Aligning Sense Inventories in Wikipedia and WordNet

Elisabeth Wolf HTTP://WWW.UKP.TU-DARMSTADT.DE
Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt,
Hochschulstraße 10, D-64289 Darmstadt, Germany

Iryna Gurevych HTTP://WWW.UKP.TU-DARMSTADT.DE
Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt,
Hochschulstraße 10, D-64289 Darmstadt, Germany

Abstract

In this paper, we study the alignment of Wikipedia articles and WordNet synsets. Therefore, we propose a method to convert Wikipedia to a sense inventory. We show that an aligned sense inventory of both resources has two major benefits: the coverage of senses can be increased and enhanced information about aligned senses can be obtained. Our study and conclusions are based on human annotations of sense alignments.

1. Introduction

In recent years, Wikipedia has turned to a valuable resource in major NLP tasks, such as automatic word sense disambiguation, semantic relatedness, or named entity recognition. Wikipedia has the advantage of being multilingual and freely available containing a tremendous amount of encyclopedic knowledge enriched with hyperlink information constantly being updated by thousands of voluntary contributors. It has also been used as a lexical semantic resource (Zesch et al., 2008) and as a source of manually sense-disambiguated text (Mihalcea, 2007). In NLP tasks, it has turned out to perform comparable or even better than standard lexical semantic resources, such as WordNet (Fellbaum, 1998).

Many researchers observed that the knowledge represented in Wikipedia is complementary to WordNet regarding the coverage of different parts-of-speech and domains (Sanderson, 2008; Toral et al., 2008; Ponzetto & Navigli, 2009). While WordNet covers nouns, adjectives, verbs, and adverbs, Wikipedia mainly represents nouns of encyclopedic interest. For example, the noun *sphinx* has 3 meanings in WordNet and refers to 12 articles in Wikipedia[1]. The meanings *"the daughter of Typhon"* and the mythological figure in general are covered in both resources. In addition, WordNet contains the meaning of *an inscrutable person who keeps his thoughts and intentions secret*, which is not covered in Wikipedia. Instead, Wikipedia contains 10 articles about named entities, such as song, film, and book titles.

Previous work explored the overlap of Wikipedia and WordNet on the term level. The goal thereby is to increase the coverage of a lexical semantic resource in an NLP application. However, the question of overlap regarding the terms' word senses has been largely ignored. For example, even though the noun *incursion* is covered by both resources the meanings contained in WordNet and Wikipedia do not overlap. WordNet contains three different meanings: (i) *the act of entering some territory or domain (often in large numbers)*, (ii) *an attack that penetrates into enemy territory*, and (iii) *the mistake of incurring liability or blame*. In contrast, Wikipedia contains only one article about a science fiction role playing game. An aligned sense inventory integrating WordNet and Wikipedia on the sense level has two major benefits. On the one hand, the coverage of senses can be increased in a complementary lexical resource. On the other hand, enhanced information about aligned senses can be obtained combining the strengths of both resources: relational information from WordNet and the rich encyclopedic and multilingual information from Wikipedia.

The study presented in this paper aims to provide major insights into the nature of sense alignment of a linguistic resource such as WordNet and a collaboratively constructed resource such as Wikipedia. Therefore, we first convert Wikipedia to a sense inventory based on the Wikipedia article graph. We further present an algorithm to extract all potential sense alignments from Wikipedia given a WordNet synset. We employ four human annotators in the sense alignment task and draw some preliminary conclusions on the aspect of complementarity of senses in both resources.

[1] Any numbers regarding Wikipedia in this paper refer to the English edition of August, 22nd 2009.

Though the results of this work are in the domain of lexical semantics, we believe they will have substantial impact upon the field of lexical semantic processing in NLP, including any sense based NLP application, such as word sense disambiguation, machine translation, or sense cluster based information retrieval.

2. Related Work

Previous research works proposed methods for the automatic alignment of the category system of Wikipedia to WordNet synsets to create a semantically enriched ontology (Suchanek et al., 2007; Toral et al., 2008; 2009; Ponzetto & Navigli, 2009). None of these approaches, however, examined the article level, which causes two major limitations. First of all, the number of Wikipedia categories (about 0.5 million) is much smaller compared to the number of Wikipedia articles (about 3.35 million). Secondly, disregarding the article level neglects the huge amount of semantically structured textual content that the articles provide. Therefore, we aim at using the Wikipedia article graph representing senses.

To our knowledge, there exist only two works addressing the issue of integrating Wikipedia article and WordNet synsets (Ruiz-Casado et al., 2005; Mihalcea, 2007). Ruiz-Casado et al. (2005) proposed a method mapping Simple English Wikipedia articles to their most similar WordNet synsets. The mapping depends on the vector-based similarity of the synsets' gloss and the article text. However, an analysis regarding the complementarity of senses in both resources is missing. Mihalcea (2007) automatically generated a sense-tagged corpus using Wikipedia as a source of sense annotations. Based on it, they show that Wikipedia sense annotations can be used to construct accurate word sense classifiers. For evaluation, they manually mapped Wikipedia articles to WordNet synsets. However, no comprehensive analysis of the sense overlap in both resources is provided.

3. Representing Senses

In this section, we provide an operational definition of sense in Wikipedia and WordNet to be used in the remainder of this paper. As almost all Wikipedia articles refer to nouns, we focus on this part-of-speech in our study.

Sense representation in Wikipedia. Wikipedia's guidelines state: *"Articles are about a person, or a group, a concept, a place, a thing, an event, etc. [...] articles rarely, if ever, contain several distinct definitions or usages of a term"*.[2] We therefore consider each Wikipedia article as a representation of a particular sense. Further, Mihalcea (2007) states that looking at one article as a sense enables manual alignment with WordNet synsets. In their study, 30 nouns are used.

To retrieve the set of senses for a given term, we select all articles fulfilling one of the following criteria: the article title string matches the term string (e.g., the article *Window* is retrieved for the term *Window*), the article title string is of the form *term_(description tag)* (e.g., *Window_(computing)* and *Window_(film)*), or the article has a redirect that matches the term string or is of the form *term_(description tag)* (e.g., *Chaff_(countermeasure)* has a redirect *Window_(codename)* and, thus, is retrieved for the term *Window*). The string comparison is case-insensitive.

Sense representation in WordNet. Terms are organized in synsets. A synset is a collection of synonymous word senses, which are described by glosses and optionally by some exemplary sentences in which the sense is used. As a term can have more than one meaning, each term can be assigned to several synsets. Semantic relations can be established between synsets, such as hypernymy and hyponymy. We consider a WordNet synset (possibly represented by more than one word sense) as a sense to be aligned with a Wikipedia article.

Extracting potential sense pairs. In order to align senses, we first retrieve all potential Wikipedia sense alignments (articles) for a given WordNet sense (synset). Therefore, let S be a WordNet synset with a set of synonyms $\{s_1, \cdots, s_n\}$ of size n. For each synonym $s \in S$, we extract all Wikipedia senses according to the criteria introduced above[3]. For example, for the synset <*article, clause: a separate section of a legal document*> we extract all Wikipedia senses for the nouns *article* and *clause* yielding four senses in total (here: two for *article* and another two for *clause*). For WordNet synsets defining living and fossil organisms we perform some terminological simplification as they are usually preceded by one of the eight major biological classification codes, e.g., *genus* or *family*, which does not always hold in Wikipedia. For those synsets, we collect all Wikipedia senses according to the synonymous words in the synset with and without the preceding codes. Table 1 lists two examples of extracted sense pairs.

Based on 82,115 noun synsets in WordNet 3.0, we retrieve more than one Wikipedia sense for 41,649 synsets (50.72%), exactly one Wikipedia sense for 27,973 synsets (34.07%) and no Wikipedia senses for the remaining 12,493 synsets (15.21%). For synsets with more than one Wikipedia sense, we extract 6.36 articles per synset on av-

[2] http://en.wikipedia.org/wiki/Wikipedia:WWIN [Last access: Feb, 09th 2010]

[3] This method automatically allows the extraction on word sense level by considering only one synonymous word.

Table 1. Examples of sense pairs (left: WordNet, right: Wikipedia)

Synset	Gloss + Examples	Article Title	First Paragraph (shortened)
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; ”I felt so bad I went to see my doctor”*	*Physician*	*A physician, medical practitioner, doctor of medicine, or medical doctor practices medicine, and is concerned with [· · ·]*
doctor, medico, doc, MD, physician, Dr.	*a licensed medical practitioner; ”I felt so bad I went to see my doctor”*	*Doctor (title)*	*Doctor (gen.: doctoris) means teacher in Latin. The word is originally an agentive noun of the verb docre (’to teach’). It has been used continuously as [· · ·]*

erage. The size of these synsets, i.e. the number of synonymous words, is 2.15 on average.

4. Sense Alignment Study

Dataset. For the annotation study, we randomly sampled 14 nouns, which are polysemous in WordNet and for which at least one Wikipedia article is retrieved: *alignment*, *article*, *bandwagon*, *bird of paradise*, *borrowing*, *bump*, *carat*, *chevalier*, *cobweb*, *configuration*, *cox*, *damper*, *detention*, and *doctor*. They yield 38 synsets, i.e. 2.71 synsets per noun on average. For 38 synsets, 297 Wikipedia articles were extracted, i.e. 7.82 articles per synset on average.

Annotation Task. The sense alignment is performed by four human annotators. According to the examples in Table 1 the annotators were provided sense pairs, each consisting of a WordNet and a Wikipedia sense generated in the previous step. The annotation task was to label each sense pair either as the same sense or not[4]. In case of uncertainty, the annotators were allowed to leave a comment instead of a binary judgment for later analysis. This helps to gain insights in difficult cases.

Inter-Annotator Agreement. Table 2 outlines the class distribution for four annotators.

Table 2. Annotations per class

Annotator	*A*	*B*	*C*	*D*	majority
# different senses	274	274	271	266	275
# same senses	23	22	26	21	22
# comments	0	1	0	10	0

The majority of sense pairs were annotated as different sense; only between 21 and 26 sense pairs were considered the same sense. It is remarkable that only annotator D made ample use of the comment option.

In order to assess the reliability of our data, we computed the observed inter-annotator agreement A_{O} and the chance-corrected agreement κ within single classes as introduced by Fleiss (1971). As the distribution of the two classes *different sense* and *same sense* is highly skewed this variant of the κ statistics ensures a more accurate analysis of the annotations. The agreement values are outlined in Table 3.

Table 3. Inter-annotator agreement

	A–B	*A–C*	*A–D*	*B–C*	*B–D*	*C–D*
A_{O}	.9932	.9764	.9630	.9764	.9596	.9495
κ - diff	.9529	.8443	.8370	.8443	.7963	.7921
κ - same	.9760	.8443	.9018	.8640	.9248	.7921

The average observed agreement A_{O} is 0.97, while the multi-κ for the *different sense* class is 0.84, and even 0.88 for the *same sense* class. These numbers confirm a high reliability of our data.

Manual analysis of disagreements shows that almost all of them are caused by the lack of instructions how to handle senses that are part of a more general sense. For example, the WordNet sense <*configuration, constellation: an arrangement of parts or elements*> was paired with the Wikipedia sense <*Configuration (mathematics): In mathematics, especially geometry, a configuration is an arrangement of points in a certain way [...]*>. Another example is the pair <*alignment: the spatial property possessed by an arrangement or position of things in a straight line or in parallel lines*> and <*Typographic alignment: In typesetting and page layout, alignment or range, is the setting of text flow or image placement [...]*>. In both examples, the Wikipedia sense is actually a hyponym of the WordNet sense as it is more specific. By analyzing the annotated data, we found out that a substantial number of Wikipedia senses follow this pattern. While the annotators A and B both judged those cases as different (yielding a high inter-annotator agreement), annotator C judged most of them as the same sense, and D used the comment option.

In future work, the annotation guidelines will be extended in a way that sense pairs where, e.g. the Wikipedia sense is a hyponym of the WordNet sense, should not be annotated to be the same sense – as the hyponym itself should be contained in WordNet as a separate sense.

Analysis. The final dataset was compiled by means of a majority decision. There were no ties in our data. Given 297 sense pairs, 275 were annotated as different sense,

[4] The annotation guideline and the final dataset are available at http://www.ukp.tu-darmstadt.de/data/lexical-resources

while 22 were annotated as the same sense. 20 synsets were aligned with one article, while one synset was aligned with two articles.

For the remaining 17 synsets, which were not aligned with any article, we searched for the corresponding senses in Wikipedia manually. Four of them could be aligned with a list entry on the corresponding disambiguation page, e.g., for the synset <*bandwagon: a large ornate wagon for carrying a musical band*> the disambiguation page *Bandwagon* contains a list entry: <*a wagon which carries a band in a parade*>. Another example is the synset <*article: one of a class of artifacts; "an article of clothing"*>. The disambiguation page *Article* contains a list entry: <*Item, as in "article of clothing"*>. Both examples demonstrate that the sense is not represented as a Wikipedia article yet, but as a list entry in a disambiguation page only. As Wikipedia is still growing, such list entries could be expanded to an article in the future.

Another two synsets could not be aligned with a Wikipedia article due to the restrictions of the algorithm for extracting sense pairs. For example, for the synset <*bandwagon: a popular trend that attracts growing support*> the corresponding Wikipedia article *Bandwagon effect* was not extracted as its title differs. However, the relaxation of our approach to include such cases, i.e. to consider longer subsuming strings, would yield a high number of false positives, making the disambiguation step harder and very time consuming. A better strategy to improve our extraction method is to include the processing of anchor text in hyperlinks. Hyperlinks in Wikipedia can be extended by a so called *anchor text*. For instance, *"Leary ultimately joined the* [[*Bandwagon effect*|*bandwagon*]] *[...]"* is an example of a link to the article *Bandwagon effect*, which appears in the text as *bandwagon*. The consideration of referenced articles in hyperlinks with anchor texts matching the given term is a subject of our future work.

An analysis of the data on the term level shows that for some nouns all senses could be aligned with at least one Wikipedia sense: *bird of paradise* (3 senses), *carat* (2), *cox* (2), *detention* (2), and *doctor* (4). For the synset <*cox, coxwain: the helmsman of a ship's boat or a racing crew*> even two Wikipedia senses were assigned, one describing the person in charge of a boat and the other one describing the person in charge of a *rowing* boat. Both senses are highly related and hard to distinguish. For the noun *bump*, however, none of its 3 senses could be assigned to either an article or a list entry in a disambiguation page, though several candidate Wikipedia articles were extracted. This, in fact, is an indication of the complementarity of senses in both resources.

5. Discussion and Conclusions

Senses in Wikipedia. The alignments of Wikipedia articles and WordNet synsets obtained in this study confirm the previous finding by Mihalcea (2007) that a Wikipedia article is an appropriate level of granularity for representing a sense. Therefore, Wikipedia articles can be aligned with WordNet synsets and the word senses therein.

Annotation task. The annotations yielded high inter-annotator agreement on sense alignment, even though the information about how to handle hyponyms was not explicitly included in the guidelines. Therefore, extended guidelines should yield even higher reliability in the future.

Sense overlap. 21 out of 38 WordNet synsets were aligned with at least one Wikipedia article by the annotators. 2 additional synsets could be aligned after the Wikipedia articles were retrieved manually and another 4 synsets could be aligned with a list entry in a disambiguation page. In total, about two thirds of the sampled synsets could be found in both resources. For all aligned senses, a merged sense inventory integrating Wikipedia and WordNet entails comprehensive information representing the sense including relational as well as encyclopedic and multilingual information.

Complementarity of senses. The remaining one third of the sampled synsets could not be aligned with Wikipedia articles, though Wikipedia article candidates have been found. This confirms our hypothesis that Wikipedia and WordNet are complementary in the coverage of senses. In particular, the number of specialized senses in Wikipedia, often hyponyms of the WordNet synset, was significant. In fact, this emphasizes the encyclopedic nature of Wikipedia containing rather factual knowledge than abstract and more general senses.

In order to further analyse both resources on the sense level, we currently construct a large dataset based on the enhanced annotation guidelines and an improved sense pair extraction method. All data will be made available to the community. Finally, we started research on automatic alignment methods using the annotations as a gold standard for evaluation.

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 colleagues Christian M. Meyer and Dr. Torsten Zesch for many fruitful discussions during the work on this paper.

References

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

Fleiss, Joseph L. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–381, November 1971.

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

Ponzetto, Simone Paolo and Navigli, Roberto. Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *Proceedings of the 21th International Joint Conference on Artificial Intelligence (IJCAI '09)*, pp. 2083–2088, Pasadena, CA, USA, 2009.

Ruiz-Casado, Maria, Alfonseca, Enrique, and Castells, Pablo. Automatic Assignment of Wikipedia Encyclopedic Entries to WordNet Synsets. In *Advances in Web Intelligence*, volume 3528 of *Lecture Notes in Computer Science*, pp. 380–386. Springer Verlag, 2005.

Sanderson, Mark. Ambiguous Queries: Test Collections Need More Sense. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '08)*, pp. 499–506, New York, NY, USA, 2008.

Suchanek, Fabian M., Kasneci, Gjergji, and Weikum, Gerhard. YAGO: A Core of Semantic Knowledge. In *16th international World Wide Web conference (WWW 2007)*, pp. 697–706, Banff, Alberta, Canada, 2007.

Toral, Antonio, Munoz, Rafael, and Monachini, Monica. Named Entity WordNet. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.

Toral, Antonio, Ferrandez, Oscar, Agirre, Eneko, and Munoz, Rafael. A study on Linking Wikipedia categories to Wordnet using text similarity. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)*, Borovets, Bulgaria, 2009.

Zesch, Torsten, Müller, Christof, and Gurevych, Iryna. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, 2008.