What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage

Christian M. Meyer and **Iryna Gurevych**
Ubiquitous Knowledge Processing Lab
Technische Universität Darmstadt
Hochschulstraße 10, 64289 Darmstadt, Germany
http://www.ukp.tu-darmstadt.de

Abstract

By today, no lexical resource can claim to be fully comprehensive or perform best for every NLP task. This caused a steep increase of resource alignment research. An important challenge is thereby the alignment of differently represented word senses, which we address in this paper. In particular, we propose a new automatically aligned resource of Wiktionary and WordNet that has (i) a very high domain coverage of word senses and (ii) an enriched sense representation, including pronunciations, etymologies, translations, etc. We evaluate our alignment both quantitatively and qualitatively, and explore how it can contribute to practical tasks.

1 Introduction

Though WordNet has been extensively used in knowledge-rich natural language processing (NLP) systems, there is no best lexical resource for all purposes. Jarmasz and Szpakowicz (2003), for example, found better results for solving word choice problems when using Roget's thesaurus instead of WordNet. There is indeed a large number of different lexical resources: The ACL Special Interest Group on the Lexicon[1] lists, for instance, more than 40 different lexical resources on their homepage that have been proposed as a source of background knowledge for different NLP tasks.

These resources typically differ in two ways: (i) They have a different coverage of words and word senses, and (ii) they encode heterogeneous types of information that is attached to their words and word senses. This heterogeneity ranges from very fundamental differences, like the distinction between lexicographic and encyclopedic knowledge to more specific ones, such as one resource encodes semantic frames, while another focuses on subsumption relations between word senses. Using WordNet without further considerations thus limits the performance of a system, since each resource has its individual advantages.

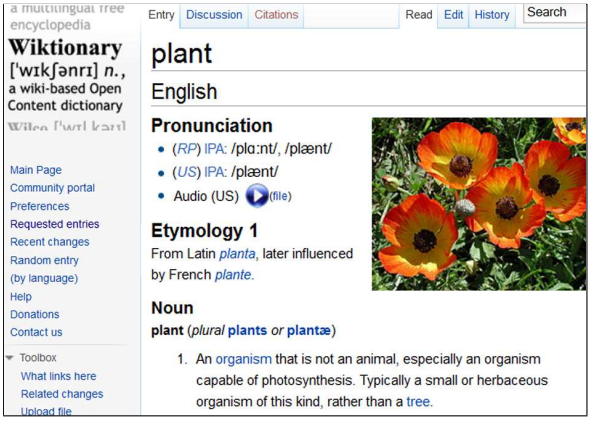

Figure 1: Wiktionary article *'plant'*

This has caused increasing research in the area of lexical resource alignment. It has been shown that aligned resources yield synergies, which lead to better performance than using the resources individually. For instance, Shi and Mihalcea (2005) improve semantic parsing using the knowledge of an aligned resource of FrameNet, WordNet, and VerbNet. Recently, Ponzetto and Navigli (2010) observed improvements for coarse-grained and domain-specific word sense disambiguation using an alignment between WordNet and Wikipedia for adding new relations to WordNet.

In another line of research, the community based online dictionary Wiktionary has been successfully applied in several NLP tasks, such as cross-lingual image retrieval (Etzioni et al., 2007), named entity recognition (Richman and Schone, 2008), or synonymy mining (Navarro et al., 2009; Sajous et al., 2010). Zesch et al. (2008b) compare different semantic relatedness measures using either WordNet, Wikipedia, or Wiktionary and find

[1] http://www.siglex.org/, accessed 2011-05-10

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 883–892,
Chiang Mai, Thailand, November 8 – 13, 2011.

the best results for Wiktionary. Besides its large coverage, Wiktionary also offers a great variety of linguistic information, such as pronunciations, etymologies, glosses, related words, translations, and many others. De Melo and Weikum (2010) exploit, for instance, alternative spellings and etymologies to enrich their lexical database.

In this work, we propose aligning WordNet and Wiktionary at the level of word senses. The resulting alignment has two important properties that go substantially beyond previous alignments: (i) increased domain coverage and (ii) enriched representation of senses. While Wiktionary is larger in size than most other previously aligned resources, such as Roget's thesaurus, Meyer and Gurevych (2010) analyze some word senses from WordNet and Wiktionary and come to the conclusion that certain domains are better covered by only one of the resources. This leads us to assume a very high domain coverage in our aligned resource.

Much work on lexical resource alignment involves Wikipedia, which contains lots of information about named entities. Wiktionary, in contrast, encodes common words and is not restricted to nouns. This opens up new possibilities for tasks including verbs, adjectives, or multiwords. Regarding the representation of senses, Wiktionary contains a great variety of linguistic information, like pronunciations or etymologies that are not found in previously aligned lexical resources.

The contributions of our work are threefold: (i) We present an automatic word sense alignment between the entire WordNet and Wiktionary, which we make publicly available. (ii) For evaluating the quality of our alignment, we introduce a new dataset based on human judgments to allow for future comparability of our results. (iii) We analyze the characteristics of our aligned resource, and how it can benefit different NLP tasks. We particularly point out that our resource has a much broader coverage of domain-specific word senses, which is important for processing real world data.

2 Notation and Lexical Resources

We first define the terminology used throughout the paper and introduce the two lexical resources WordNet and Wiktionary that are the subjects of our word sense alignment.

Lexical resources. By *lexical resource*, we mean a list of words and word senses. Our notion of *word* also includes multiwords, idioms, inflected forms, etc. Each word can have multiple *word senses*, which is one of multiple possible meanings for a word. A good illustration for this notion of words are the headwords in dictionaries, whereby the different meanings of the headword correspond to our notion of word sense.[2]

WordNet (Fellbaum, 1998) is a lexical resource for the English language that has been created by psycholinguists at the Princeton University. The resource is organized in synsets (i.e., sets of synonymous words) that are connected in a clear-cut subsumption hierarchy. The latest version 3.0 encodes 117,659 synsets. Each synset is represented by a gloss that is often followed by a short usage example. The synset {*plant, works, industrial plant*} is, for instance, represented by the gloss *"buildings for carrying on industrial labor"*.

Wiktionary[3] is a freely available, multilingual online dictionary. Similar to Wikipedia, the contents in Wiktionary can be edited by every Web user, which causes the resource to grow very quickly: by April 2010, the English Wiktionary contained over 1,700,000 article pages with linguistic knowledge about words in over 100 languages. For each word, multiple word senses can be encoded. Like in WordNet, they are represented by a gloss and example sentences illustrating the usage of a word sense. Additionally, there are hyperlinks to synonyms, hypernyms, meronyms, etc. Figure 1 shows the Wiktionary article *'plant'* as an example. For extracting the knowledge from Wiktionary, we use the Java-based Wiktionary Library (Zesch et al., 2008a). Using a Wiktionary dump of April 3, 2010, we counted 335,748 English words and 421,847 word senses.

Word sense alignment. A *word sense alignment*,[4] or *alignment* for short, is a list of pairs of word senses from two lexical resources. A pair of word senses that are aligned in a word sense alignment denote the same meaning. In WordNet, there is, for instance, a synset *"buildings for carrying on industrial labor"* for the word *'plant'*, which denotes the same meaning as the Wiktionary word

[2] Note that there is no commonly accepted standardized terminology in the field. Our notion of word is thus sometimes called *lemma* or *lexeme*; a *word sense* is also called *lexical unit*; whereas a lexical resource is also referred to as *sense inventory* or *(computational) lexicon*.

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

[4] Other terms for *(word sense) alignment* are *mapping* or *matching*. This notion of alignment is not to be mixed up with *word alignment* or *sentence alignment*, which are used for processing parallel texts as in machine translation.

sense *"a factory or other industrial or institutional building or facility"*. Another Wiktionary sense *"an organism that is not an animal* [...]*"*, however, clearly denotes a different meaning and should thus not be aligned to the WordNet synset.

3 Related Work

In the last twenty years, there have been many works on aligning lexical resources at the level of word senses. Almost all alignment approaches for the English language include WordNet, which is the *de facto* standard resource in the field. Early works address the alignment of WordNet with: Roget's thesaurus and the Longman Dictionary of Contemporary English (Kwong, 1998) [K98], the HECTOR corpus (Litkowski, 1999) [L99], the Unified Medical Language System (Burgun and Bodenreider, 2001) [BB01], CYC (Reed and Lenat, 2002) [RL02], VerbNet and FrameNet (Shi and Mihalcea, 2005) [SM05], as well as the Oxford Dictionary of English (Navigli, 2006) [N06].

The great potential of the collaborative resource Wikipedia in many NLP applications, such as semantic relatedness (Gabrilovich and Markovitch, 2007; Milne and Witten, 2008), word sense disambiguation (Mihalcea, 2007; Ponzetto and Navigli, 2010), or named entity recognition (Bunescu and Paşca, 2006), motivates aligning WordNet and Wikipedia to benefit from the advantages of both these resources. One line of research is thereby the alignment of WordNet synsets and Wikipedia categories, which has been done based on the shared taxonomic structure (Toral et al., 2008) [T08], textual entailment and semantic relatedness methods (Toral et al., 2009) [T09], as well as graph algorithms (Ponzetto and Navigli, 2009) [PN09].

Since the vast majority of knowledge is encoded in the Wikipedia article pages, also those have been aligned to WordNet synsets. The first work in this direction has been carried out by Ruiz-Casado et al. (2005) [R05] for the Simple Wikipedia, which is a smaller version of the full Wikipedia. Most of the published work, however, focuses on the articles in the full Wikipedia and their alignment to WordNet synsets. This task has been done based on: human judgments (Mihalcea, 2007) [M07], giving preference to WordNet's first sense (Suchanek et al., 2008) [S08], word overlap (de Melo and Weikum, 2010; Navigli and Ponzetto, 2010) [MW10,NP10], and using semantic relatedness measures (Niemann and Gurevych, 2011) [NG11]. Each approach has been evaluated on a separate, manually annotated dataset: De Melo and Weikum (2010) report a precision of $P = .85$, Navigli and Ponzetto (2010) observe $F_1 = .79$, and the alignment described by Niemann and Gurevych (2011) evaluates to $F_1 = .78$. It should be noted that these numbers are not comparable to each other, since they are based on different datasets and annotation schemes.

Work	Method	Resource	Full
[K98]	overlap	LDOCE & Roget	–
[L99]	syntax	HECTOR	–
[BB01]	overlap	UMLS	–
[RL02]	manual	CYC	–
[SM05]	structure	VerbNet & FrameNet	√
[N06]	relatedness	Oxford Dictionary	√
[T08]	structure	Wikipedia categories	√
[T09]	relatedness	Wikipedia categories	√
[PN09]	structure	Wikipedia categories	√
[R05]	overlap	Simple Wikipedia art.	√
[M07]	manual	Wikipedia articles	–
[S08]	mfs	Wikipedia articles	√
[MW10]	overlap	Wikipedia articles	√
[NP10]	overlap	Wikipedia articles	√
[NG11]	relatedness	Wikipedia articles	√
[MG10]	manual	Wiktionary senses	–
This work	relatedness	Wiktionary senses	√

Table 1: Previous work on aligning WordNet

Recently, also Wiktionary has been found to be a very promising resource for NLP tasks. So far, Wiktionary knowledge has been used for image search (Etzioni et al., 2007), calculating semantic relatedness (Zesch et al., 2008b), information retrieval (Müller and Gurevych, 2009), and synonymy detection (Navarro et al., 2009). An alignment has been done manually for a small number of word senses shared by Wiktionary and WordNet (Meyer and Gurevych, 2010) [MG10], but to the best of our knowledge, there is yet no word sense alignment covering the full resources. For applying an aligned resource in a practical system, such as word sense disambiguation, we, however, need a full alignment of the two resources. This is the subject of our work.

Table 1 shows an overview of related work on aligning WordNet with different lexical resources. Besides the resource that it is aligned to and whether the full resources have been processed, the table shows the utilized methods, which we classified into methods: aligning the first sense [mfs], counting weighted or normalized word overlaps (including the cosine measure) [overlap], using syntactic patterns [syntax], considering the (graph) structure of the resource [structure], uti-

lizing measures of semantic relatedness, such as semantic vectors or personalized PageRank [relatedness], and aligning senses manually [manual].

4 Word Sense Alignment

Most previous alignments are based on a one-to-one alignment assumption – i.e., that each sense is aligned with exactly one sense in the other resource. Niemann and Gurevych (2011), however, argue that there are senses requiring none, one, or multiple aligned senses.

This also holds for alignments of Wiktionary and WordNet. For example, the Wiktionary word sense *"the people who decide on the verdict; the judiciary"* for the word *'bench'* can be aligned to the two WordNet synsets *"persons who administer justice"* and *"the magistrate or judge or judges* [...]*"*. Accordingly, the Wiktionary word sense *"the bottom part of a sand casting mold"* for the noun *'drag'* is not covered by any WordNet synset and should thus not be aligned.

Therefore, we follow the alignment approach by Niemann and Gurevych (2011), which includes a state-of-the-art word sense disambiguation method by Agirre and Soroa (2009) that is known to outperform word overlap based measures. The method consists of the two steps (i) candidate extraction and (ii) candidate alignment that we briefly review in the following.

In the *candidate extraction* step, the algorithm iterates over all word senses in one lexical resource and extracts suitable candidates within the other resource that *might* form a valid alignment. In our case, we iterate over all synsets in WordNet and extract all word senses from Wiktionary that are encoded for one of the synset's synonymous words. For example, we extract all 9 Wiktionary word senses from the article *'plant'* and all 4 word senses from *'works'* for the WordNet synset {*plant, works, industrial plant*}. The word *'industrial plant'* is not encoded in Wiktionary. In the *candidate alignment* step, each candidate is then scored with two similarity measures:

(i) The *cosine similarity* (COS) calculates the cosine of the angle between a vector representation of the two senses s_1 and s_2:

$$\mathrm{COS}(s_1, s_2) = \frac{\mathrm{BoW}(s_1) \cdot \mathrm{BoW}(s_2)}{||\mathrm{BoW}(s_1)||\,||\mathrm{BoW}(s_2)||}$$

To represent a sense as a vector, we use a bag-of-words approach – i.e., a vector $\mathrm{BoW}(s)$ containing the term frequencies of all words in the definition of s. Note that there are different options for choosing the definition of sense s: For WordNet, the gloss of the synset can be used alone or in combination with its hyponyms and/or hypernyms. For Wiktionary, we can choose between gloss, usage examples, and related words of the word sense. We will discuss the best configuration during our evaluation in the following section.

(ii) The *personalized PageRank based measure* (PPR) estimates the semantic relatedness between two word senses s_1 and s_2 by representing them in a semantic vector space and comparing these semantic vectors $\mathbf{Pr}_{s_1}$ and $\mathbf{Pr}_{s_2}$ by computing

$$\mathrm{PPR}(s_1, s_2) = 1 - \sum_i \frac{(\mathbf{Pr}_{s_1,i} - \mathbf{Pr}_{s_2,i})^2}{\mathbf{Pr}_{s_1,i} + \mathbf{Pr}_{s_2,i}},$$

which is a χ^2 variant introduced in Niemann and Gurevych (2011). The main idea of choosing $\mathbf{Pr}$ is to use the personalized PageRank algorithm for identifying those synsets that are central for describing a sense's meaning. The sense *"buildings for carrying on industrial labor"* is, for instance, well represented by the WordNet noun synsets {*plant, works, industrial plant*}, {*building complex, complex*}, or the adjective synset {*industrial*}. These synsets should have a high centrality (i.e., a high PageRank score), which is calculated as

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

with the damping factor c controlling the random walk, the transition matrix M of the underlying semantic graph, and the probabilistic vector $\mathbf{v}$, whose i^{th} component $\mathbf{v}_i$ denotes the probability of randomly jumping to node i in the next iteration step.[5] Unlike in the traditional PageRank algorithm, the components of the jump vector $\mathbf{v}$ are not uniformly distributed, but personalized to the sense s by choosing $\mathbf{v}_i = \frac{1}{m}$ if at least one synonymous word of synset i occurs in the definition of sense s, and $\mathbf{v}_i = 0$ otherwise. The normalization factor m is set to the total number of synsets that share a word with the sense definition, which is required for obtaining a probabilistic vector.

Having calculated the similarity scores, we add the pair of the WordNet synset and the Wiktionary

[5] We use the publicly available UKB software (Agirre and Soroa, 2009) for calculating the PageRank scores and utilize the WordNet 3.0 graph augmented with the Princeton Annotated Gloss Corpus as M. The damping factor c is set to 0.85.

```
1  function ALIGN(WordNet, Wiktionary)
2    alignment := ∅;
3    for each synset ∈ WordNet.getSynsets() do
4      // Candidate extraction
5      candidates := ∅;
6      for each word ∈ synset.getWords() do
7        candidates := candidates
               ∪ Wiktionary.getWordSenses(word);
8      // Candidate alignment
9      for each candidate ∈ candidates do
10       sim_cos := COS(synset, candidate);
11       sim_ppr := PPR(synset, candidate);
12       if sim_cos ≥ τ_cos ∧ sim_ppr ≥ τ_ppr then
13         alignment := alignment
                 ∪ (synset, candidate);
14   return alignment;
15 end.
```

Figure 2: Pseudo code of the alignment algorithm

sense to our alignment if both similarity scores are above a certain threshold τ_{cos} and τ_{ppr}. We learned these thresholds in a 10 fold cross validation on our dataset that is explained in the following section. The optimal thresholds have been determined independently from each other using a simple binary split of the fold's items. The final thresholds are $\tau_{cos} = .13$ and $\tau_{ppr} = .49$.

Figure 2 shows the alignment algorithm in pseudo-code. Further details can be found in (Niemann and Gurevych, 2011).

5 Evaluation

To evaluate our WordNet–Wiktionary alignment, we follow the methodology of previous approaches and compare the result of our automatic alignment algorithm with human judgments. Therefore, we create a new manually annotated dataset, as we are not aware of any other datasets that could be used for this task. Our dataset is publicly available for future work on aligning WordNet and Wiktionary.

Dataset creation. Niemann and Gurevych (2011) introduce a well-balanced dataset for the alignment of WordNet and Wikipedia. Their sampled WordNet synsets are uniformly distributed in the number of synonyms, distance to the root node, and unique beginners. This way, a quantitative judgment of the alignment quality is as unbiased as possible. Since lexical resources are known to be very diverse (e.g., in terms of domain coverage (Burgun and Bodenreider, 2001; Meyer and Gurevych, 2010)), this is very important to get an impression about the alignment in general. Therefore, we reuse 320 synsets from their dataset as a primer for our evaluation dataset. For each synset, we extract all possible Wiktionary senses according to the candidate extraction step introduced in the previous section. This results in 2,423 sense pairs.

We asked 10 annotators to rate each sense pair as describing the same meaning (class 1) or describing a different meaning (class 0). The annotators are students in computer science, math, or linguistics, whereby two of them had previous experience with annotation studies. We described the annotation task in an annotation guidebook[6] and trained the annotators with some example cases.

Inter-rater agreement. To ensure the reliability of our annotated dataset, we calculate the inter-rater agreement between the annotators using the measures described by Artstein and Poesio (2008). The average observed agreement is $A_O = .93$ and the multi-rater chance-corrected agreement is $\kappa = .70$. Table 2 shows the pairwise κ for each pair of raters. The annotators C and F have the lowest inter-rater agreement between each other (.58) and with all other raters (.62 and .65). These two raters are thus on the opposite sides of the scale. Further analysis reveals that C is biased towards class 0 (different meaning) and F is biased towards class 1 (same meaning). We removed the annotations of these two raters, which yields an inter-rater agreement of $\kappa = .74$.

A dataset with such an agreement is considered reliable and allows to draw tentative conclusions (Krippendorff, 1980), although its agreement is lower than for WordNet–Wikipedia alignment datasets. More precisely, Niemann and Gurevych (2011) report $\kappa = .87$ and Navigli and Ponzetto (2010) measure $\kappa = .9$. Since even the two skilled annotators I and J only obtained an agreement of .80, we conclude that the alignment task of WordNet and Wiktionary is harder than the alignment of WordNet and Wikipedia. This does not come as a surprise, because Wikipedia contains encyclopedic knowledge that is largely complementary to the linguistic knowledge in WordNet and thus does not require to make fine-grained sense distinctions. WordNet and Wiktionary, however, both encode lexicographic knowledge about common words of the English language and thus require the distinction of very subtle differences in

[6] Available from our homepage: `http://www.ukp.tu-darmstadt.de/data/sense-alignment/`

κ	A	B	C	D	E	F	G	H	I	J
B	.72									
C	.60	.64								
D	.72	.75	.60							
E	.73	.72	.63	.74						
F	.64	.65	**.58**	.65	.68					
G	.75	.72	.66	.73	.75	.64				
H	.67	.72	.60	.72	.68	.64	.68			
I	.75	.74	.64	.77	.76	.67	.79	.73		
J	.72	.75	.62	.77	.77	.67	.76	.73	**.80**	
∅	.70	.71	**.62**	.72	.72	**.65**	.72	.69	.74	.73

Table 2: Pairwise κ of our annotation study

Method	A	P	R	F_1
RAND	.662	.212	.594	.313
MFS	.802	.329	.508	.399
COS only	.901	.598	**.703**	.646
PPR only	**.915**	**.684**	.636	.659
COS&PPR	**.914**	.674	.649	**.661**

Table 3: Performance of our alignment algorithm

the word sense definitions. We will discuss some examples during our error analysis.

Alignment quality. From our annotated data, we create a gold standard using majority vote of the remaining 8 annotators. An additional rater is asked to break the 27 ties. Following Navigli and Ponzetto (2010), we compare our automatic sense alignment with the gold standard using accuracy A, precision P, recall R, and the $F_1 = \frac{2PR}{P+R}$ score.

As baseline approaches, we implemented a first sense heuristic (MFS) and a method making a random selection (RAND). Table 3 shows the results of these baselines as well as our COS and PPR measures and their combination (COS&PPR). As noted in the previous section, there are multiple options for representing a sense. For WordNet, the synonyms, the gloss of the synset, and its direct hypernym and hyponyms have been tried as features. For Wiktionary, we experimented with the word, its gloss, usage examples, and synonyms. We tried all possible combinations and found the best result for using the synonyms and the gloss of the WordNet synset and its hypernym together with all four Wiktionary features. The table shows only the results for these features.

Our COS, PPR, and COS&PPR methods outperform the baseline by far. The difference is statistically significant at the 1% level in each case.[7] While COS has the highest recall and PPR has the highest precision, COS&PPR is a reasonable trade-off yielding the highest F_1 score. The difference of PPR and COS&PPR over COS is again statistically significant at the 1% level. The difference between PPR and COS&PPR is not statistically significant, which leads us to the conclusion that the PPR and COS&PPR methods perform equally well for our alignment task.

When analyzing the dataset, we observed a lower inter-rater agreement than for WordNet–Wikipedia alignments. This effect also becomes visible in our evaluation results: While Niemann and Gurevych (2011) measure an F_1 score of .53 for their MFS baseline and .78 for their COS&PPR method, the results are between .12 to .14 lower for the WordNet–Wiktionary alignment, which again shows that the word sense alignment between WordNet and Wiktionary is a more complex task than for WordNet and Wikipedia.

Error analysis. We carried out a detailed error analysis to identify the main types of errors made by our algorithm. Of the 2,423 sense pairs in the dataset, our COS&PPR algorithm yields 98 false positives and 110 false negatives.

Regarding the false negatives (i.e., the sense pairs that the method could not align, although they represent the same meaning), we found three main error classes: (i) The sense definitions were very different in their choice of words, such as in *"good discernment"* and *"ability to notice what others might miss"* for the word *'eye'*. These errors are hard to resolve, as they require a deep understanding and world knowledge. (ii) The sense definitions are very similar (e.g., *"any of various plants of the genus Centaurea* [...]*"* and *"any of various common weeds of the genus Centaurea"* for the word *'knapweed'*), but the similarity scores of the two measures were slightly below the chosen thresholds. These errors are caused by our choice of fixed similarity thresholds, which could, for instance, be improved by using machine learning for aligning the sense pairs. (iii) References to derived words occur in the sense definitions. An example is the word *'pacification'*, which is described as *"the process of pacifying"* and thus refers to the definition of *'pacifying'*. Such errors might be alleviated by taking the definitions of the derived words into account. This, however, raises again a word sense disambiguation challenge for finding the correct word sense of the derived word.

Amongst the false positives (i.e., the automatically aligned sense pairs with different meanings), we mainly found (i) highly related senses, such

[7] We use McNemar's test with Yates' correction.

as *"a computer that provides client stations with access to files and printers as shared resources to a computer network"* and *"any computer attached to a network"* for *'host'*, which are clearly related, but differ in their specification. The latter word sense does not require the host to provide file access or resources, but the former does. Although these two senses do not represent exactly the same meaning, their alignment is very useful for many NLP applications; e.g., for a semantic information retrieval system, which usually does not require to make subtle sense distinctions when searching relevant documents. Future work could distinguish between sense alignments sharing the same meaning and sharing a highly related meaning. (ii) Another large class of errors is due to an erroneous interpretation of a definition's meaning. Consider again the computing related sense of *'host'*. This sense is also aligned to *"any organization that provides resources and facilities for a function or event"*, because the words *resource, facility, function,* and *event* also frequently occur in the computer science domain. These errors are hard to tackle, but we plan to further investigate the influence of a sense's position in the taxonomy of a lexical resource.

6 Characteristics of the Wiktionary–WordNet Alignment

Aligning lexical resources is only one side of the coin. Another one is the question, how the aligned resource can be applied in practice and which NLP tasks can benefit from it. Our alignment of Wiktionary and WordNet yields a new resource with (i) increased coverage and (ii) an enriched representation of word senses.

Increased coverage. Coverage is crucial for almost every NLP task. Our final Wiktionary–WordNet alignment consists of 315,583 candidates, of which 56,970 pairs are marked as alignments. For 60,707 WordNet synsets there has been no corresponding word sense found in Wiktionary, and, vice versa, there are 371,329 Wiktionary word senses that have not been aligned with any WordNet synset. The word *'devisor'* is, for instance, only found within WordNet, and *'libero'* merely has an entry in Wiktionary. The new aligned lexical resource contains 488,988 word senses.

Table 4 shows the number of word senses per part of speech (POS) that are shared by both resources and that have no alignment with the respective other resource. The high number of word senses only occurring in Wiktionary can be explained by the 106,328 inflected word forms that are not encoded by WordNet. While the vast majority of encoded senses are nouns, also the coverage of other parts of speech benefits from the alignment of the two resources. This is a clear advantage over Wikipedia–WordNet alignments, which usually focus on nouns only. Besides verbs, adjectives, and adverbs that are also encoded by WordNet, Wiktionary additionally contains pronouns, phrases, idioms, sayings, etc.

	Overlap	only Wiktionary	only WordNet
Nouns	34,464	158,085	47,651
Verbs	8,252	29,119	5,515
Adj./Adv.	14,236	60,977	7,541
Other POS	0	16,778	0
Inflected Forms	0	106,328	0
Biology	4,465	4,067	12,869
Chemistry	2,561	8,260	2,268
Engineering	1,108	940	1,080
Geology	2,287	2,898	2,479
Humanities	4,949	2,700	5,060
IT	439	3,032	557
Linguistics	1,249	1,011	1,576
Math	615	2,747	483
Medicine	3,613	3,728	3,058
Military	574	426	585
Physics	1,246	2,835	1,252
Religion	733	1,154	781
Social Sciences	3,745	2,907	4,458
Sport	905	2,821	807

Table 4: POS and domains of our aligned resource

Pantel and Lin (2002) note that manually compiled lexicons are often missing domain-specific word senses, which is an important aspect for domain-aware NLP tools. In their manual Wiktionary–WordNet alignment, Meyer and Gurevych (2010) come to the conclusion that WordNet has a focus on humanities and social sciences, while Wiktionary has a higher coverage of natural sciences and sports. Their findings are, however, limited to a very small set of word senses and thus might not hold for the entire resources. Therefore, we analyze the encoded domains for the whole aligned resource. To identify the domain of a sense, we use WordNet Domains (Bentivogli et al., 2004) to classify the WordNet synsets into 157 domains (e.g., *'biology'*). For Wiktionary, we use the domain markers encoded in the glosses. An example is the sense *"(snooker) A play in which the cue ball knocks one (usually red) ball onto another* [...]" of the word *'plant'*, labeled with the *'snooker'* domain. We count 714

different labels in Wiktionary.[8] For being able to relate WordNet's and Wiktionary's labels, we manually grouped them into 14 general classes listed in Table 4. The specialized domains *'genetics'* and *'botany'*, for instance, have been grouped together in a more general domain *'biology'*. For each of these general domains, we count the number of word senses that are either overlapping between the resources or found in only one of them.

In the analysis, we confirm the findings of Meyer and Gurevych (2010): WordNet encodes a larger number of word senses from humanities and social sciences. About twice as many senses are only found in WordNet compared to the respective number in Wiktionary. Moreover, word senses from natural sciences and information sciences are, in general, better represented by Wiktionary. In particular, word senses related to chemistry, math, and IT are almost exclusively found in Wiktionary. Examples are the computer science related sense of *'host'* discussed above or the chemistry related sense *"an intramolecular valence bond, atom or chain of atoms that connects two different parts of a molecule"* of *'bridge'*, which both have no counterpart in WordNet. The situation is, however, different for the biology domain. WordNet covers the entire taxonomy of plants and animals, which is only fragmentarily found in Wiktionary. A high overlap between the two resources can be observed for linguistics and medicine. Aligning Wiktionary and WordNet hence allows for fast adaptation to a certain domain and fosters the development of high quality cross-domain applications.

Enriched sense representation. Besides its coverage, Wiktionary is also very rich in its lexical semantic information, which includes etymologies, alternative spellings, pronunciations, glosses, related words, translations, and many more. De Melo and Weikum (2010) exploit, for example, alternative spellings and etymologies for enriching their lexical database. They, however, do not align their resource with Wiktionary and thus cannot make use of the semantic information contained in glosses or related words. WordNet, on the other hand, is known for its rigid subsumption hierarchy and contains a large number of synonyms that proved useful for many NLP tasks.

[8] To avoid noise, we only consider labels occurring at least 10 times and manually filter register or style labels, such as *'poetic'* or *'archaic'*.

Applicability. The potential of aligned resources has been previously shown by many researchers: Shi and Mihalcea (2005), for instance, align FrameNet and VerbNet with WordNet and obtain improved results for semantic parsing. A similar approach has been followed by Loper et al. (2007), who align VerbNet and PropBank for improving semantic role labeling. Recently, Ponzetto and Navigli (2010) have used their Wikipedia–WordNet alignment to improve a knowledge-based word sense disambiguation system.

Our alignment of Wiktionary and WordNet now allows for further work in these directions by (i) exploiting the high coverage of our aligned resource, and (ii) using the enriched representation of senses. Apart from semantic parsing and word sense disambiguation noted above, also semantic relatedness is an interesting task, since Zesch et al. (2008b) found very good results using Wiktionary alone. This might be even surmounted by using our aligned resource. In our future work, we plan to investigate these applications in greater detail.

7 Conclusion

In this paper, we propose a novel word sense alignment between the entire WordNet and the collaborative online dictionary Wiktionary. This work goes beyond previous research efforts in aligning WordNet with Wikipedia, FrameNet, VerbNet and similar lexical resources, as Wiktionary allows for (i) an increased coverage of word senses and (ii) an enriched representation of senses, including pronunciations, etymologies, translations, etc. In our analysis, we particularly found a higher coverage of technical domains in Wiktionary and of humanities and social sciences in WordNet, which are consolidated in our aligned resource. For our alignment, we follow the method by Niemann and Gurevych (2011). We create a well-balanced evaluation dataset, which we make publicly available together with the entire aligned resource.[9]

Acknowledgments. This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806. We thank Elisabeth Niemann, Christian Kirschner, Christian Wirth, and Dr. Judith Eckle-Kohler for their contributions to this paper, and the IXA group at the University of the Basque Country for sharing their UKB software.

[9] `http://www.ukp.tu-darmstadt.de/data/sense-alignment/`

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.

Ron Artstein and Massimo Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

Luisa Bentivogli, Pamela Forner, Bernardo Magnini, and Emanuele Pianta. 2004. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proceedings of the COLING '04 Workshop on 'Multilingual Linguistic Resources'*, pages 101–108, Geneva, Switzerland.

Razvan Bunescu and Marius Paşca. 2006. Using Encyclopedic Knowledge for Named Entity Disambiguation. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy.

Anita Burgun and Olivier Bodenreider. 2001. Comparing Terms, Concepts and Semantic Classes in WordNet and the Unified Medical Language System. In *Proceedings of the NAACL '01 Workshop 'WordNet and Other Lexical Resources: Applications, Extensions and Customizations'*, pages 77–82, Pittsburgh, PA, USA.

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

Oren Etzioni, Kobi Reiter, Stephen Soderland, and Marcus Sammer. 2007. Lexical Translation with Application to Image Search on the Web. In *Proceedings of Machine Translation Summit XI*, Copenhagen, Denmark.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*. Cambridge, MA: MIT Press.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Mario Jarmasz and Stan Szpakowicz. 2003. Roget's Thesaurus and Semantic Similarity. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, pages 212–219, Borovets, Bulgaria.

Klaus Krippendorff. 1980. *Content Analysis: An Introduction to Its Methodology*. Thousand Oaks, CA: Sage Publications.

Oi Yee Kwong. 1998. Aligning WordNet with Additional Lexical Resources. In *Proceedings of the COLING-ACL '98 Workshop 'Usage of WordNet in Natural Language Processing Systems'*, pages 73–79, Montreal, QC, Canada.

Kenneth C. Litkowski. 1999. Towards a Meaning-Full Comparison of Lexical Resources. In *Proceedings of the ACL Special Interest Group on the Lexicon Workshop on Standardizing Lexical Resources*, pages 30–37, College Park, MD, USA.

Edward Loper, Szu-ting Yi, and Martha Palmer. 2007. Combining Lexical Resources: Mapping Between PropBank and VerbNet. In *Proceedings of the Seventh International Workshop on Computational Semantics*, Tilburg, The Netherlands.

Christian M. Meyer and Iryna Gurevych. 2010. How Web Communities Analyze Human Language: Word Senses in Wiktionary. In *Proceedings of the Second Web Science Conference*, Raleigh, NC, USA.

Rada Mihalcea. 2007. Using Wikipedia for Automatic Word Sense Disambiguation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pages 196–203, Rochester, NY, USA.

David Milne and Ian H. Witten. 2008. Learning to Link with Wikipedia. In *Proceedings of the 17th ACM International Conference on Information and Knowledge Management*, pages 509–518, Napa Valley, CA, USA.

Christof Müller and Iryna Gurevych. 2009. Using Wikipedia and Wiktionary in Domain-Specific Information Retrieval. In *Evaluating Systems for Multilingual and Multimodal Information Access: Proceedings of the 9th Workshop of the Cross-Language Evaluation Forum*, volume 5706 of *Lecture Notes in Computer Science*, pages 219–226. Berlin/Heidelberg: Springer.

Emmanuel Navarro, Franck Sajous, Bruno Gaume, Laurent Prévot, ShuKai Hsieh, Ivy Kuo, Pierre Magistry, and Chu-Ren Huang. 2009. Wiktionary and NLP: Improving synonymy networks. In *Proceedings of the ACL '09 Workshop 'The People's Web Meets NLP: Collaboratively Constructed Semantic Resources'*, pages 19–27, Suntec, Singapore.

Roberto Navigli and Simone Paolo Ponzetto. 2010. BabelNet: Building a Very Large Multilingual Semantic Network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden.

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.

Elisabeth Niemann and Iryna Gurevych. 2011. The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet. In *Proceedings of the Ninth International Conference on Computational Semantics*, pages 205–214, Oxford, UK.

Patrick Pantel and Dekang Lin. 2002. Discovering Word Senses from Text. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, AB, Canada.

Simone Paolo Ponzetto and Roberto Navigli. 2009. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, pages 2083–2088, Pasadena, CA, USA.

Simone Paolo Ponzetto and Roberto Navigli. 2010. Knowledge-Rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1522–1531, Uppsala, Sweden.

Stephen L. Reed and Douglas B. Lenat. 2002. Mapping Ontologies into Cyc. In *Proceedings of the AAAI '02 Workshop 'Ontologies and the Semantic Web'*, pages 1–6, Edmonton, AB, Canada.

Alexander E. Richman and Patrick Schone. 2008. Mining Wiki Resources for Multilingual Named Entity Recognition. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1–9, Columbus, OH, USA.

Maria Ruiz-Casado, Enrique Alfonseca, and Pablo Castells. 2005. Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence: Proceedings of the Third International Atlantic Web Intelligence Conference*, volume 3528 of *Lecture Notes in Computer Science*, pages 380–386. Berlin/Heidelberg: Springer.

Franck Sajous, Emmanuel Navarro, Bruno Gaume, Laurent Prévot, and Yannick Chudy. 2010. Semi-automatic Endogenous Enrichment of Collaboratively Constructed Lexical Resources: Piggybacking onto Wiktionary. In *Advances in Natural Language Processing: Proceedings of the 7th International Conference on NLP*, volume 6233 of *Lecture Notes in Artificial Intelligence*, pages 332–344. Berlin/Heidelberg: Springer.

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 Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2008. YAGO — A Large Ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3):203–217.

Antonio Toral, Rafael Muñoz, and Monica Monachini. 2008. Named Entity WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 741–747, Marrakech, Morocco.

Antonio Toral, Óscar Ferrández, Eneko Agirre, and Rafael Muñoz. 2009. A study on Linking Wikipedia categories to Wordnet synsets using text similarity. In *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*, pages 449–454, Borovets, Bulgaria.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008a. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, pages 1646–1652, Marrakech, Morocco.

Torsten Zesch, Christof Müller, and Iryna Gurevych. 2008b. Using Wiktionary for Computing Semantic Relatedness. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 861–867, Chicago, IL, USA.