Word Sense Disambiguation For Information Retrieval

Boris Katz, Ozlem Uzuner & Deniz Yuret

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: Despite the increasing importance of Information Retrieval (IR) systems as data retrieval tools, the performance of most of these systems has not yet reached a satisfactory level. Word sense ambiguity is one of the reasons for their poor performance. Overcoming this problem may improve IR performance.

Motivation: Documents related to an IR query sometimes contain only the synonyms of the query words instead of the query words themselves. A simple IR system with no knowledge of synonyms fails to recognize the relevance of these documents to the query. So, we can improve recall of IR systems by considering the synonyms of the query words as a part of the IR query. However, only relevant synonyms of the query words in the given context contribute useful information to the query. We can identify these relevant synonyms with the help of a disambiguation algorithm.

Previous Work: So far, there has been conflicting information on the effect of WSD on IR. While Schutze and Pedersen [4] describe a sense disambiguator that improves the precision of an IR system by 4%, Sanderson [3] presents results which show that disambiguation (unless at least 90% accurate) makes IR performance worse.

Approach: We use the local context[1] of a word to identify its sense. Due to this definition of context, words used in the same context (called *selectors*) most of the time have similar or related meanings. That is, an occurrence of a word and its synonym often belong to the same sense if they have similar local contexts.

We use WordNet [2] and selectors extracted from Associated Press Articles [7] to find the appropriate synset of a word in its context. The figure below shows some of the selectors of the word "charge" in two different sentences. The examples show that contexts play an important role in finding selectors which enable us to identify the correct sense of an ambiguous word.

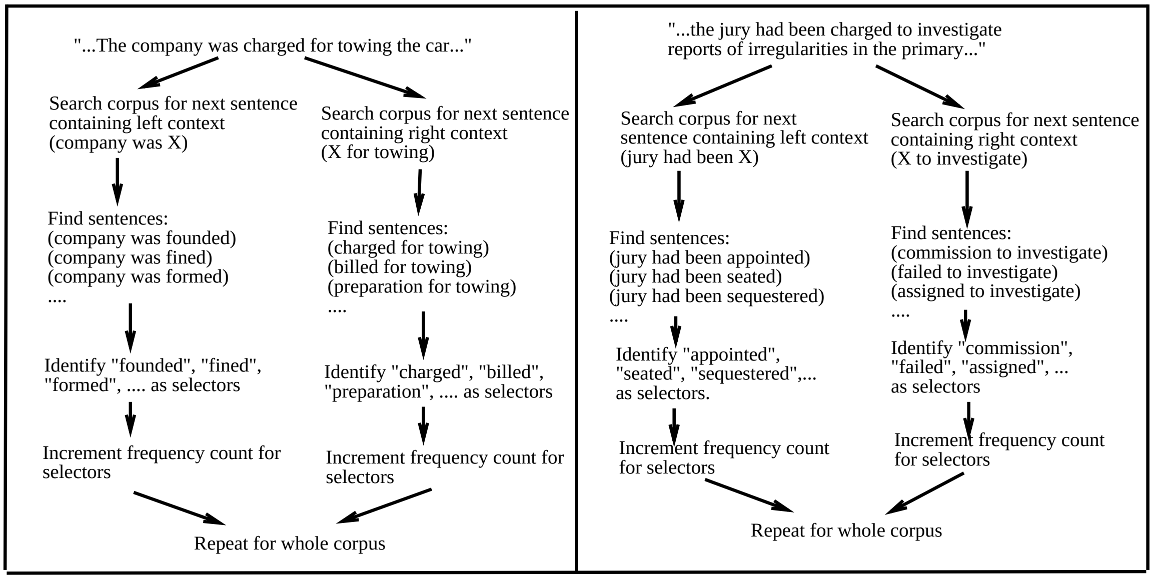
Difficulty: Correctly disambiguating words is a difficult problem. When restricted to available on-line dictionaries like WordNet, it is sometimes impossible even for human beings to pick the right sense for words. Expecting a machine to resolve such ambiguities is not reasonable. But, a good online dictionary with example uses of words in each of their possible senses can allow a machine to disambiguate words accurately. Such dictionaries are not yet available.

Impact: The disambiguation algorithm was tested on the Semcor corpus where each word is tagged with its correct part-of-speech and sense number from WordNet. On this corpus, the accuracy of our disambiguator was almost 60% excluding words which have only one sense. When incorporated into the IR system SMART [1], the disambiguation did not improve performance. Although in some cases the expansion of the query with synonyms helped, especially for short queries the disambiguation accuracy was low. Incorrect disambiguation not only excludes correct synonyms from the query but it also introduces incorrect information to it reducing retrieval performance [6, 3]. Although 60% accuracy is not insignificant for an unsupervised algorithm which tries to disambiguate *any* content word in a context, the performance of this disambiguator can be improved with the use of better online dictionaries with less fine-grained sense distinctions. Improving disambiguation performance can help IR.

Future Work: Previous research suggests that using cross-lingustic information for disambiguation performs better than single language disambiguation. There is a lot of contextual information which is lost by trying to disambiguate an ambiguous word whose context is also ambiguous. Cross linguistic information can, to a certain extent, disambiguate the context of the ambiguous word and help the disambiguation of the word itself. Our future work will focus on development of such a system which, we expect, will significantly improve performance.

[1] Local context is the ordered list of words from the closest context 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".

Disambiguation of “charged” in two different contexts:


"...The company was charged for towing the car..."
Search corpus for next sentence containing left context (company was X)
Search corpus for next sentence containing right context (X for towing)
Find sentences: (company was founded) (company was fined) (company was formed) ....
Find sentences: (charged for towing) (billed for towing) (preparation for towing) ....
Identify "founded", "fined", "formed", ... as selectors
Identify "charged", "billed", "preparation", ... as selectors
Increment frequency count for selectors
Increment frequency count for selectors
Repeat for whole corpus
"...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
Increment frequency count for selectors
Repeat for whole corpus

Final tally of selector frequencies:

SELECTOR	FREQUENCY
vessel	2
equipped	1
billed	1
charged	1
...	...

SELECTOR	FREQUENCY
appointed	52
assigned	28
established	20
hired	16
...	...

Some senses of *charge* as they appear in WordNet:

Sense 1	Sense 2	Sense 3	Sense 4	Sense 5	...
charge, bear down	charge, accuse	charge, bill	appoint, charge	charge	...

Comparing the selectors of the input word against the WordNet synsets matches input sense 1 to WordNet sense 3 and input sense 2 to WordNet sense 4; the algorithm has selected the most appropriate WordNet senses.

Research Support: This research is funded by DARPA under contract number F30602-00-1-0545 and administered by the Air Force Research Laboratory.

References:

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

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

[3] M. Sanderson. Word-sense Disambiguation and Information Retrieval. *In proceedings of ACM-SIGIR*,1994.

[4] H. Schutze and J. Pedersen. Information Retrieval based on Word Senses. In *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval*, U. of Nevada at Las Vegas. 1995.

[5] Ozlem Uzuner. Word-sense Disambiguation Applied to Information Retrieval. M.Eng thesis, 1998, MIT.

[6] E. M. Voorhees. Using WordNet to Disambiguate Word Senses for Text Retrieval. In *SIGIR ’93, Proceedings of the 16th Annual International ACM SIGIR conference of Research and Development in Information Retrieval*, pages 171-180. 1993.

[7] Deniz Yuret. Discovery of Linguistic Relations Using Lexical Attraction. Ph.D. dissertation, 1998, MIT.