Word Sense Disambiguation for Information Retrieval

Ozlem Uzuner, Boris Katz, Deniz Yuret

MIT Artificial Intelligence Laboratory
545 Technology Sq Cambridge MA 02139
{ozlem,boris,deniz}@ai.mit.edu

Despite their increasing importance as data retrieval tools, most Information Retrieval (IR) systems do not have high precision and recall. Lack of disambiguation power is one reason for the poor performance of these systems. Correctly disambiguating and expanding a query only with intended synonyms before retrieval may improve their performance.

We use the local context[1] of a word to identify its sense. Words used in the same context (called selectors) often 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 *charge* in a given sentence. The final tally of identified selectors is shown in Table 1.

Selector	Appointed	Assigned	Established	Hired
Frequency	52	28	20	16

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

Once the selectors are extracted, the appropriate WordNet synset is selected by comparing the selectors against the ambiguous word's WordNet synsets.

Semcor, a subset of Brown corpus, is commonly used for disambiguation evaluation. In Semcor, each word is tagged with its correct part of speech and sense number taken from WordNet.

The "most frequent heuristic" is accepted as the baseline for measuring performance of WSD algorithms. When tested only on words with more than one sense, the accuracy of the "most frequent" heuristic on Semcor was approximately 54%. In comparison, our algorithm achieved an accuracy of 45%.

To evaluate the effect of disambiguation on IR, we tested the performance of Smart (Buckley et. al., 1995). These tests were done in two ways: In the first, the original queries were expanded with the identified potential synonyms. In the second test, the queries were replicated by replacing only the target word with one of its identified synonyms. This was done for all content words in the query. Retrieval tests were done on CACM, CISI and CRAN collections. In all cases, the performance of the system became worse.

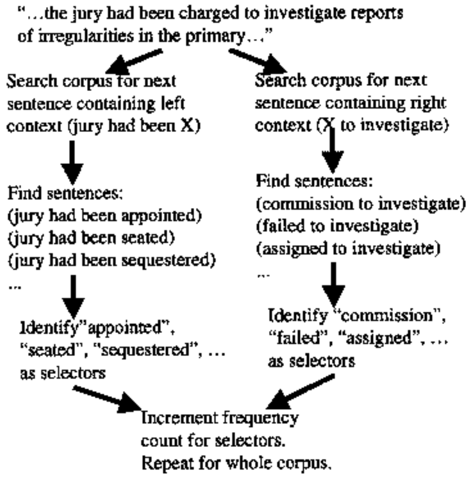
"...the jury had been charged to investigate reports of irregularities in the primary..."
Search corpus for next sentence containing left context (jury had been X)
Search corpus for next sentence containing right context (X to investigate)
Find sentences: (jury had been appointed) (jury had been seated) (jury had been sequestered) ...
Find sentences: (commission to investigate) (failed to investigate) (assigned to investigate) ...
Identify"appointed", "seated", "sequestered", ... as selectors
Identify "commission", "failed", "assigned", ... as selectors
Increment frequency count for selectors. Repeat for whole corpus.

Figure 1: Identification of selectors in a given context

Low disambiguation performance is probably the main cause of poor IR performance (Voorhees 1993, Sanderson 1994). Improving disambiguation performance through use of a different lexical source, or through use of different context definitions can improve IR performance as well.

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. *Int'l 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 EECS, MIT.

Voorhees, E. M. 1993. Using WordNet to Disambiguate Word Senses for Text Retrieval. *Proceedings of ACM SIGIR Conference.,* pages 171-180.

Yuret, D. 1998. Discovery of Linguistic Relations Using Lexical Attraction. Ph.D. , Dept. of Comp. Sci., MIT.

[1] Local context of a word is the ordered list of words from the closest content word on each side up to the target word 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".