WordNet–Wikipedia–Wiktionary: Construction of a Three-way Alignment

Tristan Miller[1], Iryna Gurevych[1,2]
[1]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[2]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for International Educational Research, Frankfurt am Main
http://www.ukp.tu-darmstadt.de/

Abstract

The coverage and quality of conceptual information contained in lexical semantic resources is crucial for many tasks in natural language processing. Automatic alignment of complementary resources is one way of improving this coverage and quality; however, past attempts have always been between pairs of specific resources. In this paper we establish some set-theoretic conventions for describing concepts and their alignments, and use them to describe a method for automatically constructing n-way alignments from arbitrary pairwise alignments. We apply this technique to the production of a three-way alignment from previously published WordNet–Wikipedia and WordNet–Wiktionary alignments. We then present a quantitative and informal qualitative analysis of the aligned resource. The three-way alignment was found to have greater coverage, an enriched sense representation, and coarser sense granularity than both the original resources and their pairwise alignments, though this came at the cost of accuracy. An evaluation of the induced word sense clusters in a word sense disambiguation task showed that they were no better than random clusters of equivalent granularity. However, use of the alignments to enrich a sense inventory with additional sense glosses did significantly improve the performance of a baseline knowledge-based WSD algorithm.

Keywords: lexical semantic resources, sense alignment, word sense disambiguation

1. Introduction

Lexical semantic resources (LSRs) are used in a wide variety of natural language processing tasks, including machine translation, question answering, automatic summarization, and word sense disambiguation. Their coverage of concepts and lexical items, and the quality of the information they provide, are crucial for the success of these tasks, which has motivated the manual construction of full-fledged electronic LSRs. However, the effort required to produce and maintain such expert-built resources is phenomenal (Briscoe, 1991). Early attempts at resolving this knowledge acquisition bottleneck focused on methods for automatically acquiring structured knowledge from unstructured knowledge sources (Hearst, 1998). More recent contributions treat the question of automatically connecting or merging existing LSRs which encode heterogeneous information for the same lexical and semantic entities, or which encode the same sort of information but for different sets of lexical and semantic entities. These approaches have until now focused on pairwise linking of resources, and in most cases are applicable only to the particular resources they align.

In this research we address the novel task of automatically aligning arbitrary numbers and types of LSRs through the combination of existing pairwise alignments, which in theory reduces the number of specialized algorithms required to find concept pairs in any n resources from as many as $\binom{n}{2} = n! \div 2(n-2)!$ to as little as $n-1$. The remainder of this paper is structured as follows: In the next section, we give some background on LSRs and alignments, both in general and for the specific ones we will be working with. Section 3 describes a technique for constructing n-way alignments and applies it to the production of a three-way alignment of concepts from WordNet, Wikipedia, and Wiktionary using existing WordNet–Wikipedia and WordNet–Wiktionary alignments. Though the technique is straightforward, this is, to our knowledge, the first time anyone has actually used it to align more than two heterogeneous LSRs at the concept level. Section 4 presents various statistical and qualitative analyses of the aligned resource. Our paper concludes with a discussion of possible applications and further evaluations of the aligned resource.

2. Background

2.1. Lexical semantic resources

The oldest type of lexical semantic resource is the *dictionary*. In its simplest form, a dictionary is a collection of *lexical items* (words, multiword expressions, *etc.*) for which the various *senses* (or *concepts*) are enumerated and explained through brief prose *definitions*. Many dictionaries provide additional information at the lexical or sense level, such as etymologies, pronunciations, example sentences, and usage notes. A *wordnet*, like a dictionary, enumerates the senses of its lexical items, and may even provide some of the same sense-related information, such as definitions and example sentences. What distinguishes a wordnet, however, is that the senses and lexical items are organized into a network by means of conceptual-semantic and lexical relations. *Encyclopædias* are similar to dictionaries, except that their concept descriptions are much longer and more detailed.

WordNet. *WordNet* (Fellbaum, 1998) is an expert-built English-language wordnet which has seen myriad applications. For each sense (in WordNet parlance, a *synset*) WordNet provides a list of synonymous lexical items, a definition, and zero or more example sentences showing use of

the lexical items.[1] Within each version of WordNet, synsets can be uniquely identified with a label, the *synset offset*, which encodes the synset's part of speech and its position within an index file. Synsets and lexical items are connected to each other by various semantic and lexical relations, respectively, in a clear-cut subsumption hierarchy. The latest version of WordNet, 3.0, contains 117 659 synsets and 206 941 lexical items.

Wiktionary. *Wiktionary*[2] is an online, free content dictionary collaboratively written and edited by volunteers. It includes a wide variety of lexical and semantic information such as definitions, pronunciations, translations, inflected forms, pictures, example sentences, and etymologies, though not all lexical items and senses have all of this information. The online edition does not provide a convenient and consistent means of directly addressing individual lexical items or their associated senses; however, the third-party API JWKTL (Zesch et al., 2008) can assign unique identifiers for these in snapshot editions downloaded for offline use. A snapshot of the English edition from 3 April 2010 contains 421 847 senses for 335 748 English lexical items.

Wikipedia. *Wikipedia*[3] is an online free content encyclopædia; like Wiktionary, it is produced by a community of volunteers. Wikipedia is organized into millions of uniquely named *articles*, each of which presents detailed, semi-structured knowledge about a specific concept. Among LSRs, encyclopædias do not have the same established history of use in NLP as dictionaries and wordnets, but Wikipedia has a number of features—particularly its network of internal hyperlinks and its comprehensive article categorization scheme—which make it a particularly attractive source of knowledge for NLP tasks (Zesch et al., 2007; Gurevych and Wolf, 2010).

2.2. Pairwise alignments

Each of the aforementioned resources has different coverage (primarily in terms of domain, part of speech, and sense granularity) and encodes different types of lexical and semantic information. There is a considerable body of prior work on connecting or combining them at the concept level in order to maximize the coverage and quality of the data; this has ranged from largely manual alignments of selected senses (Meyer and Gurevych, 2010; Dandala et al., 2012) to minimally supervised or even fully automatic alignment of entire resource pairs (Ruiz-Casado et al., 2005; de Melo and Weikum, 2009; Niemann and Gurevych, 2011; Navigli and Ponzetto, 2012; Meyer and Gurevych, 2011; Matuschek and Gurevych, 2013; Hartmann and Gurevych, 2013).[4] In our work, we use the alignments from Meyer and Gurevych (2011) and Matuschek and Gurevych (2013), which were among the few that were publically available in a transparent, documented format at the time of our study.

Meyer and Gurevych (2011) describe a text similarity–based technique for automatically aligning English Wiktionary senses with WordNet synsets. The versions of WordNet and Wiktionary they use contain 117 659 and 421 847 senses, respectively, for 206 941 and 335 748 lexical items, respectively. Their published alignment file consists of 56 952 aligned pairs, but as the same Wiktionary sense is sometimes paired with multiple WordNet synsets, the set of aligned pairs can be reduced mathematically (see §3) to 50 518 n:1 sense mappings, where $1 \leq n \leq 7$. Alignments for a well-balanced sample of 320 WordNet synsets (Niemann and Gurevych, 2011) were compared with human judgments, and were found to greatly outperform the random and MFS baselines, with $F_1 = 0.66$.

Dijkstra-WSA (Matuschek and Gurevych, 2013) is a state-of-the-art graph-based technique which was applied to align WordNet with a snapshot of the English edition of Wikipedia containing 3 348 245 articles, resulting in 42 314 aligned pairs. Here, too, the set of aligned pairs can be mathematically reduced to 30 857 n:1 mappings, where $1 \leq n \leq 20$. The alignment achieved $F_1 = 0.67$ on the aforementioned well-balanced reference dataset.

3. Construction of the three-way alignment

Since synonymy is reflexive, symmetric, and transitive (Edmundson, 1967), we can define an equivalence relation $\sim$ on a set of arbitrary sense identifiers $T = \{t_1, t_2, \ldots\}$ such that $t_i \sim t_j$ if t_i and t_j are synonyms (*i.e.*, if the senses they refer to are equivalent in meaning). The *synonym set* of an identifier $t \in T$, denoted $[t]_T$, is the equivalence class of t under $\sim$: $[t]_T = \{u \in T | u \sim t\}$. The set of all such equivalence classes is the quotient set of T by $\sim$: $T/\sim = \{[t]_T | t \in T\}$. For any pair of disjoint sets U and V such that $T = U \cup V$ and there exist some $u \in U$ and some $v \in V$ for which $u \sim v$, we say that u and v are an *aligned pair* and that $A_f(\{U,V\}) = T/\sim$ is a *full alignment* of the *sources* $\{U, V\}$. More generally, for any set of disjoint sets $W = \{W_1, W_2, \ldots\}$ where $T = \bigcup W$ and there exist distinct $W_i, W_j \in W : \exists u \in W_i, v \in W_j : u \sim v$, we say that $A_f(W) = T/\sim$ is a full alignment of W.

Full alignments may include synonym sets which do not contain at least one identifier from each of their sources. The *conjoint alignment* which excludes these synonym sets is defined as $A_c(W) = \{[t]_T | t \in T, \forall W_i \in W : \exists u \in W_i \cap [t]_T\}$.

The cardinality of a full or conjoint alignment is a count of its synonym sets. The number of individual identifiers referenced in an alignment $A(W)$ can also be computed: $\|A(W)\| = |\bigcup A(W)|$. If $\|A(W)\| = |T|$ then $A(W)$ must be a full alignment.

Given a set of identifiers and a set of aligned pairs, finding all the synonym sets is analogous to computing the connected components in a graph. Hopcroft and Tarjan (1973) describe an algorithm for this which requires time and space proportional to the greater of the number of identifiers or the number of aligned pairs.

Let *WKT*, *WN*, and *WP* be disjoint sets of unique sense identifiers from Wiktionary, WordNet, and Wikipedia, re-

[1] In this paper we use the term *sense* in a general way to refer to the concepts or meanings described by an LSR. This is in contrast to the WordNet documentation, where it refers to the pairing of a lexical item with a synset.

[2] `https://www.wiktionary.org/`

[3] `https://www.wikipedia.org/`

[4] A different approach with some of the same benefits is to provide a unified interface for accessing multiple LSRs in the same application (Garoufi et al., 2008; Gurevych et al., 2012).

spectively; the combined set of all their identifiers is $T = WKT \cup WN \cup WP$. The Dijkstra-WSA data corresponds to a set of ordered pairs $(n, p) \in WN \times WP$ where $n \sim p$. This data was sufficient for us to employ the connected component algorithm to compute $A_c(\{WN, WP\})$, the conjoint alignment between WordNet and Wikipedia. We reconstructed the full alignment, $A_f(\{WN, WP\})$, by adding the unaligned identifiers from the original Wikipedia and WordNet databases. Similarly, the Meyer and Gurevych (2011) data contains a set of pairs $(n,k) \in WN \times WKT$ such that $n \sim k$, but it also contains a list of unaligned singletons from both *WN* and *WKT*. We therefore directly computed both $A_f(\{WN, WKT\})$ and $A_c(\{WN, WKT\})$ using the connected component algorithm.

4. Analysis

The conjoint three-way alignment of WordNet, Wiktionary, and Wikipedia is a set of 15 953 synonym sets relating 63 771 distinct sense identifiers (27 324 WordNet synsets, 19 916 Wiktionary senses, and 16 531 Wikipedia articles). Of the synonym sets, 9987 (63%) contain exactly one identifier from each source; Table 1 gives further details on synonym set sizes. Since our WordNet–Wikipedia alignment is for nouns only, the synonym sets in the conjoint three-way alignment consist entirely of nouns. The full three-way alignment groups all 3 887 751 identifiers from the original sources into 3 789 065 synonym sets: 69 259 of these are described by adjectives, 3 613 514 by nouns, 12 415 by adverbs, 76 992 by verbs, and 16 885 by other parts of speech.
Coverage of lexical items is not as straightforward to analyze owing to how Wikipedia treats them. Concepts in Wikipedia are canonically identified by an article title, which is typically a lexical item optionally followed by a parenthetical description which serves to disambiguate the concept from others which would otherwise share the same title. Lexical synonyms for the concept, however, are not explicitly and consistently encoded as they are in WordNet synsets. These synonyms are sometimes given in the unstructured text of the article, though identifying these requires sophisticated natural language processing. Many redirect page titles[5] and incoming hyperlink texts—which are much easier to compile—are also synonyms, but others are anaphora or circumlocutions, and Wikipedia does not distinguish between them.
If we make no attempt to identify lexical synonyms from Wikipedia other than the article title, we find that the three-way conjoint alignment covers at least 44 803 unique lexical items, 42 165 of which are found in WordNet, 17 939 in Wiktionary, and 16 365 in Wikipedia. Moreover 20 609 of these lexical items are unique to WordNet and 2638 to Wikipedia. (There are no lexical items unique to Wiktionary.) We can also calculate the *word sense distribution* $d(k)$ of the conjoint alignment—that is, the percentage of lexical items which have a given number of senses k. Table 2 shows this distribution for WordNet, Wiktionary, and the conjoint three-way alignment; and also the average ($\bar{\omega}$) and maximum ($\hat{\omega}$) number of senses per lexical item. We observe that while the distributions for the unaligned resources are similar, the conjoint alignment demonstrates a marked shift towards monosemy. Though Zipf's law of meaning (Zipf, 1949) suggests that this might be the result of poor coverage of very high frequency lexical items, we found that the conjoint alignment actually covers 97 of the 100 most common (and 934 of the 1000 most common) nouns occurring in the British National Corpus.
Informal spot checks of synonym sets show them to be generally plausible, which is to be expected given the accuracy of the source alignments. However, the incidence of questionable or obviously incorrect mappings seems disproportionately higher in larger synonym sets. For example, one synonym set of cardinality 21 reasonably groups together various equivalent or closely related senses of the noun "hand", but also includes senses for "Palm OS" and "left-wing politics", since in the two-way alignments they had been mistakenly aligned with the anatomical senses for "palm" and "left hand", respectively. It appears that such errors are not only propagated but actually exaggerated by our algorithm, resulting in noisy data.

5. Evaluation

There are several different ways in which sense alignments can be formally evaluated. The conceptually simplest is comparison with human judgments as Meyer and Gurevych (2011) and Matuschek and Gurevych (2013) did with their pairwise alignments. However, there are many reasons why this sort of evaluation is not appropriate for an alignment of more than two resources: First, it disregards the transitive nature of synonymy. That is, if the two-way alignment contains the pairs (n_1,k) and (n_2,k), then those two pairs are considered in the evaluation, but not the implied pair (n_1,n_2). This was perhaps more acceptable for the two-way alignments where only a small minority of the mappings are not 1:1, but our three-way alignments rely more heavily on the transitive property; indeed, in the conjoint alignment 100% of the synonym sets were produced by exploiting it. Second, even if we modify the evaluation setup such that the implied pairs are also considered, since the number of identifiers per synonym set is much higher in the three-way alignment, there is a combinatorial explosion in the number of candidate pairs for the judges to consider. Finally, considering sense pairs in isolation may not be the most appropriate way of evaluating what are essentially *clusters* of ostensibly synonymous sense descriptions.
We could therefore reduce the problem to one of evaluating clusters of senses from a single resource—that is, for every synonym set in the full alignment, we remove sense identifiers from all but one resource, and treat the remainder as a coarse-grained clustering of senses. Established intrinsic or extrinsic sense cluster evaluation techniques can then be applied. An example of the former would be computing the entropy and purity of the clusters with respect to a human-produced gold standard (Zhao and Karypis, 2003). However, while such gold standards have been constructed for early versions of WordNet (Agirre and Lopez de Lacalle, 2003; Navigli, 2006), they have not, to our knowledge, been produced for the more recent version used in our alignment.

[5] In Wikipedia parlance, a *redirect page* is an empty pseudo-article which simply refers the visitor to a different article. They are analogous to "see" cross-references in indices (Booth, 2001).

alignment	2	3	4	5	6	7	8	9	≥ 10	total
$A_c(\{WN, WP\})$	23 737	4 801	1 355	492	234	112	54	28	44	30 857
$A_c(\{WN, WKT\})$	45 111	4 601	656	99	35	12	4	0	0	50 518
$A_c(\{WN, WP, WKT\})$	0	9 987	2 431	1 666	654	441	209	164	401	15 953

Table 1: Distribution of synonym sets by cardinality in the two- and three-way conjoint alignments

resource	$d(1)$	$d(2)$	$d(3)$	$d(4)$	$d(\geq 5)$	$\bar{\omega}$	$\hat{\omega}$
WN	83.4%	10.4%	3.1%	1.3%	1.8%	1.32	59
WKT	85.2%	9.4%	2.8%	1.1%	1.3%	1.26	58
$A_c(\{WN, WP, WKT\})$	91.0%	6.4%	1.6%	0.6%	0.5%	1.14	16

Table 2: Word sense distribution in WordNet, Wiktionary, and the three-way conjoint alignment

A possible extrinsic cluster evaluation would be to take the sense assignments of a state-of-the-art word sense disambiguation (WSD) system and rescore them on clustered versions of the gold standard (Navigli, 2006; Snow et al., 2007). That is, the system is considered to disambiguate a term correctly not only if it chooses the gold-standard sense, but also if it chooses any other sense in that sense's cluster. The improvement for using a given sense clustering is measured relative to a computed random clustering of equivalent granularity.

Cluster evaluations are appropriate if constructing the alignment is simply a means of decreasing the granularity of a single sense inventory. However, they do not measure the utility of the alignment as an LSR in its own right, which calls for extrinsic evaluations in scenarios where unaligned LSRs are normally used. One previous study (Ponzetto and Navigli, 2010) demonstrated marked improvements in accuracy of two different knowledge-based WSD algorithms when they had access to additional definition texts or semantic relations from a WordNet–Wikipedia alignment. Conventional wisdom in WSD is that for knowledge-based approaches, more data is always better, so a three-way alignment which provides information from Wiktionary as well could boost performance even further. A complication with this approach is that our alignment method produces coarse-grained synonym sets containing multiple senses from the same resource, and so without additional processing a WSD algorithm would not distinguish between them. For use with existing fine-grained data sets, such synonym sets could either be removed from the alignment, or else the WSD algorithm would need to be written in such a way that if it selects such a synonym set as the correct one, it performs an additional, finer-grained disambiguation within it.

In this study we performed two of the aforementioned types of WSD-based evaluations. The first evaluation is a cluster-based one where we rescore the results of existing WSD systems using the clusters induced by our three-way alignment; we describe this and present the results in §5.1. In our second evaluation, we use our three-way alignment to enrich WordNet glosses with those from aligned senses in the other two resources, and then use our enriched sense inventory with a knowledge-based WSD algorithm; this is covered in §5.2. For both evaluations we use the freely available DKPro WSD framework (Miller et al., 2013).

5.1. Clustering of WSD results

In this evaluation, we follow the approach of Snow et al. (2007). Specifically, we take the raw sense assignments made by existing word sense disambiguation systems on a standard data set and then rescore them according to a given clustering. A system is considered to have correctly disambiguated a term not only if it chose the correct sense specified by the data set's answer key, but also if it chose any other sense in the same cluster as the correct one. Of course, any clustering whatsoever is likely to increase accuracy, simply by virtue of there being fewer senses for systems to choose among. To account for this, we measure the accuracy obtained with each clustering relative to that of a random clustering of equivalent granularity.

Like Snow et al. (2007), we use the raw sense assignments of the three top-performing systems in the Senseval-3 English all-words WSD task (Snyder and Palmer, 2004): GAMBL (Decadt et al., 2004), SenseLearner (Mihalcea and Faruque, 2004), and the Koç University system (Yuret, 2004). While other datasets would be equally applicable, we use this one as it ensures comparability to the previous work. The scores for our random clusterings are determined computationally: For a given clustering, let C be the set of clusters over the N senses of a given term. Then the expectation that the correct sense and an incorrectly chosen sense will have been clustered together is

$$\frac{\sum_{c \in C} |c| \, (|c| - 1)}{N \, (N - 1)},$$

where $|c|$ is the number of senses in the cluster c. Note that all of the Senseval systems we rescore attempt to disambiguate every item in the data set, so coverage is always 100%. This means that in this evaluation, recall, precision, and F-score are always equivalent; we refer to these collectively simply as "accuracy" and report them as percentages.

Whereas our alignment uses WordNet 3.0, the Senseval-3 data set uses WordNet 1.7.1, so we use the WN-Map mappings (Daudé et al., 2003) to convert the WordNet 1.7.1 synset offsets to WordNet 3.0 synset offsets. Furthermore, because some of the WordNet synset clusters induced by our alignment contain no one common lexical item, we "purify" these clusters by splitting them into smaller ones

system	base	MFF	random	Δ
GAMBL	65.21	69.13	68.88	+0.25
SenseLearner	64.72	68.10	68.47	−0.37
Koç	64.23	67.76	67.54	+0.22
average	64.72	68.33	68.30	+0.03

Table 3: Senseval-3 WSD accuracy using our MFF-purified clusters and random clustering of equivalent granularity

system	base	LFF	random	Δ
GAMBL	65.21	68.99	68.70	+0.28
SenseLearner	64.72	67.96	68.22	−0.26
Koç	64.23	67.71	67.43	+0.28
average	64.72	68.22	68.12	+0.10

Table 4: Senseval-3 WSD accuracy using our LFF-purified clusters and random clustering of equivalent granularity

such that each synset in the cluster shares at least one lexical item with all the others. We tested two cluster purification approaches: in the first, we create a new cluster by taking from the original cluster all synsets containing its most common lexical item, and repeat this until the original cluster is empty. We refer to this technique as *most-frequent first*, or MFF. The second approach (*least-frequent first*, or LFF) works similarly, except that new clusters are constructed according to the *least* common lexical item.

The results of this evaluation using MFF and LFF clusters are shown in Tables 3 and 4, respectively. The table columns show, in order, the systems' original accuracy scores,[6] the accuracies rescored according to the WordNet clustering induced by our full three-way alignment, the accuracies rescored according to a random clustering of equivalent granularity, and the improvement of our clustering relative to the random one. As can be seen, the effect of our clusters on system performance is practically indistinguishable from using the random clusterings. By comparison, Snow et al. (2007) report a modest but presumably significant average improvement of 3.55 percentage points.

5.2. Enriched sense inventory for knowledge-based WSD

In this evaluation we attempted to measure the contribution of additional sense information from aligned senses to knowledge-based word sense disambiguation. First, we enriched the glosses of WordNet senses with those from their aligned Wiktionary and Wikipedia senses. (In the case of Wikipedia, we used the first paragraph of the article.) We then ran a popular knowledge-based WSD baseline, the simplified Lesk algorithm (Kilgarriff and Rosenzweig, 2000), on the aforementioned Senseval-3 data set. This algorithm selects a sense for the target word solely on the basis of how many words the sense gloss and target word context have in common, so additional, accurate gloss information should help close the lexical gap and therefore increase both coverage and accuracy.

[6]The slight difference in scores with respect to those reported in Snyder and Palmer (2004) is an artifact of the conversion from WordNet 1.7.1 to WordNet 3.0.

glosses	coverage	precision	recall	F_1
standard	26.85	69.23	18.59	29.30
enriched	29.17	67.26	19.62	30.38

Table 5: Senseval-3 WSD accuracy using simplified Lesk, with and without alignment-enriched sense glosses

glosses	coverage	precision	recall	F_1
standard	98.61	53.46	52.71	53.08
enriched	98.76	51.07	50.44	50.75

Table 6: Senseval-3 WSD accuracy using simplified extended Lesk with 30 lexical expansions, with and without alignment-enriched sense glosses

The results of this evaluation are shown in Table 5. As predicted, coverage increased somewhat. The overall increase in recall was modest but statistically significant (corrected McNemar's $\chi^2 = 6.22$, df $= 1$, $\chi^2_{1,0.95} = 3.84$).

The fact that our enriched sense representations boosted the accuracy of this simple baseline motivated us to repeat the experiment with a state-of-the-art knowledge-based WSD system. For this we used the system described in Miller et al. (2012), a variant of the simplified extended Lesk algorithm (Banerjee and Pedersen, 2002) which enriches the context and glosses with lexical items from a distributional thesaurus. However, as can be seen in Table 6, recall decreased by 2.27 percentage points; this difference was also statistically significant (corrected McNemar's $\chi^2 = 6.51$, df $= 1$, $\chi^2_{1,0.95} = 3.84$). It seems, therefore, that the additional gloss information derived from our alignment is not compatible with the lexical expansion technique.

To gain some insight as to why, or at least when, this is the case, we compared the instances incorrectly disambiguated when using the standard glosses but not when using the enriched glosses against the instances incorrectly disambiguated when using the enriched glosses but not when using the standard glosses. Both sets had about the same POS distribution. However, the words represented in the latter set were much rarer (an average of 178 occurrences in SemCor (Miller et al., 1994), versus 302 for the former set) and more polysemous (7.8 senses on average versus 6.5). The correct disambiguations in the latter set were also more likely to be the most frequent sense (MFS) for the given word, as tabulated in SemCor (71.6% MFS versus 63.3%). Using the enriched sense glosses seems to be slightly worse for shorter contexts—the corresponding second set of misclassified instances had an average sentence length of 123 tokens compared to the other's 127. (By comparison, the average sentence lengths where both methods correctly or incorrectly disambiguated the target word were 135 and 131, respectively.)

6. Conclusion and future work

In this paper we described a straightforward technique for producing an n-way alignment of LSRs from arbitrary pair-

wise alignments, and applied it to the production of a three-way alignment of WordNet, Wikipedia, and Wiktionary. We examined the characteristics of this alignment and identified various approaches to formally evaluating it, along with their particular suitabilities and drawbacks. Informal examination of the synonym sets in our conjoint alignment show them to be generally correct, though in many cases existing errors in the source alignments were magnified. Extrinsic evaluation of our full alignment in WSD settings gave mixed results: whereas using the alignment to enrich sense definitions proved useful for a baseline WSD algorithm, the same enriched definitions confounded a more sophisticated approach and significantly decreased its performance. Similarly, use of the alignment to cluster WordNet senses did not show any measurable improvement over a random baseline.

Given these inconsistent results, future work could be directed to refinement of the alignment technique to reduce the noise in the synonym sets. This could involve, for example, filtering outlier senses using text similarity measures similar to those used in the construction of the WordNet–Wiktionary alignment. Alternatively, we could try applying our original technique to pairwise alignments which are known to be more accurate (i.e., with higher precision), as this would reduce the incidence of error cascades. We might also try other ways of using the alignment for knowledge-based WSD—in this evaluation we made use of the resources' glosses only, though of course each resource provides much richer lexical and semantic information which can be exploited.

As a service to the research community, we make our full and conjoint three-way alignments publically available at `https://www.ukp.tu-darmstadt.de/data/`.

7. Acknowledgments

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg Professorship Program under grant № I/82806 and by the German Ministry of Education and Research under grant № 01IS10054G.

8. References

Agirre, E. and Lopez de Lacalle, O. (2003). Clustering WordNet word senses. In Angelova, G., Bontcheva, K., Mitkov, R., Nicolov, N., and Nikolov, N., editors, *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 11–18, September.

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In Gelbukh, A., editor, *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 2002)*, volume 2276 of *Lecture Notes in Computer Science*, pages 136–145. Springer, February.

Booth, P. F. (2001). *Indexing: The Manual of Good Practice*. K. G. Saur, Munich, Germany.

Briscoe, T. (1991). Lexical issues in natural language processing. In Klein, E. and Veltman, F., editors, *Proceedings of the Symposium on Natural Language and Speech*, ESPRIT Basic Research Series, pages 39–68, Berlin, November. Springer.

Dandala, B., Mihalcea, R., and Bunescu, R. (2012). Word sense disambiguation using Wikipedia. In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing. Springer.

Daudé, J., Padró, L., and Rigau, G. (2003). Validation and tuning of WordNet mapping techniques. In *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP 2003)*, pages 117–123.

de Melo, G. and Weikum, G. (2009). Towards a universal WordNet by learning from combined evidence. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522.

Decadt, B., Hoste, V., Daelemans, W., and van den Bosch, A. (2004). GAMBL, genetic algorithm optimization of memory-based WSD. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 108–112, July.

Edmundson, H. P. (1967). Axiomatic characterization of synonymy and antonymy. In *Proceedings of the 2nd International Conference on Computational Linguistics (COLING 1967)*, pages 1–11.

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

Garoufi, K., Zesch, T., and Gurevych, I. (2008). Representational interoperability of linguistic and collaborative knowledge bases. In *Proceedings of the KONVENS Workshop on Lexical-semantic and Ontological Resources – Maintenance, Representation, and Standards*, October.

Gurevych, I. and Wolf, E. (2010). Expert-built and collaboratively constructed lexical semantic resources. *Language and Linguistics Compass*, 4(11):1074–1090, November.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY – A large-scale unified lexical-semantic resource. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2012)*, pages 580–590, April.

Hartmann, S. and Gurevych, I. (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.

Hearst, M. A. (1998). Automated discovery of WordNet relations. In Fellbaum, C., editor, *WordNet: An electronic lexical database*, pages 131–152. MIT Press, Cambridge, MA, USA.

Hopcroft, J. and Tarjan, R. (1973). Algorithm 447: Efficient algorithms for graph manipulation. *Communications of the ACM*, 16(6):372–378, June.

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and results for English SENSEVAL. *Computers and the Humanities*, 34:15–48.

Matuschek, M. and Gurevych, I. (2013). Dijkstra-WSA: A graph-based approach to word sense alignment. *Transactions of the Association for Computational Linguistics*, 1:151–164, May.

Meyer, C. M. and Gurevych, I. (2010). How web communities analyze human language: Word senses in Wiktionary. In *Proceedings of the 2nd Web Science Conference (WebSci10)*, April.

Meyer, C. M. and Gurevych, I. (2011). What psycholinguists know about chemistry: Aligning Wiktionary and WordNet for increased domain coverage. In *Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP 2011)*, pages 883–892, November.

Mihalcea, R. and Faruque, E. (2004). SenseLearner: Minimally supervised word sense disambiguation for all words in open text. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 155–158, July.

Miller, G. A., Chodorow, M., Landes, S., Leacock, C., and Thomas, R. G. (1994). Using a semantic concordance for sense identification. In *Proceedings of the 7th Human Language Technologies Conference (HLT 1994)*, pages 240–243.

Miller, T., Biemann, C., Zesch, T., and Gurevych, I. (2012). Using distributional similarity for lexical expansion in knowledge-based word sense disambiguation. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)*, pages 1781–1796, December.

Miller, T., Erbs, N., Zorn, H.-P., Zesch, T., and Gurevych, I. (2013). DKPro WSD: A generalized UIMA-based framework for word sense disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (System Demonstrations) (ACL 2013)*, pages 37–42, August.

Navigli, R. and Ponzetto, S. P. (2012). An overview of BabelNet and its API for multilingual language processing. In Gurevych, I. and Kim, J., editors, *The People's Web Meets NLP: Collaboratively Constructed Language Resources*, Theory and Applications of Natural Language Processing. Springer.

Navigli, R. (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 (ACL-COLING 2006)*, pages 105–112.

Niemann, E. and Gurevych, I. (2011). The people's Web meets linguistic knowledge: Automatic sense alignment of Wikipedia and WordNet. In *Proceedings of the 9th International Conference on Computational Semantics (IWCS 2011)*, pages 205–214, January.

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

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

Snow, R., Prakash, S., Jurafsky, D., and Ng, A. Y. (2007). Learning to merge word senses. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP–CoNLL 2007)*, pages 1005–1014, June.

Snyder, B. and Palmer, M. (2004). The English all-words task. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 41–43, July.

Yuret, D. (2004). Some experiments with a naive Bayes WSD system. In Mihalcea, R. and Edmonds, P., editors, *Proceedings of the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text (Senseval-3)*, pages 265–268, July.

Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. In *Data Structures for Linguistic Resources and Applications*, pages 197–205. Narr, Tübingen, Germany, April.

Zesch, T., Müller, C., and Gurevych, I. (2008). Extracting lexical semantic knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2008)*, pages 1646–1652.

Zhao, Y. and Karypis, G. (2003). Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55(3):311–331.

Zipf, G. K. (1949). *Human Behaviour and the Principle of Least Effort*. Addison–Wesley, Cambridge, MA.