From: AAAI-99 Proceedings. Copyright © 1999, AAAI (www.aaai.org). All rights reserved.

Word Sense Disambiguation for Information Retrieval

Ozlem Uzuner, Boris Katz, Deniz Yuret

Massachusetts Institute of Technology
Artificial Intelligence Laboratory
545 Technology Sq. NE43-825
Cambridge, MA 02139
{ozlem,boris,deniz}@ai.mit.edu

Despite their increasing importance as data retrieval tools, most Information Retrieval (IR) systems are deficient in precision and recall. Lack of disambiguation power is one reason for the poor performance of these systems. Correctly disambiguating and expanding a query with intended synonyms before retrieval may improve the performance.

We use the local context of a word to identify its sense. In our case, the local context of a word is the ordered list of words from the closest content word on each side of the target word up to the target word which is expressed as a placeholder. For example, in "the jury had been charged to investigate reports of irregularities in the primary…" the right-side local context of "charged" is "X to investigate". Due to this definition of context, words used in the same context (called selectors) most of the time have related senses. So, *an occurrence of a word and its synonym belong to the same sense if they have similar local contexts*.

We use WordNet (Miller 1990) and selectors extracted from Associated Press articles (Yuret 1998) for disambiguation. Selectors help us find the right WordNet synset (synonyms of only one sense) of a word in its context. Figure 1 shows the process of extracting selectors of the word *charge* in the sentence fragment "the jury had been charged to investigate reports of irregularities in the primary…". The final tally of selectors for this fragment is shown in Table 1.

Figure 2 shows the same process for the sentence fragment "the company was charged for towing the car…". The final tally of selectors for this fragment is shown in Table 2.

Selector	Frequency
Appointed	52
Assigned	28
Established	20
Hired	16
…	…

Table 1: **Final tally of selector frequencies for Figure 1.**

Copyright © 1999, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Once the selectors are extracted, the appropriate WordNet synset is selected by comparing the selectors against the ambiguous word's WordNet synsets. This comparison matches *charge* in the first context (Figure 1) to WordNet sense 4 (Table 3) and in the second context (Figure 2) to WordNet sense 3, thus correctly identifying the intended WordNet senses for this word in each of the given contexts.

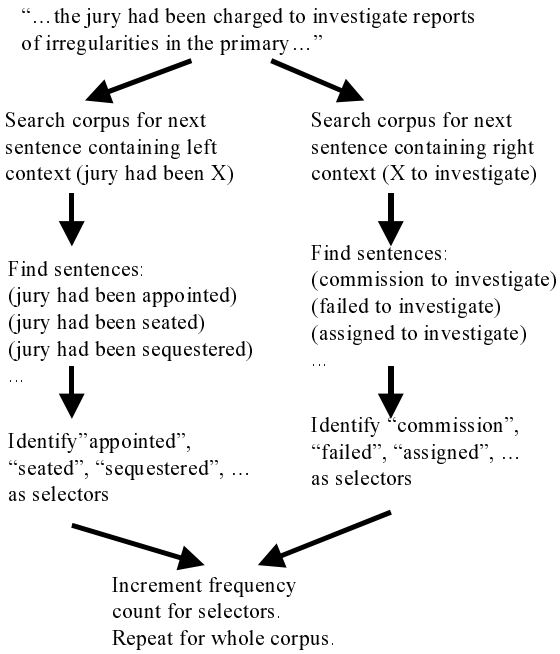

Figure 1: Identification of selectors in the given context.

Selector	Frequency
Vessel	2
Equipped	1
Billed	1
Charged	1
…	…

Table 2: Final tally of selector frequencies for Figure 2.

One of the most commonly used data sets for disambiguation evaluation is the Semcor, a subset of the

Brown corpus. In Semcor, each word in a sentence is tagged with its correct part-of-speech and sense number taken from WordNet. We use WordNet senses of the words in evaluating the performance of our system.

Sense Number	Sense
Sense 1	Charge, bear down
Sense 2	Charge, accuse
Sense 3	Charge, bill
Sense 4	Appoint, charge
…	…

Table 3: Some senses of *charge* as they appear in WordNet.

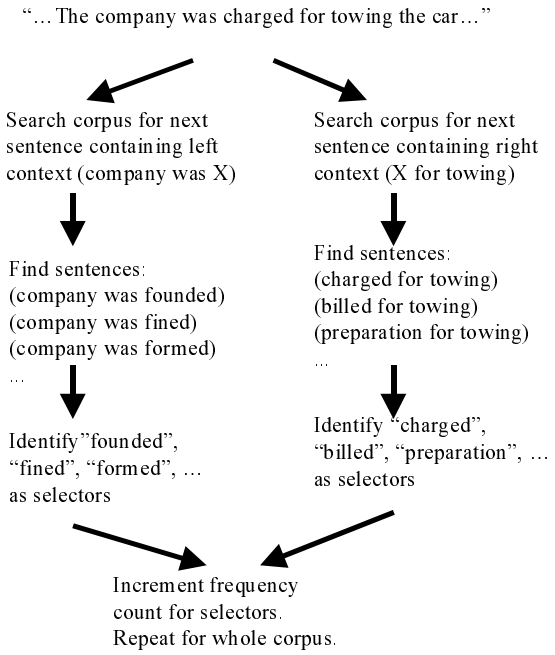

Figure 2: Identification of selectors in the given context

Disambiguation Performance

The "most frequent heuristic" has been accepted as the baseline for measuring performance of WSD algorithms. Since WordNet orders the synonyms of each word from the most to the least frequent, the performance of the "most frequent" heuristic is easily evaluated by assigning each ambiguous content word sense 1 of the first part-of-speech that it appears in when the parts of speech are checked in the order: noun, verb, adjective, and adverb. When tested only on words with more than one sense, the accuracy of the "most frequent" heuristic on this corpus was approximately 54%. In comparison, guessing only the senses of words with more than one sense, our algorithm achieved an accuracy of 45%.

Performance of SMART

To evaluate the effect of disambiguation on IR, we tested the performance of Smart (Buckley et. al., 1995) with and without disambiguation. These tests were done in two different ways: In the first, the original queries were expanded with the identified potential synonyms. So, each query was replaced by a much longer version of it. In the second test, the queries were replicated and reproduced by replacing only the target word by one of its identified synonyms. This was done for all content words in the query, creating hundreds of queries from one. The retrieval was performed on CACM, CISI and CRAN collections. In both cases, the performance of the system was worse.

Conclusion

The disambiguation performance of our system can be improved in several ways. First of all, using another disambiguation source instead of WordNet will help us avoid the problems caused by fine-grained word senses present in this dictionary. In addition, changing the definition of context and using informativeness of selectors as a weighting criterion can improve the disambiguation performance.

These results of IR tests showed that although in some cases the expansion of a query with synonyms helps, especially for short queries the disambiguation accuracy is low. Naturally, retrieval results are directly influenced by the disambiguation performance. This is because incorrect disambiguation not only excludes correct synonyms from the query but it also introduces incorrect information to it. This has deleterious effects on retrieval performance (Voorhees 1993). So, low disambiguation performance is probably the main cause of poor IR performance (Voorhees 1993, Sanderson 1994).

References

Buckley C., Singhal A., Mitra M., Salton G., 1995. New Retrieval Approaches Using SMART: TREC 4. *Proceedings of the 3rd Text Retrieval Conference,* NIST Special Publ.

Miller, G. A. 1990. WordNet: An online lexical database. *International Journal of Lexicography*, 3(4):235-312.

Sanderson, M. 1994. Word disambiguation and information retrieval. *Proceedings of ACM SIGIR Conference*.

Uzuner, O. 1998. Word-sense Disambiguation Applied to Information Retrieval. M.Eng Thesis, Dept. of Electrical Engineering and Computer Science, MIT.

Voorhees, E. M. 1993. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *SIGIR '93, Proceedings of the sixteenth annual international ACM SIGIR conference on Research and Development in Information Retrieval,* pages 171-180.171-180.

Yuret, D. 1998. Discovery of Linguistic Relations Using Lexical Attraction. Ph.D. diss., Dept. of Computer Science, MIT.