High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity

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 a machine learning approach for word sense alignment (WSA) which combines distances between senses in the graph representations of lexical-semantic resources with gloss similarities. In this way, we significantly outperform the state of the art on each of the four datasets we consider. Moreover, we present two novel datasets for WSA between Wiktionary and Wikipedia in English and German. The latter dataset in not only of unprecedented size, but also created by the large community of Wiktionary editors instead of expert annotators, making it an interesting subject of study in its own right as the first crowdsourced WSA dataset. We will make both datasets freely available along with our computed alignments.

1 Introduction

Lexical-semantic resources (LSRs) are an important foundation for numerous natural language processing (NLP) tasks such as word sense disambiguation (WSD) or information extraction (IE). However, large-scale LSRs are only available for a few languages. The Princeton WordNet (Fellbaum, 1998) is commonly used for English, but for most languages such resources are small or missing altogether. Another problem is that, even for English, there is no single LSR which is suitable for all different application scenarios, because the resources contain different words, senses or even information types. Recently, it has been argued that collaboratively constructed resources (e.g. Wiktionary (Meyer and Gurevych, 2012))) are a viable alternative, especially for smaller languages (Matuschek et al., 2013), but there are still considerable drawbacks in coverage which make their usage challenging.

These observations have led to the insight that word sense alignment (WSA), i.e. linking at the level of word senses, is key for the efficient exploitation of LSRs, and it was shown that the usage of linked resources can indeed yield performance improvements. Examples include WSD using aligned WordNet and Wikipedia (Navigli and Ponzetto, 2012a), semantic role labeling using PropBank, VerbNet and FrameNet (Palmer, 2009), the construction of a semantic parser using FrameNet, WordNet, and VerbNet (Shi and Mihalcea, 2005) and IE using WordNet and Wikipedia (Moro et al., 2013). Cholakov et al. (2014) address the special task of verb sense disambiguation. They use the large-scale resource UBY (Gurevych et al., 2012) which contains nine resources in two languages, mapped to a uniform representation using the LMF standard for interoperability (Eckle-Kohler et al., 2012), and also (among others) sense alignments between WordNet, FrameNet, VerbNet and Wiktionary which are exploited in their approach.

However, WSA is challenging because of word ambiguities, different sense granularities and information types (Navigli, 2006), so that past efforts mostly focused on specific resources or applications, where expert-built resources such as WordNet played a central role in most cases. Approaches which aim at being more generic (i.e. applicable to a wider range of LSRs) usually focused on only one information source for the alignment (e.g. glosses or graph structures) without combining them in an elaborate way.

In this paper, we want to go beyond this previous work in two ways: i) For the first time, we present an alignment between the large-scale collaboratively constructed resources Wiktionary and Wikipedia. While both LSRs have been extensively used in NLP and especially WSA (see Section 2), no attempt has

been made to combine them, although Wiktionary was explicitly designed to complement the encyclopedic knowledge in Wikipedia with linguistic knowledge. Apart from already established tasks like WSD, the strong multilingual focus of both resources makes their combination especially promising for applications such as knowledge-based machine translation or computer-assisted translation where additional background knowledge and translation options can be crucial (Matuschek et al., 2013). To fill this gap in the body of research, we present two new evaluation datasets for English and German, where the latter is not only of remarkable size, but also directly extracted from Wiktionary in a novel approach, making it the first crowdsourced WSA dataset. ii) Also for the first time, we jointly model different aspects of sense similarity by applying machine learning techniques to WSA. However, unlike previous approaches, we do not engineer our features towards a specific resource pair, rendering the approach powerful but proprietary. Instead, we aim to combine generic features which are applicable to a variety of resources, and we show that combining them leads to state-of-the-art WSA performance. In particular, we employ distances calculated with Dijkstra-WSA (Matuschek and Gurevych, 2013), an algorithm which works on graph representations of resources, as well as gloss similarity values. This lets us take advantage of both (orthogonal) ways of identifying equivalent senses and yields a very robust and flexible WSA framework.

The rest of this paper is structured as follows: In Section 2 we discuss related work, in Section 3 we describe our approach and introduce the resources and datasets we use in our experiments, in Section 4 we evaluate our results, and we conclude in Section 5 with some directions for future work.

2 Related Work

There are two main approaches to WSA which have been applied: Similarity-based and graph-based ones. To our knowledge, there exists no previous work which effectively combines both approaches in a unified framework, and only few works which combine both kinds of features for different purposes.

2.1 Similarity-based Approaches

WordNet was aligned to Wikipedia (Niemann and Gurevych, 2011) and Wiktionary (Meyer and Gurevych, 2011) using a framework based on gloss similarity, in spirit of the earliest work in WSD presented by Lesk (1986). In both cases, cosine and personalized PageRank (PPR) similarity (Agirre and Soroa, 2009) were calculated, and a simple machine learning approach was used to classify each pair of senses (see Section 3.3). This idea was also applied to cross-lingual alignment between WordNet and the German part of OmegaWiki (Gurevych et al., 2012), using machine translation as an intermediate component. Henrich et al. (2011) use a similar approach for aligning GermaNet and Wiktionary, but with word overlap as the similarity measure. De Melo and Weikum (2010) report an alignment of WordNet synsets to Wikipedia articles which is also based on word overlap. We later report results based on gloss similarity as one of our baselines (Tables 2 and 3).

2.2 Graph-based Approaches

In one of the earliest structure-based works, Daudé et al. (2003) map different versions of WordNet based on the synset hierarchy. Navigli (2009) disambiguates WordNet glosses, i.e. sense markers are assigned to all non-stopwords in each WordNet gloss. The approach is based on finding circles in the WordNet relation graph to identify disambiguations. In later work, this idea was applied to the disambiguation of translations in a bilingual dictionary (Flati and Navigli, 2012). While this "alignment" of dictionary entries is related to our problem, it was not discussed how this idea could be applied to word sense alignment of two resources. Laparra et al. (2010) use a shortest path algorithm (SSI-Dijkstra+) to align FrameNet lexical units (LUs) with WordNet synsets. They align monosemous LUs first and then search for the closest synset in WordNet for the other LUs in the same frame. The LUs are, however, considered as mere texts to be disambiguated; there is no attempt made to exploit the graph structure of FrameNet. Ponzetto and Navigli (2009) use a graph-based method for aligning WordNet synsets and Wikipedia categories. Using semantic relations, they build subgraphs of WordNet for each category and then align senses to categories based on the structural features. In our own previous work, we presented Dijkstra-

WSA, a graph-based approach working with shortest paths (Matuschek and Gurevych, 2013). It achieves state-of-the-art precision, but recall is an issue if the graphs are sparse (i.e. in case of only few semantic relations). As Dijkstra-WSA distances are one of the features we use for our machine learning approach, we will present this approach in more detail in section 3.2.2 and also report results for Dijkstra-WSA on our evaluation datasets for comparison.

2.3 Hybrid Approaches

In later work, Navigli and Ponzetto (2012a) also align WordNet with the full Wikipedia. Besides using bag-of-words overlap to compute gloss similarity, they also build a graph structure for the senses in both resources by using WordNet semantic relations. The goal is to determine which WordNet sense is closest to the Wikipedia sense to be aligned. However, the graph structure of Wikipedia is disregarded, as is the global structure of WordNet, as just a locally restricted subset of WordNet relations is used. In the same context of BabelNet, Navigli and Ponzetto (2012b) also present BabelRelate, an approach which relies on translations to compute cross-lingual semantic similarity; however, they do not apply it to WSA. Dijkstra-WSA was enhanced by using a backoff, by means of performing a graph-based alignment first, and in cases where no alignment target sense can be found, a decision is made based on the similarity of glosses (Matuschek and Gurevych, 2013). While this simple two-step approach increases recall substantially, it comes at the expense of lower precision. However, the overall F-measure achieved state-of-the-art performance on every considered dataset (0.65–0.87). We also report the results for this hybrid approach as a baseline (Tables 2 and 3). De Melo and Weikum (2008) use a machine learning approach with a combination of structural and content-based features of WordNet, but for building new wordnets in other languages, not aligning existing ones.

In summary, the different approaches to compute similarity have mostly been used in isolation, or combined in a shallow or restricted way. More complex approaches usually require resource-specific feature engineering, which makes their transferability to other resources or languages difficult. Thus, we present a framework which combines different similarity measures in a generic and flexible way and enables state-of-the-art WSA performance on a variety of resources with modest effort.

3 The Alignment Procedure

The basic steps of our alignment algorithm are:

1. For each sense in one resource, all possible candidates in the other resource are retrieved. Candidates are senses which have the same attached lemma and part of speech. For instance, for the *programming* sense of *Java* in one resource, their might exist senses for *programming*, *island* or *coffee* in the other one which are all possible alignment targets.

2. For each candidate pair, we calculate a set of features describing their similarity in different ways.

3. For a set of word senses (the gold standard), the alignment decision is made by human annotators.

4. A machine learning classifier is trained on this gold standard, and an alignment decision is made for the remainder of the candidate pairs to produce a complete alignment of the resources. In our setup, we use 10-fold cross validation to train the classifier.

The different datasets and steps of the algorithm are explained in more detail in the following sections.

3.1 Resources and Datasets

We use four different WSA evaluation datasets, two of which are presented for the first time. To ensure compatibility with previous work, we use the same versions of the resources as reported in (Gurevych et al., 2012) and (Matuschek and Gurevych, 2013).

Pair	Pos.	Neg.	Polysemy	One cand.	F_1	A_0	Composition
WordNet-OmegaWiki	210	473	1.50	75.2%	0.84	0.85	random
WordNet-Wiktionary	313	2 110	4.76	18.6%	0.78	0.93	manual
Wiktionary-Wikipedia (En)	75	292	1.27	87.6%	0.79	0.95	automatic
Wiktionary-Wikipedia (De)	21 855	9 953	1.47	77.6%	0.85	0.89	crowd

Table 1: Characteristics of the gold standards used in the evaluation. The degree of polysemy (i.e. the number of possible alignment targets per sense) hints towards the difficulty of the task, as does the number of senses with only one alignment candidate. WordNet-Wiktionary stands out as it was manually composed and is not representative of the full alignment (Meyer and Gurevych, 2011). The inter-annotator agreements A_0 and F_1 can be considered as upper bounds for automatic alignment accuracy and F-measure. Note that for the Wiktionary-Wikipedia datasets, due to the nature of their creation, the agreement was originally not available; we estimated it by manually re-annotating a sample of 100 examples with two annotators.

3.1.1 Resources

WordNet (Fellbaum, 1998) is a computational lexicon for English created at Princeton University. It is organized in sets of synonyms (synsets), each expressing a distinct concept. Synsets are represented by textual definitions (so-called glosses). A hierarchical organization is encoded via semantic relations such as hyponymy.

Wikipedia is a collaboratively created online encyclopedia available in almost 300 languages. The current English version contains around 4 400 000 articles, and the German one around 1 700 000 articles, each usually describing a particular concept. Due to its encyclopedic nature, Wikipedia mostly covers nouns, while the other LSRs discussed also cover verbs, adjectives, etc. Articles are connected via hyperlinks in the article text (implying a graph structure), and the first paragraph usually gives a short summary of the topic, serving as a gloss for our purposes. Articles are also linked to the equivalent articles in other languages.

Wiktionary is a dictionary "side project" of Wikipedia, available in over 500 languages. Currently, the English Wiktionary contains over 500 000 lexical entry pages, while the German one contains around 350 000 ones. For a word, multiple senses can be encoded, and these are usually represented by glosses. Wiktionary also contains hyperlinks to synonyms, hypernyms, etc. and translations into other languages.

OmegaWiki is a freely editable online dictionary like Wiktionary. However, instead of distinct language editions, OmegaWiki contains language-independent concepts ("Defined Meanings") which carry lexicalizations in different languages. These concepts are connected via semantic relations. OmegaWiki contains over 46 000 concepts and lexicalizations in almost 500 languages.

3.1.2 Datasets

WordNet–OmegaWiki: The first alignment between these LSRs based on the German part of OmegaWiki was reported in (Gurevych et al., 2012). As OmegaWiki Defined Meanings are multilingual, we used the same dataset for monolingual WSA in later work (Matuschek and Gurevych, 2013). Table 1 presents details about this and the other evaluation datasets.

WordNet–Wiktionary: Meyer and Gurevych (2011) originally used this dataset for similarity-based alignment. While we could not improve upon this using Dijkstra-WSA on its own (Matuschek and Gurevych, 2013), the backoff approach yielded a significant improvement. This dataset was manually composed according to specific criteria, hence it differs from the others and is not fully representative of the full alignment.

Wiktionary–Wikipedia (English): No evaluation dataset (let alone a full alignment) has been reported for this resource pair yet. However, as the datasets for WordNet-Wiktionary (Meyer and Gurevych, 2011) and WordNet-Wikipedia (Niemann and Gurevych, 2011) are lexically overlapping, we were able to automatically create a gold standard for Wiktionary-Wikipedia by exploiting the transitivity of the alignment relation, i.e. by using WordNet as a pivot. Note that, unlike Wiktionary, Word-

Net synsets have multiple lexicalizations for the same meaning, introducing alignment candidates from Wikipedia which might not be applicable to a particular Wiktionary sense. Hence, we decided to filter the examples where the lexeme of the Wiktionary sense and the Wikipedia article title did not match. An effect of this process was that words not contained in all three resources were filtered out, and many examples were left with few or only one candidate, leading to a low polysemy. We also manually checked the derived gold standard and corrected a small number of wrong annotations introduced through the automatic process. The resulting dataset is thus considerably smaller than the others, but it still turned out to be sufficient for machine learning experiments.

Wiktionary–Wikipedia (German): Same as for the English editions, neither a gold standard nor an alignment was previously reported for this pair. We were able to create a gold standard in a novel way by exploiting the fact that many German Wiktionary senses contain links to the corresponding Wikipedia articles, inducing a sense alignment between the two LSRs manually validated by the Wiktionary community. However, we were unable to extract such an alignment for English, as Wikipedia articles are attached to the lexical entry page in this version and not to a specific sense.

In the German Wiktionary, a large portion of the senses is linked in this way, and even after aggressively filtering out invalid link targets (e.g. disambiguation pages or pages with a non-matching title), we retained over 20 000 alignments between Wiktionary senses and Wikipedia pages, a sample of which we manually confirmed to be correct. Of course, this only yields positive examples; to also include cases of non-alignment, we extracted the other candidate (i.e. lexically matching) Wikipedia articles for each aligned Wiktionary sense, assuming that Wiktionary editors also considered and discarded them before eventually creating a link. Interestingly, the number of negative examples derived in this way is relatively low in comparison to the other datasets. An analysis revealed that a large fraction of the linked Wiktionary senses are either scientific terms (e.g. from biology) or named entities such as cities. Both types of senses tend to have few alternative candidates in Wikipedia due to their specificity, and it seems logical that Wiktionary users predominantly link these senses to the explanatory Wikipedia articles which are not familiar to the majority of users.

In the end, this process yielded a WSA dataset with unprecedented characteristics: It was not only created and validated by a crowd of editors rather than a handful of annotators, but it is also an order of magnitude larger than previously reported datasets (Table 1). This enables us to assess the performance of our WSA approach in a scenario which is close in size to a full alignment task, allowing a more well-grounded statement about its effectiveness.

3.2 Feature Engineering

The selection of features for our machine learning approach was driven by the premise to keep the framework as generic and resource-agnostic as possible, in order to ensure applicability to many different LSRs without additional engineering effort. Thorough analysis of existing resources and approaches revealed that two types of information are available for the vast majority of LSRs: i) Glosses, or more general, textual descriptions of concepts, and ii) Relationships between concepts inducing a graph, given through semantic relations, links, or other means. We also evaluated some features which are specific to a smaller subset of resources (see Section 3.2.3).

3.2.1 Gloss Similarity

Cosine similarity (COS) calculates the cosine of the angle between a vector representation of two senses s_1 and s_2. For the vector representation of a sense, we use a bag-of-words approach, i.e., a vector $\mathrm{BoW}(s)$ contains the term frequencies of all words in the description of s. In this work, we only rely on the textual definition of a sense to keep the approach as generic as possible, while the usage of example sentences, related words, synonyms etc. would also be possible.

Personalized PageRank similarity (PPR) (Agirre and Soroa, 2009) measures the semantic relatedness between two word senses s_1 and s_2 by comparing semantic vectors which can be derived in different ways; we utilize the variant introduced by Niemann and Gurevych (2011). The idea is to identify senses of words in a sense's gloss which are central for describing its meaning. These senses (represented in a graph derived from an LSR such as WordNet) should have a high PageRank score (i.e. a high centrality).

3.2.2 Dijkstra-WSA Distance

Dijkstra-WSA (Matuschek and Gurevych, 2013) is the graph-based WSA algorithm we use to calculate a distance-based similarity measure between word senses. We will briefly explain its two steps.

Graph construction: The *resource graph* is comprised of a set of nodes V which represents the senses of an LSR and a set of edges $E \subseteq V \times V$ which expresses semantic relatedness between them. One can use semantic relations, hyperlinks, or other relatedness indicators. For sparse LSRs, it is advisable to add edges between senses s_1 and s_2 if a monosemous term t with sense s_2 is included in the gloss of s_1. For example, one can link a sense of *Java* to *programming language* if the latter term is included in the former's definition text. This *monosemous linking* enhances the graph density (and hence, the recall) significantly.

Computing sense alignments: First, trivial alignments between the two resource graphs A and B are created. Alignments are trivial if two senses have the same attached lexeme in A and B and this lexeme is also unique in either resource. Intuitively, these alignments serve as "bridges" between highly related regions of A and B. Next, for each remaining sense $s \in A$, the set of possible target senses $T \subset B$ is retrieved in a similar fashion as for our approach, and for each of them the shortest path is computed using Dijkstra's algorithm (Dijkstra, 1959). While Dijkstra-WSA then goes on to directly align the sense which is closest to the source sense, we save the distance for each candidate sense and directly use it as a feature, expressing semantic relatedness based on the structure of both underlying resources. When no distance can be computed (in case of a disconnected graph), we assume infinite distance.

3.2.3 Other Features

We also experimented with other features which were accessible directly from the resources, i.e. without the need for external knowledge or extensive computational effort; these were usually not available for every resource pair. Features we tried were the part of speech (Wiktionary, OmegaWiki, WordNet), the sense index, i.e. the position in the sense list for a lexeme (WordNet, Wiktionary), similarity of example sentences (WordNet, Wiktionary), overlap of translations into other languages (Wikipedia, Wiktionary, cf. (Bond and Foster, 2013)) and overlap of domain labels (Wikipedia, Wiktionary, WordNet, OmegaWiki). However, for none of these features we could observe any significant[1] impact on the results, mostly due to sparsity of the respective features. Thus, we do not report them, but on the other hand we consider this an indicator that gloss similarity and distance in the resource graph already sufficiently capture the similarity between senses.

3.3 Machine Learning Classifiers

We experimented with different machine learning classifiers using WEKA (Hall et al., 2009). While a detailed discussion of these classifiers is beyond the scope of this work, we will at least give a short description of the ones we eventually used. For more details, please refer to textbooks such as (Murphy, 2012). We used WEKA's standard configuration in every case.

Threshold-based classifiers work by simply trying to learn a numeric boundary value which separates positive examples from negative ones. Although this approach is rather naive, it has been successfully used in previous WSA efforts (Meyer and Gurevych, 2011; Niemann and Gurevych, 2011).

A **Naive Bayes** classifier assumes that features are independent (i.e. the value of one feature is unrelated to any other feature), and is thus able to learn reliable classification probabilities on relatively small training sets. While the independence assumption can be considered an oversimplification, the algorithm is widely used due to its efficiency and good precision.

Bayesian Networks (or *belief networks*) also classify based on probabilities learned from training data, however, they offer the advantage of modeling dependencies between features, hence allowing a more accurate representation of the data. Technically, such a network is a directed acyclic graph modeling the conditional dependencies between variables.

A **Perceptron** is a classifier which maps a real-valued input vector to a binary output, by means of an artificial neural network. It is commonly used for pattern recognition, also in NLP (Collins, 2002).

[1] All significance claims in this paper are based on McNemar's test at a confidence level of 1%.

Support Vector Machines (SVMs) construct a hyperplane in a multi-dimensional space which yields a good separation between positive and negative training examples, represented as data points.

Decision Trees are built from training input by iteratively splitting the set of samples based on attribute values so that the resulting subset is as homogeneous as possible with regard to the class label. Unseen examples can be classified by testing the attribute values and following different branches of the tree. One of the main advantages (e.g. in comparison to SVMs) is that this approach is easily interpretable.

4 Experimental Results and Analysis

Baselines For reference, we report six different baselines: i) *Random*: A random sense from the set of candidates is chosen in each case, ii) *1:1*: An alignment is always made if and only if there is exactly one candidate, iii) *1st*: The first of the candidate senses is always selected[2], iv) *SIM*: A similarity threshold is learned for gloss similarity values as suggested by Meyer and Gurevych (2011), cf. Section 3.2.1, v) *DWSA*: The closest candidate sense in the resource graph is aligned as we suggested in (Matuschek and Gurevych, 2013), cf. Section 3.2.2, vi) *HYB*: A hybrid approach of using *DWSA* first and then *SIM* as a backoff, also suggested by us (Matuschek and Gurevych, 2013). The latter approach represents state-of-the-art performance for WSA. Note that for the two Wiktionary-Wikipedia datasets, no previous results were available, so we created similarity-based and Dijkstra-WSA alignments ourselves, based on the same versions of the resources as in the previous work. For the other datasets, we used the numbers reported in the original papers (Matuschek and Gurevych, 2013; Meyer and Gurevych, 2011).

Overview Tables 2 and 3 present the results for all setups. Although the best classifiers for each dataset always outperform the previous state of the art and the baselines by a significant margin, there is no consistent pattern in the results across different LSRs and classifiers. One reason for this is that the range of feature values varies substantially between different datasets. For instance, Dijkstra-WSA distances tend to be greater when Wikipedia is involved simply by its virtue of being larger than the other LSRs, and gloss similarities also differ depending on the average length of the glosses and the language. Another factor are the gold standards, which are quite different in terms of size and composition (see Table 1). Thus, no classifier is the undisputed "winner", but Bayesian Networks proved most robust in our experiments, showing competitive results in every case. As training them is also computationally cheap (compared to SVMs, for instance), we would generally recommend this kind of classifier for WSA tasks. In the following, we also provide a more detailed discussion of the results for each individual dataset.

WordNet-OmegaWiki In this case, the precision of the alignment is satisfactory for every classifier, while both previously reported approaches struggle for different reasons (Gurevych et al., 2012; Matuschek and Gurevych, 2013). The strength of the machine learning becomes apparent especially in comparison with the *HYB* approach: While the latter merely combines independent alignment decisions, hence achieving better recall but failing to improve precision (cf. Section 2.3), the joint usage of features leads to a massive improvement. Analysis of the decision tree classifier shows that, as we suspected, the "edge cases" are explicitly reflected in the learned model, i.e. examples with high gloss similarity but also a high Dijkstra-WSA distance (or vice versa) are ruled out with higher confidence. This observation generally also holds for the other datasets. As an example, the two senses of *genome* in biology (*"The non-redundant genetic information stored in DNA sequences that defines an individual organism"*) and algorithmics (*"In the context of a genetic algorithm, the information that defines an individual entity"*) have similar glosses; they are, however, quite far apart in the graph and thus not aligned. The Bayesian Network achieves the best results as it comprehensively models this interdependence of features. The SVM achieves the best precision, but the distribution of feature values does not lend itself well to linear separation in this case, leading to unsatisfactory recall.

WordNet-Wiktionary For this dataset, the results look similar to WordNet-OmegaWiki as far as the improvement of precision is concerned, as the joint usage of features helps to make a correct decision on

[2]While this corresponds to the most frequent sense baseline in other setups, note that no explicit frequency information is available for OmegaWiki, Wiktionary and Wikipedia, so that the first sense baseline is only a rough approximation.

	WordNet-OmegaWiki				WordNet-Wiktionary			
	P	R	F_1	A	P	R	F_1	A
Random	0.46	0.35	0.40	0.51	0.21	0.59	0.31	0.67
1:1	0.36	0.64	0.46	0.55	0.68	0.19	0.30	0.88
1st	0.34	0.80	0.48	0.47	0.33	0.51	0.40	0.80
SIM	0.55	0.53	0.54	0.73	0.67	0.65	0.66	0.91
DWSA	0.56	0.69	0.62	0.74	0.68	0.27	0.39	0.89
HYB	0.57	**0.75**	0.65	0.75	0.68	0.71	0.69	0.92
SVM	**0.95**	0.32	0.48	0.79	**0.82**	0.61	0.70	0.93
Naive Bayes	0.73	0.62	0.67	0.82	0.71	0.79	0.75	0.92
Bayesian Network	0.75	0.72	**0.74**	**0.84**	0.70	**0.84**	**0.77**	**0.94**
Perceptron	0.73	0.58	0.65	0.81	0.74	0.72	0.73	0.92
Decision Tree	0.68	0.63	0.66	0.80	0.78	0.66	0.72	0.93
Agreement	-	-	0.84	0.85	-	-	0.78	0.93

Table 2: Alignment results for WordNet-OmegaWiki and WordNet-Wiktionary: Using baselines (top), approaches from previous work (middle) and different machine learning classifiers (bottom). We report precision, recall, F-measure (the harmonic mean of both) and accuracy. Best results for each value and dataset are marked in bold. The inter-annotator agreements A_0 and F_1 are given as upper bounds.

borderline examples. However, in this case the recall is also substantially improved, especially for the Bayesian classifiers. This was an issue in the original Dijkstra-WSA results (Matuschek and Gurevych, 2013) due to the low connectivity of the English Wiktionary graph. The combination of distances and gloss similarities is able to alleviate this shortcoming of Wiktionary to some extent, as examples with missing Dijkstra-WSA distance can still be aligned in case of sufficient gloss similarity. SVMs also show the best precision here, but are challenged by the suboptimal separability of the feature space.

Wiktionary-Wikipedia (English) The low connectivity of Wiktionary is not as much an issue here as for WordNet-Wiktionary, mostly due to the different composition of the gold standard – higher-frequency words tended to be retained (see Section 3.1.2), which in turn are better connected within Wiktionary. This leads to reasonable results for Dijkstra-WSA alone. The hybrid approach reaches the best recall, but due to the relatively low precision of the *SIM* alignment, the overall result leaves room for improvement. This improvement is again achieved via joint modeling of features. As for the datasets discussed above, the precision is improved significantly; this is especially true for the Bayesian Network classifier. Precision and recall for the SVM classifier are also satisfactory in this case (due to the better linear separability of the feature space), making it the best overall classifier along with the Perceptron.

Wiktionary-Wikipedia (German) On this dataset, the naive baselines are very strong, due to the disproportionately large number of positive examples – this is especially true for the *1:1* setup which reaches perfect precision. In other words, whenever there is only one alignment candidate, it is already the correct one. The *HYB* approach also yields good results thanks to the high precision of its two components, but recall is an issue for gloss similarity due to the richer morphology and different formation of compounds in German. We did not use a compound splitter (an obvious extension for future work), so that, for instance *"Kinderspiel"* and *"Spiel für Kinder"* (both meaning *"a game for children"*) could not be lexically matched. However, when machine learning is applied, the recall can again be significantly improved at only a negligible expense of precision. Here, as for the WordNet-Wiktionary dataset, the joint modeling of distance and gloss similarity allows to correctly align more borderline examples. While the strong bias towards positive examples might make this dataset not fully representative of a full alignment task (which is the eventual goal of WSA), the results still beat the strong baselines in terms of F-measure and thus indicate that WSA, and especially our approach, works well on such a large-scale dataset.

	Wiktionary-Wikipedia (En)				Wiktionary-Wikipedia (De)			
	P	R	F_1	A	P	R	F_1	A
Random	0.41	0.49	0.45	0.48	0.68	0.40	0.51	0.46
1:1	0.17	0.56	0.26	0.33	**1.0**	0.63	0.77	0.75
1st	0.23	0.88	0.36	0.37	0.93	0.66	0.78	0.74
SIM	0.60	0.67	0.63	0.84	0.85	0.46	0.60	0.57
DWSA	0.78	0.55	0.65	0.87	0.85	0.61	0.71	0.66
HYB	0.62	**0.79**	0.70	0.86	0.90	0.72	0.80	0.75
SVM	0.82	0.70	**0.76**	0.92	0.76	0.84	0.80	0.71
Naive Bayes	0.79	0.69	0.73	0.92	0.85	0.54	0.66	0.62
Bayesian Network	**0.91**	0.63	0.74	**0.93**	0.86	0.81	0.83	0.77
Perceptron	0.82	0.70	**0.76**	0.92	0.75	**0.92**	0.82	0.73
Decision Tree	0.79	0.69	0.73	0.92	0.87	0.81	**0.84**	**0.78**
Agreement	-	-	0.79	0.95	-	-	0.85	0.89

Table 3: Results for Wiktionary-Wikipedia alignment in English and German: Using baselines (top), approaches from previous work (middle) and different machine learning classifiers (bottom). We report precision, recall, F-measure (the harmonic mean of both) and accuracy. Best results for each value and dataset are marked in bold. The inter-annotator agreements A_0 and F_1 are given as upper bounds.

Error analysis Error sources for our system are mostly the same as for the previously reported approaches – if equivalent concepts are described very differently (known as the "lexical gap", e.g. the senses *"divulge confidential information"* and *"to confess under interrogation"* of the verb *to sing*) and happen to be not very close in the resource graph, i.e. both similarity measures fail at once, they are likely not aligned (false negatives). On the other hand, false positives occur for examples such as *Brand*, which is the name of districts in two different German cities (Aachen and Zwickau). The sense descriptions are very much alike, and the senses are also located in similar regions of the resource graphs (roughly speaking, *German geography*), which makes the distinction hard. Addressing these issues might be possible by computing more sophisticated gloss similarity measures (e.g. using lexical expansion (Iida et al., 2008)) or enhancing the graph construction process. In general, however, there are no discernible systematic errors made by our system.

5 Conclusions and future work

We have shown that through joint modeling of different similarity measures for WSA the overall alignment quality in terms of F-measure can be significantly improved over the state of the art for each and every of the considered four datasets. This proves that such a joint usage of global structure as well as the content of the LSRs is indeed preferable over using either of them in isolation or combining them in a simple backoff approach, since it effectively utilizes both ways of calculating similarity.

Apart from substantially improving WSA performance, we also present two new datasets for Wiktionary-Wikipedia alignment in English and German which fill a considerable gap in the previous work on WSA. One of Wiktionary's explicit purposes is to complement the knowledge in Wikipedia, so that an alignment between these widely used resources seems a natural and important extension to the body of work in this field. Especially for (semi-) automatic translation tasks, this resource combination seems extremely promising due to the abundant multilingual content in both resources (see Section 3.1.1). We suggested a comparable combination of Wiktionary and OmegaWiki in the past (Matuschek et al., 2013), but the much larger Wikipedia is bound to hold even more potential. Moreover, the German dataset is of unprecedented size, allowing more credible statements about the performance of WSA algorithms in a full alignment scenario. Another interesting aspect is that this dataset was derived from links created by the crowd of Wiktionary editors, not by expert annotators; thus, it can be considered the first crowdsourced WSA dataset. This type of dataset creation is also one aspect of future work. We want to investigate in more detail to what extent these alignments are trustworthy, what steps are necessary

to improve the dataset's size and quality, and how negative examples (i.e. non-alignments) can be more reliably derived. We also plan to find out if such datasets could be created for other Wiktionary language editions.

The fact that the achieved results are close to the human agreement suggests that, for the datasets considered, there is not much room for improvement. Thus, we plan to apply and adapt the algorithm to LSRs with different properties than the ones considered here, such as the more syntax-focused FrameNet (Ruppenhofer et al., 2010) which only recently has received research attention in automatic WSA (Hartmann and Gurevych, 2013). The usage of syntactic features to express sense similarity has not been thoroughly explored yet, and it seems a promising direction to make further progress in WSA. Usage of more elaborate textual similarity features (e.g. covering semantic similarity or using lexical expansion) as it was suggested for text reuse detection (Bär et al., 2012) would be another direction worth exploring.

Inspired by the semi-automatic construction of the Wiktionary-Wikipedia gold standard for English from existing datasets, we also want to investigate whether an alignment of more than two resources at once (n-way alignment) is feasible, using joint knowledge from all LSRs involved. For instance, the information that two senses in resources A and B share a strong resemblance to a sense in another resource C could be expressed by an additional feature.

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 also like to 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.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2012. Text Reuse Detection Using a Composition of Text Similarity Measures. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 167–184, Mumbai, India, December.

Francis Bond and Ryan Foster. 2013. Linking and Extending an Open Multilingual Wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362, Sofia, Bulgaria, August.

Kostadin Cholakov, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Automated verb sense labelling based on linked lexical resources. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)*, pages 68–77, Gothenburg, Sweden, April.

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, pages 1–8, Philadelphia, USA.

Jordi Daudé, Lluís Padró, and German Rigau. 2003. Validation and tuning of wordnet mapping techniques. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*, Borovets, Bulgaria.

Gerard De Melo and Gerhard Weikum. 2008. A Machine Learning Approach to Building Aligned Wordnets. In *Proceedings of the First International Conference on Global Interoperability for Language Resources*, pages 163–170, Hong Kong.

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.

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, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA.

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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. volume 11, pages 10–18.

Silvana Hartmann and Iryna Gurevych. 2013. FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, volume 1, pages 1363–1373, August.

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.

Ryu Iida, Diana McCarthy, and Rob Koeling. 2008. Gloss-based semantic similarity metrics for predominant sense acquisition. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP '08)*, pages 561–568.

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 Lesk. 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, Toronto, Canada.

Michael Matuschek and Iryna Gurevych. 2013. Dijkstra-WSA: A Graph-Based Approach to Word Sense Alignment. *Transactions of the Association for Computational Linguistics (TACL)*, 1:151–164, May.

Michael Matuschek, Christian Meyer, and Iryna Gurevych. 2013. Multilingual Knowledge in Aligned Wiktionary and OmegaWiki for Translation Applications. *Translation: Computation, Corpora, Cognition*, 3(1):87–118.

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 2011)*, 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.

Andrea Moro, Hong Li, Sebastian Krause, Feiyu Xu, Roberto Navigli, and Hans Uszkoreit. 2013. Semantic rule filtering for web-scale relation extraction. In *Proceedings of the* 12^{th} *International Semantic Web Conference (ISWC 2013)*, pages 347–362, Sydney, Australia.

Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning series)*. The MIT Press, August.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012b. BabelRelate! A Joint Multilingual Approach to Computing Semantic Relatedness. In *Proceedings of the 26th AAAI Conference on Artificial Intelligence*, Toronto, Canada, July.

Roberto Navigli. 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 105–112, Sydney, Australia.

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 Conference (GenLex-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.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2010. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, CA, September.

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.