Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment

Michael Matuschek [‡] **and Iryna Gurevych** [†‡]
† Ubiquitous Knowledge Processing Lab (UKP-DIPF),
German Institute for Educational Research and Educational Information
Schloßstr. 29, 60486 Frankfurt, Germany
[‡]Ubiquitous Knowledge Processing Lab (UKP-TUDA),
Department of Computer Science, Technische Universität Darmstadt
Hochschulstr. 10, 64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

In this paper, we present Dijkstra-WSA, a novel graph-based algorithm for word sense alignment. We evaluate it on four different pairs of lexical-semantic resources with different characteristics (WordNet-OmegaWiki, WordNet-Wiktionary, GermaNet-Wiktionary and WordNet-Wikipedia) and show that it achieves competitive performance on 3 out of 4 datasets. Dijkstra-WSA outperforms the state of the art on every dataset if it is combined with a back-off based on gloss similarity. We also demonstrate that Dijkstra-WSA is not only flexibly applicable to different resources but also highly parameterizable to optimize for precision or recall.

1 Introduction

Lexical-semantic resources (LSRs) are a cornerstone for many Natural Language Processing (NLP) applications such as word sense disambiguation (WSD) and information extraction. However, the growing demand for large-scale resources in different languages is hard to meet. The Princeton WordNet (WN) (Fellbaum, 1998) is widely used for English, but for most languages corresponding resources are considerably smaller or missing. Collaboratively constructed resources like Wiktionary (WKT) and OmegaWiki (OW) provide a viable option for such cases and seem especially suitable for smaller languages (Matuschek et al., 2013), but there are still considerable gaps in coverage which need to be filled. A related problem is that there usually does not exist a single resource which works best for all purposes, as different LSRs cover different words, senses and information types.

These considerations have sparked increasing research efforts in the area of word sense alignment (WSA). It has been shown that aligned resources can indeed lead to better performance than using the resources individually. Examples include semantic parsing using FrameNet (FN), WN, and VerbNet (VN) (Shi and Mihalcea, 2005), word sense disambiguation using an alignment of WN and Wikipedia (WP) (Navigli and Ponzetto, 2012) and semantic role labeling using a combination of PropBank, VN and FN in the *SemLink* project (Palmer, 2009). Some of these approaches to WSA either rely heavily on manual labor (e.g. Shi and Mihalcea (2005)) or on information which is only present in few resources such as the most frequent sense (MFS) (Suchanek et al., 2008). This makes it difficult to apply them to a larger set of resources.

In earlier work, we presented the large-scale resource UBY (Gurevych et al., 2012). It contains nine resources in two languages which are mapped to a uniform representation using the LMF standard (Eckle-Kohler et al., 2012). They are thus structurally interoperable. UBY contains pairwise sense alignments between a subset of these resources, and this work also presented a framework for creating alignments based on the similarity of glosses (Meyer and Gurevych, 2011). However, it is not clear to what extent this approach can be applied to resources which lack this kind of information (see Section 3).

In summary, aligning senses is a key requirement for semantic interoperability of LSRs to increase the

coverage and effectiveness in NLP tasks. Still, existing efforts are mostly focused on specific types of resources (most often requiring glosses) or application scenarios. In this paper, we propose an approach to alleviate this and present Dijkstra-WSA, a novel, robust algorithm for word sense alignment which is applicable to a wide variety of resource pairs and languages. For the first time, we apply a graph-based algorithm which works on full graph representations of both resources to word sense alignment. This enables us to take a more abstract perspective and reduce the problem of identifying equivalent senses to the problem of matching nodes in these graphs. Also for the first time, we comparatively evaluate a WSA algorithm on a variety of different datasets with different characteristics.

The key properties of Dijkstra-WSA are:

Robustness The entities within the LSRs which are to be aligned (usually senses or synsets) are modeled as nodes in the graph. These nodes are connected by an edge if they are semantically related. While, for instance, semantic relations lend themselves very well to deriving edges, different possibilities for graph construction are equally valid as the algorithm is agnostic to the origin of the edges.

Language-independence No external resources such as corpora or other dictionaries are needed; the graph construction and alignment only rely on the information from the considered LSRs.

Flexibility The graph construction as well as the actual alignment are highly parameterizable to accommodate different requirements regarding precision or recall.

The rest of this paper is structured as follows: In Section 2 we give a precise problem description and introduce the resources covered in our experiments, in Section 3 we discuss some related work, while our graph-based algorithm Dijkstra-WSA is presented in Section 4. We describe an evaluation on four datasets with different properties, including an error analysis, in Section 5 and conclude in Section 6, pointing out directions for future work.

2 Notation and Resources

2.1 Problem Description

A *word sense alignment*, or *alignment* for short, is formally defined as a list of pairs of senses from two LSRs. A pair of aligned senses denote the same meaning. E.g., the two senses of *letter* "The conventional characters of the alphabet used to represent speech" and "A symbol in an alphabet, bookstave" (taken from WN and WKT, respectively) are clearly equivalent and should be aligned.

2.2 Evaluation Resources

For the evaluation of Dijkstra-WSA, we align four pairs of LSRs used in previous work, namely WN-OW (Gurevych et al., 2012), WN-WKT (Meyer and Gurevych, 2011), GN-WKT (Henrich et al., 2011) and WN-WP (Niemann and Gurevych, 2011). Our goal is to cover resources with different characteristics: Expert-built (WN, GN) and collaboratively constructed LSRs (WP, WKT, OW), resources in different languages (English and German) and also resources with few sense descriptions (GN) or semantic relations (WKT). We contrastively discuss the results of the Dijkstra-WSA algorithm on these different datasets and relate the results to the properties of the LSRs involved. Moreover, using existing datasets ensures comparability to previous work which discusses only one dataset at a time.

WordNet (WN) (Fellbaum, 1998) is a lexical resource for the English language created at Princeton University. The resource is organized in sets of synonymous words (synsets) which are represented by glosses (sometimes accompanied by example sentences) and organized in a hierarchy. The latest version 3.0 contains 117,659 synsets.

Wikipedia (WP) is a freely available, multilingual online encyclopedia. WP can be edited by every Web user, which causes rapid growth: By February 2013 the English WP contained over 4,000,000 article pages. Each article usually describes a distinct concept, and articles are connected by hyperlinks within the article texts.

Wiktionary (WKT) is the dictionary pendant to WP. By February 2013 the English WKT contained over 3,200,000 article pages, while the German edition contained over 200,000 ones. For each word, multiple senses can be encoded. Similar to WN, they are represented by a gloss and usage examples. There also exist hyperlinks to synonyms, hypernyms, meronyms etc. The targets of these relations are not senses, however, but merely lexemes (i.e. the relations are not disambiguated).

	LSRs	$P/R/F_1/Acc.$	Approach
Meyer and Gurevych (2011)	WN-WKT	0.67/0.65/0.66/0.91	Gloss similarity + Machine learning
Niemann and Gurevych (2011)	WN-WP	0.78/0.78/0.78/0.95	Gloss similarity + Machine learning
Henrich et al. (2011)	GN-WKT	0.84/0.85/0.84/0.94	Pseudo-gloss overlap
de Melo and Weikum (2010)	WN-WP	0.86/NA/NA/NA	Gloss/article overlap
Laparra et al. (2010)	FN-WN	0.79/0.79/0.79/NA	Dijkstra-SSI+ (WSD algorithm)
Navigli (2009)	WN	0.64/0.64/0.64/NA	Graph-based WSD of WN glosses
Ponzetto and Navigli (2009)	WN-WP	NA/NA/NA/0.81	Graph-based, only for WP categories
Navigli and Ponzetto (2012)	WN-WP	0.81/0.75/0.78/0.83	Graph-based WSA using WN relations

Table 1: Summary of various approaches to WSA. "NA" stands for "Not Available".

OmegaWiki (OW) is a freely editable online dictionary like WKT. However, there do not exist distinct language editions as OW is organized in language-independent concepts ("Defined Meanings") to which lexicalizations in various languages are attached. These can be considered as multilingual synsets, and they are interconnected by unambiguous relations just like WN. As of February 2013, OW contains over 46,000 of these concepts and lexicalizations in over 400 languages.

GermaNet (GN) is the German counterpart to WN (Hamp and Feldweg, 1997). It is also organized in synsets (around 70,000 in the latest version 7.0) which are connected via semantic relations.

3 Related Work

The are two strands of closely related work: Similarity-based and graph-based approaches to word sense alignment. To our knowledge, there exists no previous work which fully represents both LSRs involved in an alignment as graphs. We give a summary of different approaches in Table 1.

3.1 Similarity-based Approaches

Niemann and Gurevych (2011) and Meyer and Gurevych (2011) created WN-WP and WN-WKT alignments using a framework which first calculates the similarity of glosses (or glosses and articles in the case of WN-WP) using either cosine or personalized page rank (PPR) similarity (Agirre and Soroa, 2009) and then learns a threshold on the gold standard to classify each pair of senses as a (non-)valid alignment. This approach was later extended to cross-lingual alignment between the German OW and WN (Gurevych et al., 2012) using a machine translation component. However, its applicability depends on the availability and quality of the glosses, which are not present in every case (e.g. for VN). Moreover, as it involves supervised machine learning, it requires the initial effort of manually annotating a sufficient amount of training data. Henrich et al. (2011) use a similar approach for aligning GN and WKT. However, they use word overlap as a similarity measure and do not require a machine learning component as they align to the candidate sense with the highest similarity regardless of the absolute value. The alignment of WP articles and WN synsets reported by de Melo and Weikum (2010) also relies on word overlap.

Although these approaches give reasonable results (with precision in the range of 0.67-0.84), they all depend on the lexical knowledge contained in the glosses, yielding low recall if there is insufficient lexical overlap (known as the "lexical gap", see for instance (Meyer and Gurevych, 2011)). Consider these two senses of *Thessalonian* in WKT and WN: "A native or inhabitant of Thessalonica" and "Someone or something from, or pertaining to, Thessaloniki". These are (mostly) identical and should be aligned, but there is no word overlap due to the interchangeable usage of the synonyms *Thessalonica* and *Thessaloniki*.

3.2 Graph-based Approaches

Laparra et al. (2010) utilize the SSI-Dijkstra+ algorithm to align FN lexical units (LUs) with WN synsets. The basic idea is to align monosemous LUs first and, based on this, find the closest synset in WN for the other LUs in the same frame. However, as SSI-Dijkstra+ is a word sense disambiguation (not alignment) algorithm, the LUs are merely considered as texts which are to be disambiguated; there is no attempt made to build a global graph structure for FN. Moreover, the algorithm solely relies on the

semantic relations found in WN and eXtended WN (Mihalcea and Moldovan, 2001). Thus, it is not applicable to other resources which have no or only few relations such as WKT.

Navigli (2009) aims at disambiguating WN glosses, i.e. assigning the correct senses to all non-stopwords in each WN gloss. His approach is to find the shortest possible circles in the WN relation graph to identify the correct disambiguation. In later work, this idea was extended to the disambiguation of translations in a bilingual dictionary (Flati and Navigli, 2012). However, there is no discussion of how this idea could be applied to word sense alignment of two or more resources. We build upon this idea of finding shortest paths (circles are a special kind of path) and extend it to multiple resources and edges other than semantic relations, in particular WP links and links to senses of monosemous lexemes appearing in glosses.

Ponzetto and Navigli (2009) propose a graph-based method to tackle the related, but slightly different problem of aligning WN synsets and WP categories (not articles). Using semantic relations, they build WN subgraphs for each WP category and then align those synsets which best match the category structure. In later work, Navigli and Ponzetto (2012) also align WN with the full WP. They build "disambiguation contexts" for the senses in both resources by using, for instance, WP redirects or WN glosses and then compute the similarity between these contexts. Again, a graph structure is built from WN semantic relations covering all possible senses in these contexts. The goal is to determine which WN sense is closest to the WP sense to be aligned. While these approaches are in some respects similar to Dijkstra-WSA, they do not take the global structure of both resources into account. Instead, they merely rely on a (locally restricted) subset of WN relations for creating the alignment. Thus, applying these approaches to resources in different languages might be difficult if WN relations are not applicable.

4 Dijkstra-WSA

In this section, we discuss our approach to aligning lexical-semantic resources based on the graph structure. This includes two steps: (i) the initial construction of the graphs using appropriate parameters, and (ii) the alignment itself.

4.1 Graph Construction

We represent the set of senses (or synsets, if applicable) of an LSR L as a set of nodes V where the set of edges $E, E \subseteq V \times V$ between these nodes represents semantic relatedness between them. We call this a *resource graph*. A WP article is considered a sense as it represents a distinct concept.

There are multiple options for deriving the edges from the resource. The most straightforward approach is to directly use the existing semantic relations (such as hyponymy), as it has been reported in previous work (Laparra et al., 2010; Navigli, 2009). For WP, we can directly use the given hyperlinks between articles as they also express a certain degree of relatedness (Milne and Witten, 2008). However, for many LSRs no or only few semantic relations exist. Consider WKT: Its relations are not sense disambiguated (Meyer and Gurevych, 2012). We thus cannot determine the correct target sense if a relation is pointing to an ambiguous word.

Our solution to this is twofold: First, for each sense s, we create an edge (s, t) for those semantic relations which have a monosemous target t, as in this case the target sense is unambiguous. This approach, however, only recovers a subset of the relations, and it is not applicable to resources where no sense relations exist at all, e.g. FN. For this case, we propose to use the glosses of senses in the LSR to derive additional edges in the following way: For each monosemous, non-stopword lexeme l (a combination of lemma and part of speech) in the gloss of a sense s_1 with a sense s_l, we introduce an edge (s_1, s_l). Moreover, if there is another sense s_2 with l in its gloss, we also introduce an edge (s_1, s_2). This technique will be called *linking of monosemous lexemes* or *monosemous linking* throughout the rest of this paper. The intuition behind this is that monosemous lexemes usually have a rather specific meaning, and thus it can be expected that the senses in whose description they appear have at least a certain degree of semantic relationship. This directly relates to the notion of "information content" (Resnik, 1995), stating that senses in an LSR which are more specific (and hence more likely to be monosemous) are more useful for evaluating semantic similarity. Note that this step requires part of speech tagging

of the glosses, which we perform as a preprocessing step. Thereby we filter out stopwords and words tagged as "unknown" by the POS tagger.

As an example, consider the gloss of *Java*: "An object-oriented programming language". Even in the absence of any semantic relations, we could unambiguously derive an edge between this sense of *Java* and the multiword noun *programming language* if the latter is monosemous, i.e. if there exists exactly one sense for this lexeme in the LSR. Also, if *programming language* appears in the gloss of one of the senses of *Python*, we can derive an edge between these senses of *Java* and *Python*, expressing that they are semantically related.

An important factor to keep in mind, however, is the density of the resulting graph. In preliminary experiments, we discovered that linking every monosemous lexeme yielded very dense graphs with short paths between most senses. In turn, we decided to exclude "common" lexemes and focus on more specific ones in order to increase the graph's meaningfulness. The indicator for this is the frequency of a lexeme in the LSR, i.e. how often it occurs in the glosses. Our experiments on small development sets (100 random samples of each gold standard) indeed show that a strict filter leads to discriminative edges resulting in high precision, while at the same time graph sparsity decreases recall. Independently of the resource pair, we discovered that setting this frequency limit value ϕ to about 1/100 of the graph size (e.g. 1,000 for a graph containing 100,000 senses) gives the best balance between precision and recall; larger values of ϕ usually led to no significant improvement[1] in recall while the precision was continuously degrading. Note that WP was excluded from these experiments as the identification and linking of monosemous lexemes in all WP articles proved too time-consuming; instead, we decided to use only the already given links (see Section 5.3).

4.2 Computing Sense Alignments

Initialization After resource graphs for both LSRs A and B are created, the trivial alignments are retrieved and introduced as edges between them. Trivial alignments are those between senses which have the same attached lexeme in A and B and where this lexeme is also monosemous within either resource. E.g., if the noun phrase *programming language* is contained in either resource and has exactly one sense in each one, we can directly infer the alignment. For WP, a lexeme was considered monosemous if there was exactly one article with this title, also counting titles with a bracketed disambiguation (e.g., *Java (programming language)* and *Java (island)* are two distinct senses of *Java*). While this method does not work perfectly, we observed a precision > 0.95 for monosemous gold standard senses, which is in line with the observations by Henrich et al. (2011).

Alignment We consider each sense $s \in A$ which has not been aligned in the initialization step. For this, we first retrieve the set of possible target senses $T \subset B$ (those with matching lemma and part of speech) and compute the shortest path to each of them with Dijkstra's shortest path algorithm (Dijk-

```
    Dijkstra-WSA(A,B)
1   ASenseSet = A.senses
2   BSenseSet = B.senses
3   UnalignableSenses = ∅
4
5   foreach sense s ∈ ASenseSet
6     if(s.isMonosemous)
7       t = findTrivialMatch(s, BSenseSet)
8       if(t != null)
9           ASenseSet.remove(s)
10          BSenseSet.remove(t)
11          createEdge(s,t)
12
13  foreach sense s' ∈ ASenseSet
14    ASenseSet.remove(s')
15    T=findCandidatesWithSameLexeme(s', B)
16    if(T!= ∅)
17      t'=findShortestPathToCandidates(s', T)
18      if(t' != null)
19          createEdge(s', t')
20      else
21          UnalignableSenses.put(s')
22    else
23      UnalignableSenses.put(s')
```

Table 2: Pseudocode of the Dijkstra-WSA algorithm.

[1] All significance statements throughout the paper are based on McNemar's test and the confidence level of 1%.

stra, 1959). The candidate $t \in T$ with the shortest distance is then assigned as the alignment target, and the algorithm continues with the next still unaligned sense in A until either all senses are aligned or no path can be found for the remaining senses. The intuition behind this is that the trivial alignments from the initialization serve as "bridges" between A and B, such that a path starting from a sense s_1 in A traverses edges to find a nearby already aligned sense s_2, "jumps" to B using a cross-resource edge leading to t_2 and then ideally finds an appropriate target sense t_1 in the vicinity of t_2. Note that with each successful alignment, edges are added to the graph so that, in theory, a different ordering of the considered senses would lead to different results. While we observed slight differences for repeated runs using the same configuration, these were in no case statistically significant. The pseudo code of this algorithm can be found in Table 2, while an example can be found in Figure 1.

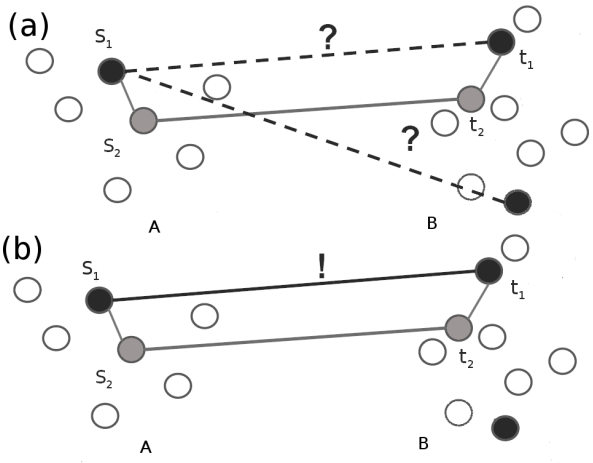

Figure 1: An example of how Dijkstra-WSA works. While there exist 2 candidates for aligning a sense $s_1 \in A$ (dashed lines) (a), the correct one $t_1 \in B$ can be determined by finding the shortest path using an already established edge between two monosemous senses $s_2 \in A$ and $t_2 \in B$ (solid line) (b).

Parameter Influence Apart from the already mentioned parameter ϕ for limiting the number of edges in the graph, another important variable is the maximum allowed path length λ of Dijkstra's algorithm. In general, allowing an unbounded search for the candidate senses is undesirable as long paths, while increasing recall, usually also lead to a decrease in precision, as the nodes which can be reached in many steps are usually also semantically distant from the source sense. In this respect, we found significant differences between the optimal configuration for individual resource pairs. However, the general observation is that short paths ($\lambda \leq 3$) lead to a very high precision, while paths longer than 10 do not increase recall significantly any more.

A modification of the algorithm is to not only align the closest target sense, but all senses which can be reached with a certain number of steps. This caters to the fact that, due to different sense granularities, one coarser sense in A can be represented by several senses in B and vice versa (see Table 3 for the fraction of 1:n alignments in the datasets). Regarding this modification, we made the observation that the recall improved (sometimes considerably), but at the same time the precision decreased, sometimes to an extent where the overall F-Measure (the harmonic mean of precision and recall) got worse. In the evaluation section, we state which setting is used for which datasets and configurations.

5 Experimental Work

5.1 Datasets and their Characteristics

WN 3.0-English OW The previous alignment between these two resources reported in Gurevych et al. (2012) is based on the German OW (database dump from 2010/01/03) and WN 3.0 and utilizes gloss similarities using machine translation as an intermediate component. This does not pose a problem since for each synset in the German part of OW, there is a translation in the English part. This makes the German-English gold standard directly usable for our purposes.[2] Table 3 presents the details about this as well as the other evaluation datasets, including the observed inter-rater agreement A_0 (where available) which can be considered as an upper bound for automatic alignment accuracy and the degree of polysemy (i.e. the number of possible alignment targets per sense) which is a hint towards the difficulty of the alignment task.

WN 3.0-English WKT We use the gold standard dataset from Meyer and Gurevych (2011) without any modification, thus for comparability to this

[2]Cross-lingual alignment is left to future work.

work, we use the same WKT dump version (from 2010/02/01) which contains around 421,000 senses.

GN 7.0-German WKT Henrich et al. (2011) aligned the German WKT (dump from 2011/04/02, 72,000 senses) and GN 7.0. This is the only existing alignment between these two resources so far, and we use their freely available dataset[3] to test Dijkstra-WSA on a language other than English. As this alignment is fairly large (see Table 3), we created a random sample as a gold standard to keep the computation time at bay. However, the datasets are still similar enough to allow direct comparison of the results. Note that no inter-annotator agreement is available for this study.

WN 3.0-English WP We use the gold standard from Niemann and Gurevych (2011). For comparability, we use the same Wikipeda dump version (from 2009/08/22) with around 2,921,000 articles.

5.2 Baselines

WN-OW We used the same configuration as in Gurevych et al. (2012) to calculate a similarity-based alignment for the monolingual case (i.e. without the translation step) as a baseline and achieved comparable results.

WN-WKT As stated above, the alignment[4] presented in Meyer and Gurevych (2011) was created by calculating the similarity of glosses and training a machine-learning classifier on the gold standard to classify each pair of senses.

GN-WKT The automatic alignment results (i.e. the outcome of the algorithm without manual post-correction) reported by Henrich et al. (2011) were unavailable for us as a baseline. Thus, we utilize the alignment approach by Meyer and Gurevych (2011) to create a similarity-based baseline, with minor modifications. Unlike the original approach, we directly align senses regardless of their similarity if the decision is trivial (see Section 4.2). We also do not train a machine learning component on a gold standard. Instead, we adapt the idea of Henrich et al. (2011) to align the most similar candidate regardless of the absolute value.

WN-WP The alignment reported in Niemann and Gurevych (2011) was created in the same way as the WN-WKT alignment described in Meyer and Gurevych (2011). Note that while the full alignment results[5] proved incomplete, the correct alignment results on the gold standard were available and thus used in our experiments.

We will henceforth mark these similarity-based results with *SB*.

5.3 System Configurations

For the construction of the resource graphs we experimented with three options:

Semantic relations only (SR) OW, WN and GN all feature disambiguated sense relations which can be directly used as edges between senses. Note that in the expert-built resources, the majority of nodes are connected by sense relations, while this is not the case for OW. For WKT, only the unambiguous semantic relations can be used (see Section 4.1), resulting in graphs less dense and with many isolated nodes. However, as we reported in Matuschek et al. (2013), the English WKT is almost 6 times as large as the German one for the versions we used in our experiments (421,000 senses vs. 72,000 senses), while it contains not even twice as many relations (720,000 vs. 430,000). This is directly reflected in the fewer isolated nodes for the German WKT. WP links are also unambiguous as they lead to a distinct article. However, intuitively not all links in an article are equally meaningful. Thus, for the SR configuration, we decided to retain only the category links and the links within the first paragraph of the article. We assume that the targets of these links are most closely related to the sense an article represents as the first paragraph usually includes a concise definition of a concept, and the category links allow determining the topic an article belongs to.

Linking of monosemous lexemes only (LM) For this configuration, the limiting parameter ϕ was set to 1/100 of the graph size for every resource except WP as described in section 4.1. As our experiments show, linking the monosemous lexemes in the glosses while disregarding semantic relations results in well-connected graphs for all resources but GN and WKT. Only about 10% of the GN senses have a gloss, thus this option was completely disre-

[3] http://www.sfs.uni-tuebingen.de/GermaNet/wiktionary.shtml

[4] Available at http://www.ukp.tu-darmstadt.de/data/lexical-resources/wordnet-wiktionary-alignment/

[5] Available at http://www.ukp.tu-darmstadt.de/data/lexical-resources/wordnet-wikipedia-alignment/

Pair	Aligned	Not Aligned	Sum	1:n Alignments %	Polysemy	Sampling	A_0
WN-OW	210	473	683	10.7%	1.50	Random	0.85
WN-WKT	313	2,110	2,423	2.7%	4.76	Balanced	0.93
GN-WKT (full)	27,127	18,509	45,636	5.6%	1.78	All	N/A
GN-WKT (sample)	1,000	751	1,751	4.8%	1.84	Random	N/A
WN-WP	227	1,588	1,815	5.2%	5.7	Balanced	0.97

Table 3: Characteristics of the gold standards used in the evaluation. A_0 is the observed inter-rater agreement which can be considered as an upper bound for alignment accuracy. The degree of polysemy (i.e. the number of possible alignment targets per sense) hints towards the difficulty of the alignment task.

garded in this case. For both WKTs, an analysis of the graphs revealed that the reason for the relatively high number of isolated nodes are very short glosses, containing many polysemous lexemes. For WP, we refrained from monosemous linking due to the prohibitive computation time. Instead, we decided to use the fully linked WP (excluding the links used for the SR configuration) in this case. The rationale is that in the majority of articles many meaningful terms link to the corresponding articles anyway, so that the resulting graph is comparable with those for the other LSRs.

Combining both (SR+LM) This configuration yields the maximum number of available edges. We report the results for GN only for this configuration and omit the SR results for the sake of brevity as the influence on the F-Measure for the GN-WKT alignment (see Section 5.4) is not statistically significant. For WKT, this configuration only increases the number of connected nodes slightly (as insufficient glosses often coincide with missing semantic relations), while for OW an almost connected graph can be constructed.

Table 4 gives an overview of the fraction of isolated nodes for each resource in every configuration.

Note again that for each alignment task (i.e. each pair of resources), we tuned the parameters on 100 random samples from each gold standard for a result balancing precision and recall as discussed above. Individual tuning of parameters was necessary for each pair due to the greatly varying properties of the LSRs (e.g. the number of senses). While it would have been ideal to train and test on disjoint sets, we calculated the overall results on the full gold standards including the development sets to ensure comparability with the previous work.

Hybrid Approach Manual inspection of the results revealed that the alignments found by Dijkstra-WSA are usually different from those based on the gloss similarity. While the latter precisely recognizes alignments with similar wording of glosses, Dijkstra-WSA is advantageous if the glosses are different but the senses are still semantically close in the graph. Section 5.5 will analyze this in greater detail. Exploiting this fact, we experimented with a hybrid approach: We perform an alignment using Dijkstra-WSA, tuned for high precision (i.e. using shorter path lengths) and fall back to using the results of the similarity-based approaches for those cases where no alignment target could be found in the graph. These results are marked with *+SB* in the result overview (Table 5).

Resource	SR	LM	SR+LM
WN	0.25	0.07	0.02
GN	0.0	0.92	0.0
WKT-en	0.98	0.32	0.30
WKT-de	0.69	0.18	0.15
OW	0.41	0.33	0.04
WP	0.06	0.05	0.04

Table 4: This table describes what percentage of nodes remains isolated (i.e. with 0 attached edges) in different graph configurations using semantic relations (SR), monosemous linking (LM) ($\phi = 1/100$) or both (SR+LM). Note that this number is highest for the English WKT as the few semantic relations and short glosses do not offer many possibilities for connecting nodes, while the German WKT and OW do not suffer from this problem as much. GN is fully linked via relations, but has only few glosses which makes monosemous linking ineffective. WN and WP are relatively well-linked in all configurations. Also note that for WP, SR means that we used category links and links from the first paragraph, while links from the rest of the article were used for the LM configuration.

	WordNet-OmegaWiki				WordNet-Wiktionary				GermaNet-Wiktionary				WordNet-Wikipedia			
	P	R	F_1	$Acc.$	P	R	F_1	$Acc.$	P	R	F_1	$Acc.$	P	R	F_1	$Acc.$
Random	0.46	0.35	0.40	0.51	0.21	0.59	0.31	0.67	0.44	0.51	0.47	0.54	0.49	0.62	0.53	0.86
SB	0.55	0.53	0.54	0.73	0.67	0.65	0.66	0.91	0.93	0.74	0.83	0.83	0.78	0.78	0.78	**0.95**
SR	**0.66**	0.45	0.53	0.76	**0.95**	0.13	0.23	0.89	**0.94**	0.65	0.77	0.78	**0.82**	0.63	0.71	0.93
LM	0.62	0.54	0.58	**0.77**	0.72	0.24	0.36	0.89	0.89	0.75	0.81	0.80	0.65	0.66	0.65	0.91
SR+LM	0.56	0.69	0.62	0.74	0.68	0.27	0.39	0.89	0.90	0.78	0.83	0.82	0.75	0.67	0.71	0.93
SR+SB	0.60	0.65	0.63	0.76	0.68	0.67	0.68	0.92	0.90	0.82	0.86	0.84	0.75	**0.87**	**0.81**	**0.95**
LM+SB	0.60	0.70	0.64	0.76	0.68	0.70	**0.69**	**0.92**	0.86	0.87	**0.87**	**0.85**	0.70	**0.87**	0.78	0.94
SR+LM+SB	0.57	**0.75**	**0.65**	0.75	0.68	**0.71**	**0.69**	**0.92**	0.87	**0.88**	**0.87**	**0.85**	0.75	**0.87**	**0.81**	**0.95**
A_0	-	-	-	0.85	-	-	-	0.93	-	-	-	N/A	-	-	-	0.97

Table 5: Alignment results for all datasets and configurations: Using semantic relations (SR), monosemous links (LM) or both (SR+LM). The similarity-based (SB) baselines, also used as a back-off for the hybrid approaches (+SB), were created using the approach reported in Gurevych et al. (2012). Note that for GN, the SR+LM configuration was always used. The different configurations given for this alignment thus only apply to WKT. For WP, SR means that only category links and links within the first paragraph were used, while LM uses links from the full article. A random baseline and the inter-annotator agreement A_0 of the gold standard annotation (if available) are given for reference.

5.4 Experimental Results

WN-OW When using only semantic relations (SR), we achieved an F-Measure of 0.53 which is comparable with the 0.54 from Gurevych et al. (2012). Notably, our approach has a high precision, while the recall is considerably worse due to the relative sparsity of the resulting OW resource graph. When adding more edges to the graph by linking monosemous lexemes (SR+LM), we can drastically improve the recall, leading to an overall F-Measure of 0.62, which is a significant improvement over our previous results (Gurevych et al., 2012). Using monosemous links only (LM), the result of 0.58 still outperforms Gurevych et al. (2012) due to the higher precision. Building a graph from glosses alone is thus a viable approach if no or only few semantic relations are available. Regarding the path lengths, $\lambda = 10$ works best when semantic relations are included in the graph, while for the LM configuration shorter paths ($\lambda \leq 5$) were more appropriate. The intuition behind this is that for semantic relations, unlike monosemous links, even longer paths still express a high degree of semantic relatedness. Also, when semantic relations are involved allowing multiple alignments increases the overall results (which is in line with the relatively high number of 1:n alignments in the gold standard), while this is not the case for the LM configuration; here, the edges again do not sufficiently express relatedness.

Using the hybrid approach (+SB), we can increase the F-Measure up to 0.65 if semantic relations and monosemous linking are combined (SR+LM) and the parameters are tuned for high precision ($\lambda \leq 3$, 1:1 alignments). This is significantly better than Dijkstra-WSA alone in any configuration. In this scenario, we also observe the best overall recall.

WN-WKT Experiments using only the semantic relations (SR) yield a very low recall - the small number of sense relations with monosemous targets in WKT leaves the graph very sparse. Nevertheless, the alignment targets which Dijkstra-WSA finds are mostly correct, with a precision greater than 0.95 even when allowing 1:n alignments. Using only monosemous links (LM) improves the recall considerably, but unlike the WN-OW alignment, it stays fairly low. Consequently, even when using semantic relations and monosemous links in conjunction (SR+LM), the recall can only be increased slightly, leading to an overall F-Measure of 0.39. As mentioned above, this is due to the WKT glosses. In many cases, they are very short, often consisting of only 3-5 words, many of which are polysemous. This leads to many isolated nodes in the graph with no or only very few connecting edges. The ideal, rather short path length λ of 2-3 stems from the relatively high polysemy of the gold standard (see Table 3). We experimented with $\lambda \geq 4$, achieving reasonable recall, but in this case the precision was so low that this configuration, in conclusion, does not increase the F-Measure. However, 1:n alignments work well with these short paths as the correct alignments are mostly in the close vicinity of a sense,

hence we achieve an increase in recall in this case without too much loss of precision.

For the hybrid approach, we achieve an F-Measure of 0.69 when using all edges (SR+LM+SB), setting the path length to 2, and also allowing 1:n alignments. This is a statistically significant improvement over Meyer and Gurevych (2011) which again confirms the effectiveness of the hybrid approach.

GN-WKT As stated above, we used the SR+LM configuration for GN in every case. For the German WKT, the much greater number of relations compared to its English counterpart is directly reflected in the results, as using the semantic relations only (SR) not only yields the best precision of 0.94 but also a good recall of 0.65. Using the semantic relations together with monosemous links (SR+LM) yields the F-Measure of 0.83, which is on par with the similarity-based (SB) approach.

In the hybrid configuration, we can increase the performance to an F-Measure of up to 0.87 (SR+LM+SB), significantly outperforming all graph-based and similarity-based configurations.

In general, results for this pair of LSRs are higher in comparison with the others. We attribute this to the fact that the German WKT and GN both are densely linked with semantic relations which is especially beneficial for the recall of Dijkstra-WSA. This is also reflected in the ideal λ of 10-12: Many high-confidence edges allow long paths which still express a considerable degree of relatedness. However, while the results for 1:n alignments are already good, restricting oneself to 1:1 alignments gives the best overall results as the precision can then be pushed towards 0.90 without hurting recall too much. An important factor in this respect is that the GN-WKT dataset has a relatively low degree of polysemy (compared to WN-WKT) and only few 1:n alignments (compared to WN-OW), two facts which make the task significantly easier.

WN-WP The SR configuration (WN relations + WP category/first paragraph links) yields the best precision (0.82), even outperforming the SB approach, and an F-Measure of 0.71. This again shows that using an appropriate parametrization ($\lambda \leq 4$ in this case) Dijkstra-WSA can detect alignments with high confidence. The relatively low recall of 0.63 could be increased by allowing longer paths, however, as hyperlinks do not express relatedness as reliably as semantic relations, this introduces many false positives and thus hurts precision considerably. This issue of "misleading" WP links becomes even more prominent when the links from the full articles are used as edges (LM); while the increase in recall is relatively small the precision drops substantially. However, using all possible links (SR+LM) allows us to balance out precision and recall to some extent, while yielding the same F-Measure as the SR configuration. Note that 1:1 alignments were enforced in any case, as the high polysemy of the dataset in conjunction with the dense WP link structure rendered 1:n alignments very imprecise.

Using the hybrid approach, we can increase the F-Measure up to 0.81 (SR+SB), outperforming the results reported in Niemann and Gurevych (2011) by a significant margin. The F-Measure for LM+SB is slightly worse due to the lower precision. Combining all edges (SR+LM+SB) does not influence the results any more, but in any case the hybrid configuration achieves the best overall recall (0.87).

In conclusion, our experiments on all four datasets consistently demonstrate that combining Dijkstra-WSA with a similarity-based approach as a back-off yields the strongest performance. The results of these best alignments will be made freely available to the research community on our website (http://www.ukp.tu-darmstadt.de).

5.5 Error Analyis

The by far most significant error source, reflected in the relatively low recall for different configurations, is the high number of false negatives, i.e. sense pairs which were not aligned although they should have been. This is especially striking for the WN-WKT alignment. As discussed earlier, WKT contains a significant number of short glosses, which in many cases also contain few or no monosemous terms. A prototypical example is the first sense of *seedling*: "A young plant grown from seed". This gloss has no monosemous words which could be linked, and as there are also no semantic relations attached to this sense which could be exploited, the node is isolated in the graph. Our experiments show that for the English WKT, even when optimizing the parameters for recall, around 30% of the senses remain isolated, i.e. without edges. This is by far the high-

est value across all resources (see Table 4). Solving this problem would require making the graph more dense, and especially finding ways to include isolated nodes as well. However, this example also shows why the hybrid approach works so well: The correct WN sense "young plant or tree grown from a seed" was recognized by the similarity-based approach with high confidence.

With regard to false positives, Dijkstra-WSA and the similarity-based approaches display very similar performance. This is because senses with very similar wording are likely to share the same monosemous words, leading to a close vicinity in the graph and the false alignment. As an example, consider two senses of *bowdlerization* in WN ("written material that has been bowdlerized") and WKT ("The action or instance of bowdlerizing; the omission or removal of material considered vulgar or indecent."). While these senses are clearly related, they are not identical and should not be aligned, nevertheless the similar wording (and especially the use of the highly specific verb "bowdlerize") results in an alignment. Similarly to the similarity-based approaches, it is an open question how this kind of error can be effectively avoided (Meyer and Gurevych, 2011).

There is a considerable number of examples where Dijkstra-WSA recognizes an alignment which similarity-based approaches do not. The two senses of *Thessalonian* from the introductory example (Section 3) contain the terms *Thessalonica* and *Thessaloniki* in their glosses which are both monosemous in WN as well as in WKT, sharing the also monosemous noun *Greece* in their glosses. This yields the bridge between the resources to find a path and correctly derive the alignment.

6 Conclusions and Future Work

In this work, we present Dijkstra-WSA, a graph-based algorithm for word sense alignment. We show that this algorithm performs competitively on 3 out of 4 evaluation datasets. A hybrid approach leads to a statistically significant improvement over similarity-based state of the art results on every dataset. Dijkstra-WSA can operate on glosses or semantic relations alone (although it is beneficial if both are combined), and it does not require any external knowledge in the form of annotated training data or corpora. Additionally, it is flexibly configurable for different pairs of LSRs in order to optimize for precision or recall.

An important task for future work is to evaluate Dijkstra-WSA on LSRs which structurally differ from the ones discussed here. It is important to determine how resources like FN or VN can be meaningfully transformed into a graph representation. Another idea is to extend the approach to cross-lingual resource alignment, which would require a machine translation component to identify sense alignment candidates with the correct lexeme.

Regarding the algorithm itself, the main direction for future work is to increase recall while keeping high precision. One possible way would be to not only link monosemous lexemes, but also to create edges for polysemous ones. Laparra et al. (2010) discuss a possibility to do this with high precision. The main idea is to focus on lexemes with a low degree of polysemy and align if one of the possible senses is clearly more similar to the source sense than the other(s). If recall is still low, more polysemous lexemes can be examined.

A weighting of edges (e.g. based on gloss similarities) has not been considered at all, but would be easily applicable to the existing framework.

A more elaborate idea would be to investigate entirely different graph-based algorithms, e.g. for matching nodes in bipartite graphs. Also, we plan to investigate if and how our approach can be extended to align more than two resources at once using the graph representations. This might improve alignment results as more information about the overall alignment topology becomes available.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant No. I/82806 and by the Hessian research excellence program "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz (LOEWE)" as part of the research center "Digital Humanities". We would like to thank Christian M. Meyer, Wolfgang Stille, Karsten Weihe and Tristan Miller for insightful discussions and comments. We also thank the anonymous reviewers for their helpful remarks.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41, Athens, Greece.

Gerard de Melo and Gerhard Weikum. 2010. Providing Multilingual, Multimodal Answers to Lexical Database Queries. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC 2010)*, pages 348–355, Valetta, Malta.

Edsger. W. Dijkstra. 1959. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271. 10.1007/BF01386390.

Judith Eckle-Kohler, Iryna Gurevych, Silvana Hartmann, Michael Matuschek, and Christian M. Meyer. 2012. UBY-LMF - A Uniform Model for Standardizing Heterogeneous Lexical-Semantic Resources in ISO-LMF. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, pages 275–282, Istanbul, Turkey.

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

Tiziano Flati and Roberto Navigli. 2012. The CQC algorithm: Cycling in graphs to semantically enrich and enhance a bilingual dictionary. *Journal of Artificial Intelligence Research (JAIR)*, 43:135–171.

Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Christian M. Meyer, and Christian Wirth. 2012. UBY - A Large-Scale Unified Lexical-Semantic Resource Based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12)*, pages 580–590, Avignon, France.

Birgit Hamp and Helmut Feldweg. 1997. Germanet - a lexical-semantic net for german. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2011. Semi-Automatic Extension of GermaNet with Sense Definitions from Wiktionary. In *Proceedings of the 5th Language and Technology Conference (LTC 2011)*, pages 126–130, Poznan, Poland.

Egoitz Laparra, German Rigau, and Montse Cuadros. 2010. Exploring the integration of WordNet and FrameNet. In *Proceedings of the 5th Global WordNet Conference (GWC'10)*, Mumbai, India.

Michael Matuschek, Christian M. Meyer, and Iryna Gurevych. 2013. Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Computer-Aided Translation. *Translation: Computation, Corpora, Cognition. Special Issue on "Language Technology for a Multilingual Europe"*, to appear.

Christian M. Meyer and Iryna Gurevych. 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)*, pages 883–892, Chiang Mai, Thailand.

Christian M. Meyer and Iryna Gurevych. 2012. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Sylviane Granger and Magali Paquot, editors, *Electronic Lexicography*, chapter 13, pages 259–291. Oxford University Press.

Rada Mihalcea and Dan I. Moldovan. 2001. eXtended WordNet: progress report. In *Proceedings of NAACL Workshop on WordNet and Other Lexical Resources*, pages 95–100, Pittsburgh, PA, USA.

David Milne and Ian H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, Chicago, IL, USA.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Roberto Navigli. 2009. Using Cycles and Quasi-Cycles to Disambiguate Dictionary Glosses. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL'09)*, pages 594–602, Athens, Greece.

Elisabeth Niemann and Iryna Gurevych. 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)*, pages 205–214, Oxford, UK.

Martha Palmer. 2009. SemLink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon ConferenceGenLex-09*, pages 9–15, Pisa, Italy.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the* 21^{st} *International Joint Conference on Artificial Intelligence*, pages 2083–2088, Pasadena, CA, USA.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *International Joint Conference for Artificial Intelligence (IJCAI-95)*, pages 448–453, Montreal, Canada.

Lei Shi and Rada Mihalcea. 2005. Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing. In *Computational Linguistics and Intelligent Text Processing: 6th International Conference*, volume 3406 of *Lecture Notes in*

Computer Science, pages 100–111. Berlin/Heidelberg: Springer.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO: A Large Ontology from Wikipedia and WordNet. *Web Semantics*, 6(3):203–217.