Increased Retrieval Performance using Word Sense Discrimination

Atelach Alemu Argaw, Lars Asker

Stockholm University, KTH
Department of Computer and Systems Sciences
{atelach ; asker}@dsv.su.se

Résumé

Nous prouvons que l'information mutuelle entre des paires de mots peut être employée avec succès pour distinguer entre différents usages des mots dans la traduction des requêtes pour la recherche d'information translinguistique. Les expérmentations sont entreprises dans le contexte de la recherche d'information translinguistique amharique-francais. Des expérmentations sont entreprises qui comparent la performance de la collection des termes des requêtes désambiguïsés et non désambiguïsés contre une collection de documents ordonnés. Les résultats montrent une amélioration de performance pour les requêtes désambiguïsées en comparison avec l'approche alternative qui emploie la collection de termes entièrement expansés.

Mots-clés : recherche d'information, désambiguïsation des mots, information mutuelle, amharique.

Abstract

We show that Mutual Information between word pairs can be successfully used to discriminate between word senses in the query translation step of Cross Language Information Retrieval. The experiment is conducted in the context of Amharic to French Cross Language Information Retrieval. We have performed a number of retrieval experiments in which we compare the performance of the sense discriminated and non-discriminated set of query terms against a ranked document collection. The results show an increased performance for the discriminated queries compared to the alternative approach, which uses the fully expanded set of terms.

Keywords: information retrieval, word sense discrimination, mutual information, amharic.

1. Introduction

A common approach in Cross Language Information Retrieval (CLIR) is to look up and translate each query term using a machine readable dictionary (MRD) and then do the retrieval step using the translated query. The use of a MRD to translate terms will on one hand function as a form of query expansion by allowing each query term to be translated into not just one, but several words with similar meaning (synonyms), in the target language. On the other hand, it may well introduce irrellevant terms in the query and thereby cause decreased performance of the retrieval system. Such irrellevant terms originate from polysemy where words with different meaning by same spelling will incorrectly be selected in the translation process. It is therefor important to find methods whereby we can automatically distinguish between correct and incorrect translations given the context of the query at hand.

We present work aimed at investigating how Mutual Information (MI) between word pairs can be used to discriminate between word senses in the query translation step of (CLIR). Section

2 below contains an overview of related work. The experiments (described in section 3), are conducted in the context of Amharic[1] to French Information Retrieval. Two Machine Readable Dictionaries (MRDs) were used to translate the Amharic query terms to French. In a second step, the MI among word pairs in the French text collection was used to discriminate between word senses. We then performed a number of retrieval experiments where we compare the performance of the sense disambiguated and non-disambiguated set of query terms against a ranked document collection. The results (presented in section 4) show an increased performance for the disambiguated queries compared the the alternative approach that uses no disambiguation.

2. Word Sense Disambiguation in Information Retrieval

Word sense is defined as the mental representation of different meanings of a word. Many words might have several possible meanings (senses) but without a certain context, the actual interpretation (or translation) is undecided. The task of disambiguation hence, is determining which of the senses a certain ambiguous word represents in a given context (Manning and Schütze, 1999).

Researchers have been investigating a wide range of approaches to the problem of word sense disambiguation. A procedural approach, where words are considered experts of their own meaning and resolve their senses by passing messages between themselves is implemented by Small and Rieger (1982). Other approaches include spreading activation and semantic networks by Hirst (1988) and Hayes (1977), word co occurrence for word sense disambiguation in the context of information retrieval (Weiss, 1973), word collocation (Brown *et al*., 1991), (Black, 1988), (Dahlgren, 1988), word collocation and syntax (Atkins, 1997), thesaurus (Slator, 1988), (Voorhees, 1993), (Sussna, 1993), machine readable dictionary (Lesk, 1986), (Wilks *et al*., 1989), (Cowie *et al*., 1992), (Black, 1988), morphological analysis (Zernik, 1990), bilingual corpora (Dagan *et al*., 1991), disambiguation using a second language corpus (Dagan and Itai, 1994), and clustering (Yarowsky, 1995), (Zernik, 1990). There has been a substantial amount of work done in the field of WSD to date though most approaches are limitted to experiments with a few hand picked words and there was no standardized scheme for evaluation of such systems until recently (*e.g*. the in vitro evaluation of SENSEVAL). See (Ide and Veronis, 1998) for a detailed review of research conducted on WSD in the past 50 years.

Word sense discrimination, which is very much related to word sense disambiguation, on the other hand is not concerned with sense labelling at all. Rather it divides the occurrences of a word into a number of classes by determining for any two occurrences whether they belong to the same sense or not. Word sense discrimination is an easier task than full disambiguation (which in many cases involves both sense labelling and discrimination) since we need only to determine which occurrences have the same meaning and not what the meaning actually is (Schütze, 1998).

In information retrieval the focus is to find representations and methods of comparison that will accurately discriminate between relevant and non-relevant documents. In most retrieval systems words are used to represent queries and documents and such a representation poses two major problems. One is that of polysemy (words with the same spelling but different meanings) that would cause the retrieval of irrelevant documents, while the other is that of synonyms (words that are spelled differently but representing the same concept) since users are not interested in retrieving documents with exactly the same words as in the query. The first problem could be

[1] Amharic is the official government language of Ethiopia and belongs to the Semitic language group.

addressed partially through the use of phrases instead of words though it's not always possible to provide phrases in which the word occurs only with the desired sense as well as it imposes a significant burden on the user. The second problem could be solved by expanding the query terms with synonyms (Krovetz and Croft, 1992).

Eliminating occurrences of words in documents where they are used in an inappropriate sense is claimed to be desirable when searching for specific keywords (c.f. (Salton and McGill, 1983), (Voorhees, 1993), (Schütze and Pedersen, 1995)). Weiss (1973) pioneered the implementation of a disambiguator in an IR system for a set of five ambiguous words and reported a performance increase of 1 %. Krovetz and Croft (1992) report experiments designed to discover the degree of lexical ambiguity in information retrieval test collections and the utility of word senses for discriminating between relevant and non-relevant documents, using sense information from a machine readable dictionary. They report that resolving sense ambiguity doesn't have a substantial effect in IR. They also claim that although clear correct senses could be assigned to words in some applications, in the information retrieval context, it may not be necessary to identify and use a single correct sense of a word, rather, ruling out as many of the incorrect word senses as possible and giving a high weight to the sense most likely to be correct may improve retrieval effectiveness.

Voorhees (1993) and Wallis (1993) performed large scale experiments and applied a disambiguator in an IR system, both reporting a drop in retrieval performance. Sanderson (1994) uses Yarowsky's (1995) pseudo-words ambiguity introduction to compare the performance of a probabilistic weighted term IR system. He concludes that the performance of IR systems is insensitive to ambiguity but very sensitive to erroneous disambiguation. On the other hand, Schütze and Pedersen (1995) show a marked improvement in retrieval (14.4 %) using a method which combines search-by-word and search-by-sense.

Translation ambiguity and target polysemy are two major problems in CLIR. The target polysemy adds extraneous senses and may thereby affect the retrieval performance (Chen *et al*., 1999). Although the importance of sense weighting and disambiguation is still a matter of debate, it is still considered one of the crucial aspects of CLIR. Ballesteros and Croft (1998), Bian and Chen (Chen *et al*., 1999), Dagan *et. al*. (1994; 1991), use MRDs and co-occurrence statistics trained from target language text collection, which is also the method employed in our experiments.

3. Experimental setup

In our experiments, we made use of two MRDs to get all the different senses of a term (word or phrase) - as given by the MRD, and a statistical collocation measure of mutual information to discriminate among these senses. One is an Amharic - French dictionary containing 12,000 Amharic entries with corresponding 36,000 French entries (Abebe, 2004) and the other is an Amharic - English dictionary with approximately 15,000 Amharic entries (Aklilu, 1981). The translations from the Amharic - French dictionary were always preferred before the Amharic - English ones and only in cases when a term had not been matched in the French dictionary was it matched against the English one. The English translations were then translated into French using an online dictionary found at WordReference (www.wordreference.com).

For evaluation, each term in the maximally expanded set of translated query terms was manually labelled as relevant or non-relevant by a human expert. The set of classified query terms was then used to evaluate the discriminator by measuring the precision and recall for various word

association threshold values.

For the experiments, we used the 50 Amharic queries from the CLEF[2] 2005 topic set. The amount of words in each query differed substantially from one query to another. After the dictionary lookup and stop word removal, the number of words in each query ranged between 8 and 71. This is due to a large difference in the number of words and in the number of stop words in each query as well as the number of senses and synonyms that are given in the dictionary for each word. In this way all the translated terms for each of the 50 queries were represented as a bag of words consisting of all possible translations for all terms in the original query.

Mutual Information between word pairs, as measured on the target language text collection was then used to discriminate word senses. (Pointwise) mutual information compares the probability of observing two events x and y together (the joint probability) with the probabilities of observing x and y independently (chance). If two (words), x and y, have probabilities P(x) and P(y), then their mutual information, I(x,y), is defined to be:

$$I(x,y) = log_2 \frac{P(x,y)}{P(x)*P(y)} = log_2 \frac{P(x/y)}{P(x)}$$

If there is a genuine association between x and y, P(x,y) will be much larger than chance P(x)* P(y), thus I(x,y) will be greater than 0. If there is no interesting relationship between x and y, P(x,y) will be approximately equal to P(x)* P(y), and thus, I(x,y) will be close to 0. And if x and y are in complementary distribution, P(x,y) will be much less than P(x)* P(y), and I(x,y) will be less than 0.

Although very widely used by researchers for different applications, MI has also been criticized by many as to its ability to capture the similarity between two events especially when there is data scarcity (Manning and Schütze, 1999). Since we had access to a very large text collection in the target language, and because of its wide implementation, we chose to use MI.

As mentioned above, the translated query terms were put in a bag of words, and the mutual information for each of the possible word pairs was calculated. When we put the expanded words we treat both synonyms and translations with a distinct sense as given in the MRD equally. Another way of handling this situation is to group synonyms before the discrimination. We chose the first approach with two assumptions: one is that even though words may be synonymous, it doesn't necessarily mean that they are all equally used in a certain context, and the other being even though a word may have distinct senses defined in the MRD, those distinctions may not necessarily be applicable in the context the term is currently used. This approach is believed to ensure that words with inappropriate senses and synonyms with less contextual usage will be removed while at the same time the query is being expanded with appropriate terms.

We used a subset of the CLEF 2005 French document collection consisting of 14,000 news articles with 4.5 million words to calculate the MI values. Both the French keywords and the document collection were lemmatized in order to cater for the different forms of each word under consideration.

Following the idea that ambiguous words can be used in a variety of contexts but collectively they indicate a single context and particular meanings, we relied on the number of association as given by MI values that a certain word has in order to determine whether the word should be removed from the query or not. Given the bag of words for each query, we calculated the mutual

[2] Cross Language Evaluation Forum, http://www.clef-campaign.org.

information for each unique pair. The next step was to see for each unique word how many positive associations it has with the rest of the words in the bag. We experimented with different levels of combining precision and recall values depending on which one of these two measures we want to give more importance to. To contrast the approach of using the maximum recall of words (no discrimination) we decided that precision should be given much more priority over recall (beta value of 0.15 - see the following section), and we set an empirical threshold value of 0.4. *i.e.* a word is kept in the query if it shows positive associations with 40 % of the words in the list, otherwise it is removed. Here, note that the mutual information values are converted to a binary 0 or 1. 0 being assigned to words that have less than or equal to 0 MI values (independent term pairs), and 1 to those with positive MI values (dependent term pairs). We are simply taking all positive MI values as indicators of association without any consideration as to how strong the association is. This is done to input as much association between all the words in the query as possible rather than putting the focus on individual pairwise association values.

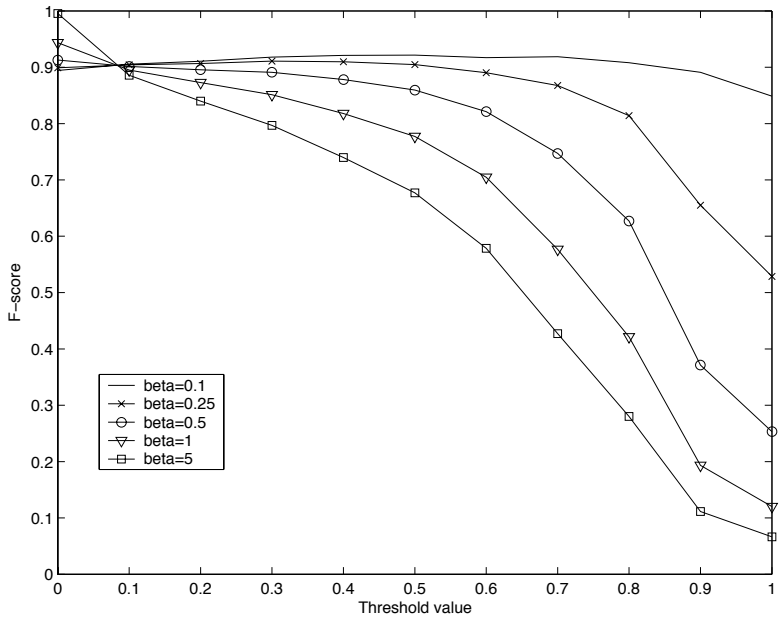

Figure 1. F-score as a function of beta values and threshold values

The purpose of the sense discrimination in this case is to determine if a certain term in a maximally expanded query terms list is relevant or not. Therefore we have used a binary (relevant, non-relevant) evaluation for the experiments. In order to evaluate the precision, recall and accuracy of the WSD, a human judged list of keywords was first prepared. A French speaker was given the original French queries and the list of translated query terms which includes all senses and synonyms of a word as given in the MRD. The stop words were removed from this list and hence it is the possible keywords list. She marked the terms in the keywords list as either relevant or non-relevant in the specific context of the query under consideration. The list of keywords picked out to be relevant according to their MI score, were then compared to the list marked by the human judge, to evaluate the performance of the system.

Figure 1 shows how the F-score value varies for different beta values and at different threshold levels. We used the F-Score in order to combine the precision and recall measures with the intent of determining a reasonable word association threshold value. The F-score is a value from 0 to 1 inclusive. Note that beta is a parameter to the F-score, and higher beta values would favour recall

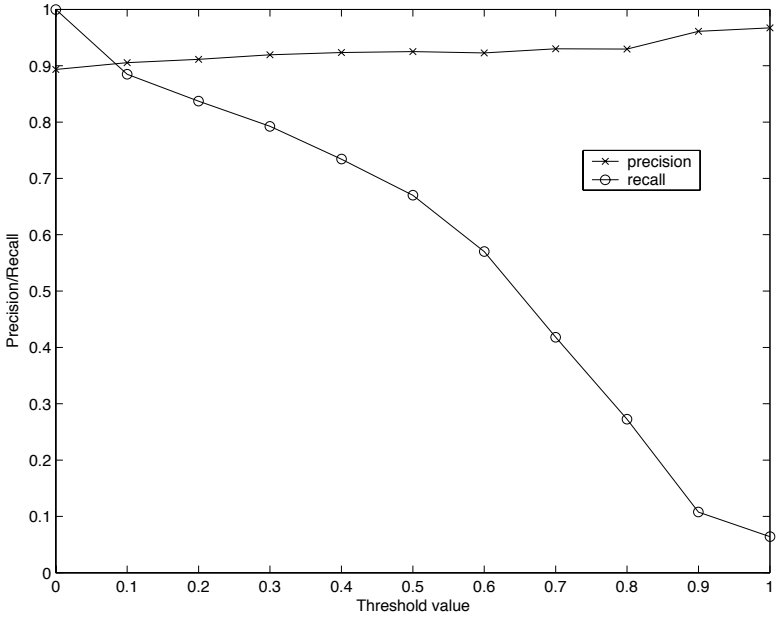

Figure 2. Precision and Recall as a function of the threshold value

over precision. Often, F(1)score is used, that is, precision and recall are given equal weights. The F-score is calculated using the formula:

$$F(\beta) = \frac{(\beta^2+1)*P*R}{(\beta^2*P)+R}$$

where $P = Precision$ and $R = Recall$.

The beta values in figure 1 vary from 0.1 which indicates a strong preference for precision, up to 5 indicating a strong preference for recall. The threshold value of 0.4 reported in these experiments reflects the view that precision is rated as more important than recall. This evaluation was a comparision of the new query terms against the manually relevance judged query terms. The new query terms are the ones that are selected by this model to be judged relevant. At a threshold value of 0.4, precision is 0.92, recall is 0.73, accuracy is 0.71 while the maximum F-score with beta (0.15) is 0.92.

Figure 2 shows the precision and recall at different threshold levels. At the threshold level 0 (zero), where all the terms are included, the recall is 1 (since all the correct terms are included) and the precision is around 0.89 since incorrect query terms are also included (approximately one in ten is incorrect). As the threshold levels increase, the proportion of correct/incorrect query terms slowly increases giving a higher precision value at the price of a reduced recall. At a threshold level of 1.0 approximately 1 term in 30 is incorrect, but the average number of terms per query has dropped from 20 (at threshold level 0) down to 1.2.

4. Results and Evaluation

We have initially performed a retrieval experiment using Searcher - an experimental search engine developed at SICS[3]. In the experiment, two runs were conducted with one of them searching for all content bearing, expanded query terms without any discrimination while the other

[3] The Swedish Institute of Computer Science.

Recall	word sense discrimination	no word sense discrimination
0.00	24.55	23.84
0.10	9.12	9.18
0.20	5.13	4.71
0.30	3.75	3.36
0.40	2.83	2.71
0.50	2.02	1.85
0.60	1.36	1.45
0.70	0.76	0.60
0.80	0.57	0.37
0.90	0.39	0.23
1.00	0.27	0.17

Table 1. Recall-Precision tables for the two runs

one searches for the discriminated set of content bearing query terms. For the first run the recall value is set to 1.0 *i.e.* include all possible translations of each word in the query. This is believed to produce a list of query terms which may contain inappropriate word senses and let the search engine implicitly perform disambiguation. The second approach is to set a high precision value and a lower recall (as given by the beta value of 0.15 in these experiments) which will remove inappropriate sense words and possibly some words with appropriate senses as well and pass the disambiguated list to the search engine. The experimental results are given in table 1. As can be seen from the results, word sense discrimination in query translation gives an overall better accuracy.

To further investigate how discrimination vs. non-discrimination will affect the performance of a retrieval engine, we performed additional experiments using Lucene (URL, 2005), an open source search toolbox, on the same document collection and for the same 50 queries. Also here we compared the performance of the discriminated versus fully expanded set of query terms in the context of different additional terms. In the previous experiments using Searcher we only used the query terms that were found in the dictionaries. In this set of experiments we added manual translations of the terms that were not found in the dictionaries to one set (see curve A in figure 3), and we added all the terms from an original French version of the query to another set (see curve B in figure 3). These two additions were made to the sense discriminated and non-discriminated sets of queries producing four distinct sets. The results of these experiments show that word sense discriminated sets of queries performed better than the non-dicriminated ones.

5. Conclusions

We have described experiments conducted to discriminate among word senses in query translation for the purpose of an Amharic to French bilingual Information Retrieval. Two Machine Readable Dictionaries were used to look up the Amharic query terms and translate them to French. The mutual information between word pairs in the target language text collection was used as the primary source of information for this task. An empirical threshold value of 0.4 was set to eliminate words of inappropriate sense from the translated terms list. A human judged list of keywords was used to evaluate the discrimination by using recall and precision measures. Depending on the threshold value set, the recall and precision measures vary.

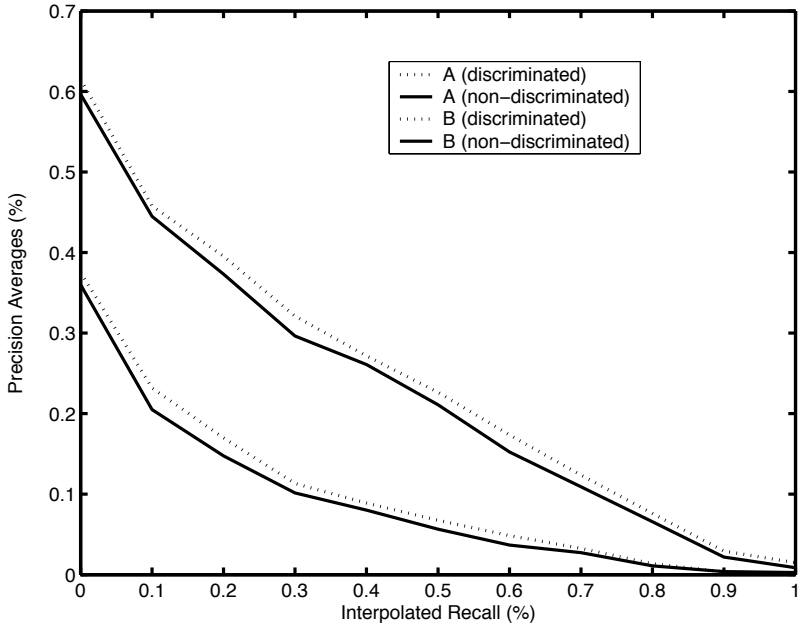

Figure 3. Precision averages versus interpolated recall over 50 queries for query set A and B. Dotted lines show the performance of the discriminated set of queries while the solid lines show the performance for the non-discriminated set of queries.

A threshold value of 0.4 gave a reasonable trade off between recall and precision in our experiments. With this threshold value, and at a recall of 73 %, the precision of the word sense discrimination was found to be 92 %.

These experiments show that the implementation of word sense discrimination during query translation showed better retrieval performance than the approach of maximally expanding query terms. The discriminated set of query terms outperforms the non-discriminated queries by between 0.28 and 0.94 percent (R-precision). Although not statistically significant, the experimental results are incouraging and we plan to investigate further comparative studies that will include the different measures of word collocation as well as perform large scale experiments with the current approach.

Acknowledgements

The copyright to the two volumes of the French-Amharic and Amharic-French dictionary (« Dictionnaire Francais-Amharique » and « Dictionnaire Amharique-Francais ») by Dr Berhanou Abebe and Eloi Fiquet is owned by the French Ministry of Foreign Affairs. We would like to thank the authors and the French embassy in Addis Ababa for allowing us to use the dictionary in this research. The content of the « English - Amharic Dictionary » is the intellectual property of Dr Amsalu Aklilu. We would like to thank Dr Amsalu as well as Daniel Yacob (yacob at geez dot org) of the Geez frontier foundation (www.geez.org) for making it possible for us to use the dictionary and other resources in this work.

References

ABEBE B. (2004). *Dictionnaire Amharique-Francais*. Shama Books, Addis-Abeba, Éthiopie.

AKLILU A. (1981). *English - Amharic Dictionary*. Mega Publishing Enterprise, Ethiopia.

ATKINS B. (1997). "Semantic ID Tags: Corpus Evidence for Dictionary Senses". In *Advances in Lexicology, Proceedings of the Third Annual Conference of the Center for the New OED*. Ontario : University of Waterloo.

BALLESTEROS L. and CROFT B. (1998). "Resolving Ambiguity for Cross-Language Retrieval". In *Proceedings of ACM Special Interest Group on Information Retreival (SIGIR)*. p. 64–71.

BLACK E. (1988). "An Experiment in Computational Discrimination of English Word Senses". In *IBM Journal of Research and Development*, 32 (2), 185–194.

BROWN P. F., PIETRA S. D., PIETRA V. J. D. and MERCER R. L. (1991). "Word-Sense Disambiguation Using Statistical Methods". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. p. 264-270.

CHEN H.-H., BIAN G.-W. and LIN W.-C. (1999). "Resolving Translation Ambiguity and Target Polysemy in Cross-Language Information Retrieval.". In *Proceednings of the 37th Annual Meeting of the Association for Computational Linguistics*.

COWIE J., GUTHRIE J. A. and GUTHRIE L. M. (1992). "Lexical Disambiguation using Simulated Annealing". In *Proceedings of COLING-92*. Nantes, France. p. 359–365.

DAGAN I. and ITAI A. (1994). "Word Sense Disambiguation Using a Second Language Monolingual Corpus.". In *Computational Linguistics*, 20 (4), 563-596.

DAGAN I., ITAI A. and SCHWALL U. (1991). "Two languages are more informative than one". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*. p. 130–137.

DAHLGREN K. (1988). *Naive Semantics for Natural Language Understanding*. Kluwer Academic Publishers, Norwell, MA, USA.

HAYES P. (1977). *Some association based techniques for lexical disambiguation by machine*. PhD thesis, University of Rochester, Department of Computer Science. Tech. Report No. 25.

HIRST G. (1988). "Resolving lexical ambiguity computationally with spreading activation and polaroid words". In S. L. Small, G. W. Cottrell and M. K. Tanenhaus(eds.), *Lexical ambiguity resolution: Perspectives from psycholinguistics, neuropsychology, and artificial intelligence*, p. 73–108. San Mateo, CA: Morgan Kaufmann.

IDE N. and VERONIS J. (1998). "Word sense disambiguation: The state of the art". In *Computational Linguistics*, 28 (1), 1–40.

KROVETZ R. and CROFT W. B. (1992). "Lexical Ambiguity and Information Retrieval". In *ACM Transactions on Information Systems*, 10 (2), 115–141.

LESK M. (1986). "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone". In *Proceedings of the 1986 SIGDOC Conference*. p. 24–26.

MANNING C. D. and SCHÜTZE H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.

SALTON G. and MCGILL M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, NY.

SANDERSON M. (1994). "Word Sense Disambiguation and Information Retrieval". In *Proceedings of SIGIR-94*. Dublin, Ireland : ACM, p. 142–151.

SCHÜTZE H. (1998). "Automatic Word Sense Discrimination". In *Computational Linguistics*, 24 (1), 97-123.

SCHÜTZE H. and PEDERSEN J. O. (1995). "Information Retrieval Based on Word Senses". In

Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval. p. 161–175.

SLATOR B. (1988). *Lexical semantics and preference semantics analysis*. PhD thesis, New Mexico State University. Ph.D. dissertaion, Report MCCs-88-143.

SMALL S. L. and RIEGER C. (1982). "Parsing and Comprehending with Word Experts". In W. G. Lehnert and M. H. Ringle(eds.), *Strategies for Natural Language Processing*, p. 89–147. Hillsdale, NJ: Lawrence Erlbaum.

SUSSNA M. (1993). "Word Sense Disambiguation for Free-text indexing using a massive semantic network". In *Proceedings of the Second International Conference on Information and Knowledge Management (CIKM-93)*. Arlington, Virginia.

URL (2005). "http://lucene.apache.org/java/docs/index.html".

VOORHEES E. M. (1993). "Using WordNet to disambiguate word senses for text retrieval". In *Proceedings of the 16th Annual International ACM SIGIR conference on Research and Development in Information Retrieval*. New York, NY, USA : ACM Press, p. 171–180.

WALLIS P. (1993). "Information Retrieval Based on Paraphrase". In *Proceedings of PACLING Conference*.

WEISS S. (1973). "Learning to disambiguate". In *Information storage and retrieval*, 9, 33–41.

WILKS Y., FASS D., GUO C.-M., MCDONALD J., PLATE T. and SLATOR B. (1989). "A Tractable Macine Dictionary as a Resource for Computational Semantics". In Boguraev and Briscoe(eds.), *Computational Lexicography for Natural Language Processing*. Longman.

YAROWSKY D. (1995). "Unsupervised word sense disambiguation rivaling supervised methods". In *Proceedings of the 33rd conference on Association for Computational Linguistics*. Morristown, NJ, USA : Association for Computational Linguistics, p. 189–196.

ZERNIK U. (1990). "Tagging word senses in corpus: the needle in the haystack revisited". In *Technical Report 90CRD198, GE R&D Center*.