Using Multiple Knowledge Sources for Word Sense Identification

Haodong Wu, Teiji Furugori
Department of Computer Science
University of Electro-Communications

1. Introduction

Word sense disambiguation (WSD) is necessary in the areas of natural language processing (NLP) such as machine translation, information retrieval, natural language understanding, word clustering, text processing and speech synthesis. The process involved in WSD is to identify the meaning of a particular word in a particular context.

We in this paper propose a word sense identification method that uses a wide-variety of information, pertinent to parts of speech, word co-occurrences, contextual similarities, syntactic relations, contained in corpora and a machine-readable dictionary (MRD). We then demonstrate our method in a disambiguation experiment.

2. Work on Word Sense Disambiguation

There have been many approaches on WSD. In early work in WSD, researchers took a pure AI approach: they manually built up a knowledge base for words and described their senses in various linguistic usages (e.g., Hirst, 1987, Small & Rieger, 1982); later, others used MRD-based approach: they selected the proper sense of a polysemous word using MRDs (e.g., Lesk, 1986; Walker, 1987); and recently, many employ corpus-based approach and use various of corpora for WSD (e.g., Gale et al. 1993; Hiro et al. 1996; Yarowsky, 1992).

Obviously, a hand-crafted knowledge base of any sort has a difficulty in scaling up and its application is restricted to small domains (e.g., Isabelle, 1984). The MRD-based approach in principle is widely applicable, but the dictionary definitions of a word are often too short and too uneven to cover necessary collocations or co-occurrences of words that are essential to capture the context for WSD. Corpus-based approaches to WSD have generality and flexibility but suffer from the sparse data problem.

3. Corpus-based Word Sense Identification

Like many recent studies, we use a corpus-based approach for WSD. Unlike others, however, we base our method on a number of syntactic and semantic cues in the disambiguation process that come from our linguistic intuitions and observations:

Word Co-occurrences or Collcations Co-occurrences or collcations in certain syntactic relations can determine the meaning of a word in context. They offer immediate and obvious predictions. For example, in *hard candy*, a syntactic relation of adjective followed by a noun, it becomes obvious from the noun *candy* as the word sense indicator that the meaning of *hard* is 'not soft' but not 'difficult'.

Contextual Similarities Contextual similarities play an important role for lexical disambiguation. Words are considered to be similar if they appear in similar contexts; contexts are considered to be similar if they contain similar words.

Parts of Speech Parts of speech (POS) have a deciding power in the meaning of a word in context. The word *round*, for instance, takes different syntactic functions and different meanings accordingly: it may be

an adjective (e.g., a *round* table), an adverb (e.g., to work the day *round*), a noun (e.g., a *round* of cheese), or a verb (e.g., to *round* out his interests).

We try to get syntactic and semantic cues, or necessary and relevant information from the EDR English Corpus[1] and a parser. Our disambiguator is to learn to discriminate the correct sense from the appearances of an ambiguous word w with its senses $s_1,\ldots,s_n$, in a given context, using the examples that contains the target word w in the EDR corpus.

No corpus-based methods are free from that the sparse data problem. Neither smoothing methods nor class-based methods are good enough for overcoming data sparseness for WSD purpose (Ide & Véronis, 1998). Here, We use a similarity-based method to solve this problem.

3.1 Computation of Similarity

Similar words tend to appear in similar contexts, and their textual proximity to the polysemous word w is indicative of the sense of w. How then can we measure the contextual similarity?

Nurse and *hospital* in an example seem conceptual similar. But semantic networks like WordNet (Miller, 1995) and the EDR Concept Classification Dictionary (EDR, 1993) are not designed to capture contextual similarity: they have no common ancestor in them and hence their similarity is 0, while *nurse* and *teacher* in them are quite similar, because they are included in a conceptual hierarchy *of professionals, human beings, animals,* and so on.

We may be able to find contextually similar words by using mutual information (MI), a common measure for word association (Church & Hanks, 1990). Assume we need to determine the degree of similarity between two words, w_1 and w_2. We consider the two words are conceptually similar if they have similar mutual information, I, with some other word w in the local context (a sentence). We define the contextual similarity between w_1 and w_2 by:

$$sim(w_1, w_2) = \frac{\sum_{w \in lexicon} (\min(I(w, w_1), I(w, w_2)) + \min(I(w_1, w), I(w_2, w)))}{\sum_{w \in lexicon} (\max(I(w, w_1), I(w, w_2)) + \max(I(w_1, w), I(w_2, w)))}$$

3.2 The Algorithm for Word Sense Identification

Given a polysemous word in text, the task of word sense identification is to choose the 'correct' meaning or sense from a set of probable candidates. An algorithm we developed for this is:

Let the ambiguous word be w.

1. Collect the example sentences that contains w from the EDR Corpus. Let it be the training data.
2. Replace personal pronouns with the word *human*.
3. Eliminate the function words (namely, the words other than nouns, verbs, adjectives and adverbs) in the test data.
4. Cluster closely related senses (or concepts) defined in EDR English Word Dictionary into its upper concept by using EDR Concept Classification Dictionary[2].

---

[1] The EDR English Corpus contains about 220,000 sentences with hand-crafted tagged information of morphologic, syntactic and semantic (EDR, 1993).

[2] The word senses defined by the EDR English Word Dictionary are often too detail for our WSD purpose. To deal with this problem, we use the EDR Concept Classification Dictionary to cluster the detail senses (concepts) into the upper class concept.

5. Assign weight for the importance of the co-occurrence *w* and wi (wi is a word in the window of 50 words[3]) according to both our linguistic and experimental observations:
   If there is a verb-object relation between *w* and wi, assign 1.0 for the weight weight(*w*,wi).
   If there is an adjective-noun relation between *w* and wi, assign 1.0 for weight(*w*,wi).
   If the preposition phrase {prep wi} is attached to *w*, assign 0.8 for weight(*w*,wi).
   If there is a verb-adverb between *w* and wi, assign 0.8 for weight(*w*,wi).
   If there is a noun-noun relation between *w* and wi, assign 0.7 for weight(*w*,wi).
   If there is a subject-verb between *w* and wi, assign 0.6 for weight(*w*,wi).
   In case of no any syntactic relation between *w* and wi: if both *w* and wi are nouns, assign 0.5 for weight(*w*,wi); if *w* is a verb and wi is a noun, assign 0.3 for weight(*w*,wi); if *w* is an adverb or an adjective, assign 0.1 for weight(*w*,wi); assign 0.2 for weight(*w*,wi) in any other case.
6. For the *jth* sense of *w*, termed *wj*, compute the strength of association between *wj* and wi. Choose a set of the N (e.g., 20) words , *S*, which are most similar to wi :

$$MI(w_j, w_i) = \frac{1}{N} \sum_{w_k \in S} \left(\frac{P(w_i, w_k)}{P(w_i)P(w_k)} \right)$$

   where P(wj) and P(wk) are the probabilities of the *jth* sense of the ambiguous word *w* and the word wk in the local context. Compute the score of the sense in the context using the following formula:

$$SP(w_j) = \sum_{i=1}^{n} (weight(w_j, w_i) MI(w_j, w_i))$$

   here *SP(wj)* is the score which denotes the strength of the *jth* sense of *w* with the context (w1,...,wn).
7. Select the k*th* sense of *w* where *SP(wk)* is the largest one as the correct sense of *w*.

4. Experiment and Result

We have tested our method on a total of 433 examples of three polysemous words: the verb *produce,* the noun *bank,* and the adjective *old* taken from the Brown Corpus and LOB corpus. Table 1 shows the results. The average success rate of our method is 89.8%. This is as good as or better than the ones in other studies (e.g., Karov & Edelman, 1998; Hiro et al. 1996; Yarowsky, 1992), although direct comparisons are impossible since the experiments vary in conditions.

It is proven that the use of contextually similar words is of great help to removing the sparse data problem and improving the success rate. When we used the contextual words without using the contextually similar words, the overall success rate was 81.3%. This number is 8.5% lower than the similarity-based approach we employed. The use of weights on different syntactic relations is effective also in our work. When we did not use the various weights in the algorithm, the success rate was 82.7%.

5. Concluding Remark

We proposed a method for word sense identification. We used an annotated corpus and a machine readable dictionary to acquire various kinds of information in syntactic relations, word co-occurrences, parts of speech, contextual similarities that can efficiently improve the performance in word sense identification. The use of the

[3] See Hiro et al. (1996) for details.

Table 1
The algorithm's performance on the three test words

Word	Senses	Sample Size	% Correct per Sense	% Correct Total
produce	bring forth	23	82.6	89.7
	make or create	87	92.0	
	bring about	35	88.6	
bank	financial institution	103	91.2	90.9
	slope land near water	82	93.9	
	things arranged in a row	13	76.9	
old	not young	43	88.2	86.6
	not new	39	87.2	
	previous	7	71.4	

contextually similar words in place of the contextual words of the ambiguous word has dramatically reduced the data sparseness and proven effective in WSD.

With its better performance, our method may be employable to many applications such as information retrieval and hypertext navigation. We believe that the performance can be improved further by using larger corpora that would contribute to find proper contextual similar words and find more features for WSD.

References

Church, K.W., and Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16, pages 22-29.

Gale, W.A., Church K.W., and Yarowsky, D. (1993). A method for disambiguating word senses in a large corpus. *Computers and Humanities*, 26, pages 415-439.

Hiro, K., Wu, H., and Furugori, T. (1996). Word-sense disambiguation with a corpus-based semantic network. *Journal of Quantitative Linguistics*, 3, pages 244-251.

Hirst, G. (1987). *Semantic interpretation and the resolution of ambiguity*. Cambridge University Press.

Ide, N. and Véronis, J. (1998). Introduction to special issue on word sense disambiguation: The state of the art. *Computational Linguistics*, 24, pages 1-40.

Isabelle, P. (1984). Machine translation at the TAUM group. In M. King (Ed.), *Machine translation today: the state of the art* (pages 247-277). Edinburgh University Press.

Japan Electronic Dictionary Research Institute, Ltd. (1993). *EDR electronic Dictionary Specifications Guide.*

Karov, Y., and Edelman, S. (1998). Similarity-based word sense disambiguation. *Computational Linguistics*, 24, pages 41-59.

Lesk, M. (1986). Automatic sense disambiguation: How to tell a pine cone from an ice cream cone. In *Proceedings of the 1986 SIGDOC Conference*. New York: Association for Computing Machinery.

Miller, G.A. (Ed.) (1995). WordNet: An on-line lexical database. Communications of ACM, 38(11).

Small, S., and Reiger, C. (1982). Parsing and comprehending with word experts (A theory and its realization). In W. Lehnert, and M. Ringle (eds.), *Strategies for natural language processing* (pages 89-147). Hillsdale NJ: Lawrence Erlbaum.

Walker, D. (1987). Knowledge resource tools for accessing large text files. In S. Nirembe (Ed.), *Machine translation: Theoretical and methodological issues*. Cambridge University Press.

Yarowsky, D. (1992). Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. In *Proceedings of COLING-92*.