senses of an ambiguous word in GermaNet are each represented by a lexical unit. In Wikipedia, the senses of an ambiguous term are summarized in a 'disambiguation page' that lists all word meanings distinguished in Wikipedia along with short descriptions of each sense. Figure 1 shows a simplified example of such a disambiguation page for the German noun *Brücke*.[4]

The disambiguation page for *Brücke* in Figure 1 lists 9 distinct senses: *Brücke* in the sense of a structure built to span physical obstacles, a sportive excercise, a charge of heraldy, a defensive stance in wrestling, a bridge as a fixed partial denture, a small carpet, an edge in a graph, a structure located on the brain stem (pons), and a bridge of a ship. Each of these senses is summarized by a short description that contains a link to the corresponding Wikipedia article. Additionally, the disambiguation page also lists the use of *Brücke* in named entities such as family names (see the four bullet points in the lower part of Figure 1). Since named entities are not modelled in GermaNet, these additional senses can be ignored and the mapping can be limited to the ordinary senses of the word.

In GermaNet, the word *Brücke* is associated with three distinct lexical units (senses) that are contained in the following synsets:

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

[4] Note that there are further senses for *Brücke* in Wikipedia that are not shown in the figure for reasons of space.

Figure 1: Disambiguation page for the word *Brücke* in Wikipedia

Sense 1 ('bridge of a ship'): *Kommandobrücke, Brücke* – Schifffahrt; hypernyms: Deck, Schiffsdeck

Sense 2 ('bridge as a structure built to span physical obstacles'): *Brücke* – ein künstlicher Weg zur Überquerung eines Flusses, eines Tales oder Ähnlichem; hypernyms: Übergang, Überweg

Sense 3 ('bridge as a fixed partial denture'): *Brücke* – Zahnmedizin: modellierte Zahnreihe zur Überwindung eines oder mehrerer fehlender Zähne; hypernyms: Zahnersatz

The mapping task between GermaNet and Wikipedia now has to associate the correct GermaNet sense with the corresponding word meaning in Wikipedia. In general, this involves an n:m mapping. In the case that there is no disambiguation page, but the term is contained in Wikipedia, i. e. the term is monosemous, the Wikipedia article itself is used as a candidate for the mapping. Even if each of the resources only lists a single sense, it cannot automatically be assumed that the two entries in question refer to the same sense. Please also note that the titles of the Wikipedia articles are not always identical to the word under consideration. For example, two of the word senses of *Brücke* link to Wikipedia articles with the titles *Pons* 'pons' and *Kommandobrücke* 'bridge of a ship'.

For the mapping between GermaNet and Wikipedia several systems were implemented which basically rely on two different algorithms: Lesk and PageRank.

Lesk: Lesk (1986) introduces a word sense disambiguation algorithm that disambiguates two words by counting the overlaps between their respective sense definitions. Applied to the task at hand, this means that given two bag of words (BOW) for a GermaNet sense s_i and a Wikipedia page p_j, the overlap between these is calculated.

PageRank: PageRank (Brin and Page, 1998) is Google's algorithm for ranking webpages. Given a graph, every node v is initialized with $v = \frac{1}{|nodes|}$. In the following iteration steps every node spreads its mass equally to its neighbour nodes. The process is repeated until the values for each node converge. The resulting PageRank vector $\mathbf{Pr}$ is equivalent to:

$$\mathbf{Pr} = cM\mathbf{Pr} + (1 - c)\mathbf{v}$$

where M is the adjacency matrix for the graph, $\mathbf{v}$ is the vector with the initial values and c is a damping factor, which controls, how much of the initial mass is infused in every iteration step. Since $\sum_i Pr_i = 1$, the resulting value Pr_i may be considered as the probability to end up with node v_i in a random walk over the graph.

Both techniques have in common, that they use bag of words (BOW) for the disambiguation. A bag of words representing a given text is just the set of lemmas occurring in the text, i.e., just the words without syntactic information. Although a BOW is a very basic data structure, it is very common in Information Retrieval to represent whole documents. In the implementation for our algorithms, two kinds of BOWs are used: one representing a Wikipedia page and one representing a sense in GermaNet. In the case of a Wikipedia page, the corresponding article is used for the BOW, in the case of a Germanet sense all synonyms of the given sense and all neighbouring words/synsets up to a certain distance are included in the BOW. There are several parameters, which allow to control which words are actually included in the BOWs (see Section 6.1 for more details about these parameters).

What follows is a detailed description of the different systems we implemented.

1. Lesk: Given two BOWs, one for a given Germanet sense s_i and one for a given Wikipedia page w_j, the overlap between the two is calculated and normalized with respect to the minimum of the two.

2. We have reimplemented the approach by Niemann and Gurevych (2011). Given the two BOWs for GermaNet sense s_i and Wikipedia page w_j, PageRank is run twice on the whole GermaNet graph, initializing only those nodes whose corresponding synsets have at least one lemma in common with both BOWs. To calculate the semantic relatedness between a sense s_i and a Wikipedia page w_j

three similarity measures were applied to the resulting two PageRank vectors $\Pr(s_i)$ and $\Pr(w_j)$: Euklidian distance, cosine, and chi^2:

$$\chi^2(Pr(s_i), Pr(w_j)) = \sum_k \frac{(Pr(s_i)_k - Pr(w_j)_k)^2}{Pr(s_i)_k + Pr(w_j)_k}$$

3. We developed a system called *TextLink*, which is an adaptation of the PageRank algorithm. It uses a special directed bipartite multigraph, which consists on one side of all Wikipedia articles and on the other side of all lemmas which function as a link in Wikipedia – see Figure 2: Wikipedia articles are shown in the upper part of the figure, the lemmas occuring as links in the lower part. For this purpose the whole Wikipedia is scanned for links. Whenever a link is found, the lemma/phrase, which is configured as a link (i.e., the link label), is added as a new node to the graph, if not already existent, connecting it with the two nodes corresponding to the interlinked Wikipedia pages (parallel edges are allowed). Note that the example in Figure 2 is a pretty small excerpt from the whole graph.

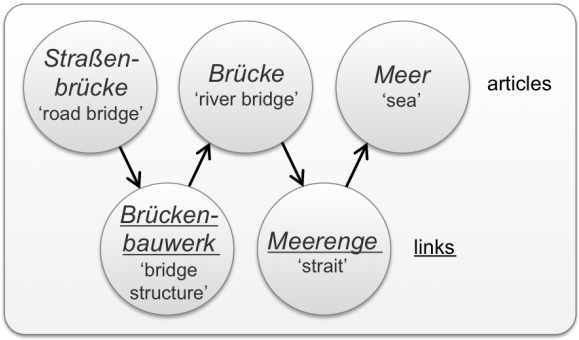

Figure 2: Bipartite graph illustration for *Brücke* ('bridge' architecture)

More formally, the definition of the graph is $G(V, A)$ with vertices $V = W + L$, $W \cap L = \emptyset$, where W is the set of all Wikipedia pages, $w_i \in W$ refers to a specific Wikipedia page w_i, L describes the set of all hyperlinks, and $l_k \in L$ is a hyperlink h with anchor text (label) l_k. Given two Wikipedia pages w_i and w_j and a hyperlink $h(l_k, w_i, w_j)$, directing from Wikipedia page w_i to page w_j, whose anchor text (label) is l_k: For every such hyperlink h we create two arcs $a_s, a_t \in A$ with $a_s = w_i l_k$ and $a_t = l_k w_j$ (parallel arcs allowed).

To better understand the construction of the graph, see the second Wikipedia article in Figure 3 (which will be described in Section 5) entitled with *Brücke* 'bridge': the mouseover symbol on the left side illustrates that the link labelled

with *Meerenge* 'strait' connects the two Wikipedia articles *Brücke* and *Meer* 'sea' with each other.

In order to calculate a mapping between the GermaNet senses s of a given lemma and the corresponding set of Wikipedia pages w, for each sense a BOW is created. Given a BOW for sense s_i all link nodes whose labels are contained as a lemma in the BOW are initialized and PageRank is run with just one iteration and a damping factor $c = 1$. Sense s_i is then mapped to the Wikipedia page w_j which maximizes the resulting value and which is above a certain threshold.

Alternatively we applied three iterations, slightly modifying the original PageRank algorithm in that we added up the values in each iteration step, so that the value for vertex $v_i = \sum_{k=1}^{3} v_i^{(k)}$. Note that this is a slight alteration of the original PageRank algorithm because we iterate exactly three times and not until the node values remain constant as it is the idea in the original PageRank algorithm. Experiments showed better results with this procedure, which can be regarded as a weighted breadth-first-search of distance three with the exception that nodes can be visited more than once.

4. Combination of two different systems: we tested, if any combination of two systems (out of the three systems described in 1., 2., and 3. above) might give better results, thus showing that the power of Lesk and PageRank lie in different fields and act to some degree in a complementary way.

For all of the algorithms just described, we use thresholding for the mapping between GermaNet senses and Wikipedia articles: a mapping is established only if the numeric value computed for a putative mapping by the WSD algorithm is above a certain threshold. This threshold has been computed by a series of test runs on the training corpus (described in Section 6.1).

5 Harvesting Corpus Examples

Once the GermaNet word senses have been mapped to Wikipedia articles, these articles need to be mined for relevant corpus examples that include the target word in question. 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" (Gale et al., 1992), multiple occurrences of a target word in a given text are all automatically assigned to the same GermaNet sense.

In a morphologically rich language like German, the automatic harvesting of example sentences requires some lexical preprocessing of the Wikipedia articles in order to be able to robustly identify the occurrences of the target word under consideration. Automatic detection of target words is performed by the software tool used by Henrich et al. (2012) for the construction of WebCAGe. This tool splits the text up into individual sentences, performs tokenization, lemmatization, and compound splitting. Apart from lemmatization, compound splitting is also necessary because the target word can be part

of a compound. Since the constituent parts of a compound are not usually separated by blank spaces or hyphens, German compounding poses a particular challenge for target word identification.

Figure 3 shows the combined result of the GermaNet to Wikipedia mapping and the harvesting of example sentences for each of the Wikipedia articles associated with the GermaNet senses of the German noun *Brücke*. The occurrences of the target words are highlighted in the running text by surrounding boxes. Because of the sense mapping between GermaNet and Wikipedia, each target word occurrence is automatically associated with a corresponding GermaNet sense.

The primary use of the harvested examples in the present study is to enrich the GermaNet lexical units by corpus examples from Wikipedia. However, an interesting and highly useful by-product of this work is the construction of a large sense-annotated corpus of Wikipedia data for German, which will be referred to as WikiCAGe (short for: *Wikipedia-Harvested Corpus Annotated with GermaNet Senses*). This by-product is particularly valuable because sense-annotated corpora for German are in short supply.

6 Evaluation

The two tasks to be solved in this research (the mapping and the harvesting) require separate evaluations. This section presents both evaluation steps: Section 6.1 evaluates the automatic mapping of word senses in GermaNet to articles in Wikipedia. The harvesting of the corpus examples, which relies on this mapping, is analysed in Section 6.2.

6.1 Evaluation of the Automatic Mapping

In order to be able to evaluate the automatic alignment of lexical units (senses) in GermaNet to articles in Wikipedia, three experienced lexicographers created two manually annotated gold standards:

1. The gold standard that was used for training, i.e., to identify the best performing systems and to fine-tune the most reliable parameter settings, consists of 30 polysemous nouns. These 30 nouns comprise a total of 862 potential sense mappings between GermaNet senses and Wikipedia articles of which 82 were manually classified as correct. The nouns were manually chosen with the goal of including examples with different numbers of senses, ranging from 2 to 6 distinct senses. On average, the 30 nouns exhibit 3.7 senses in GermaNet. This degree of polysemy is considerably higher compared to the average number of 2.3 word senses of polysemous nouns in GermaNet. The reason for choosing a set of nouns with a higher than average degree of polysemy for training was deliberate so as to provide ample data for a fine-grained adjustment of the parameter and threshold settings with respect to all classifiers used for the GermaNet-Wikipedia mapping.

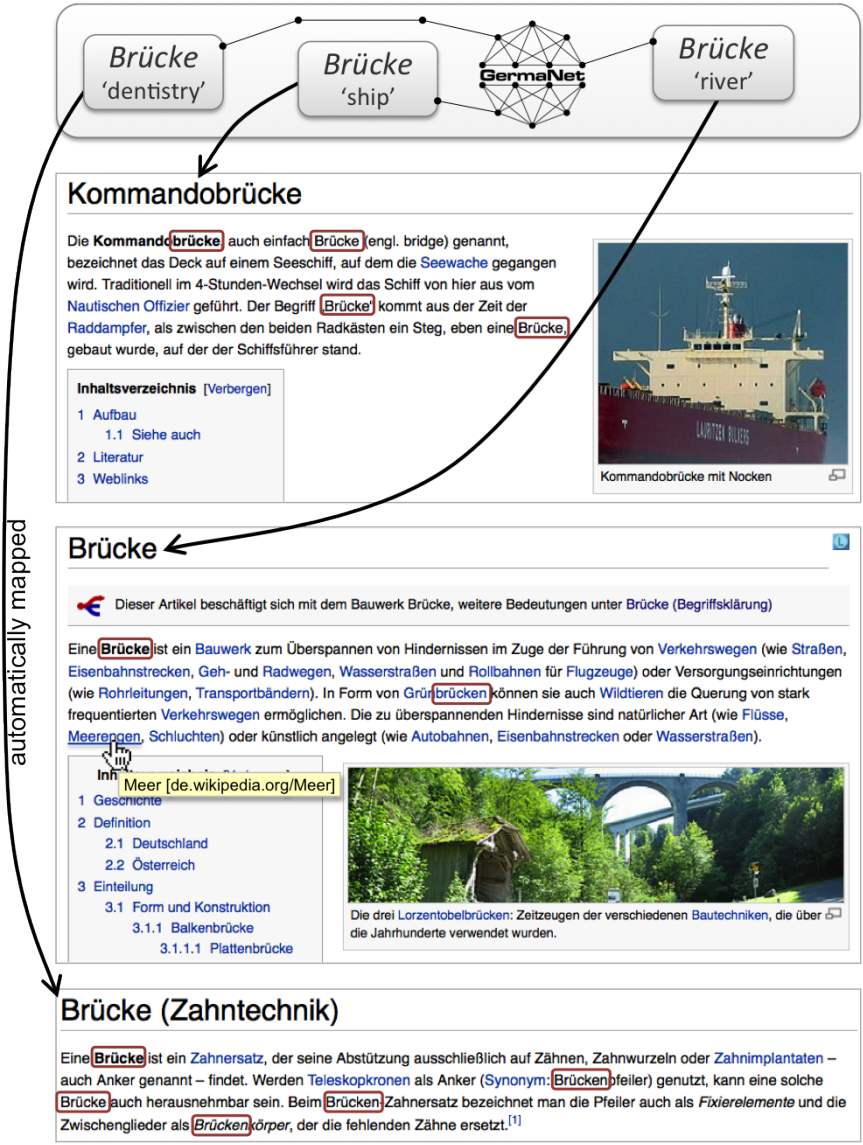

Figure 3: Mapping example for the word *Brücke* with corpus examples

2. For testing the best algorithm setup, another gold standard of 270 randomly chosen polysemous nouns with an average of 2.4 senses was created.[5] These 270 nouns comprise a total of 4308 sense mappings of which 446 were classified as correct.

The first gold standard has been used to identify the best performing systems and to fine-tune the most reliable parameter settings. All systems that are evaluated use one or two bag of words for the disambiguation. Which words are actually included in the BOW is a matter of parameter setting. In the case of Wikipedia, the choices are the following: (i) whether the BOW consists of the title and the first paragraph of an article or of the entire page, (ii) whether to include in the BOW the Wikipedia categories linked to the article or not, and (iii) whether the anchor words of ingoing resp. outgoing links should be included in the BOW or not.

The experiments with the training corpus show constantly better results when the BOW representing a Wikipedia page consists of the title and the first paragraph instead of the entire page. This is not surprising since the first paragraph of a Wikipedia article usually serves as a short definition of the presented concept. Further, the results are much better when the anchor words of the links are included in the BOW of a Wikipedia page. This can be explained by the fact that a term, which is configured as a link directing to that page, is usually semantically closely related to the term described on the page.

In the case of GermaNet, the BOW includes all synonyms from the target word synset and can be expanded to include synsets that are linked to the target word by conceptual or lexical relations. This expansion is again a matter of parameter setting and includes the following choices: (i) the graph distance between the target word and the candidate synset, (ii) a weighting parameter that is proportional to the graph distance, and (iii) whether to include or exclude the hyponymy relation among the conceptual relations used for expansion.

The parameter settings just described determine the strength of association between a GermaNet sense and a Wikipedia article. This numerical score can then be used for thresholding. That is, the association strength is considered a match only if the score is above a given threshold.

The mapping algorithm follows a maximal matching strategy of the GermaNet-Wikipedia bipartite graph. Another choice point concerns the interaction of thresholding and maximal match calculation. Thresholding can either be incorporated into the maximal match calculation in the sense that candidate matches below a given threshold are discarded when the overall optimal mapping is calculated or thresholding can be applied after maximal match calculation. In the latter scenario, which empirically turned out to be superior, thresholding is in effect used to prune individual sense mappings from the maximal match result.

[5] By choosing the set of 270 polysemous nouns at random, we ensure that the degree of ambiguity closely matches the average number of 2.3 word senses per polysemous noun in GermaNet.

Since our primary goal is to extract example sentences in an automated way, the priority is the optimization of precision, neglecting recall. Therefore we focused on configurations which resulted in a precision of 0.85 or better.

Table 1 gives an overview of the results for the three individual mapping algorithms introduced in Section 4 (shown in rows 1 to 3) as well as for all pairwise combinations of the three individual algorithms (shown in rows 4 to 6). Precision is determined as the ratio between correctly identified mappings (i.e., true positives) and the overall number of automatically proposed mappings (i.e., true positives plus false positives). Recall is the ratio of true positives compared to the overall number of correct mappings in the gold standard (i.e., true positives and false negatives). F-score represents the harmonic mean between recall and precision. Among the individual algorithms, *Niemann/Gurevych* yields the best precision (0.85) for the test corpus and performs best in terms of F-score for both the training and the test corpora.[6]

Table 1: Evaluation results

System	Training			Testing		
	Prec.	**Rec.**	**F**	**Prec.**	**Rec.**	**F**
Lesk	**0.95**	0.25	0.40	0.81	0.27	0.41
Niemann/Gurevych	0.91	**0.29**	**0.44**	**0.85**	**0.30**	**0.44**
Textlink	0.90	0.22	0.35	0.79	0.23	0.36
Lesk + Niemann/Gur.	0.96	**0.32**	**0.48**	0.88	**0.35**	**0.50**
Lesk + Textlink	**1.00**	0.25	0.40	**0.90**	0.21	0.34
Niemann + TextLink	0.94	0.23	0.37	0.87	0.22	0.35

In order to test whether the three individual algorithms may yield better results when they are combined with one another, all pairwise combinations were evaluated as well. Here, the combination of the *Lesk* and the *Niemann/Gurevych* algorithms achieved the best F-score for both training and test corpora. It is therefore this combined algorithms that was used as the basis for the automatic harvesting of corpus examples.

6.2 Evaluation of the Automatic Harvesting of Corpus Examples

The algorithm for harvesting corpus examples is evaluated in terms of precision- and recall and an error analysis is provided. We also assess the effectiveness of our harvesting approach by comparing the overall size of WikiCAGe to existing sense-annotated corpora for German.

[6] Note that we have also conducted experiments with PageRank itself as in the approach by Agirre and Soroa (2009), but as these experiments – surprisingly – perform worse than the Lesk algorithm, we have not included the results in the table. For the task at hand, the results for PageRank are in an acceptable range only in combination with error measures well-known in the area of Information Retrieval as in the account of *Niemann/Gurevych*.

In order to inspect the quality of the harvested corpus examples, 261 automatically annotated Wikipedia articles were manually verified and, where required, post-corrected. We will make this manually verified excerpt of WikiCAGe freely available on the web. A precision of 0.89 with a recall of 0.91 prove the viability of the proposed method for automatic harvesting of sense-annotated data. In practise, this means that human post-correction is needed on average only for one out of ten harvested corpus examples in order to eliminate the remaining noise in the annotated data.

An analysis of those harvested corpus examples that are tagged with a wrong GermaNet word sense shows three predominant error types: (i) errors that are caused by an erroneous mapping between GermaNet and Wikipedia, (ii) errors that clash with the heuristic "one sense per discourse", and (iii) errors that are due to the software tool used for the detection of the target words. Erroneous mappings between word senses in GermaNet and articles in Wikipedia make up 6.0% of the total errors. An inspection of the "one sense per discourse" heuristic shows that this heuristic is violated by 3.3% of all marked target word occurrences. The last identified error type, i.e., errors that are due to the identification of the target word in the text, make up 3.0%.

Altogether, the presented approach has mapped 1 030 polysemous nouns from GermaNet to Wikipedia. Since GermaNet contains a total of 4 358 polysemous nouns, this amounts to a coverage of 23.6% for all such nouns and of 30.8% for all polysemous nouns that occur both in GermaNet and Wikipedia.

The successful mappings yield a total of 24 344 tagged word tokens occurring in 18 868 example sentences. This means that for each of the 1 030 nouns approximately 18 examples sentences are harvested on average. The large number of 18 868 harvested example sentences also leads to a sizable corpus of sense-annotated data. Table 2 shows a comparison of WikiCAGe to other existing sense-annotated corpora for German, i.e., to the manually constructed resources by Broscheit et al. (2010) and Raileanu et al. (2002) and the automatically created resource WebCAGe by Henrich et al. (2012). The number of sense tagged words that are listed separately per word class show that WebCAGe and the corpus by Broscheit et al. contain occurrences for words of all the three word classes of adjectives, nouns, and verbs, whereas WikiCAGe and the corpus by Raileanu et al. are limited to nouns only. By comparison, the overall number of sense-tagged words in WikiCAGe (24 344) is considerably larger than in all the other corpora.

7 Conclusion and Future Work

In this paper, we have presented an automatic method for enriching GermaNet senses with example sentences from Wikipedia. This method has the desirable side-effect of yielding a sense-annotated corpus for German, which we refer to by the name WikiCAGe, at the same time. We plan to make the excerpt of WikiCAGe, that was already

Table 2: Comparing WikiCAGe to other sense-tagged corpora of German.

		WikiCAGe	**WebCAGe**	**Broscheit et al., 2010**	**Raileanu et al., 2002**
Sense tagged words	adj. (a)	0	211	6	0
	nouns (n)	1 030	1 499	18	25
	verbs (v)	0	897	16	0
	a/n/v	1 030	2 607	40	25
Number of tagged word tokens		24 344	10 750	approx. 800	2 421
Domain independent		yes	yes	yes	medical domain

manually post-corrected for the evaluation of the presented algorithm, available to the larger research community.[7]

The algorithms used for the GermaNet to Wikipedia mapping and for the automatic harvesting of corpus examples were optimized for precision, resulting in an enrichment of 23.6% of all polysemous nouns in GermaNet. The motivation for optimizing on precision is to minimize the noise in the harvested data. The precision of 89% achieved for the automatic automatic harvesting of Wikipedia examples is sufficient to use the WikiCAGe corpus as is for NLP applications such as word sense disambiguation and statistical machine translation, whose statistical models are robust enough to cope with noisy training data. In future work, we plan to explore the precision vs. recall trade-off in order to increase the coverage of the methods described in this paper. This will increase the need for manual post-inspection of the harvested examples. However, since this post-inspection will not require any editing but just discarding of examples that do not match the candidate word sense, the amount of noise in the data does not have to be as tightly controlled. This in turn means that there is a priori no tight restriction on boosting recall and thus coverage.

Another direction for future work concerns the selection of those examples that best illustrate the use of a particular GermaNet word sense. As noted in Section 6.2, an average of 18 examples is harvested for each polysemous noun in GermaNet. In order to be able to select the most appropriate example(s) one needs to formulate clear criteria for what counts as a good example. Here we intend to build on the work of Kilgarriff et al. (2008). They specify the following properties of a good example: (i) it should represent a typical, exhibiting frequent and well-dispersed pattern of usage, (ii) it should be informative, helping to elucidate the definition, and (iii) it should be intelligible to learners, avoiding gratuitously difficult lexis and structures, puzzling or distracting

[7] `http://www.sfs.uni-tuebingen.de/en/wikicage.shtml`

names, anaphoric references or other deictics which cannot be understood without access to the wider context. Kilgarriff et al. further describe how these properties can be applied in practise to given example sentences, e.g., by using features such as the length of a sentence or the frequencies of words in a sentence.

Acknowledgments

The research reported in this paper was jointly funded by the SFB 833 grant of the DFG and by the CLARIN-D grant of the BMBF. We would like to thank Marie Hinrichs as well as the anonymous JLCL reviewers for their helpful comments on earlier versions of this paper. We are very grateful to Reinhild Barkey, Valentin Deyringer, Sarah Schulz, and Johannes Wahle for their help with the evaluation reported in Section 6.

References

Agirre, E., Aldabe, I., Lersundi, M., Martínez, D., Pociello, E., and Uria, L. (2004). The basque lexical-sample task. In Mihalcea, R. and Edmonds, P., editors, *Senseval-3: 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 1–4, Barcelona, Spain. Association for Computational Linguistics.

Agirre, E. and Lacalle, O. L. D. (2004). Publicly available topic signatures for all WordNet nominal senses. In *Proceedings of the 4th International Conference on Languages Resources and Evaluations*, LREC '04, pages pp. 1123–1126.

Agirre, E., Màrquez, L., and Wicentowski, R., editors (2007). *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Agirre, E. and Soroa, A. (2009). Personalizing pagerank for word sense disambiguation. *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, (April):33–41.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *7th International World-Wide Web Conference*, WWW '98.

Broscheit, S., Frank, A., Jehle, D., Ponzetto, S. P., Rehl, D., Summa, A., Suttner, K., and Vola, S. (2010). Rapid bootstrapping of word sense disambiguation resources for German. In *Proceedings of the 10. Konferenz zur Verarbeitung Natürlicher Sprache*.

Erk, K. and Strapparava, C., editors (2010). *SemEval '10: Proceedings of the 5th International Workshop on Semantic Evaluation*, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fellbaum, C., editor (1998). *WordNet: an Electronic Lexical Database*. MIT Press.

Fernando, S. and Stevenson, M. (2012). Mapping WordNet synsets to Wikipedia articles. In Chair), N. C. C., Choukri, K., Declerck, T., Doğan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation*, LREC '12, pages 590–596. European Language Resources Association (ELRA).

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One sense per discourse. In *DARPA Speech and Natural Language Workshop*.

Henrich, V. and Hinrichs, E. (2010). GernEdiT – The GermaNet Editing Tool. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation*, LREC '10, pages 2228–2235. European Language Resources Association (ELRA).

Henrich, V., Hinrichs, E., and Vodolazova, T. (2011). Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. In *Proceedings of 5th Language & Technology Conference*, LTC '11, pages 126–130.

Henrich, V., Hinrichs, E., and Vodolazova, T. (2012). 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 '12, pages 387–396.

Henrich, V., Hinrichs, E., and Vodolazova, T. (to appear). An automatic method for creating a sense-annotated corpus harvested from the web. In *International Journal of Computational Linguistics and Applications*, volume 3.2.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., and Rychly, P. (2008). *GDEX: Automatically finding good dictionary examples in a corpus*, pages 425–432. Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Koeva, S., Lesseva, S., and Todorova, M. (2006). Bulgarian sense tagged corpus. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, Italy*, pages 79–86.

Kunze, C. and Lemnitzer, L. (2002). GermaNet – representation, visualization, application. In *Proceedings of the 3rd International Language Resources and Evaluation*, LREC '02, pages 1485–1491.

Leacock, C., Miller, G. A., and Chodorow, M. (1998). Using corpus statistics and WordNet relations for sense identification. *Comput. Linguist.*, 24(1):147–165.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, SIGDOC '86, pages 24–26, New York, NY, USA. ACM.

Meyer, C. M. and Gurevych, I. (2011). 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 '11, pages 883–892.

Mihalcea, R. and Moldovan, D. I. (1999). An automatic method for generating sense tagged corpora. In *Proceedings of the American Association for Artificial Intelligence*, AAAI '99, pages 461–466, Menlo Park, CA, USA. American Association for Artificial Intelligence.

Niemann, E. and Gurevych, I. (2011). The people's web meets linguistic knowledge: automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics*, IWCS '11, pages 205–214, Stroudsburg, PA, USA. Association for Computational Linguistics.

Palmer, M., Dang, H. T., and Fellbaum, C. (2007). Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13(2):137–163.

Ponzetto, S. P. and Navigli, R. (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI '09, pages 2083–2088, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich word sense disambiguation rivaling supervised systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1522–1531, Stroudsburg, PA, USA. Association for Computational Linguistics.

Raileanu, D., Buitelaar, P., Vintar, S., and Bay, J. (2002). Evaluation corpora for sense disambiguation in the medical domain. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, LREC '02.

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2005). Automatic assignment of Wikipedia encyclopedic entries to WordNet synsets. In *In: Proceedings of the Atlantic Web Intelligence Conference, AWIC-2005. Volume 3528 of Lecture Notes in Computer Science*, pages 380–386. Springer Verlag.

Santamaría, C., Gonzalo, J., and Verdejo, F. (2003). Automatic association of web directories with word senses. *Comput. Linguist.*, 29(3):485–502.

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA. ACM.

Toral, A., Ferrández, O., Agirre, E., and Muñoz, R. (2009). A study on linking Wikipedia categories to WordNet synsets using text similarity. In *Proceedings of the International Conference*, RANLP '09, pages 449–454, Borovets, Bulgaria. Association for Computational Linguistics.

Wolf, E. and Gurevych, I. (2010). Aligning sense inventories in wikipedia and wordnet. In *Proceedings of the First Workshop on Automated Knowledge Base Construction*, pages 24–28.

Wu, Y., Jin, P., Zhang, Y., and Yu, S. (2006). A chinese corpus with word sense annotation. In *Proceedings of the 21st international conference on Computer Processing of Oriental Languages: beyond the orient: the research challenges ahead*, ICCPOL'06, pages 414–421, Berlin, Heidelberg. Springer-Verlag.

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the Conference on Language Resources and Evaluation*, LREC '08.