Using Wikipedia for Automatic Word Sense Disambiguation

Rada Mihalcea
Department of Computer Science
University of North Texas
rada@cs.unt.edu

Abstract

This paper describes a method for generating sense-tagged data using Wikipedia as a source of sense annotations. Through word sense disambiguation experiments, we show that the Wikipedia-based sense annotations are reliable and can be used to construct accurate sense classifiers.

1 Introduction

Ambiguity is inherent to human language. In particular, word sense ambiguity is prevalent in all natural languages, with a large number of the words in any given language carrying more than one meaning. For instance, the English noun *plant* can mean *green plant* or *factory*; similarly the French word *feuille* can mean *leaf* or *paper*. The correct sense of an ambiguous word can be selected based on the context where it occurs, and correspondingly the problem of *word sense disambiguation* is defined as the task of automatically assigning the most appropriate meaning to a polysemous word within a given context.

Among the various knowledge-based (Lesk, 1986; Galley and McKeown, 2003; Navigli and Velardi, 2005) and data-driven (Yarowsky, 1995; Ng and Lee, 1996; Pedersen, 2001) word sense disambiguation methods that have been proposed to date, supervised systems have been constantly observed as leading to the highest performance. In these systems, the sense disambiguation problem is formulated as a supervised learning task, where each sense-tagged occurrence of a particular word is transformed into a feature vector which is then used in an automatic learning process. Despite their high performance, these supervised systems have an important drawback: their applicability is limited to those few words for which sense tagged data is available, and their accuracy is strongly connected to the amount of labeled data available at hand.

To address the sense-tagged data bottleneck problem, different methods have been proposed in the past, with various degrees of success. This includes the automatic generation of sense-tagged data using monosemous relatives (Leacock et al., 1998; Mihalcea and Moldovan, 1999; Agirre and Martinez, 2004), automatically bootstrapped disambiguation patterns (Yarowsky, 1995; Mihalcea, 2002), parallel texts as a way to point out word senses bearing different translations in a second language (Diab and Resnik, 2002; Ng et al., 2003; Diab, 2004), and the use of volunteer contributions over the Web (Chklovski and Mihalcea, 2002).

In this paper, we investigate a new approach for building sense tagged corpora using Wikipedia as a source of sense annotations. Starting with the hyperlinks available in Wikipedia, we show how we can generate sense annotated corpora that can be used for building accurate and robust sense classifiers. Through word sense disambiguation experiments performed on the Wikipedia-based sense tagged corpus generated for a subset of the SENSEVAL ambiguous words, we show that the Wikipedia annotations are reliable, and the quality of a sense tagging classifier built on this data set exceeds by a large margin the accuracy of an informed baseline that selects the most frequent word sense by default.

The paper is organized as follows. We first pro-

Proceedings of NAACL HLT 2007, pages 196–203,
Rochester, NY, April 2007. ©2007 Association for Computational Linguistics

vide a brief overview of Wikipedia, and describe the view of Wikipedia as a sense tagged corpus. We then show how the hyperlinks defined in this resource can be used to derive sense annotated corpora, and we show how a word sense disambiguation system can be built on this dataset. We present the results obtained in the word sense disambiguation experiments, and conclude with a discussion of the results.

2 Wikipedia

Wikipedia is a free online encyclopedia, representing the outcome of a continuous collaborative effort of a large number of volunteer contributors. Virtually any Internet user can create or edit a Wikipedia webpage, and this "freedom of contribution" has a positive impact on both the quantity (fast-growing number of articles) and the quality (potential mistakes are quickly corrected within the collaborative environment) of this online resource. Wikipedia editions are available for more than 200 languages, with a number of entries varying from a few pages to more than one million articles per language.[1]

The basic entry in Wikipedia is an *article* (or *page*), which defines and describes an entity or an event, and consists of a hypertext document with hyperlinks to other pages within or outside Wikipedia. The role of the hyperlinks is to guide the reader to pages that provide additional information about the entities or events mentioned in an article.

Each article in Wikipedia is uniquely referenced by an identifier, which consists of one or more words separated by spaces or underscores, and occasionally a parenthetical explanation. For example, the article for *bar* with the meaning of *"counter for drinks"* has the unique identifier *bar (counter)*.[2]

The hyperlinks within Wikipedia are created using these unique identifiers, together with an *anchor text* that represents the surface form of the hyperlink. For instance, *"Henry Barnard, [[United States|American]] [[educationalist]], was born in [[Hartford, Connecticut]]"* is an example of a sentence in Wikipedia containing links to the articles *United States*, *educationalist*, and *Hartford, Connecticut*. If the surface form and the unique identifier of an article coincide, then the surface form can be turned directly into a hyperlink by placing double brackets around it (e.g. *[[educationalist]]*). Alternatively, if the surface form should be hyperlinked to an article with a different unique identifier, e.g. link the word *American* to the article on *United States*, then a piped link is used instead, as in *[[United States|American]]*.

One of the implications of the large number of contributors editing the Wikipedia articles is the occasional lack of consistency with respect to the unique identifier used for a certain entity. For instance, the concept of *circuit (electric)* is also referred to as *electronic circuit*, *integrated circuit*, *electric circuit*, and others. This has led to the so-called *redirect pages*, which consist of a redirection hyperlink from an alternative name (e.g. *integrated circuit*) to the article actually containing the description of the entity (e.g. *circuit (electric)*).

Finally, another structure that is particularly relevant to the work described in this paper is the *disambiguation page*. Disambiguation pages are specifically created for ambiguous entities, and consist of links to articles defining the different meanings of the entity. The unique identifier for a disambiguation page typically consists of the parenthetical explanation *(disambiguation)* attached to the name of the ambiguous entity, as in e.g. *circuit_(disambiguation)* which is the unique identifier for the disambiguation page of the entity *circuit*.

3 Wikipedia as a Sense Tagged Corpus

A large number of the concepts mentioned in Wikipedia are explicitly linked to their corresponding article through the use of links or piped links. Interestingly, these links can be regarded as *sense annotations* for the corresponding concepts, which is a property particularly valuable for entities that are ambiguous. In fact, it is precisely this observation that we rely on in order to generate sense tagged corpora starting with the Wikipedia annotations.

For example, ambiguous words such as e.g. *plant*, *bar*, or *chair* are linked to different Wikipedia articles depending on their meaning in the context where they occur. Note that the links are *manually* created by the Wikipedia users, which means that they are most of the time accurate and referencing

[1] In the experiments reported in this paper, we use a download from March 2006 of the English Wikipedia, with approximately 1 million articles, and more than 37 millions hyperlinks.

[2] The unique identifier is also used to form the article URL, e.g. http://en.wikipedia.org/wiki/Bar_(counter)

the correct article. The following represent five example sentences for the ambiguous word *bar*, with their corresponding Wikipedia annotations (links):

In 1834, Sumner was admitted to the **[[bar (law)|bar]]** at the age of twenty-three, and entered private practice in Boston.

It is danced in 3/4 time (like most waltzes), with the couple turning approx. 180 degrees every **[[bar (music)|bar]]**.

Vehicles of this type may contain expensive audio players, televisions, video players, and **[[bar (counter)|bar]]**s, often with refrigerators.

Jenga is a popular beer in the **[[bar (establishment)|bar]]**s of Thailand.

This is a disturbance on the water surface of a river or estuary, often cause by the presence of a **[[bar (landform)|bar]]** or dune on the riverbed.

To derive sense annotations for a given ambiguous word, we use the links extracted for all the hyperlinked Wikipedia occurrences of the given word, and map these annotations to word senses. For instance, for the *bar* example above, we extract five possible annotations: *bar (counter), bar (establishment), bar (landform), bar (law),* and *bar (music)*.

Although Wikipedia provides the so-called disambiguation pages that list the possible meanings of a given word, we decided to use instead the annotations collected directly from the Wikipedia links. This decision is motivated by two main reasons. First, a large number of the occurrences of ambiguous words are not linked to the articles mentioned by the disambiguation page, but to related concepts. This can happen when the annotation is performed using a concept that is similar, but not identical to the concept defined. For instance, the annotation for the word *bar* in the sentence *"The blues uses a rhythmic scheme of twelve 4/4 [[measure (music)|bars]]"* is *measure (music)*, which, although correct and directly related to the meaning of *bar (music)*, is not listed in the disambiguation page for *bar*.

Second, most likely due to the fact that Wikipedia is still in its incipient phase, there are several inconsistencies that make it difficult to use the disambiguation pages in an automatic system. For example, for the word *bar*, the Wikipedia page with the identifier *bar* is a disambiguation page, whereas for the word *paper*, the page with the identifier *paper* contains a description of the meaning of paper as *"material made of cellulose,"* and a different page *paper_(disambiguation)* is defined as a disambiguation page. Moreover, in other cases such as e.g. the entries for the word *organization*, no disambiguation page is defined; instead, the articles corresponding to different meanings of this word are connected by links labeled as "alternative meanings."

Therefore, rather than using the senses listed in a disambiguation page as the sense inventory for a given ambiguous word, we chose instead to collect all the annotations available for that word in the Wikipedia pages, and then map these labels to a widely used sense inventory, namely WordNet.[3]

3.1 Building Sense Tagged Corpora

Starting with a given ambiguous word, we derive a sense-tagged corpus following three main steps:

First, we extract all the paragraphs in Wikipedia that contain an occurrence of the ambiguous word as part of a link or a piped link. We select paragraphs based on the Wikipedia paragraph segmentation, which typically lists one paragraph per line.[4] To focus on the problem of word sense disambiguation, rather than named entity recognition, we explicitly avoid named entities by considering only those word occurrences that are spelled with a lower case. Although this simple heuristic will also eliminate examples where the word occurs at the beginning of a sentence (and therefore are spelled with an upper case), we decided nonetheless to not consider these examples so as to avoid any possible errors.

Next, we collect all the possible labels for the given ambiguous word by extracting the leftmost component of the links. For instance, in the piped link *[[musical_notation|bar]]*, the label *musical_notation* is extracted. In the case of simple links (e.g. *[[bar]]*), the word itself can also play the role of a valid label if the page it links to is not determined as a disambiguation page.

Finally, the labels are manually mapped to their corresponding WordNet sense, and a sense tagged

[3] Alternatively, the Wikipedia annotations could also play the role of a sense inventory, without the mapping to WordNet. We chose however to perform this mapping for the purpose of allowing evaluations using a widely used sense inventory.

[4] The average length of a paragraph is 80 words.

Word sense	Labels in Wikipedia	Wikipedia definition	WordNet definition
bar (establishment)	bar_(establishment), nightclub gay_club, pub	a retail establishment which serves alcoholic beverages	a room or establishment where alcoholic drinks are served over a counter
bar (counter)	bar_(counter)	the counter from which drinks are dispensed	a counter where you can obtain food or drink
bar (unit)	bar_(unit)	a scientific unit of pressure	a unit of pressure equal to a million dynes per square centimeter
bar (music)	bar_(music), measure_music musical_notation	a period of music	musical notation for a repeating pattern of musical beats
bar (law)	bar_association, bar_law law_society_of_upper_canada state_bar_of_california	the community of persons engaged in the practice of law	the body of individuals qualified to practice law in a particular jurisdiction
bar (landform)	bar_(landform)	a type of beach behind which lies a lagoon	a submerged (or partly submerged) ridge in a river or along a shore
bar (metal)	bar_metal, pole_(object)	-	a rigid piece of metal or wood
bar (sports)	gymnastics_uneven_bars, handle_bar	-	a horizontal rod that serves as a support for gymnasts as they perform exercises
bar (solid)	candy_bar, chocolate_bar	-	a block of solid substance

Table 1: Word senses for the word *bar*, based on annotation labels used in Wikipedia

corpus is created. This mapping process is very fast, as a relatively small number of labels is typically identified for a given word. For instance, for the dataset used in the experiments reported in Section 5, an average of 20 labels per word was extracted.

To ensure the correctness of this last step, for the experiments reported in this paper we used two human annotators who independently mapped the Wikipedia labels to their corresponding WordNet sense. In case of disagreement, a consensus was reached through adjudication by a third annotator. In a mapping agreement experiment performed on the dataset from Section 5, an inter-annotator agreement of 91.1% was observed with a kappa statistics of κ=87.1, indicating a high level of agreement.

3.2 An Example

As an example, consider the ambiguous word *bar*, with 1,217 examples extracted from Wikipedia where *bar* appeared as the rightmost component of a piped link or as a word in a simple link. Since the page with the identifier *bar* is a disambiguation page, all the examples containing the single link *[[bar]]* are removed, as the link does not remove the ambiguity. This process leaves us with 1,108 examples, from which 40 different labels are extracted. These labels are then manually mapped to nine senses in WordNet. Figure 1 shows the labels extracted from the Wikipedia annotations for the word *bar*, the corresponding WordNet definition, as well as the Wikipedia definition (when the sense was defined in the Wikipedia disambiguation page).

4 Word Sense Disambiguation

Provided a set of sense-annotated examples for a given ambiguous word, the task of a word sense disambiguation system is to automatically learn a disambiguation model that can predict the correct sense for a new, previously unseen occurrence of the word.

We use a word sense disambiguation system that integrates local and topical features within a machine learning framework, similar to several of the top-performing supervised word sense disambiguation systems participating in the recent SENSEVAL evaluations (http://www.senseval.org).

The disambiguation algorithm starts with a preprocessing step, where the text is tokenized and annotated with part-of-speech tags. Collocations are identified using a sliding window approach, where a collocation is defined as a sequence of words that forms a compound concept defined in WordNet.

Next, local and topical features are extracted from the context of the ambiguous word. Specifically, we use the current word and its part-of-speech, a local context of three words to the left and right of the ambiguous word, the parts-of-speech of the surrounding words, the verb and noun before and after the ambiguous words, and a global context implemented through sense-specific keywords determined as a list of at most five words occurring at least three times in the contexts defining a certain word sense.

This feature set is similar to the one used by (Ng and Lee, 1996), as well as by a number of state-of-the-art word sense disambiguation systems participating in the Senseval-2 and Senseval-3 evaluations. The features are integrated in a Naive Bayes classifier, which was selected mainly for its performance in previous work showing that it can lead to a state-of-the-art disambiguation system given the features we consider (Lee and Ng, 2002).

5 Experiments and Results

To evaluate the quality of the sense annotations generated using Wikipedia, we performed a word sense disambiguation experiment on a subset of the ambiguous words used during the Senseval-2 and Senseval-3 evaluations. Since the Wikipedia annotations are focused on nouns (associated with the entities typically defined by Wikipedia), the sense annotations we generate and the word sense disambiguation experiments are also focused on nouns.

Starting with the 49 ambiguous nouns used during the Senseval-2 (29) and Senseval-3 (20) evaluations, we generated sense tagged corpora following the process outlined in Section 3.1. We then removed all those words that have only one Wikipedia label (e.g. *detention*, which occurs 58 times, but appears as a single link *[[detention]]* in all the occurrences), or which have several labels that are all mapped to the same WordNet sense (e.g. *church*, which has 2,198 occurrences with several different labels such as *Roman church*, *Christian church*, *Catholic church*, which are all mapped to the meaning of *church, Christian church* as defined in WordNet). This resulted in a set of 30 words that have their Wikipedia annotations mapped to at least two senses according to the WordNet sense inventory.

Table 2 shows the disambiguation results using the word sense disambiguation system described in Section 4, using ten-fold cross-validation. For each word, the table also shows the number of senses, the total number of examples, and two baselines: a simple informed baseline that selects the most frequent sense by default,[5] and a more refined baseline that implements the corpus-based version of the Lesk algorithm (Kilgarriff and Rosenzweig, 2000).

			baselines		word sense
word	#s	#ex	MFS	LeskC	disambig.
argument	2	114	70.17%	73.63%	**89.47%**
arm	3	291	61.85%	69.31%	**84.87%**
atmosphere	3	773	54.33%	56.62%	**71.66%**
bank	3	1074	**97.20%**	**97.20%**	**97.20%**
bar	10	1108	47.38%	68.09%	**83.12%**
chair	3	194	67.57%	65.78%	**80.92%**
channel	5	366	51.09%	52.50%	**71.85%**
circuit	4	327	85.32%	85.62%	**87.15%**
degree	7	849	58.77%	73.05%	**85.98%**
difference	2	24	**75.00%**	**75.00%**	**75.00%**
disc	3	73	52.05%	52.05%	**71.23%**
dyke	2	76	77.63%	82.00%	**89.47%**
fatigue	3	123	66.66%	70.00%	**93.22%**
grip	3	34	44.11%	**77.00%**	70.58%
image	2	84	69.04%	74.50%	**80.28%**
material	3	223	**95.51%**	**95.51%**	**95.51%**
mouth	2	409	94.00%	94.00%	**95.35%**
nature	2	392	**98.72%**	**98.72%**	98.21%
paper	5	895	**96.98%**	**96.98%**	**96.98%**
party	3	764	68.06%	68.28%	**75.91%**
performance	2	271	**95.20%**	**95.20%**	**95.20%**
plan	3	83	77.10%	81.00%	**81.92%**
post	5	33	54.54%	**62.50%**	51.51%
restraint	2	9	**77.77%**	**77.77%**	**77.77%**
sense	2	183	**95.10%**	**95.10%**	**95.10%**
shelter	2	17	**94.11%**	**94.11%**	**94.11%**
sort	2	11	81.81%	**90.90%**	**90.90%**
source	3	78	55.12%	81.00%	**92.30%**
spade	3	46	60.86%	**81.50%**	80.43%
stress	3	565	53.27%	54.28%	**86.37%**
Average	3.31	316	72.58%	78.02%	**84.65%**

Table 2: Word sense disambiguation results, including two baselines (MFS = most frequent sense; LeskC = Lesk-corpus) and the word sense disambiguation system. Number of senses (#s) and number of examples (#ex) are also indicated.

6 Discussion

Overall, the Wikipedia-based sense annotations were found reliable, leading to accurate sense classifiers with an average relative error rate reduction of 44% compared to the most frequent sense baseline, and 30% compared to the Lesk-corpus baseline.

There were a few exceptions to this general trend. For instance, for some of the words for which only a small number of examples could be collected from Wikipedia, e.g. *restraint* or *shelter*, no accuracy improvement was observed compared to the most frequent sense baseline. Similarly, several words in the

[5] Note that this baseline assumes the availability of a sense tagged corpus in order to determine the most frequent sense of a word. The baseline is therefore "informed," as compared to a random, "uninformed" sense selection.

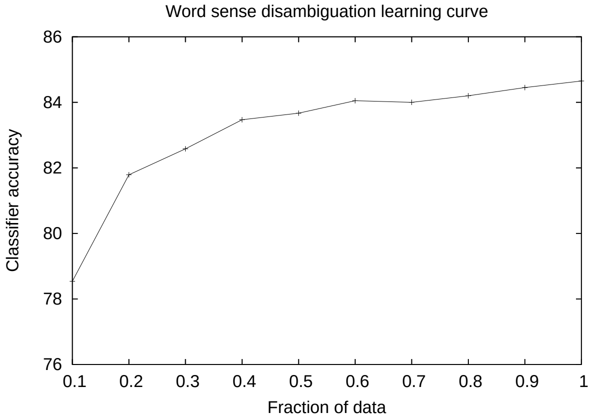

Figure 1: Learning curve on the Wikipedia data set.

data set have highly skewed sense distributions, such as e.g. *bank*, which has a total number of 1,074 examples out of which 1,044 examples pertain to the meaning of *financial institution*, or the word *material* with 213 out of 223 examples annotated with the meaning of *substance*.

One aspect that is particularly relevant for any supervised system is the learning rate with respect to the amount of available data. To determine the learning curve, we measured the disambiguation accuracy under the assumption that only a fraction of the data were available. We ran ten fold cross-validation experiments using 10%, 20%, ..., 100% of the data, and averaged the results over all the words in the data set. The resulting learning curve is plotted in Figure 1. Overall, the curve indicates a continuously growing accuracy with increasingly larger amounts of data. Although the learning pace slows down after a certain number of examples (about 50% of the data currently available), the general trend of the curve seems to indicate that more data is likely to lead to increased accuracy. Given that Wikipedia is growing at a fast pace, the curve suggests that the accuracy of the word sense classifiers built on this data is likely to increase for future versions of Wikipedia.

Another aspect we were interested in was the correlation in terms of sense coverage with respect to other sense annotated data currently available. For the set of 30 nouns in our data set, we collected all the word senses that were defined in either the Wikipedia-based sense-tagged corpus or in the Senseval corpus. We then determined the percentage covered by each sense with respect to the entire data set available for a given ambiguous word. For instance, the noun *chair* appears in Wikipedia with senses #1 (68.0%), #2 (31.9%), and #4(0.1%), and in Senseval with senses #1 (87.7%), #2 (6.3%), and #3 (6.0%). The senses that do not appear are indicated with a 0% coverage. The correlation is then measured between the relative sense frequencies of all the words in our dataset, as observed in the two corpora. Using the Pearson (r) correlation factor, we found an overall correlation of $r = 0.51$ between the sense distributions in the Wikipedia corpus and the Senseval corpus, which indicates a medium correlation. This correlation is much lower than the one observed between the sense distributions in the training data and in the test data in the Senseval corpus, which was measured at a high $r = 0.95$. This suggests that the sense coverage in Wikipedia follows a different distribution than in Senseval, mainly reflecting the difference between the genres of the two corpora: an online collection of encyclopedic pages as available from Wikipedia, versus the manually balanced British National Corpus used in Senseval. It also suggests that using the Wikipedia-based sense tagged corpus to disambiguate words in the Senseval data or viceversa would require a change in the distribution of senses as previously done in (Agirre and Martinez, 2004).

			baselines		word sense
Dataset	#s	#ex	MFS	LeskC	disambig.
Senseval	4.60	226	51.53%	58.33%	68.13%
Wikipedia	3.31	316	72.58%	78.02%	84.65%

Table 3: Average number of senses and examples, most frequent sense and Lesk-corpus baselines, and word sense disambiguation performance on the Senseval and Wikipedia datasets.

Table 3 shows the characteristics of the Senseval and the Wikipedia datasets for the nouns listed in Table 2. The table also shows the most frequent sense baseline, the Lesk-corpus baseline, as well as the accuracy figures obtained on each dataset using the word sense disambiguation system described in Section 4.[6]

[6] As a side note, the accuracy obtained by our system on the Senseval data is comparable to that of the best participating systems. Using the output of the best systems: the JHU$_R$ system on the Senseval-2 words, and the HLTS$_3$ system on the

Overall the sense distinctions identified in Wikipedia are fewer and typically coarser than those found in WordNet. As shown in Table 3, for the set of ambiguous words listed in Table 2, an average of 4.6 senses were used in the SENSEVAL annotations, as compared to about 3.3 senses per word found in Wikipedia. This is partly due to a different sense coverage and distribution in the Wikipedia data set (e.g. the meaning of *ambiance* for the ambiguous word *atmosphere* does not appear at all in the Wikipedia corpus, although it has the highest frequency in the SENSEVAL data), and partly due to the coarser sense distinctions made in Wikipedia (e.g. Wikipedia does not make the distinction between the act of grasping and the actual hold for the noun *grip*, and occurrences of both of these meanings are annotated with the label *grip_(handle)*).

There are also cases when Wikipedia makes different or finer sense distinctions than WordNet. For instance, there are several Wikipedia annotations for *image* as *copy*, but this meaning is not even defined in WordNet. Similarly, Wikipedia makes the distinction between *dance performance* and *theatre performance*, but both these meanings are listed under one single entry in WordNet (*performance* as *public presentation*). However, since at this stage we are mapping the Wikipedia annotations to WordNet, these differences in sense granularity are diminished.

7 Related Work

In word sense disambiguation, the line of work most closely related to ours consists of methods trying to address the sense-tagged data bottleneck problem.

A first set of methods consists of algorithms that generate sense annotated data using words semantically related to a given ambiguous word (Leacock et al., 1998; Mihalcea and Moldovan, 1999; Agirre and Martinez, 2004). Related non-ambiguous words, such as monosemous words or phrases from dictionary definitions, are used to automatically collect examples from the Web. These examples are then turned into sense-tagged data by replacing the non-ambiguous words with their ambiguous equivalents.

Another approach proposed in the past is based on the idea that an ambiguous word tends to have different translations in a second language (Resnik and Yarowsky, 1999). Starting with a collection of parallel texts, sense annotations were generated either for one word at a time (Ng et al., 2003; Diab, 2004), or for all words in unrestricted text (Diab and Resnik, 2002), and in both cases the systems trained on these data were found to be competitive with other word sense disambiguation systems.

The lack of sense-tagged corpora can also be circumvented using bootstrapping algorithms, which start with a few annotated seeds and iteratively generate a large set of disambiguation patterns. This method, initially proposed by (Yarowsky, 1995), was successfully evaluated in the context of the SENSEVAL framework (Mihalcea, 2002).

Finally, in an effort related to the Wikipedia collection process, (Chklovski and Mihalcea, 2002) have implemented the Open Mind Word Expert system for collecting sense annotations from volunteer contributors over the Web. The data generated using this method was then used by the systems participating in several of the SENSEVAL-3 tasks.

Notably, the method we propose has several advantages over these previous methods. First, our method relies exclusively on monolingual data, thus avoiding the possible constraints imposed by methods that require parallel texts, which may be difficult to find. Second, the Wikipedia-based annotations follow a natural Zipfian sense distribution, unlike the equal distributions typically obtained with the methods that rely on the use of monosemous relatives or bootstrapping methods. Finally, the grow pace of Wikipedia is much faster than other more task-focused and possibly less-engaging activities such as Open Mind Word Expert, and therefore has the potential to lead to significantly higher coverage.

With respect to the use of Wikipedia as a resource for natural language processing tasks, the work that is most closely related to ours is perhaps the name entity disambiguation algorithm proposed in (Bunescu and Pasca, 2006), where an SVM kernel is trained on the entries found in Wikipedia for ambiguous named entities. Other language processing tasks with recently proposed solutions relying on Wikipedia are co-reference resolution using Wikipedia-based measures of word similarity (Strube and Ponzetto, 2006), enhanced text classification using encyclopedic knowledge (Gabrilovich

SENSEVAL-3 words, an average accuracy of 71.31% was measured (the output of the systems participating in SENSEVAL is publicly available from http://www.senseval.org).

and Markovitch, 2006), and the construction of comparable corpora using the multilingual editions of Wikipedia (Adafre and de Rijke, 2006).

8 Conclusions

In this paper, we described an approach for using Wikipedia as a source of sense annotations for word sense disambiguation. Starting with the hyperlinks available in Wikipedia, we showed how we can generate a sense annotated corpus that can be used to train accurate sense classifiers. Through experiments performed on a subset of the SENSEVAL words, we showed that the Wikipedia sense annotations can be used to build a word sense disambiguation system leading to a relative error rate reduction of 30–44% as compared to simpler baselines.

Despite some limitations inherent to this approach (definitions and annotations in Wikipedia are available almost exclusively for nouns, word and sense distributions are sometime skewed, the annotation labels are occasionally inconsistent), these limitations are overcome by the clear advantage that comes with the use of Wikipedia: large sense tagged data for a large number of words at virtually no cost.

We believe that this approach is particularly promising for two main reasons. First, the size of Wikipedia is growing at a steady pace, which consequently means that the size of the sense tagged corpora that can be generated based on this resource is also continuously growing. While techniques for supervised word sense disambiguation have been repeatedly criticized in the past for their limited coverage, mainly due to the associated sense-tagged data bottleneck, Wikipedia seems a promising resource that could provide the much needed solution for this problem. Second, Wikipedia editions are available for many languages (currently about 200), which means that this method can be used to generate sense tagged corpora and build accurate word sense classifiers for a large number of languages.

References

S. F. Adafre and M. de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In *Proceedings of the EACL Workshop on New Text*, Trento, Italy.

E. Agirre and D. Martinez. 2004. Unsupervised word sense disambiguation based on automatically retrieved examples: The importance of bias. In *Proceedings of EMNLP 2004*, Barcelona, Spain, July.

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL 2006*, Trento, Italy.

T. Chklovski and R. Mihalcea. 2002. Building a sense tagged corpus with Open Mind Word Expert. In *Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions"*, Philadelphia, July.

M. Diab and P. Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of ACL 2002*, Philadelphia.

M. Diab. 2004. Relieving the data acquisition bottleneck in word sense disambiguation. In *Proceedings of ACL 2004*, Barcelona, Spain.

E. Gabrilovich and S. Markovitch. 2006. Overcoming the brittleness bottleneck using wikipedia: Enhancing text categorization with encyclopedic knowledge. In *Proceedings of AAAI 2006*, Boston.

M. Galley and K. McKeown. 2003. Improving word sense disambiguation in lexical chaining. In *Proceedings of IJCAI 2003*, Acapulco, Mexico.

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

C. Leacock, M. Chodorow, and G.A. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

Y.K. Lee and H.T. Ng. 2002. An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *Proceedings of EMNLP 2002*, Philadelphia.

M.E. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the SIGDOC Conference 1986*, Toronto, June.

R. Mihalcea and D.I. Moldovan. 1999. An automatic method for generating sense tagged corpora. In *Proceedings of AAAI 1999*, Orlando.

R. Mihalcea. 2002. Bootstrapping large sense tagged corpora. In *Proceedings of LREC 2002*, Canary Islands, Spain.

R. Navigli and P. Velardi. 2005. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 27.

H.T. Ng and H.B. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. In *Proceedings of ACL 1996*, New Mexico.

H.T. Ng, B. Wang, and Y.S. Chan. 2003. Exploiting parallel texts for word sense disambiguation: An empirical study. In *Proceedings of ACL 2003*, Sapporo, Japan.

T. Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of NAACL 2001*, Pittsburgh.

P. Resnik and D. Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–134.

M. Strube and S. P. Ponzetto. 2006. Wikirelate! computing semantic relatedeness using Wikipedia. In *Proceedings of AAAI 2006*, Boston.

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of ACL 1995*, Cambridge.