Word Sense Disambiguation using Conceptual Density

Eneko Agirre*
Lengoaia eta Sistema Informatikoak saila. Euskal Herriko Universitatea.
p.k. 649, 200800 Donostia. Spain. jibagbee@si.heu.es

German Rigau**
Departament de Llenguatges i Sistemes Informàtics. Universitat Politècnica de Catalunya.
Pau Gargallo 5, 08028 Barcelona. Spain. g.rigau@lsi.upc.es

Abstract.

This paper presents a method for the resolution of lexical ambiguity of nouns and its automatic evaluation over the Brown Corpus. The method relies on the use of the wide-coverage noun taxonomy of WordNet and the notion of conceptual distance among concepts, captured by a Conceptual Density formula developed for this purpose. This fully automatic method requires no hand coding of lexical entries, hand tagging of text nor any kind of training process. The results of the experiments have been automatically evaluated against SemCor, the sense-tagged version of the Brown Corpus.

1 Introduction

Much of recent work in lexical ambiguity resolution offers the prospect that a disambiguation system might be able to receive as input unrestricted text and tag each word with the most likely sense with fairly reasonable accuracy and efficiency. The most extended approach use the context of the word to be disambiguated together with information about each of its word senses to solve this problem.

Interesting experiments have been performed in recent years using preexisting lexical knowledge resources: [Cowie et al. 92], [Wilks et al. 93] with LDOCE, [Yarowsky 92] with Roget's International Thesaurus, and [Sussna 93], [Voorhees 93], [Richardson et al. 94], [Resnik 95] with WordNet.

Although each of these techniques looks promising for disambiguation, either they have been only applied to a small number of words, a few sentences or not in a public domain corpus. For this reason we have tried to disambiguate all the nouns from real texts in the public domain sense tagged version of the Brown corpus [Francis & Kucera 67], [Miller et al. 93], also called Semantic Concordance or SemCor for short[1]. The words in SemCor are tagged with word senses from WordNet, a broad semantic taxonomy for English [Miller 90][2]. Thus, SemCor provides an appropriate environment for testing our procedures and comparing among alternatives in a fully automatic way.

The automatic decision procedure for lexical ambiguity resolution presented in this paper is based on an elaboration of the conceptual distance among concepts: Conceptual Density [Agirre & Rigau 95]. The system needs to know how words are clustered in semantic classes, and how semantic classes are hierarchically organised. For this purpose, we have used WordNet. Our system tries to resolve the lexical ambiguity of nouns by finding the combination of senses from a set of contiguous nouns that maximises the Conceptual Density among senses.

The performance of the procedure was tested on four SemCor texts chosen at random. For comparison purposes two other approaches, [Sussna 93] and [Yarowsky 92], were also tried. The results show that our algorithm performs better on the test set.

Following this short introduction the Conceptual Density formula is presented. The main procedure to resolve lexical ambiguity of nouns using Conceptual Density is sketched on section 3. Section 4 describes extensively the experiments and its results. Finally, sections 5 and 6 deal with further work and conclusions.

*Eneko Agirre was supported by a grant from the Basque Goverment. Part of this work is included in projects 141226-TA248/95 of the Basque Country University and PI95-054 of the Basque Government.

**German Rigau was supported by a grant from the Ministerio de Educación y Ciencia.

[1] Semcor comprises approximately 250,000 words. The tagging was done manually, and the error rate measured by the authors is around 10% for polysemous words.

[2] The senses of a word are represented by synonym sets (or synsets), one for each word sense. The nominal part of WordNet can be viewed as a tangled hierarchy of hypo/hypernymy relations among synsets. Nominal relations include also three kinds of meronymic relations, which can be paraphrased as member-of, made-of and component-part-of. The version used in this work is WordNet 1.4, The coverage in WordNet of senses for open-class words in SemCor reaches 96% according to the authors.

2 Conceptual Density and Word Sense Disambiguation

Conceptual distance tries to provide a basis for measuring closeness in meaning among words, taking as reference a structured hierarchical net. Conceptual distance between two concepts is defined in [Rada et al. 89] as the length of the shortest path that connects the concepts in a hierarchical semantic net. In a similar approach, [Sussna 93] employs the notion of conceptual distance between network nodes in order to improve precision during document indexing. [Resnik 95] captures semantic similarity (closely related to conceptual distance) by means of the information content of the concepts in a hierarchical net. In general these approaches focus on nouns.

The measure of conceptual distance among concepts we are looking for should be sensitive to:

- the length of the shortest path that connects the concepts involved.
- the depth in the hierarchy: concepts in a deeper part of the hierarchy should be ranked closer.
- the density of concepts in the hierarchy: concepts in a dense part of the hierarchy are relatively closer than those in a more sparse region.
- the measure should be independent of the number of concepts we are measuring.

We have experimented with several formulas that follow the four criteria presented above. The experiments reported here were performed using the Conceptual Density formula [Agirre & Rigau 95], which compares areas of subhierarchies.

To illustrate how Conceptual Density can help to disambiguate a word, in figure 1 the word W has four senses and several context words. Each sense of the words belongs to a subhierarchy of WordNet. The dots in the subhierarchies represent the senses of either the word to be disambiguated (W) or the words in the context. Conceptual Density will yield the highest density for the subhierarchy containing more senses of those, relative to the total amount of senses in the subhierarchy. The sense of W contained in the subhierarchy with highest Conceptual Density will be chosen as the sense disambiguating W in the given context. In figure 1, sense2 would be chosen.

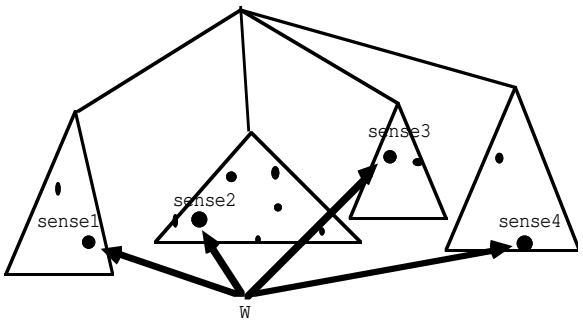

Figure 1: senses of a word in WordNet

Given a concept *c*, at the top of a subhierarchy, and given *nhyp* (mean number of hyponyms per node), the Conceptual Density for *c* when its subhierarchy contains a number *m* (marks) of senses of the words to disambiguate is given by the formula below:

$$CD(c,m) = \frac{\sum_{i=0}^{m-1} nhyp^{i^{0.20}}}{descendants_c} \qquad (1)$$

Formula 1 shows a parameter that was computed experimentally. The 0.20 tries to smooth the exponential *i*, as *m* ranges between 1 and the total number of senses in WordNet. Several values were tried for the parameter, and it was found that the best performance was attained consistently when the parameter was near 0.20.

3 The Disambiguation Algorithm Using Conceptual Density

Given a window size, the program moves the window one noun at a time from the beginning of the document towards its end, disambiguating in each step the noun in the middle of the window and considering the other nouns in the window as context. Non-noun words are not taken into account.

The algorithm to disambiguate a given noun w in the middle of a window of nouns W (c.f. figure 2) roughly proceeds as follows:

```
(Step 1) tree := compute_tree(words_in_window)
         loop
(Step 2)   tree := compute_conceptual_distance(tree)
(Step 3)   concept := selecct_concept_with_highest_weigth(tree)
           if  concept = null then exitloop
(Step 4)   tree := mark_disambiguated_senses(tree,concept)
         endloop
(Step 5) output_disambiguation_result(tree)
```

Figure 2: algorithm for each window

First, the algorithm represents in a lattice the nouns present in the window, their senses and hypernyms (step 1). Then, the program computes the Conceptual Density of each concept in WordNet according to the senses it contains in its subhierarchy (step 2). It selects the concept c with highest Conceptual Density (step 3) and selects the senses below it as the correct senses for the respective words (step 4).

The algorithm proceeds then to compute the density for the remaining senses in the lattice, and continues to disambiguate the nouns left in W (back to steps 2, 3 and 4). When no further disambiguation is possible, the senses left for w are processed and the result is presented (step 5).

Besides completely disambiguating a word or failing to do so, in some cases the disambiguation algorithm returns several possible senses for a word. In the experiments we considered these partial outcomes as failure to disambiguate.

4 The Experiments

4.1 The texts

We selected four texts from SemCor at random: br-a01 (where a stands for gender "Press: Reportage"), br-b20 (b for "Press: Editorial"), br-j09 (j means "Learned: Science") and br-r05 (r for "Humour"). Table 1 shows some statistics for each text.

text	words	nouns	nouns in WN	monosemous
br-a01	2079	564	464	149 (32%)
br-ab20	2153	453	377	128 (34%)
br-j09	2495	620	586	205 (34%)
br-r05	2407	457	431	120 (27%)
total	9134	2094	1858	602 (32%)

Table 1: data for each text

An average of 11% of all nouns in these four texts were not found in WordNet. According to this data, the amount of monosemous nouns in these texts is bigger (32% average) than the one calculated for the open-class words from the whole SemCor (27.2% according to [Miller et al. 94]).

For our experiments, these texts play both the role of input files (without semantic tags) and (tagged) test files. When they are treated as input files, we throw away all non-noun words, only leaving the lemmas of the nouns present in WordNet.

4.2 Results and evaluation

One of the goals of the experiments was to decide among different variants of the Conceptual Density formula. Results are given averaging the results of the four files. Partial disambiguation is treated as failure to disambiguate. Precision (that is, the percentage of actual answers which were correct) and recall (that is, the percentage of possible answers which were correct) are given in terms of polysemous nouns only. Graphs are drawn against the size of the context[3] .

• **meronymy does not improve performance as expected.** A priori, the more relations are taken in account (e.i. meronymic relations, in addition to the hypo/hypernymy relation) the better density would capture semantic relatedness, and therefore better results can be expected.

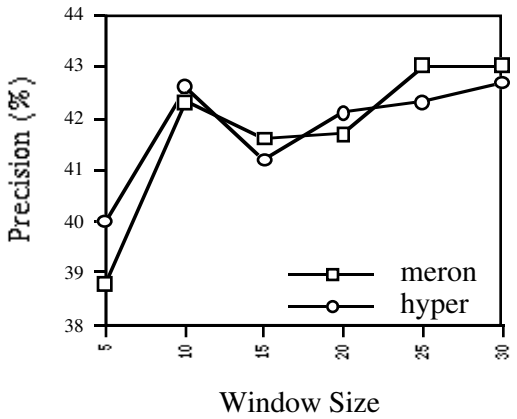

Figure 3: meronymy and hyperonymy

The experiments (see figure 3) showed that there is not much difference; adding meronymic information does not improve precision, and raises coverage only 3% (approximately). Nevertheless, in the rest of the results reported below, meronymy and hypernymy were used.

• **global nhyp is as good as local nhyp.** The average number of hyponyms or *nhyp* (c.f. formula 1) can be approximated in two ways. If an independent *nhyp* is computed for every concept in WordNet we call it *local nhyp*. If instead, a unique *nhyp* is computed using the whole hierarchy, we have *global nhyp*.

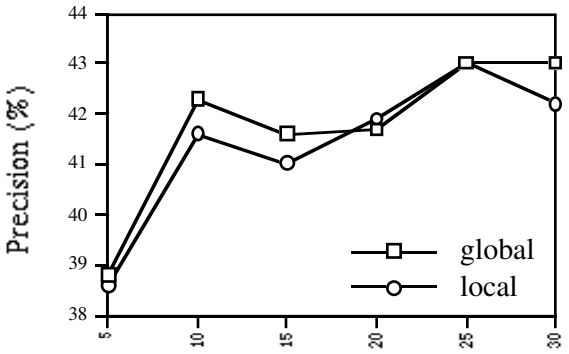

Figure 4: *local nhyp* vs. *global nhyp*

[3] context size is given in terms of nouns.

While *local nhyp* is the actual average for a given concept, *global nhyp* gives only an estimation. The results (c.f. figure 4) show that *local nhyp* performs only slightly better. Therefore *global nhyp* is favoured and was used in subsequent experiments.

• **context size: different behaviour for each text.** One could assume that the more context there is, the better the disambiguation results would be. Our experiments show that each file from SemCor has a different behaviour (c.f. figure 5) while br-b20 shows clear improvement for bigger window sizes, br-r05 gets a local maximum at a 10 size window, etc.

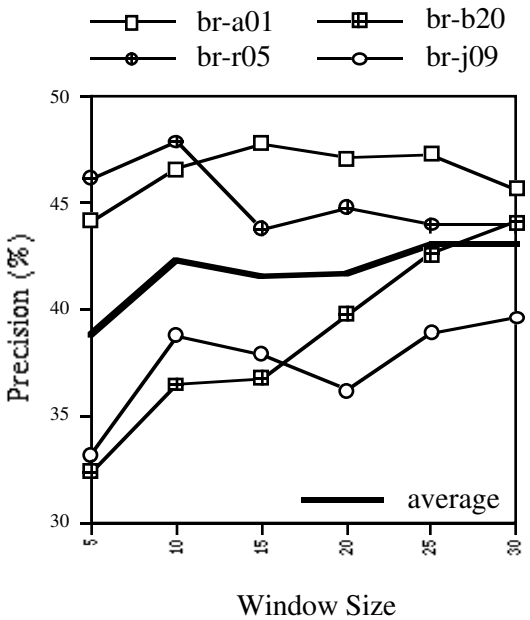

Figure 5: context size and different files

As each text is structured a list of sentences, lacking any indication of headings, sections, paragraph endings, text changes, etc. the program gathers the context without knowing whether the nouns actually occur in coherent pieces of text. This could account for the fact that in br-r05, composed mainly by short pieces of dialogues, the best results are for window size 10, the average size of this dialogue pieces. Likewise, the results for br-a01, which contains short journalistic texts, are best for window sizes from 15 to 25, decreasing significatly for size 30.

In addition, the actual nature of each text is for sure an important factor, difficult to measure, which could account for the different behavoiur on its own. In order to give an overall view of the performance, we consider the average behaviour.

• **file vs. sense.** WordNet groups noun senses in 24 lexicographer's files. The algorithm assigns a noun both an specific sense and a file label. Both file matches and sense matches are interesting to count. While the sense level gives a fine graded measure of the algorithm, the file level gives an indication of the performance if we were interested in a less sharp level of disambiguation. The granularity of the sense distinctions made in [Hearst, 91], [Yarowsky 92] and [Gale et al. 93] also called homographs in [Guthrie et al. 93], can be compared to that of the file level in WordNet.

For instance, in [Yarowsky 92] two homographs of the noun bass are considered, one characterised as MUSIC and the other as ANIMAL, INSECT. In WordNet, the 6 senses of bass related to music appear in the following files: ARTIFACT, ATTRIBUTE, COMMUNICATION and PERSON. The 3 senses related to animals appear in the files ANIMAL and FOOD. This means that while the homograph level in [Yarowsky 92] distinguishes two sets of senses, the file level in WordNet distinguishes six sets of senses, still finer in granularity.

Figure 6 shows that, as expected, file-level matches attain better performance (71.2% overall and 53.9% for polysemic nouns) than sense-level matches.

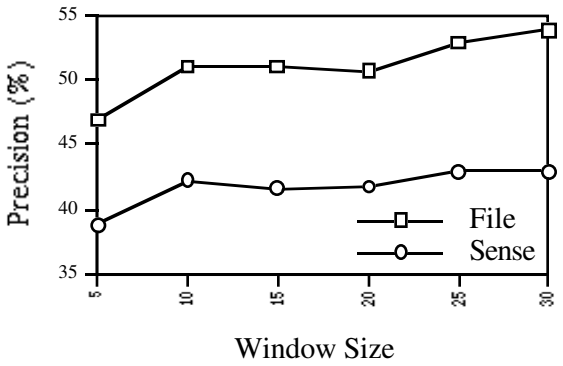

Figure 6: sense level vs. file level

• **evaluation of the results** Figure 7 shows that, overall, coverage over polysemous nouns increases significantly with the window size, without losing precision. Coverage tends to get stabilised near 80%, getting little improvement for window sizes bigger than 20.

The figure also shows the guessing baseline, given by selecting senses at random. This baseline was first calculated analytically and later checked experimentally. We also compare the performance of our algorithm with that of the "most frequent" heuristic. The frequency counts for each sense were collected using the rest of SemCor, and then applied to the four texts. While the precision is similar to that of our algorithm, the coverage is 8% worse.

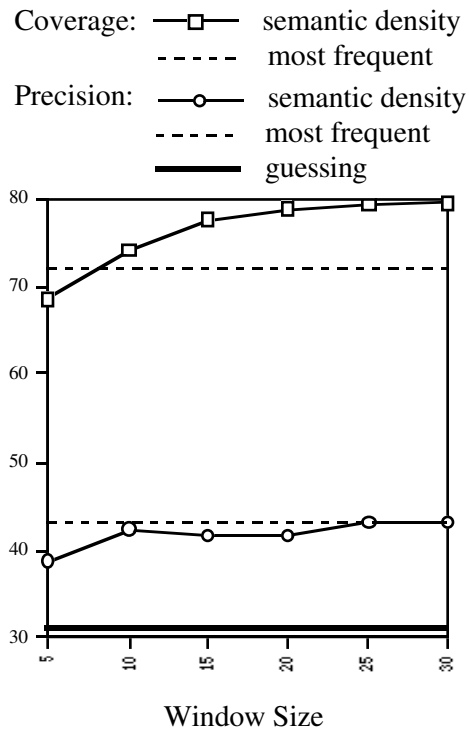

Figure 7: precision and coverage

All the data for the best window size can be seen in table 2. The precision and coverage shown in all the preceding graphs were relative to the polysemous nouns only. Including monosemic nouns precision raises, as shown in table 2, from 43% to 64.5%, and the coverage increases from 79.6% to 86.2%.

%	w=30	Cover.	Prec.	Recall
overall	File	86.2	71.2	61.4
	Sense		64.5	55.5
polysemic	File	79.6	53.9	42.8
	Sense		43	34.2

Table 2: overall data for the best window size

4.3 Comparison with other works

The raw results presented here seem to be poor when compared to those shown in [Hearst 91], [Gale et al. 93] and [Yarowsky 92]. We think that several factors make the comparison difficult. Most of those works focus in a selected set of a few words, generally with a couple of senses of very different meaning (coarse-grained distinctions), and for which their algorithm could gather enough evidence. On the contrary, we tested our method with **all** the nouns in a subset of an unrestricted public domain corpus (more than 9.000 words), making fine-grained distinctions among all the senses in WordNet.

An approach that uses hierarchical knowledge is that of [Resnik 95], which additionally uses the information content of each concept gathered from corpora. Unfortunately he applies his method on a different task, that of disambiguating sets of related nouns. The evaluation is done on a set of related nouns from Roget's Thesaurus tagged by hand. The fact that some senses were discarded because the human judged them not reliable makes comparison even more difficult.

In order to compare our approach we decided to implement [Yarowsky 92] and [Sussna 93], and test them on our texts. For [Yarowsky 92] we had to adapt it to work with WordNet. His method relies on cooccurrence data gathered on Roget's Thesaurus semantic categories. Instead, on our experiment we use saliency values[4] based on the lexicographic file tags in SemCor. The results for a window size of 50 nouns are those shown in table 3[5]. The precision attained by our algorithm is higher. To compare figures better consider the results in table 4, were the coverage of our algorithm was easily extended using the version presented below, increasing recall to 70.1%.

%	Cover.	Prec.	Recall
C.Density	86.2	71.2	61.4
Yarowsky	100.0	64.0	64.0

Table 3: comparison with [Yarowsky 92]

From the methods based on Conceptual Distance, [Sussna 93] is the most similar to ours. Sussna disambiguates several documents from a public corpus using WordNet. The test set was tagged by hand, allowing more than one correct senses for a single word. The method he uses has to overcome a combinatorial explosion[6] controlling the size of the window and "freezing" the senses for all the nouns preceding the noun to be disambiguated. In order to freeze the winning sense Sussna's algorithm is forced to make a unique choice. When Conceptual Distance is not able to choose a single sense, the algorithm chooses one at random.

Conceptual Density overcomes the combinatorial explosion extending the notion of conceptual distance from a pair of words to n words, and therefore can yield more than one correct sense for a word. For comparison, we altered our algorithm to also make random choices when unable to choose a single sense. We applied the algorithm Sussna considers best,

[4] We tried both mutual information and association ratio, and the later performed better.

[5] The results of our algorithm are those for window size 30, file matches and overall.

[6] In our replication of his experiment the mutual constraint for the first 10 nouns (the optimal window size according to his experiments) of file br-r05 had to deal with more than 200,000 synset pairs.

discarding the factors that do not affect performance significantly[7], and obtain the results in table 4.

%		Cover.	Prec.
C.Density	File	100.0	70.1
	Sense		60.1
Sussna	File	100.0	64.5
	Sense		52.3

Table 4: comparison with [Sussna 93]

A more thorough comparison with these methods could be desirable, but not possible in this paper for the sake of conciseness.

5 Further Work

We would like to have included in this paper a study on whether there is or not a correlation among correct and erroneous sense assignations and the degree of Conceptual Density, that is, the actual figure held by formula 1. If this was the case, the error rate could be further decreased setting a certain threshold for Conceptual Density values of winning senses. We would also like to evaluate the usefulness of partial disambiguation: decrease of ambiguity, number of times correct sense is among the chosen ones, etc.

There are some factors that could raise the performance of our algorithm:

•Work on coherent chunks of text. Unfortunately any information about discourse structure is absent in SemCor, apart from sentence endings The performance would gain from the fact that sentences from unrelated topics would not be considered in the disambiguation window.

• Extend and improve the semantic data. WordNet provides sinonymy, hypernymy and meronyny relations for nouns, but other relations are missing. For instance, WordNet lacks cross-categorial semantic relations, which could be very useful to extend the notion of Conceptual Density of nouns to Conceptual Density of words. Apart from extending the disambiguation to verbs, adjectives and adverbs, cross-categorial relations would allow to capture better the relations among senses and provide firmer grounds for disambiguating.

These other relations could be extracted from other knowledge sources, both corpus-based or MRD-based. If those relations could be given on WordNet senses, Conceptual Density could profit from them. It is our belief, following the ideas of [McRoy 92] that full-fledged lexical ambiguity resolution should combine several information sources. Conceptual Density might be only one of a number of complementary evidences of the plausibility of a certain word sense.

Furthermore, WordNet 1.4 is not a complete lexical database (current version is 1.5).

• Tune the sense distinctions to the level best suited for the application. On the one hand the sense distinctions made by WordNet 1.4 are not always satisfactory. On the other hand, our algorithm is not designed to work on the file level, e.g. if the sense level is unable to distinguish among two senses, the file level also fails, even if both senses were from the same file. If the senses were collapsed at the file level, the coverage and precision of the algorithm at the file level might be even better.

6 Conclusion

The automatic method for the disambiguation of nouns presented in this paper is ready-usable in any general domain and on free-running text, given part of speech tags. It does not need any training and uses word sense tags from WordNet, an extensively used lexical data base.

Conceptual Density has been used for other tasks apart from the disambiguation of free-running test. Its application for automatic spelling correction is outlined in [Agirre et al. 94]. It was also used on Computational Lexicography, enriching dictionary senses with semantic tags extracted from WordNet [Rigau 94], or linking bilingual dictionaries to WordNet [Rigau and Agirre 96].

In the experiments, the algorithm disambiguated four texts (about 10,000 words long) of SemCor, a subset of the Brown corpus. The results were obtained automatically comparing the tags in SemCor with those computed by the algorithm, which would allow the comparison with other disambiguation methods. Two other methods, [Sussna 93] and [Yarowsky 92], were also tried on the same texts, showing that our algorithm performs better.

Results are promising, considering the difficulty of the task (free running text, large number of senses per word in WordNet), and the lack of any discourse structure of the texts. Two types of results can be obtained: the specific sense or a coarser, file level, tag.

Acknowledgements

This work, partially described in [Agirre &Rigau 96], was started in the Computing Research Laboratory in New Mexico State University. We wish to thank all the staff of the CRL and specially Jim Cowie, Joe Guthtrie, Louise Guthrie and David Farwell. We would also like to thank Xabier Arregi, Jose mari Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Kepa Sarasola and Aitor Soroa from the Computer Science Faculty of EHU and Francesc Ribas, Horacio Rodríguez and Alicia Ageno from the Computer Science Department of UPC.

[7]Initial mutual constraint size is 10 and window size is 41. Meronymic links are also considered. All the links have the same weigth.

References

Agirre E., Arregi X., Diaz de Ilarraza A. and Sarasola K. 1994. *Conceptual Distance and Automatic Spelling Correction.* in Workshop on Speech recognition and handwriting. Leeds, England.

Agirre E., Rigau G. 1995. *A Proposal for Word Sense Disambiguation using conceptual Distance*, International Conference on Recent Advances in Natural Language Processing. Tzigov Chark, Bulgaria.

Agirre, E. and Rigau G. 1996. *An Experiment in Word SenseDisambiguation of the Brown Corpus Using WordNet*. Memoranda in Computer and Cognitive Science, MCCS-96-291, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico.

Cowie J., Guthrie J., Guthrie L. 1992. *Lexical Disambiguation using Simulated annealing,* in proceedings of DARPA WorkShop on Speech and Natural Language, New York. 238-242.

Francis S. and Kucera H. 1967. *Computing analysis of present-day American English*, Providenc, RI: Brown University Press, 1967.

Gale W., Church K. and Yarowsky D. 1993. *A Method for Disambiguating Word Sense sin a Large Corpus*, in Computers and the Humanities, n. 26.

Guthrie L., Guthrie J. and Cowie J. 1993. *Resolving Lexical Ambiguity*, in Memoranda in Computer and Cognitive Science MCCS-93-260, Computing Research Laboratory, New Mexico State University. Las Cruces, New Mexico.

Hearst M. 1991. *Towards Noun Homonym Disambiguation Using Local Context in Large Text Corpora*, in Proceedings of the Seventh Annual Conference of the UW Centre for the New OED and Text Research. Waterloo, Ontario.

McRoy S. 1992. *Using Multiple Knowledge Sources for Word Sense Discrimination*, Computational Linguistics, vol. 18, num. 1.

Miller G. 1990. *Five papers on WordNet,* Special Issue of International Journal of Lexicogrphy 3(4). 1990.

Miller G. Leacock C., Randee T. and Bunker R. 1993. *A Semantic Concordance,* in proceedings of the 3rd DARPA Workshop on Human Language Technology, 303-308, Plainsboro, New Jersey.

Miller G., Chodorow M., Landes S., Leacock C. and Thomas R. 1994. *Using a Semantic Concordance for sense Identification,* in proceedings of ARPA Workshop on Human Language Technology, 232-235.

Rada R., Mili H., Bicknell E. and Blettner M. 1989. *Development an Applicationof a Metric on Semantic Nets,* in IEEE Transactions on Systems, Man and Cybernetics, vol. 19, no. 1, 17-30.

Resnik P. 1995. *Disambiguating Noun Groupings with Respect to WordNet Senses,* in Proceedings of the Third Workshop on Very Large Corpora, MIT.

Richardson R., Smeaton A.F. and Murphy J. 1994. *Using WordNet as a Konwledge Base for Measuring Semantic Similarity between Words*, in Working Paper CA-1294, School of Computer Applications, Dublin City University. Dublin, Ireland.

Rigau G. 1994. *An experiment on Automatic Semantic Tagging of Dictionary Senses,* WorkShop "The Future of Dictionary", Aix-les-Bains, France. published as Research Report LSI-95-31-R. Computer Science Department. UPC. Barcelona.

Rigau G. and Agirre E. 1996. *Linking Bilingual Dictionaries to WordNet*, in proceedings of the 7th Euralex International Congress on Lexcography (Euralex'96), Gothenburg, Sweden, 1996.

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. Arlington, Virginia.

Voorhees E. 1993. *Using WordNet to Disambiguate Word Senses for Text Retrival*, in proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Developement in Information Retrieval, pages 171-180, PA.

Wilks Y., Fass D., Guo C., McDonal J., Plate T. and Slator B. 1993. *Providing Machine Tractablle Dictionary Tools,* in Semantics and the Lexicon (Pustejovsky J. ed.), 341-401.

Yarowsky, D. 1992. *Word-Sense Disambiguation Using Statistical Models of Roget's Categories Trained on Large Corpora,* in proceedings of the 15th International Conference on Computational Linguistics (Coling'92). Nantes, France.