Sense Clustering Using Wikipedia

Bharath Dandala, Chris Hokamp and Rada Mihalcea
University of North Texas
bharathdandala@gmail.com
chris.hokamp@gmail.com
rada@cs.unt.edu

Razvan C. Bunescu
Ohio University
bunescu@ohio.edu

Abstract

In this paper, we propose a novel method for generating a coarse-grained sense inventory from Wikipedia using a machine learning framework. Structural and content-based features are employed to induce clusters of articles representative of a word sense. Additionally, multilingual features are shown to improve the clustering accuracy, especially for languages that are less comprehensive than English. We show the effectiveness of our clustering methodology by testing it against both manually and automatically annotated datasets.

1 Introduction

The granularity of word sense repositories has been recognized as an important factor in the development of annotated datasets for Word Sense Disambiguation (WSD) (Snow et al., 2007), with significant impacts upon both the performance of automatic WSD systems and their utility for downstream applications. Previous work on manual sense annotations with respect to WordNet has revealed low levels of agreement between human annotators, ranging between 65% (Chklovski and Mihalcea, 2002) and 72% (Snyder and Palmer, 2004), which is a clear indicator of very fine-grained word senses that are difficult to differentiate, even for humans.

To achieve the sense granularity appropriate for WSD, word senses that are closely related in meaning are grouped together in a sense clustering step. While this task was originally defined in relation to more traditional sense inventories, such as WordNet (Hovy et al., 2006; Mihalcea and Moldovan, 2001) or the Oxford dictionary (Navigli, 2006), newer user-contributed sense inventories such as Wikipedia or Wiktionary are also quickly expanding and refining the senses defined for a word, thus pointing to the need of sense clustering for coarser word sense distinctions.

In this paper, we specifically focus on the task of sense clustering over Wikipedia senses. Wikipedia has been recently recognized as a rich resource for WSD (Bunescu and Pasca, 2006; Mihalcea, 2007; Mihalcea and Csomai, 2007; Milne and Witten, 2008), offering a significantly increased coverage of word meanings relative to established repositories such as WordNet or Roget. At the same time, WSD systems using Wikipedia have been shown to obtain comparable or even increased disambiguation precision. While earlier work on WSD using the 2007 version of Wikipedia reported an average of three senses per word for a dataset of 30 nouns (Mihalcea, 2007), more recent work on the same dataset using the 2012 version of Wikipedia has shown a significant increase to an average of nine senses per word (Dandala et al., 2012). For instance, the noun “paper”, which used to have five different senses, now has ten senses; similarly, the noun “bar”, which previously had ten senses, now has 23 senses. The accuracy of a WSD system on the same set of 30 nouns dropped from an average of 85% when using Wikipedia 2007 to 62% when using Wikipedia 2012 (Dandala et al., 2012). Thus, the rapid growth of Wikipedia over the recent years has brought benefits, such as increased word and sense coverage, but it has also led to complications, such as finer sense granularity, resulting in a markedly reduced performance of WSD systems.

Related work on lexical resources, such as WordNet, has demonstrated the benefit of sense

Proceedings of Recent Advances in Natural Language Processing, pages 164–171,
Hissar, Bulgaria, 7-13 September 2013.

clustering. For example, work on mapping WordNet senses to the coarser Oxford dictionary (Navigli, 2006; Navigli et al., 2007) has resulted in improved WSD performance. The OntoNotes project, a large-scale effort to cluster and supplement word senses in WordNet in order to produce a high-quality dataset for automatic WSD (Hovy et al., 2006), has also been beneficial for other language processing tasks such as discourse analysis, coreference resolution, and semantic parsing. Coarser sense inventories also make it easier to identify synonyms or translations of selected words in context, which can lead to improvements in information retrieval (Zhong and Ng, 2012), semantic indexing (Gonzalo et al., 1998), and machine translation (Chan et al., 2007).

In this paper, we address two main research questions. First, can we build an accurate method to automatically cluster the fine-grained senses in Wikipedia? We describe a set of structural and content features that are integrated in a machine learning framework in order to automatically predict when two Wikipedia senses are close in meaning and should be clustered together. Second, can we use the multilingual links in Wikipedia to derive additional multilingual features to enhance this clustering? We rely upon the interlingua links in Wikipedia, and upon features that can be obtained from sense representations in other languages, in order to enrich the feature space and improve clustering accuracy.

In the following sections, we first briefly review Wikipedia as a large encyclopedic resource, focusing on the specific representation of word senses and groups of related word senses. We then introduce several novel datasets for sense clustering, which we use in our evaluations. Several structural and content features are described next, followed by a description of the experiments that we ran in order to evaluate the utility of these features. We conclude the paper with a discussion of the results and a presentation of related work.

2 Senses and Sense Clusters in Wikipedia

The basic entry in Wikipedia is an *article* (or, for the purpose of this paper, *word sense*[1]), which defines and describes a concept, 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. Articles are organized into *categories*, which in turn are organized into category hierarchies. For instance, the article on ALAN TURING is included in the category BRITISH CRYPTOGRAPHERS, which in turn has a parent category named BRITISH SCIENTISTS, and so forth.

Each article in Wikipedia is uniquely referenced by an identifier, consisting of one or more words separated by spaces or underscores, and occasionally a parenthetical explanation. For example, the article for the entity Turing that refers to the *"English computer scientist"* has the unique identifier ALAN TURING, whereas the article on Turing with the *"stream cipher"* meaning has the unique identifier TURING (CIPHER).

The disambiguation pages and the internal link graph of Wikipedia are a source of metadata, which can be exploited to transform the flat encyclopaedic format of Wikipedia into a rich Ontology. A structure that is particularly relevant to the work described in this paper is that of the *disambiguation pages*, which 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. SENSE_(DISAMBIGUATION), which is the unique identifier for the disambiguation page of the noun "sense". Disambiguation pages, if well-curated, can provide good clues about the set of senses defined in Wikipedia for a word, as well as the possible clusters over these senses, through the headings that group articles along named semantic axes generally corresponding to mid-level nodes in the Wikipedia category hierarchy.

Finally, also relevant for the work described in this paper are the *interlingual links*, which explicitly connect articles in different languages. For instance, the English article for the noun SENSE is connected, among others, to the Spanish article SENTIDO (PERCEPCIÓN) and the Latin article SENSUS (BIOLOGIA). On average, about half of the articles in any Wikipedia version include interlingual links to articles in other languages. The number of interlingual links per article varies

[1] The terms "article" and "word sense" are interchangeably used in this paper. Note that we are excluding articles that refer to named entities.

from an average of 5 in the English Wikipedia, to 10 in the Spanish Wikipedia, to 23 in the Arabic Wikipedia. Wikipedia editions are available for more than 280 languages, which vary widely in size. We use four of these Wikipedias in this work, namely the English, Spanish, German, and Italian versions.

3 Datasets for Sense Clustering

To evaluate our automatic sense clustering method, we build four datasets: two that are generated automatically through a set of heuristics applied on clusters extracted from existing disambiguation pages in English or Spanish, and two that are obtained through manual annotations. Additionally, we create a dataset obtained from clustering a set of Semeval word senses. All datasets follow the same format, and consist of pairs of articles annotated as either positive or negative, depending on whether they should be grouped together under one sense or not.

3.1 Automatically Extracted Datasets

We first create two large datasets using the clusters already available in some of the disambiguation pages in Wikipedia. We specifically selected only disambiguation pages that have at least five subheadings, a requirement that ensures that the word is polysemous and that also indicates that the disambiguation page is well-curated and likely to be trustworthy. After resolving redirects, we removed any duplicate senses. We then removed those senses that have less than three mentions in Wikipedia. Finally, since one of our goals is to experiment with multilingual features, we also removed senses that do not exist in all four target languages.

From the set of disambiguation pages obtained after applying all of these heuristics, we generate a dataset as follows: all of the senses that are listed under the same subheading (except for the OTHER, SEE ALSO, and MISCELLANEOUS headings) are used to create pairs of senses that are labeled as positive (i.e., they should be clustered together). All of the senses that are listed under different headings, while still on the same disambiguation page, are used to create pairs of senses that are labeled as negative (i.e., they should not be clustered). From the resulting list of pairs, we first exclude all named entities, since our work is primarily concerned with word sense clustering rather than named entity clustering. Additionally, the groupings of the named entities in the Wikipedia disambiguation pages are too coarse; for instance, in the disambiguation page for "Newton," the articles "Isaac Newton" and "Newton (surname)" are listed under the same heading "People." As mentioned above, we exclude those senses that do not have interlingua links with the other three languages of interest (i.e., a word sense in our dataset has to be represented in all languages English, Spanish, German, Italian). This constraint is applied so that we have a complete multilingual representation for our dataset, which allows us to test our hypothesis concerning the usefulness of multilingual features.

Using this approach, we automatically create two datasets, one for English and one for Spanish. Starting with the English Wikipedia disambiguation pages, from all the sense pairs obtained using the heuristics above, we randomly select a set of 3,000 positive examples and their corresponding 3,106 negative examples extracted from the same disambiguation pages, for a total of 6,106 examples.

We then use the same strategy to automatically extract a Spanish sense clustering dataset, this time starting with the Spanish Wikipedia disambiguation pages. Here, we obtain 3,270 positive examples and their corresponding 1,730 negative examples, for a total of 5,000 examples. Our goal with this second dataset is to determine to what extent the sense clustering method can be effectively applied to a language that has fewer articles and contributors than to the English Wikipedia.

3.2 Manually Annotated Datasets

We also create two smaller datasets of 500 examples each, again for English and Spanish, which were manually annotated. The sense pairs (250 positive and 250 negative pairs) were uniformly sampled from sense clusters obtained using the same automatic method described above, excluding the sense pairs that were included in the automatically created datasets. In other words, there is no overlap between the 500 sense pairs in the manually annotated datasets, and the 6,106 (5,000) sense pairs in the automatically created datasets. Annotators were asked to determine whether each pair used the same sense of the target word, or different senses. To help them in this task, an interface was created so that annotators could view each pair of pages side-by-side, in order to decide whether the pair was a positive or a negative example of senses that could be clustered together. Annotators were also given an unknown option to use in cases where they were unsure whether to label a pair as positive or negative.

Two annotators independently labeled the 500 pairs in each of the datasets. The pairwise Pearson correlation between the two annotators was measured at 0.77 and 0.83 for English and Spanish respectively, which represents a high agreement. All disagreements between annotators were resolved through adjudication by a third annotator. The final label distribution was 254 positive pairs and 246 negative pairs in the English dataset, and 212 positive pairs and 288 negative pairs in the Spanish dataset.

3.3 Semeval Dataset

Finally, we also create a dataset using a set of highly ambiguous nouns drawn from the Semeval evaluations, which was previously used in WSD experiments on Wikipedia (Mihalcea, 2007). As before, the sense pairs were labeled as either positive or negative, which resulted in 763 sense pairs marked as negative and 162 sense pairs labeled as positive, for a total of 925 examples. This dataset is built to test our system in a more realistic setting that does not follow all the constraints that we used during the construction of the manually annotated datasets. The only constraint that we placed on this dataset is the removal of named entities, for the reasons outlined above.

4 Structural and Content Features for Sense Clustering

To characterize the similarity of two word senses, we extract two types of features: *structural features*, which exploit the link structure of Wikipedia articles, and *content features* that capture vector space similarities between articles or lexical contexts. We obtain a total of 13 features for each pair of articles in each language.

4.1 Structural Features

Two well-established metrics are used to measure the similarity between the link structures of the senses in each pair. For each pair of articles, we derive four graph-based similarity features using Pointwise Mutual Information (PMI) and Google Similarity Distance (GSD) (Cilibrasi and Vitanyi, 2007). PMI and GSD features are calculated between the sets of outgoing links and between the sets of incoming links. Thus, there are four features that indicate the similarity between the sets of pages that link to the articles, and the sets of pages that are linked to by the articles. These features exploit the link structure of Wikipedia to measure the pages' relative positions in the link graph.

Two features are added to indicate whether the articles have direct links to each other. The first takes a value of 1 if both articles have a link to each other in the first paragraph, and a value of 0.5 if one of the articles links to the other in the first paragraph (0 otherwise). The second feature extends the context to the entire articles, using the same values to indicate whether one or both of the articles contain a direct link to the other anywhere on the page.

One feature is also included to indicate whether an article's template uses the {{main * <other_article>}} syntax to point to the other article in the pair. The weighting of this feature is the same as that of the direct link features.[2]

Since links between pages are very common in Wikipedia, structural features can provide a good measure of the semantic closeness of two articles, and since our data only contains pairs of articles that are potential disambiguations of a certain word, two articles that have similar link structures are likely to be good candidates for clustering.

4.2 Content Features

The ubiquitous *tf.idf* method for measuring content similarity is used to obtain four additional features. For each article in each language, we created two tf.idf indexes: one for the actual content, and one for the aggregated context of all the inlinks to the page. To construct the aggregated inlink context, the sentences containing a link to the article are globbed into one index, representative of the contexts in which this sense is used across the encyclopedia. Obtaining tf.idf scores for the articles required construction of a global Inverse Document Frequency (idf) index for each language, which was accomplished using Hadoop[3] and Apache Pig.[4] For each pair of senses, we generate four tf.idf features using each possible com-

[2] Note that it is unlikely, though not impossible, that each article could point to the other as its main article

[3] http://hadoop.apache.org/

[4] http://pig.apache.org/

bination of the indexes.

We also use the Stanford Dependency Parser(De Marneffe et al.,) to extract the head noun from each article's title, adding a binary feature that indicates whether the article titles share the same head noun.

Finally, we add a feature for the cosine similarity between the labels for each page. The set of labels for a page is obtained from the anchor text of all inlinks to the page across Wikipedia versions. We remove all occurrences of the target word from the list of labels to prevent unintended bias. For example, if the word in question is "bar" we remove the label "bar". When we move across languages to calculate this feature, the target word is obtained using Google Translate.[5] This set of keywords represents all possible labels for the particular article, and forms a "bag of labels" for that article, to be used in the calculation of the cosine similarity.

4.3 Multilingual Features

The intuition that multilingual features may improve the accuracy of sense clustering is a major inspiration for this work. With this in mind, we calculate the same set of features for the parallel sense pairs in all four languages. This allows evaluation of each language's contribution to the result of sense clustering in a particular language. We do not average the features across languages by creating a centroid vector, preferring instead to append features as languages are added.

5 Experiments and Evaluations

The WEKA toolkit (Witten and Frank, 2005) was used for all experiments. The classifiers were trained using the SMO implementation of Support Vector Machines provided by WEKA, with a quadratic kernel.

5.1 Evaluation on the Automatically Extracted Datasets

In the first experiment, we use the automatically extracted datasets to evaluate the accuracy of the sense clustering classifier, as well as the role of the multilingual features in this classification. We perform cross-validation on the automatically extracted datasets. We use the English and Spanish datasets described in Section 3.1, which include positive and negative examples of sense pairs along with their corresponding senses in three other languages. For each sense pair, and for each language, we generate the structural and content features described above.

Language(s)	Acc.	Avg. Acc.
Monolingual(English)	84.5%	
English+German	92.0%	92.5%
English+Italian	**93.2%**	
English+Spanish	92.3%	
English+Spanish+German	**93.8%**	93.03%
English+Spanish+Italian	93.2%	
English+German+Italian	92.1%	
English+Spanish+German+Italian	93.6%	**93.6%**

Table 1: Classification accuracy on the automatically extracted English dataset.

Tables 1 and 2 show the results obtained during these experiments, using one, two, three, or four languages at a time. The results indicate that sense clustering can be effectively performed, and the performance improves consistently as more languages are added. The overall improvements are significant over the most frequent class baseline of 50.8% for English and 65.4% for Spanish.

Language(s)	Acc.	Avg. Acc.
Monolingual (Spanish)	68.3%	
Spanish+English	**74.0%**	73.0%
Spanish+German	73.8%	
Spanish+Italian	71.1%	
Spanish+German+Italian	**75.7%**	75.5%
Spanish+Italian+English	75.5%	
Spanish+German+English	75.4%	
Spanish+English+German+Italian	**76.2%**	**76.2%**

Table 2: Classification accuracy on the automatically generated Spanish dataset.

5.2 Evaluation on Manually Created Datasets

We also perform evaluations on the English and Spanish manually annotated datasets, described in Section 3.2. Here, we use the automatically generated datasets to train the sense clustering classifiers, which we then test on the manually labeled data. Tables 3 and 4 show the results obtained in these experiments, again for one, two, three, and four languages at a time.

As before, the sense clustering classifiers improve over the most frequent class baseline of

[5] http://translate.google.com/

Language(s)	Acc.	Avg. Acc.
Monolingual(English)	77.4%	
English+Spanish	85.6%	85.1%
English+German	84.8%	
Spanish+Italian	**85.4%**	
English+German+Italian	**86.0%**	85.2%
English+Italian+Spanish	84.4%	
English+German+Spanish	85.4%	
Spanish+English+German+Italian	**84.4%**	**84.4%**

Table 3: Classification accuracy on manually annotated English dataset.

Language(s)	Acc.	Avg. Acc.
Monolingual(Spanish)	83.7%	
Spanish+English	88.4%	88.7%
Spanish+German	87.1%	
Spanish+Italian	**90.5%**	
Spanish+German+Italian	89.6%	90.9%
Spanish+Italian+English	**92.2%**	
Spanish+German+English	90.9%	
Spanish+English+German+Italian	**95.6%**	**95.6%**

Table 4: Classification accuracy on manually annotated Spanish dataset.

50.8% on the English dataset and 57.6% on the Spanish dataset,[6] and the inclusion of features drawn from additional languages improves the performance of the monolingual classifier significantly.

5.3 Evaluation on Semeval Dataset

The final evaluation is performed on the sense clusters derived from the set of 30 Semeval nouns, as described in Section 3.3. The most frequent class baseline for this dataset is 82.5%, obtained by assigning by default a negative label to all the sense pairs in the dataset. Using the automatically labeled data for training, the monolingual classifier yields an accuracy of 83.5%, and improves to 85.5% when the multilingual features are added. For this dataset, which includes highly ambiguous words and follows a more realistic distribution of positive versus negative sense pairs, the distribution is very skewed, so we also calculate the ROC area, measured at 76.6 for the monolingual classifier, and 79.1 for the multilingual classifier.

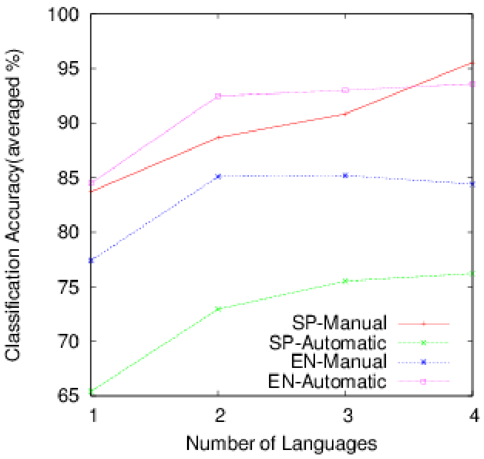

Figure 1: Using automatically and manually created English and Spanish datasets, how the sense clusters benefit from incorporating more languages

6 Discussion

The monolingual sense clustering algorithm leads to significant improvements over the most frequent class baseline, with error rate reductions of 68.5% and 8.3% obtained in the evaluations on the automatically created datasets for English and Spanish respectively, and 54.8% and 67.4%, obtained from the evaluations on the manually-created English and Spanish datasets. On the Semeval dataset, we obtained an error rate reduction of 5.7%.

An even more important result is the role played by the multilingual features in improving the sense clustering method. The incremental addition of new languages leads to steady increases in clustering accuracy. The highest accuracy is obtained when features drawn from all four languages are used, with the following error rate reductions from with the multilingual classifier relative to the monolingual classifier: 58.7% for the English automatic dataset; 24.9% for the Spanish automatic dataset; 30.9% for the English manual dataset; 73.0% for the Spanish manual dataset; and 12.1% for the Semeval dataset. To illustrate the effect of adding more languages graphically, Figure 1 shows how the performance of the Spanish sense clustering benefits from the addition of multilingual features.

The improved performance observed for all possible language groupings is good evidence that the clustering improves consistently as features from a language are supplemented with features

[6] These baselines are obtained from the distribution of positive and negative examples in the manual annotation of these datasets.

from other languages. Even for English, which is a major language with significant resources, we observe improvements when multilingual features are added.These results support our hypothesis that multilingual features can improve the accuracy of sense clustering, even in a more realistic setting where we do not have corresponding sense pairs in all languages. In such cases, when trying to cluster a sense pair from e.g. Spanish, even if features from a more resourceful language such as English are not available, the feature space can still be adjusted with sense pairs from other languages such as German or Italian.

7 Related Work

A large number of techniques have been proposed for clustering the collection of fine-grained senses available in WordNet. One of the early approaches was the automatic system of (Peters et al., 1998), in which two senses are clustered together based on a set of relational cues extracted from WordNet. (Mihalcea and Moldovan, 2001) extend the collection of WordNet relational features and propose a set of semantic and probabilistic rules for either collapsing synsets very similar in meaning or removing synsets that are very rarely used. (McCarthy, 2006) defines vector profiles for WordNet senses based on *neighboring words*, where the distributional similarity between neighbors is computed from statistics over grammatical relations extracted from the British National Corpus corpus. Similarity between two senses is then computed as the Spearman rank correlation of their corresponding vector profiles. The OntoNotes project (Hovy et al., 2006) uses a corpus-based iterative approach for sense clustering in which a sample of 50 sentences is annotated with a preliminary set of coarse senses. If the inter-annotator agreement is too low, the sense clusters are revised, and the annotation process is repeated until the agreement passes 90%. Also related is the work of (Navigli, 2006), who generates coarse senses over WordNet by mapping the WordNet senses into the more coarse-grained Oxford dictionary.

Similar to our approach, (Snow et al., 2007) train an SVM classifier to make binary "merge" vs. "not-merge" decisions. Their WordNet sense pairs are represented using a diverse set of features derived from WordNet structure, corpus-based evidence, and other lexical resources. Furthermore, the binary sense merging classifier is integrated into a model for sense clustering that takes into account taxonomic constraints that arise when merging senses in a hierarchical structures.

Another closely related work is that of (Pedersen et al., 2005), which describes an unsupervised method for discriminating ambiguous names by clustering contexts, and relies upon features found in corpora obtained for a language with more resources.

The major aim of the coarse-grained all-words WSD task at Semeval-2007 was to determine whether a more accurate WSD system can enable sense-aware applications, such as information retrieval, question answering, or machine translation.

Finally, in recent work, Erk and McCarthy (Erk and McCarthy, 2009) also considered the sense granularity issue, and introduced the idea of graded WSD, in which they relax the single sense assignment and allow for multiple sense assignments for a particular target word.

8 Conclusion

Wikipedia's sense inventory is constantly growing, and the sense distinctions in this inventory are becoming finer-grained, which means that robust methods for sense clustering are needed in order to maintain its usefulness for WSD. In this paper, we described an approach to automatically cluster senses in Wikipedia using data obtained from disambiguation pages, utilizing the multilingual data available in Wikipedia to create a rich feature space for sense clustering.

The automatic sense clustering method significantly outperforms the most frequent baseline, and these results are consistent for several datasets and several languages. Moreover, the integration of multilingual information into the clustering method was found to improve significantly over the monolingual models, with consistent improvements as features from new languages are added. Wikipedia editions are available for a large number of languages, which means that this method can be used to generate sense hierarchies and build accurate word sense clustering classifiers for many languages, even in cases where the disambiguation pages for a particular language are not well-curated.

The sense clustering datasets created during this work are publicly available at http://lit.csci.unt.edu

Acknowledgments

This material is based in part upon work supported by the National Science Foundation IIS awards #1018613 and #1018590 and CAREER award #0747340. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

R. Bunescu and M. Pasca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the European Conference of the Association for Computational Linguistics*, Italy.

Y.S. Chan, H.T. Ng, and D. Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the Association for Computational Linguistics*, Prague, Czech Republic.

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.

Rudi L Cilibrasi and Paul MB Vitanyi. 2007. The google similarity distance. *Knowledge and Data Engineering, IEEE Transactions on*, 19(3):370–383.

B. Dandala, R. Mihalcea, and R. Bunescu. 2012. Towards building a multilingual semantic network: Identifying interlingual links in wikipedia. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, Montreal, Canada.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. Generating typed dependency parses from phrase structure parses.

K. Erk and D. McCarthy. 2009. Graded word sense assignment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

J. Gonzalo, F. Verdejo, I. Chugur, and J. Cigarran. 1998. Indexing with WordNet synsets can improve text retrieval. In *Proceedings of COLING-ACL '98 Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, Canada, August.

E. Hovy, M. Marcus, M. Palmer, L. Ramshaw, and R. Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, New York City.

D. McCarthy. 2006. Relating wordnet senses for word sense disambiguation. In *Proceedings of ACL Workshop on Making Sense of Sense*.

R. Mihalcea and A. Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, Lisbon, Portugal.

R. Mihalcea and D. Moldovan. 2001. Automatic generation of a coarse grained wordnet. In *NAACL 2001 Workshop on WordNet and Other Lexical Resources: applications, extensions and customizations*, pages 35–41, Pittsburgh, June.

R. Mihalcea. 2007. Using Wikipedia for automatic word sense disambiguation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, New York, April.

D. Milne and I. Witten. 2008. Learning to link with Wikipedia. In *Proceedings of the Seventeenth ACM Conference on Information and Knowledge Management*.

R. Navigli, K. Litkowski, and O. Hargraves. 2007. Semeval-2007 task 07: Coarse-grained english all-words task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, Czech Republic, June.

R. Navigli. 2006. Meaningful clustering of senses helps boost word sense disambiguation performance. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia.

T. Pedersen, A. Purandare, and A. Kulkarni. 2005. Name discrimination by clustering similar contexts. In *Proceedings of the Conference on Computational Linguistics and Intelligent Text Processing*. Springer.

W. Peters, I. Peters, and P. Vossen. 1998. Automatic sense clustering in EuroWordNet. In *Proceedings of the First International Conference on Language Resources and Evaluation*, pages 409–416, Granada.

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

B. Snyder and M. Palmer. 2004. The English all-words task. In *Proceedings of ACL/SIGLEX Senseval-3*, Spain.

I. Witten and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Z. Zhong and H. T. Ng. 2012. Word sense disambiguation improves information retrieval. In *Proceedings of the Association for Computational Linguistics*, Jeju Island, Korea.