The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet

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

Abstract

We propose a method to automatically align WordNet synsets and Wikipedia articles to obtain a sense inventory of higher coverage and quality. For each WordNet synset, we first extract a set of Wikipedia articles as alignment candidates; in a second step, we determine which article (if any) is a valid alignment, i.e. is about the same sense or concept. In this paper, we go significantly beyond state-of-the-art word overlap approaches, and apply a threshold-based Personalized PageRank method for the disambiguation step. We show that WordNet synsets can be aligned to Wikipedia articles with a performance of up to 0.78 F_1-Measure based on a comprehensive, well-balanced reference dataset consisting of 1,815 manually annotated sense alignment candidates. The fully-aligned resource as well as the reference dataset is publicly available.[1]

1 Introduction

Lexical semantic resources often used as sense inventories are a prerequisite in automatic processing of human language. In the last few years, there has been a rise in research aligning different resources to overcome the knowledge acquisition bottleneck and coverage problems pertinent to any single resource. In this paper, we address the task of aligning WordNet noun synsets and Wikipedia articles to obtain a sense inventory of higher coverage and quality. WordNet, a lexical database for English, is extensively used in the NLP community and is a de-facto standard resource in many NLP tasks, especially in current WSD research (Fellbaum, 1998). WordNet's manually defined comprehensive taxonomy motivates many researchers to utilize it. However, as WordNet is maintained by only a small group of experts, it is hard to cope with neologisms, named entities, or rare usages on a large scale (Agirre and Edmonds, 2006; Meyer and Gurevych, 2010). In order to compensate for WordNet's lack of coverage, Wikipedia has turned out to be a valuable resource in the NLP community. Wikipedia has the advantage of being constantly updated by thousands of voluntary contributors. It is multilingual and freely available containing a tremendous amount of encyclopedic knowledge enriched with hyperlink information.

In the past, researchers have explored the alignment of Wikipedia categories and WordNet synsets (e.g., Toral et al. (2008); Ponzetto and Navigli (2009)). However, using the categories instead of the articles causes three limitations: First, the number of Wikipedia categories (about 0.5 million in the English edition) is much smaller compared to the number of articles (about 3.35 million). Secondly, the category system in Wikipedia is not structured consistently (Ponzetto and Navigli, 2009). And finally, disregarding the article level neglects the huge amount of textual content provided by the articles.

Therefore, attempts to align WordNet synsets and Wikipedia articles (instead of categories) have been recently made. This has three major benefits. First of all, as WordNet and Wikipedia were found to be partly complementary on the word sense level, an aligned resource would increase the coverage of

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

senses (Wolf and Gurevych, 2010). Second, word senses contained in both resources can then be represented by relational information from WordNet and encyclopedic information from Wikipedia in a multilingual manner yielding an enriched knowledge representation. And finally, the third major benefit of the alignment is the ability to automatically acquire sense-tagged corpora in a mono- and multilingual fashion. For each WordNet synset, the text of the aligned Wikipedia article (or all sentences or paragraphs in Wikipedia that contain a link to the article) can be automatically extracted similar to the approach proposed by Mihalcea (2007). Automatically generated sense-tagged corpora can be used to, e.g., counter the bottleneck of supervised WSD methods that rely on such sense-tagged text collections, which are rare. Further, due to the cross-lingual links in Wikipedia, also corpora in different languages can be constructed easily.

Our contribution to this paper is two-fold. First, we propose a novel two-step approach to align WordNet synsets and Wikipedia articles. We model the task as a word sense disambiguation problem applying the Personalized PageRank algorithm proposed by Agirre and Soroa (2009) as it is state-of-the-art in WSD and combine it with a word overlap measure, which increases the overall performance. Second, we generate and introduce a well-balanced reference dataset for evaluation consisting of 1,815 manually annotated sense alignment candidates. WordNet synsets and their corresponding Wikipedia article candidates are sampled along their distinctive properties such as synset size, domain, or the location in the WordNet taxonomy. An evaluation on this dataset let us generalize the performance to a full alignment between WordNet and Wikipedia, which is publicly available for further research activities.

2 Related work

The alignment of WordNet and Wikipedia has been an active area of research for several years with the goal of creating an enriched ontology. One of the first attempts proposed a new resource YAGO integrating WordNet and Wikipedia consisting of more than 1 million entities and 5 million facts (Suchanek et al., 2007). The set of entities contains all WordNet synsets and Wikipedia articles with titles that are not represented as terms in WordNet. Thus, they ignore ambiguous entities, e.g., the British rock band *Queen* is not covered as the term *queen* is already contained in WordNet.

Other approaches automatically align WordNet with the categories of Wikipedia instead of the articles. Toral et al. (2008) enrich WordNet with named entities mined from Wikipedia. Therefore, the noun *is-a* hierarchy of WordNet is mapped to the Wikipedia categories determining the overlap of articles belonging to the category and the instances for each of the senses of a polysemous word in WordNet.

Ponzetto and Navigli (2009) applied a knowledge-rich method which maximizes the structural overlap between the WordNet taxonomy and the category graph extracted from Wikipedia. Based on the mapping information, the taxonomy automatically generated from the Wikipedia category graph is restructured to enhance the quality. Toral et al. (2009) disambiguate WordNet noun synsets and Wikipedia categories using multiple text similarity measures similar to our approach. A Wikipedia category is thereby represented by its main article or an article, which has the same title string as the category. Wu and Weld (2008) integrate the Wikipedia's infobox information with WordNet to build a rich ontology using statistical-relational learning.

Ruiz-Casado et al. (2005) proposed a method to align WordNet synsets and Wikipedia articles (instead of categories). They align articles of the *Simple* English Wikipedia to their most similar WordNet synsets depending on the vector-based similarity of the synset's gloss and the article text. Recently, Ponzetto and Navigli (2010) presented a method based on a conditional probability $p(s|w)$ of selecting the WordNet sense s given the Wikipedia article w, whereas the conditional probability relies on a normalized word overlap measure of the textual sense representation. Both approaches, however, have the following two major drawbacks: first, the algorithms are modeled such that they always assume a counterpart in WordNet for a given Wikipedia article, which does not hold for the English Wikipedia (see Section 4). Second, the algorithms always assign the most likely WordNet synset to a Wikipedia article, not allowing multiple alignments. However, due to the different sense granularities in WordNet and Wikipedia, some Wikipedia articles might be assigned to more than one WordNet synset. Based on these observations, there is a need for a better approach yielding none, one, or more than one alignment for a given synset or article. We will describe a novel idea to tackle this in the next section.

3 Methodology

Automatic sense alignment aims to match senses of different resources that have the same meaning.[2] In general, one sense is given and the task is to find a correspondent within another resource, in case one exists. Thereby, automatic sense alignment meets two subgoals. At first, all potential alignment candidate senses for a given sense have to be extracted. Secondly, these extracted candidates have to be scored to select the sense(s) that match in meaning. For example, given the WordNet synset wn =<*schooner: sailing vessel used in former times*> and the two Wikipedia alignment candidate articles wp_1 =<*Schooner: A schooner is a type of sailing vessel ...*> and wp_2 =<*Schooner (glass): A schooner is a type of glass used for ...*>; the article wp_1 should be aligned with the synset wn, while the second should not be aligned. The recall of the extraction step can highly influence the performance of the whole alignment process. If a sense is not extracted in the first step, it cannot be selected in the alignment step either.
In Section 3.1, we state how we extract Wikipedia alignment candidate articles for a given synset. In the subsequent Section 3.2, we describe how we determine the article that is aligned to the synset (if any at all). As almost all Wikipedia articles refer to nouns, we focus on this part-of-speech.

3.1 Candidate extraction

In order to extract Wikipedia articles for a given WordNet synset, we follow the procedure introduced by Wolf and Gurevych (2010). We shortly summarize this method here: Let wn be a WordNet synset with a set of synonyms $\{s_1, \cdots, s_n\}$ of size n. For each synonym $s \in wn$, we extract all Wikipedia articles $wp \in WP_{wn}$ that match one of the following constraints:

a) the article title matches s, e.g., the article *Window* is retrieved for the synonym term *Window*,
b) the article title is of the form *s_(description tag)*, e.g., *Window_(computing)*,
c) the article has a redirect that matches s or is of the form *s_(description tag)*, e.g., *Chaff_(countermeasure)* has a redirect *Window_(codename)* and, thus, is retrieved for the synonym term *Window*,
d) the article is linked in a hyperlink, in which the link anchor text matches s, e.g., the article *Bandwagon effect* is retrieved for the term *bandwagon*, as there exist a hyperlink of the form [[*Bandwagon effect*|*bandwagon*]]. Only hyperlinks that occur in at least 3 different articles are taken into account in order to reduce noise.

3.2 Candidate alignment

Given the set of Wikipedia candidates WP_{wn} extracted for synset wn, we have to classify each Wikipedia article $wp \in WP_{wn}$ as being a valid alignment or not with respect to wn. Therefore, we first calculate similarities between synset–article pairs of a given training set. In a second step, we learn a threshold corresponding to the minimum similarity a sense pair should have to be aligned. This threshold is then used to fully align WordNet and Wikipedia.

Sense similarity. The basis of our new approach for sense alignment is the PageRank algorithm (Brin and Page, 1998) relying on a lexical-semantic knowledge base, which is modeled as a graph $G = (V, E)$. As knowledge base we use WordNet 3.0 extended with manually disambiguated glosses from the "Princeton Annotated Gloss Corpus"[3]. The vertices $v \in V$ represent the synsets; the edges (undirected and unweighted) represent semantic relations between synsets, such as hyponym and hypernym relations.

[2] We do not differentiate between the terms *sense* and *concept* in this paper as they both refer to the same 'artifact' and only differ in representation. Concepts in WordNet are described by the entire synset, e.g. the synset <*design, plan*>. Senses, however, are words tagged with a sense number, e.g. *design_N_#2*, which means the word *check* as a noun in its second sense.

[3] http://wordnet.princeton.edu/glosstag.shtml

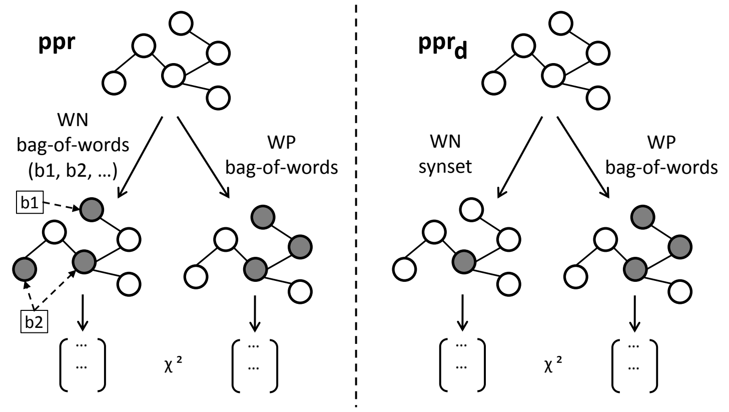

Figure 1: Schematic illustration of the basic ppr (left) and direct $\texttt{ppr}_d$ (right) approach.

The PageRank algorithm ranks the vertices in a graph according to their importance within the set. Let M be a $(n \times n)$ transition probability matrix, where $M_{ji} = \frac{1}{\text{outdegree}_i}$, if there exist a link from vertex i to vertex j. Then, the PageRank vector **pr** over the graph G is equivalent to resolve:

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

whereas c is a damping factor between 0 and 1, and v is an n-dimensional vector whose elements are $\frac{1}{n}$. An element of the PageRank vector denotes the probability for the corresponding vertex that a jumper, randomly following the edges in the graph, ends at that vertex, i.e. the importance of that vertex.

Now, vector v can be *personalized* by assigning stronger initial probabilities to certain vertices in the graph. This personalized version of the PageRank algorithm (Agirre and Soroa, 2009) is used in our approach in two different ways (see Figure 1):

In the *basic* version **ppr** , we represent both, Wikipedia articles and WordNet synsets as bag-of-words (abbreviated as b in the following). The textual representation is tokenized and lemmatized using the TreeTagger (Schmid, 1994); standard stopword removal is applied. For a given synset–article pair, we calculate two Personalized PageRank vectors. For each Personalized PageRank vector, we initialize vector v depending on the terms occurring in b:

$$v_i = \begin{cases} \frac{1}{m} & \text{if} \quad \text{a synonymous word of synset}_i \text{ in WordNet occurs in } b \\ 0 & \text{else}\,, \end{cases} \tag{2}$$

where m is the number of synsets with a synonymous word occurring in b. For example, given the WordNet synset <*payment, defrayal, defrayment: the act of paying money*> with its bag-of-words *(payment, defrayal, defrayment, act, paying, money)*, we assign each synset, i.e. vertex in the graph, a weight, for which at least one of its synonymous words occurs in the bag-of-words. Then, the PageRank vector is a semantic representation over all WordNet synsets for the given bag-of-words.

In the *direct* version **ppr $_\mathbf{d}$**, the WordNet synset is directly represented in v by assigning a weight of 1 to the corresponding vector element. It induces that the WordNet synset is already disambiguated and thus, motivates the use of the Personalized PageRank algorithm on the WordNet graph. Only for the Wikipedia article, the vector v is built up according to Eq. 2.

Given two Personalized PageRank vectors $\texttt{ppr}_{wn}$ and $\texttt{ppr}_{wp}$ for the WordNet synset wn and the Wikipedia article wp, we calculate their similarity using the χ^2 measure.[4]

$$sim_{ppr}(wn, wp) = 1 - \chi^2(ppr_{wn}, ppr_{wp}) = 1 - \sum_i \frac{(ppr_{wn_i} - ppr_{wp_i})^2}{ppr_{wn_i} + ppr_{wp_i}} \tag{3}$$

[4] This vector distance measure has shown the best overall performance compared to the cosine and euclidean distance in our experiments.

Learning classifier. Based on the similarity, the sense pair has to be classified as alignment (class 1) or non-alignment (class 0) formally defined as:

$$c(wn, wp) = \begin{cases} 1 & \text{if} \quad sim(wn, wp) > t \\ 0 & \text{else}\,, \end{cases} \tag{4}$$

where $sim(wn, wp)$ is the similarity of a WordNet synset and a Wikipedia article, and t is a real valued threshold. We apply 10-fold cross-validation to determine the threshold. We measure the performance of classification by means of F_1-Measure (see Section 5) and iteratively search (from 0 to 1 in 0.001 steps) for a threshold that maximizes the performance on the training fold. A threshold-based classification scheme induces that a WordNet synset can be aligned to none, one, or more than one Wikipedia article, which is the main potential of our approach compared to existing methods. However, in the scope of this paper, we assign at most one Wikipedia article (if any) to a WordNet synset (the one with the highest similarity above the threshold) as this yields the best performance (see Section 5).

Word overlap measure. For comparison, we also applied the standard cosine word overlap similarity measure **`cos`** used in existing sense alignment approaches (e.g., Ruiz-Casado et al. (2005)). We determine the similarity of the bag-of-words vectors of the WordNet synset and Wikipedia article calculating the cosine between them. According to Eq. 4 we also learn a classifier based on the cosine similarity.

Combination of the classifiers' output. Finally, we experiment with a heuristic, classifying only those synset–article pairs as alignment, for which the Personalized PageRank-based classifier and the cosine-based classifier, i.e. c_{ppr} and c_{cos}, or c_{ppr_d} and c_{cos}, return an alignment to further increase the precision.

Baselines. We implemented two different baselines. The baseline **`rand`** randomly selects a Wikipedia article from the extracted candidate set for each synset. The baseline **`mfs`** (most frequent sense) assigns always the most frequently linked Wikipedia article of the candidate set defined as the article with the highest number of incoming links. For example, for the synset wn =<*tree: a tall perennial woody plant having a main trunk* [...]> suppose we extract the two Wikipedia articles, namely wp_1 =<*Tree: A tree is a perennial woody plant.*> and wp_2 =<*Tree (data structure)*>. In this case, the sense wp_1 is aligned to the synset wn as it has 4,339 inlinks, about 4,000 more than the article wp_2. Both, the `rand` and `mfs` baseline always return a one-to-one alignment.

4 Well-balanced reference dataset

Publicly available evaluation datasets as provided by Fernando and Stevenson (2010) and Wolf and Gurevych (2010), are either quite small or follow a different annotation scheme. Others consist of randomly sampled synsets, which do not properly represent the distribution of synsets in WordNet following specific properties. For example, the dataset used in (Ponzetto and Navigli, 2010) consists of only 2 sense pairs, whose lemmas are monosemous in WordNet and Wikipedia (e.g. the lemma *specifier* corresponds to one synset in WordNet and one article in Wikipedia). As this property holds for one-third of all WordNet noun synsets, it is crucial for the choice of the alignment method and thus, should be represented in the evaluation dataset adequately. Therefore, our goal in this paper is to compile a well-balanced dataset to cover different domains and properties.

Synsets can be characterized with respect to their so-called assigned *Unique Beginner*, their synset size, and their location within the WordNet taxonomy. The *Unique Beginners* group synsets in semantically related fields (Fellbaum, 1998) such as *entity* (subsuming animals, persons, plants, artifacts, body and food related synsets), *abstraction*, *psychological features*, *shapes*, *states*, and *locations*. The synset size refers to the number of synonymous word senses in the synset. A synset can further be characterized by its location within the WordNet taxonomy defined as the shortest path between the given synset and the synset *entity*, which is the root element of all noun synsets. In addition, we distinguish between

Property		# synsets in WordNet	# sampled synsets	# manually aligned synsets
Synset size	=1	42,054	160	110
	> 1	40,061	160	111
Path length to root	0-5	8,586	60	33
	6-10	67,082	200	143
	11-16	6,447	60	45
Unique Beginner	*Entity*	47,330	160	118
	Non-Entity	34,785	160	103
# extracted WP candidates	=1	23,991	160	108
	> 1	46,569	160	113
Total #		82,115	320	221

Table 1: Sampling by properties and # manual alignments

Annotator	A	B	C	majority
# non-alignments	1,586	1,571	1,605	1,588
# alignments	229	244	210	227

Table 2: Annotations per class

	A–B	A–C	B–C
A_{O}	.9697	.9741	.9724
κ	.8663	.8782	.8742

Table 3: Inter-annotator agreement

synsets for which more than one Wikipedia candidate article is returned. In summary, for example, the synset <*article, clause: a separate section of a legal document*> has a synset size of 2, is assigned to the Unique Beginner *communication*, has a shortest path to the root element of length 6, and has 5 extracted Wikipedia candidate articles.

Based on these distinctive properties, we sampled 320 noun synsets yielding 1,815 sense pairs to be annotated, i.e. 5.7 Wikipedia articles per synset on average. The exact proportion of synsets with respect to their properties is detailed in Table 1 in the first four columns.

The manual sense alignment is performed by three human annotators. The annotators were provided sense alignment candidate pairs, each consisting of a WordNet synset and a Wikipedia article. The annotation task was to label each sense pair either as alignment or not. Table 2 outlines the class distribution for three annotators and the majority decision.

The most sense alignment candidates were annotated as non-alignments; only between 210 and 244 sense pairs were considered as alignments (extracted for 320 WordNet synsets). To assess the reliability of the annotators' decision, we computed the pairwise observed inter-annotator agreement A_{O} and the chance-corrected agreement κ (Artstein and Poesio, 2008)[5]. The agreement values are shown in Table 3. The average observed agreement A_{O} is 0.9721, while the multi-κ is 0.8727 indicating high reliability. The final dataset was compiled by means of a majority decision. Given 1,815 sense alignment candidate pairs, 1,588 were annotated as non-alignments, while 227 were annotated as alignments. 215 synsets were aligned with one article, while 6 synsets were aligned with two articles. Interesting to note is that the aligned samples are uniformly distributed among the different sampling dimensions as shown in Table 1 (right column). It demonstrates that WordNet synsets of different properties are contained in Wikipedia. On the other side, 99 synsets, i.e. approx. 1/3 of the sampled synsets, could not be aligned. Most of them are not contained in Wikipedia at all, e.g. the synset <*dream (someone or something wonderful)*> or <*outside, exterior (the region that is outside of something)*>. Others are not explicitly encoded on the article level such as the synset <*quatercentennial, quatercentenary (the 400th anniversary (or the celebration of it))*>, which is part of the more general Wikipedia article <*Anniversary*>.

5 Experiments

In our experiments, we represent a WordNet synset either by itself (in the direct version $\texttt{ppr}_d$) or by its set of synonymous word senses and its gloss and examples (in the basic version $\texttt{ppr}$). Optionally, we include hyponym and hypernym synsets to extend the sense representation of a synset: (SYN): the

[5] Note: "As the class distribution is highly skewed, the test for reliability in such cases is the ability to agree on the rare categories [. . .]" (Artstein and Poesio, 2008). This, in fact, is the category/class, in which we are most interested in.

WordNet	Wikipedia	cos		ppr$_d$		ppr$_d$ + cos		ppr		ppr + cos	
		F$_1$	Acc	F$_1$	Acc	F$_1$	Acc	F$_1$	Acc	F$_1$	Acc
SYN	P+T	.691	.907	.719	.921	.726	.923	.707	.914	.727	.927
+HYPO	P+T	.694	.908	.701	.916	.716	.931	.700	.912	.718	.926
+HYPER	P+T	.726	.921	.708	.918	.737	.935	.756	.928	.774	.940
+HYP2	P+T	.725	.927	.713	.920	.720	.937	.741	.923	.756	.940
SYN	P+T+R	.684	.907	.721	.921	.738	.936	.707	.913	.725	.926
+HYPO	P+T+R	.689	.910	.711	.918	.729	.936	.698	.910	.721	.927
+HYPER	P+T+R	.719	.917	.724	.928	.748	.937	.762	.938	.755	.940
+HYP2	P+T+R	.727	.920	.729	.929	.739	.937	.747	.932	.761	.940
SYN	P+T+C	.698	.909	**.754**	**.930**	**.756**	**.937**	.726	.918	.743	.931
+HYPO	P+T+C	.702	.910	.739	.927	.747	.938	.722	.917	.740	.930
+HYPER	P+T+C	**.738**	**.925**	**.752**	**.931**	.765	.943	.765	.935	**.781**	**.945**
+HYP2	P+T+C	.732	.923	.739	.928	.757	.942	.746	.930	.769	.942
SYN	P+T+R+C	.699	.912	.736	.926	.752	.939	.719	.916	.734	.929
+HYPO	P+T+R+C	.695	.911	.736	.926	.735	.936	.711	.914	.727	.928
+HYPER	P+T+R+C	.718	.917	.744	.930	.758	.940	**.776**	**.940**	.772	.943
+HYP2	P+T+R+C	.724	.918	.751	.932	.756	.939	.762	.936	.769	.942
rand	–	.527	.857								
mfs	–	.534	.860								

Table 4: Results for the automatic alignment

given synset; (HYPER): all hypernym synsets of the given synset; (HYPO): all hyponym synsets of the given synset; (HYP2): all hypernym and hyponym synsets of the given synset.

A Wikipedia article is represented by either its first paragraph[6] as it usually contains a compact description of the article or its whole article text. The article title and additional assigned information such as categories or redirects can also be taken into account: (P): first paragraph of Wikipedia article (with a minimum length of 200 characters[7]); (TXT): the whole article text; (T): article title; (C): all categories assigned to the article; (R): all redirects assigned to the article.

Table 4 lists the performance of our approach for different experimental settings.[8] We evaluate our approach in terms of F$_1$-Measure ($F_1 = \frac{2*P*R}{P+R}$), where P is the precision and R the recall. The precision P determines the ratio of correct alignments to all alignments assigned by the algorithm. The recall R identifies the number of correct alignments to the total number of correct alignments in the gold standard. Further, we provide an accuracy measure Acc, which denotes the percentage of the correctly identified alignments and non-alignments.

Similarity measure. Overall, the Personalized PageRank approach outperforms the cosine similarity. cos achieves an F$_1$-Measure of 0.738, while ppr$_d$ reaches 0.754 and ppr even 0.776, which is a performance gain of 2.1% and 5.1%, respectively. This, in fact, strengthens our motivation to employ semantic relatedness based approaches instead of a simple word overlap approach. For example, the synset *<Johannesburg>* and its corresponding Wikipedia article is not aligned based on the cosine approach as only three terms overlap. However, the ppr and ppr$_d$ approach classify the synset–article pair as alignment as there exists semantic relatedness between "large economy" and "commercial center" occurring in the textual sense representations.

The performance differences between ppr$_d$ and ppr correlate with the synset representation. On the one hand, utilizing the SYN representation, ppr$_d$ outperforms the ppr approach. This shows the effect of disambiguating the WordNet synset beforehand. On the other hand, when presenting the synset together with its hypernym or both, hypernyms and hyponyms, ppr yields the best performance. This might be due to the fact that a Wikipedia article often contains more general terms, i.e. hypernym concepts, especially within the first paragraph of a Wikipedia article.

All combinations yield higher performance compared to the stand-alone classifiers. For example, for the setting SYN+HYPER and P+T+C, cos yields 0.738, ppr 0.765, and the combination of both 0.781

[6] Extracted with JWPL (Zesch et al., 2008) and some additional post-processing steps.

[7] We have not optimized this value for this task.

[8] As all experimental settings, in which the Wikipedia article was represented with its first paragraph instead of the whole article text, yield higher performance, we report only these numbers here.

Measure	A	B	C
`cos`	.688	**.692**	.676
`ppr`$_d$	**.711**	**.711**	.690
`ppr`$_d$ + `cos`	**.724**	. 714	.716
`ppr`	**.737**	.718	.716
`ppr` + `cos`	**.740**	.730	.728

Table 5: Agreement (κ) between automatic and human annotators

		automatic	
		alignment	non-alignment
manual	alignment	178	**49**
	non-alignment	**51**	1,537

Table 6: Confusion matrix (Setting: `ppr` + `cos` , SYN+HYPER, P+T+C)

performance, which is an improvement of 5.8% and 2.1% compared to the `cos` and `ppr` approach, respectively. The performance gain originates from higher precision.

Sense representation. All similarity measures yield better performance representing the WordNet synset together with their hypernym synsets regardless of the representation of the Wikipedia article. As stated before, this might be due to the fact that Wikipedia articles often contain hypernym concepts in their textual representation. Further, each synset has exactly one direct hypernym concept, while the number of hyponym concepts is not limited. This can cause a very noisy description of a synset, not focusing on the textual representation of the actual sense. When representing the Wikipedia sense, the categories always boost the performance, while redirects are not helpful and can yield even a performance drop. The reason might be that redirects contain much noisy information, e.g. spelling variations.

Baselines. The `rand` and the `mfs` baselines achieve an F$_1$-Measure of 0.527 and 0.534, respectively. They always assign a sense even only 221 of 320 synsets can be aligned to Wikipedia. If we only consider the 221 synsets for which an alignment exist, the `mfs` baseline achieves an F$_1$-Measure of 0.76, i.e. for 146 out of 221 synsets the aligned Wikipedia article is the most frequent sense as we defined it in Section 3.2.

Upper bound. The human annotators show a pairwise agreement κ between 0.866 and 0.878, which serves as an upper bound for this task. For each measure and its best performing experimental setting as listed in Table 4, we calculate the agreement with the annotators' alignments (see Table 5). The combined approach `ppr` + `cos` achieves the highest agreement values κ, between 0.728 and 0.740. These values show that the automatic annotation is fairly reliable.

5.1 Error analysis

We manually analyzed the alignments generated by the best performing experimental setup (`ppr` + `cos`, SYN+HYPER, P+T+C). For synsets corresponding to more than one extracted Wikipedia candidate, the average number of Wikipedia candidates is around 10, which, indeed, makes the alignment step very challenging for some synsets. For example, for the synset <*mission, military mission (an operation that is assigned by a higher headquarters)*> 30 Wikipedia candidates were extracted in total, whereas only the article with the title <*Military operation*> was aligned manually. 10 out of the 30 are articles about space flight missions and Christian missionary. Most of the remaining 19 refer to city names, song titles, and other named entities. Our approach returns the highest similarity for the article <*Military operation*>, which demonstrates that the alignment works well in this example.

As listed in Table 6, the best performing experimental setup correctly aligned 178 of the 227 manual alignments. The remaining 49 manual alignments were not assigned. Instead, 51 additional sense candidate pairs were incorrectly considered as alignment. It is noticeable that the errors are almost equally distributed among the distinctive properties a synset can have as defined in Section 4. We could not observe that a specific synset property causes the majority of errors.

Most of the 51 false positives are due to highly related sense alignment candidates, e.g. *(cottonseed, cottonseed oil)*, *(electroretinogram, electroretrinography)*, or *(insulin shock, insulin shock therapy)*. These sense alignment candidates have either the same stem but different suffixes or one part is a holonym or meronym of the other part. This knowledge can be used to apply additional post-processing steps to boost the performance. Further, even if they are non-aligned manually as they do not describe the same sense, the concepts are highly related, and thus, the alignment might be useful in specific tasks.

Most of the 49 manual alignments that could not be aligned automatically are due to the differences how senses are defined in WordNet and Wikipedia. For example, the WordNet synset <*payment, defrayal, defrayment: the act of paying money*> and the manually aligned Wikipedia article <*Payment: A payment is the transfer of wealth from one party (such as person or company) to another ...*> could not be aligned automatically. In this example, the textual similarity or relatedness is not sufficient to classify them as a valid alignment. This fact shows that other types of knowledge should be additionally integrated in the alignment approach, such as structural or taxonomic knowledge.

6 Conclusions

We have presented a novel two-step approach to automatically align English Wikipedia articles and WordNet synsets. We have shown that a threshold-based method models the task properly yielding none, one, or more than one alignment of a Wikipedia article for a given WordNet synset. This is different to previous sense alignment approaches. Further, we have shown that it is important to employ semantic relatedness measuring the similarity of textual sense representations. Our approach to the automatic alignment shows an encouraging performance of 0.78 F_1-Measure and 94.5% accuracy based on a comprehensive, well-balanced reference dataset consisting of 1,815 manually annotated sense alignment candidates.

We have created a fully-aligned resource with our best performing setting (`ppr` + `cos` , SYN+HYPER, P+T+C, threshold: 0.439 for `ppr` , 0.048 for `cos`), in which two-thirds of all WordNet noun synsets are aligned with one article from the English Wikipedia. On the one hand, this fact supports our assumption and overall motivation that both resources are partly complementary at the sense level (one-third of all noun synsets are not in Wikipedia). On the other hand, for the two-thirds of WordNet noun synsets, the alignment yields relational information from WordNet and encyclopedic information from Wikipedia.

We believe that this new resource and the enhanced knowledge therein can boost the performance of various NLP systems that previously had to rely on a single resource only. We already started research on integrating the aligned resource in WSD and semantic relatedness tasks. The fully-aligned resource as well as the reference dataset are publicly available at `http://www.ukp.tu-darmstadt.de/data/sense-alignment` for further research activities.

Acknowledgments

This work has been supported by the Emmy Noether Program of the German Research Foundation (DFG) under the grant No. GU 798/3-1, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under the grant No. I/82806. We thank our colleague Christian M. Meyer for many fruitful discussions during the work on this paper and our students Yevgen Chebotar and Christian Kirschner for their help on the experiments. Further, we thank the IXA group at the University of the Basque Country for making their code on the Personalized PageRank method based on WordNet available online.[9]

[9] `http://ixa2.si.ehu.es/ukb`

References

Agirre, E. and P. G. Edmonds (2006). *Word Sense Disambiguation: Algorithms and Applications*. Springer.

Agirre, E. and A. Soroa (2009). Personalizing PageRank for Word Sense Disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, Athens, Greece, pp. 33–41.

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

Brin, S. and L. Page (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks 30*(1-7), 107–117.

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

Fernando, S. and M. Stevenson (2010). Aligning WordNet Synsets and Wikipedia Articles. In *Proceedings of the AAAI Workshop on Collaboratively-built Knowledge Sources and Artificial Intelligence*, Atlanta, GA, USA.

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

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

Ponzetto, S. P. and R. Navigli (2009). Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence*, Pasadena, CA, USA, pp. 2083–2088.

Ponzetto, S. P. and R. Navigli (2010). Knowledge-rich Word Sense Disambiguation rivaling supervised system. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, pp. 1522–1531.

Ruiz-Casado, M., E. Alfonseca, and P. Castells (2005). Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence*, Volume 3528 of *LNCS*, pp. 380–386. Springer Verlag.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, United Kingdom, pp. 44–49.

Suchanek, F. M., G. Kasneci, and G. Weikum (2007). Yago: A Core of Semantic Knowledge. In *Proceedings of the 16th International World Wide Web Conference*, Banff, Canada, pp. 697–706.

Toral, A., O. Ferrandez, E. Agirre, and R. Munoz (2009). A study on Linking Wikipedia categories to Wordnet using text similarity. In *Proceedings of Recent Advances in Natural Language Processing*, Borovets, Bulgaria, pp. 449–454.

Toral, A., R. Munoz, and M. Monachini (2008). Named Entity WordNet. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Marocco, pp. 741–747.

Wolf, E. and I. Gurevych (2010). Aligning Sense Inventories in Wikipedia and WordNet. In *Proceedings of the 1st Workshop on Automated Knowledge Base Construction*, Grenoble, France, pp. 24–28.

Wu, F. and D. S. Weld (2008). Automatically Refining the Wikipedia Infobox Ontology. In *Proceedings of the 17th International Conference on World Wide Web*, Beijing, China, pp. 635–644.

Zesch, T., C. Müller, and I. Gurevych (2008). Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Marocco, pp. 1646–1652.