Word Sense Disambiguation for Cross-Language Information Retrieval

Mary Xiaoyong Liu, Ted Diamond, and **Anne R. Diekema**

School of Information Studies
Syracuse University
Syracuse, NY 13244
xliu03@mailbox.syr.edu
tdiamon1@twcny.rr.com
diekemar@mailbox.syr.edu

Abstract

We have developed a word sense disambiguation algorithm, following Cheng and Wilensky (1997), to disambiguate among WordNet synsets. This algorithm is to be used in a cross-language information retrieval system, CINDOR, which indexes queries and documents in a language-neutral concept representation based on WordNet synsets. Our goal is to improve retrieval precision through word sense disambiguation. An evaluation against human disambiguation judgements suggests promise for our approach.

1 Introduction

The CINDOR cross-language information retrieval system (Diekema et al., 1998) uses an information structure known as "conceptual interlingua" for query and document representation. This conceptual interlingua is a hierarchically organized multilingual concept lexicon, which is structured following WordNet (Miller, 1990). By representing query and document terms by their WordNet synset numbers we arrive at essentially a language neutral representation consisting of synset numbers representing concepts. This representation facilitates cross-language retrieval by matching term synonyms in English as well as across languages. However, many terms are polysemous and belong to multiple synsets, resulting in spurious matches in retrieval. The noun *figure* for example appears in 13 synsets in WordNet 1.6. This research paper describes the early stages[1] of our efforts to develop a word sense disambiguation (WSD) algorithm aimed at improving the precision of our cross-language retrieval system.

2 Related Work

To determine the sense of a word, a WSD algorithm typically uses the context of the ambiguous word, external resources such as machine-readable dictionaries, or a combination of both. Although dictionaries provide useful word sense information and thesauri provide additional information about relationships between words, they lack pragmatic information as can be found in corpora. Corpora contain examples of words that enable the development of statistical models of word senses and their contexts (Ide and Veronis, 1998; Leacock and Chodorow, 1998).

There are two general problems with using corpora however; 1) corpora typically do not come pre-tagged with manually disambiguated senses, and 2) corpora are often not large nor diverse enough for all senses of a word to appear often enough for reliable statistical models (data sparseness). Although researchers have tried sense-tagging corpora automatically by using either supervised or unsupervised training methods, we have adopted a WSD algorithm which avoids the necessity for a sense-tagged training corpus.

[1] Please note that the disambiguation research described in this paper has not yet been extended to multiple language areas.

$$P(\text{synset}|\text{context}(w)) = \frac{P(\text{context}(w) \mid \text{synset})\, P(\text{synset})}{P(\text{context}(w))} \quad (1)$$

The problem of data sparseness is usually solved by using either smoothing methods, class-based methods, or by relying on similarity-based methods between words and co-occurrence data. Since we are using a WordNet-based resource for retrieval, using class-based methods seems a natural choice. Appropriate word classes can be formed by synsets or groups of synsets. The evidence of a certain sense (synset) is then no longer dependent on one word but on all the members of a particular synset.

Yarowsky (1992) used Rogets Thesaurus categories as classes for WSD. His approach was based on selecting the most likely Roget category for nouns given their context of 50 words on either side. When any of the category indicator words appeared in the context of an ambiguous word, the indicator weights for each category were summed to determine the most likely category. The category with the largest sum was then selected.

A similar approach to that of Yarowsky was followed by Cheng and Willensky (1997) who used a training matrix of associations of words with a certain category. Their algorithm was appealing to us because it requires no human intervention, and more importantly, it avoids the use of sense-tagged data. Our methodology described in the next section is therefore based on Cheng and Wilensky's approach.

Methods to reduce (translation) ambiguity in cross-language information retrieval have included using part-of-speech taggers to restrict the translation options (Davis 1997), applying pseudo-relevance feedback loops to expand the query with better terms aiding translation (Ballesteros and Croft 1997), using corpora for term translation disambiguation (Ballesteros and Croft, 1998), and weighted Boolean models which tend to have a self-disambiguating quality (Hull, 1997; Diekema et al., 1999; Hiemstra and Kraaij, 1999).

3 Methodology

To disambiguate a given word, we would like to know the probability that a sense occurs in a given context, i.e., P(sense|context). In this study, WordNet synsets are used to represent word senses, so P(sense|context) can be rewritten as P(synset|context), for each synset of which that word is a member. For nouns, we define the context of word w to be the occurrence of words in a moving window of 100 words (50 words on each side) around w[2].

By Bayes Theorem, we can obtain the desired probability by inversion (see equation (1)). Since we are not specifically concerned with getting accurate probabilities but rather relative rank order for sense selection, we ignore P(context(w)) and focus on estimating P(context(w)|synset)P(synset). The event space from which "context(w)" is drawn is the set of sets of words that ever appear with each other in the window around w. In other words, w induces a partition on the set of words. We define "context(w)" to be true whenever any of the words in the set appears in the window around w, and conversely to be false whenever none of the words in the set appears around w. If we assume independence of appearance of any two words in a given context, then we get:

$$P(\text{synset}) \times \left(1 - \prod_{w_i \in context} (1 - P(w_i \mid \text{synset}))\right) \quad (2)$$

Due to the lack of sense-tagged corpora, we are not able to directly estimate P(synset) and $P(w_i|\text{synset})$. Instead, we introduce "noisy estimators" ($P_e(\text{synset})$ and $P_e(w_i|\text{synset})$) to approximate these probabilities. In doing so, we make two assumptions: 1) The presence of any word w_k that belongs to synset s_i signals the presence of s_i; 2) Any word w_k belongs to all its synsets simultaneously, and with equal probability. Although the assumptions underlying the "noisy estimators" are not strictly true, it is our belief that the "noisy estimators" should work reasonably well if:

- The words that belong to synset s_i tend to appear in similar contexts when s_i is their intended sense;
- These words do not completely overlap with the words belonging to some synset s_j ($i \neq j$) that partially overlaps with s_i;

[2] For other parts of speech, the window size should be much smaller as suggested by previous research.

- The common words between s_i and s_j appear in different contexts when s_i and s_j are their intended senses.

4 The WSD Algorithm

We chose as a basis the algorithms described by Yarrowsky (1992) and by Cheng and Wilensky (1997). In our variation, we use the synset numbers in WordNet to represent the senses of a word. Our algorithm learns associations of WordNet synsets with words in a surrounding context to determine a word sense. It consists of two phases.

During the training phase, the algorithm reads in all training documents in collection and computes the distance-adjusted weight of co-occurrence of each word with each corresponding synset. This is done by establishing a 100-word window around a target word (50 words on each side), and correlating each synset to which the target word belongs with each word in the surrounding window. The result of the training phase is a matrix of associations of words with synsets.

In the sense prediction phase, the algorithm takes as input randomly selected testing documents or sentences that contain the polysemous words we want to disambiguate and exploits the context vectors built in the training phase by adding up the weighted "votes". It then returns a ranked list of probability values associated with each synset, and chooses the synset with the highest probability as the sense of the ambiguous word.

Figure 1 and Figure 2 show an outline of the algorithm.

In this algorithm, "noisy estimators" are employed in the sense prediction phase. They are calculated using following formulas:

$$P_e(w_i|x) = \frac{M[w_i][x]}{\sum_{w \in W} M[w][x]} \quad (3)$$

where w_i is a stem, x is a given synset, M[w][x] is a cell in the correlation matrix that corresponds to word w and synset x, and

$$P_e(x) = \frac{\sum_{w \in W} M[w][x]}{\sum_{w \in W, y \in Y} M[w][y]} \quad (4)$$

where w is any stem in the collection, x is a given synset, y is any synset ever occurred in collection.

```
For each document d in collection
        read in a noun stem w from d
        for each synset s in which w occurs
                get the column b in the association matrix M that corresponds to s if the column already
                                        exists; create a new column for s otherwise
                for each word stem j appearing in the 100-word window around w
                        get the row a in M that corresponds to j if the row already exists; create a new
                                        row for j otherwise
                        add a distance-adjusted weight to M[a][b]
```

Figure 1: WSD Algorithm: the training phase

```
Set value = 1
For each word w to be disambiguated
        get synsets of w
        for each synset x of w
                for each w_i in the context of w (within the 100-window around w)
                        calculate P_e(w_i|x)
                          value *= ( 1 - P_e(w_i|x))
                 P(context(w)|x) = 1 - value
                Calculate p_e(x)
                P(x|context(w)) = p_e(x)* P(context(w)|x)
        display a ranked list of the synsets arranged according to their P(x|context(w)) in decreasing
order
```

Figure 2: WSD Algorithm: the sense prediction phase

5 Evaluation

As suggested by the WSD literature, evaluation of word sense disambiguation systems is not yet standardized (Resnik and Yarowsky, 1997). Some WSD evaluations have been done using the Brown Corpus as training and testing resources and comparing the results against SemCor[3], the sense-tagged version of the Brown Corpus (Agirre and Rigau, 1996; Gonzalo et al., 1998). Others have used common test suites such as the 2094-word *line* data of Leacock et al. (1993). Still others have tended to use their own metrics. We chose an evaluation with a user-based component that allowed a ranked list of sense selection for each target word and enabled a comprehensive comparison between automatic and manual WSD results. In addition we wanted to base the disambiguation matrix on a corpus that we use for retrieval. This approach allows for a much richer evaluation than a simple hit-or-miss test. For validation purpose, we will conduct a fully automatic evaluation against SemCor in our future efforts.

We use *in vitro* evaluation in this study, i.e. the WSD algorithm is tested independent of the retrieval system. The population consists of all the nouns in WordNet, after removal of monosemous nouns, and after removal of a problematic class of polysemous nouns.[4] We drew a random sample of 87 polysemous nouns[5] from this population.

In preparation, for each noun in our sample we identified all the documents containing that noun from the Associated Press (AP) newspaper corpus. The testing document set was then formed by randomly selecting 10 documents from the set of identified documents for each of the 87 nouns. In total, there are 867 documents in the testing set. The training document set consists of all the documents in the AP corpus excluding the above-mentioned 867 documents. For each noun in our sample, we selected all its corresponding WordNet noun synsets and randomly selected 10 sentence occurrences with each from one of the 10 random documents.

After collecting 87 polysemous nouns with 10 noun sentences each, we had 870 sentences for disambiguation. Four human judges were randomly assigned to two groups with two judges each, and each judge was asked to disambiguate 275 word occurrences out of which 160 were unique and 115 were shared with the other judge in the same group. For each word occurrence, the judge put the target word's possible senses in rank order according to their appropriateness given the context (ties are allowed).

Our WSD algorithm was also fed with the identical set of 870 word occurrences in the sense prediction phase and produced a ranked list of senses for each word occurrence.

Since our study has a matched-group design in which the subjects (word occurrences) receive both the treatments and control, the measurement of variables is on an ordinal scale, and there is no apparently applicable parametric statistical procedure available, two nonparametric procedures -the Friedman two-way analysis of variance and the Spearman rank correlation coefficient -were originally chosen as candidates for the statistical analysis of our results. However, the number of ties in our results renders the Spearman coefficient unreliable. We have therefore concentrated on the Friedman analysis of our experimental results. We use the two-alternative test with α=0.05.

The first tests of interest were aimed at establishing inter-judge reliability across the 115 shared sentences by each pair of judges. The null hypothesis can be generalized as "There is no difference in judgments on the same word occurrences between two judges in the same group". Following general steps of conducting a Friedman test as described by Siegel (1956), we cast raw ranks in a two-way table having 2 conditions/columns (K = 2) with each of the human judges in the pair serving as one condition and 365 subjects/rows (N = 365) which are all the senses of the 115 word occurrences that were judged by both human judges. We then ranked

---

[3] SemCor is a semantically sense-tagged corpus comprising approximately 250, 000 words. The reported error rate is around 10% for polysemous words.

[4] This class of nouns refers to nouns that are in synsets in which they are the sole word, or in synsets whose words were subsets of other synsets for that noun. This situation makes disambiguation extremely problematic. This class of noun will be dealt with in a future version of our algorithm but for now it is beyond the scope of this evaluation.

[5] A polysemous noun is defined as a noun that belongs to two or more synsets.

	N	K	X_r^2	df	Rejection region	Reject H_0?
First pair of judges	365	2	.003	1	3.84	No
Second pair of judges	380	2	2.5289	1	3.84	No

Figure 3: Statistics for significance tests of inter-judge reliability (α=.05, 2-alt. Test)

	N	K	X_r^2	df	Rejection region	Reject H_0?
Auto WSD vs man. WSD vs sense pooling	2840	3	73.217	2	5.99	Yes
Auto WSD vs man. WSD	2840	2	3.7356	1	3.84	No
Auto WSD vs sense pooling	2840	2	5.9507	1	3.84	Yes
Man. WSD vs sense pooling	2840	2	126.338	1	3.84	Yes

Figure 4: Statistics for significance tests among automatic WSD, manual WSD, and sense pooling (α=.05, 2-alt. Test)

the scores in each row from 1 to K (in this case K is 2), summed the derived ranks in each column, and calculated X_r^2 which is .003. For α=0.05, degrees of freedom df = 1 (df = K –1), the rejection region starts at 3.84. Since .003 is smaller than 3.84, the null hypothesis is not rejected. Similar steps were used for analyzing reliability between the second pair of judges. In both cases, we did not find significant difference between judges (see Figure 3).

Our second area of interest was the comparison of automatic WSD, manual WSD, and "sense pooling". Sense pooling equates to no disambiguation, where each sense of a word is considered equally likely (a tie). The null hypothesis (H_0) is "There is no difference among manual WSD, automatic WSD, and sense pooling (all the conditions come from the same population)". The steps for Friedman analysis were similar to what we did for the inter-judge reliability test while the conditions and subjects were changed in each test according to what we would like to compare. Test results are summarized in Figure 4. In the three-way comparison shown in the first row of the table, we rejected H_0 so there was at least one condition that was from a different population. By further conducting tests which examined each two of the above three conditions at a time we found that it was sense pooling that came from a different population while manual and automatic WSD were not significantly different. We can therefore conclude that our WSD algorithm is better than no disambiguation.

6 Concluding Remarks

The ambiguity of words may negatively impact the retrieval performance of a concept-based information retrieval system like CINDOR. We have developed a WSD algorithm that uses all the words in a WordNet synset as evidence of a given sense and builds an association matrix to learn the co-occurrence between words and senses. An evaluation of our algorithm against human judgements of a small sample of nouns demonstrated no significant difference between our automatic ranking of senses and the human judgements. There was, however, a significant difference between human judgement and rankings produced with no disambiguation where all senses were tied.

These early results are such as to encourage us to continue our research in this area. In our future work we must tackle issues associated with the fine granularity of some WordNet sense distinctions, synsets which are proper subsets of other synsets and are therefore impossible to distinguish, and also extend our evaluation to multiple languages and to other parts of speech. The next step in our work will be to evaluate our WSD algorithm against the manually sense-tagged SemCor Corpus for validation, and then integrate our WSD algorithm into CINDOR's processing and evaluate directly the impact on retrieval performance. We hope to verify that word sense disambiguation leads to improved precision in cross-language retrieval.

Acknowledgements

This work was completed under a research practicum at MNIS-TextWise Labs, Syracuse, NY. We thank Paraic Sheridan for many useful discussions and the anonymous reviewers for constructive comments on the manuscript.

References

Agirre, E., and Rigau, G. (1996). Word sense disambiguation using conceptual density. In: *Proceedings of the 16th International Conference on Computational Linguistics*, Copenhagen,1996.

Ballesteros, L., and Croft, B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-Language Information Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 20th International Conference on Research and Development in Information Retrieval*; 1997 July 25-31; Philadelphia, PA. New York, NY: ACM, 1997. 84-91.

Ballesteros, L., and Croft, B. (1998). Resolving Ambiguity for Cross-language Retrieval. In: *Proceedings of the Association for Computing Machinery Special Interest Group on Information Retrieval (ACM/SIGIR) 21st International Conference on Research and Development in Information Retrieval*; 1998 August 24-28; Melbourne, Australia. New York, NY: ACM, 1998. 64-71.

Cheng, I., and Wilensky, R. (1997). An Experiment in Enhancing Information Access by Natural Language Processing. UC Berkeley Computer Science Technical Report UCB/CSD UCB//CSD-97-963.

Davis, M. (1997). New Experiments in Cross-Language Text Retrieval at NMSU's Computing Research Lab. In: D.K. Harman, Ed. *The Fifth Text Retrieval Conference (TREC-5).* 1996, November. National Institute of Standards and Technology (NIST), Gaithersburg, MD.

Diekema, A., Oroumchian, F., Sheridan, P., and Liddy, E. D. (1999). TREC-7 Evaluation of Conceptual Interlingua Document Retrieval (CINDOR) in English and French. In: E.M. Voorhees and D.K. Harman (Eds.) *The Seventh Text REtrieval Conference (TREC-7).* 1998, November 9-11; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 169-180.

Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J. (1998). Indexing with WordNet synsets can improve text retrieval. In: *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*, Montreal, 1998.

Hiemstra, D., and Kraaij, W. (1999). Twenty-One at TREC-7: Ad-hoc and Cross-language Track. In: E.M. Voorhees and D.K. Harman (Eds.) *The Seventh Text REtrieval Conference (TREC-7).* 1998, November 9-11; National Institute of Standards and Technology (NIST), Gaithersburg, MD. 227-238.

Hull, D. A. (1997). Using Structured Queries for Disambiguation in Cross-Language Information Retrieval. In: *American Association for Artificial Intelligence (AAAI) Symposium on Cross-Language Text and Speech Retrieval*; 1997 March 24-26; Palo Alto, CA 1997. 84-98.

Ide, N., and Veronis, J. (1998). Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, Vol. 24, No. 1, 1-40.

Leacock, C., and Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In: Christiane Fellbaum (Eds.) *WordNet: An Electronic Lexical Database.* Cambridge, MA: MIT Press.

Leacock, C., Towell, G., and Voorhees, E. (1993). Corpus-based Statistical Sense Resolution. In: *Proceedings, ARPA Human Language Technology Workshop*, Plainsboro, NJ. 260-265.

Miller, G. (1990). WordNet: An On-line Lexical Database. *International Journal of Lexicography*, Vol. 3, No. 4, Special Issue.

Resnik, P., and Yarowsky, D. (1997). A Perspective on Word Sense Disambiguation Methods and Their Evaluation, position paper presented at *the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What, and How?,* held April 4-5, 1997 in Washington, D.C., USA in conjunction with ANLP-97.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill, 1956.

Yarowsky, D. (1992). Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora. In: *Proceedings of the Fourteenth International Conference on Computational Linguistics.* Nantes, France. 454-460.