Word Sense Disambiguation Based on Better-Way Principle

Manabu Okumura and **Shigeru Fujiwara**
School of Information Science,
Japan Advanced Institute of Science and Technology
Tatsunokuchi, Ishikawa, 923-1292, Japan
Tel: +81-761-51-1216, Fax: +81-761-51-1149
E-mail: {oku,shigeru}@jaist.ac.jp

Graeme Hirst
Department of Computer Science,
University of Toronto
Toronto, Canada M5S 3G4
E-mail: gh@cs.toronto.edu

Abstract

Word sense disambiguation has recently received more attention, and many approaches have been proposed. In fact, their strategies are varied, most of the previous works use only the information from the sentence to be analyzed and the knowledge of the system(reader/hearer) itself. Taking into account the writer's/speaker's viewpoint, however, might provide a new approach for analyzing a sentence and improve the performance of the disambiguation. As Sperber and Wilson pointed out, the writer would aim to minimize the number of possible interpretations, if he/she observes the 'cooperative principle.' Therefore, with the assumption that the writer obeys the principle, we can adopt the approach for word sense disambiguation that any senses can be eliminated if they are irrelevant from the writer's viewpoint. In this paper, we present such an approach for word sense disambiguation. It is based on the principle that we call 'Better-Way Principle.' We also report the results of preliminary experiments with our naive implementation and show what degree the performance will be improved.

1 Introduction

Word sense disambiguation has recently received more attention, and many approaches have been proposed(Hirst, 1987; Gale et al., 1992; Yarowsky, 1992; Charniak, 1993; Resnik, 1995; Ng and Lee, 1996). In fact, their strategies are varied(example-based, statistics-based, or knowledge-intensive), most of the previous works use only the information from the sentence to be analyzed and the knowledge of the system(reader/hearer) itself.

Taking into account the writer's/speaker's viewpoint, however, might provide a new approach for analyzing a sentence and improve the performance of the disambiguation.

As Sperber and Wilson(Sperber and Wilson, 1986) pointed out, the writer would aim to minimize the number of possible interpretations, if he/she observes the 'cooperative principle'(in 'maxims of conversation')(Grice, 1975). He/She has to choose the most relevant from possible words or expressions. The writer would make it as easy as possible for the reader to understand him/her. The reader can eliminate any interpretations that are incompatible with the assumption that the writer is obeying the principle.

Therefore, with the assumption that the writer obeys the principle[1], we can adopt the approach for word sense disambiguation that any senses can be eliminated if they are irrelevant from the writer's viewpoint. This approach for analyzing a sentence can be considered as a counterpart of revision(Cline and Nutter, 1994; Robin and McKeown, 1996; Inui et al., 1996; Callaway and Lester, 1997), where the reader's viewpoint is taken into account in text generation.

In this paper, we present such an approach for word sense disambiguation. It is based on the principle that we call 'Better-Way Principle.' In the next section, we explain what the idea of 'Better-Way Principle' is. In section three, we describe a naive implementation of 'Better-Way Principle' with empirical methods. We also report the results of preliminary experiments with our naive implementation and show what degree the performance will be improved.

2 Better-Way Principle

Consider, first, the following Sperber and Wilson's example(Sperber and Wilson, 1986).

George has a big cat.

In this sentence, the word 'cat' is ambiguous in that it can refer either to the domestic cat or to

[1]We make the assumption, not in the sense that the writer never violates the Gricean maxims, but in the sense that he/she does not intentionally violate them.

any animal of the species. Therefore, the sentence has (at least) the following two interpretations.

(1) George has a big domestic cat.
(2) George has an animal such as a tiger, a lion, etc.

However, since more relevant expression such as the following can be found to express (2), Sperber and Wilson consider the first interpretation to be better with their principle of relevance.

George has a tiger.

Hirst(Hirst, 1992) shows the following scope-ambiguous sentence,

Two examiners marked six scripts.

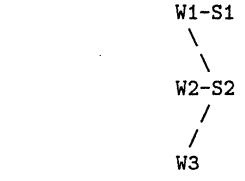
and calls it an 'icky' sentence, since any interpretations of the sentence can be readily expressed with an unambiguous sentence, as follows.

Two examiners each marked six scripts.
Two examiners marked a total of six scripts.
Two examiners marked six scripts between them.

Therefore, the above sentence is said to be never the most natural way to express either interpretation. Hirst also presents a principle of ambiguity resolution, the 'Better-Way Principle': assume that the speaker avoids icky sentences. He said the system can sometimes reduce much word sense ambiguity if it can assume that a sentence is not icky.

As the above two illustrations indicate, Sperber and Wilson, and Hirst discuss the same point. Even if a sentence has multiple interpretations, the interpretations that can be expressed with better expressions can be considered as less appropriate, and the ambiguity can be reduced. Using Hirst's term, we also call this principle for ambiguity resolution, the 'Better-Way Principle.' Hereafter, we concentrate on the cases of word sense disambiguation.

What senses are then inappropriate with the 'Better-Way Principle'? Consider, for example, the following simple situation that word W1 has two senses S1 and S2, but sense S2 can be also expressed with other words W2 and W3.



In case the correct sense of word W1 should be identified, even if senses S1 and S2 are almost equally appropriate from the co-occurrence with senses of other words in the sentence, S2 might be less appropriate if W1 is used less frequently to express S2[2].

Note that in this example, the correct disambiguation might be difficult if two senses are almost equally appropriate with the information in a sentence and the knowledge within the system. However, taking into account the writer's viewpoint enables the disambiguation with the 'Better-Way Principle.'

In the next section, we present our approach for word sense disambiguation that is based on a naive implementation of 'Better-Way Principle' with empirical methods.

3 A Naive Model of Better-Way Principle

If most writers obey the 'cooperative principle' in Gricean terms, the statistics in the text data can be considered to reflect such a tendency. Therefore, we think we can implement our 'Better-Way Principle' using the statistics from the corpus.

As we mentioned in the last section, the frequency of a word as a sense can be considered to reflect how appropriate the word is as the sense. The more frequent a word is used as a sense, the more appropriate the word is as the sense. Therefore, given the word W to be disambiguated, we think incorporating the probability of the word given a sense($p(W|S_i)$) into the score for selecting a sense can be the first implementation of our 'Better-Way Principle'[3].

Many scores have been proposed for word sense disambiguation(Gale et al., 1992; Leacock et al., 1993; Bruce and Wiebe, 1994; Ng and Lee, 1996; Yarowsky, 1992; Rigau et al., 1997; Resnik, 1993; Ribas, 1995). Among them, as the first score to combine with the above conditional probability, we adopt one of the concept(sense) co-occurrence score(Resnik, 1993; Ribas, 1995), mutual information. The mutual information between two concepts n and v is defined as follows:

$$I(n;v) = log\frac{p(n,v)}{p(n) \times p(v)},$$

where $p(n,v)$ is the probability of the co-occurrence of n and v.

Given the simple sentence that consists of a verb Wv and nouns Wn_i modifying the verb, we

[2]It is because the only way to express S1 is to use W1. We think that the low frequency of a word as a sense is originated in the inappropriateness of the word as the sense in this example. In more strictly, however, which sense is more appropriate with the 'Better-Way Principle' cannot be decided before the relevance of co-occurrence of words to express the combination of senses in a sentence should be taken into account.

[3]We admit that the implementation is naive. However, we think more elaborate implementation with empirical methods is now difficult because of the 'data sparseness' of the available corpus.

select the senses Sv_j for the verb and Sn_{ik} for the i-th noun that maximize the following score.

$$(3) Score = \prod_i I(Sn_{ik}; Sv_j) \times p(Wv|Sv_j) \times \prod_i p(Wn_i|Sn_{ik})$$

The formula (3) can be said to be similar to the one that is used for part-of-speech tagging with Hidden Markov Model(HMM)(Charniak, 1993):

$$\prod_{i=1}^{n} p(T_i|T_{i-1}) \times p(W_i|T_i).$$

This formula can be obtained by approximating the numerator of the conditional probability of a sequence of POS(part-of-speech) tags $T_1, T_2, \ldots, T_n$, given a sequence of words $W_1, W_2, \ldots, W_n$. They are similar in that they both consist of the co-occurrence score between tags(POS, sense) and the conditional probability of tags given words, because they originate in the common idea that the generation model for a sentence(a sequence of words) is used for analyzing.

In case where a Naive-Bayes classifier(Duda and Hart, 1973) is used for word sense disambiguation, the sense S_i with the maximum value of the following formula will be selected for a word W.

$$(4) p(S_i) \times \prod_j p(V_j|S_i)$$

This formula can be obtained by approximating the numerator of the conditional probability of a sense S_i, given a conjunction of feature values(clues) V_j in the context of W.

Regarding the occurrence of the word W itself as a clue for the disambiguation, the traditional Naive-Bayes classifier seems to already take into account the conditional probability of the word W given a sense S_i. We think, however, the previous works with the Naive-Bayes classifier have not used the occurrence of the word itself as a clue, because it has been thought to be meaningless. The conditional probability of a word given a sense can be useful for the disambiguation only in cases a common sense tag is used for senses of multiple words. We think those cases have been more and more because more corpora have been tagged with synsets of WordNet(Miller, 1990) or concept identifiers of EDR conceptual dictionary(EDR, 1996).

We think, however, that the 'Better-Way Principle' is a weak heuristic in that we assume most writers obey the 'cooperative principle'. Therefore, we normally use the mere mutual information to select senses, and use the formula (3) only when the mutual information cannot surely select senses[4]. We think such a usage of the formula (3) be effective because the correct disambiguation might be difficult in case senses cannot be surely selected, and the 'Better-Way Principle' can be a last resort for the disambiguation.

[4]We think, here, that senses cannot be surely selected when the difference of the score among the top sense and others is small.

4 Experiments

We make a preliminary experiment with our naive implementation mentioned in the last section. Both the test data for word sense disambiguation and the training data for calculating the score are extracted from EDR Japanese corpus(EDR, 1996), which word senses are manually tagged. To simplify the case, we extract 499 sentences to be disambiguated(the test data) that consist of only a noun and a verb. The training data consists of 321,712 sentences. The average number of senses for nouns and verbs in the test data is 3.61 and 9.02 respectively. Therefore, the average number of ambiguities for sentences is 32.56.

We select the pair of senses that maximizes the score(mutual information, or the formula (3)). We judge a sentence is correctly disambiguated when both senses are correctly identified. We regard the difference of the score(mutual information) among the top sense and others as small, if the difference is less than one third of the score of the top sense. In that case, the formula (3) will be used instead.

The accuracy with the mutual information for 499 sentences is 74.4%. The number of the sentences where the mutual information cannot surely select senses is 111 out of 499. As for these 111 sentences, the accuracy with the mere mutual information is 29.7%. By using the formula (3) instead, the accuracy can be improved up to 45.9%[5]. Therefore, using the formula (3) only when the mutual information cannot surely select senses can improve the accuracy from 74.4% to 78.2% for 499 sentences. Unfortunately, using the formula (3) for all of 499 sentences cannot improve the accuracy(72.6%).

From these results, we can say that our naive implementation of 'Better-Way Principle' is effective when the mutual information is not sure about the selection, though the test data is rather small.

5 Conclusion

We presented the approach for word sense disambiguation that is based on the principle that we call 'Better-Way Principle.' We also reported the results of preliminary experiments with our naive implementation of 'Better-Way Principle,' and showed what degree the performance is improved.

We think our method is promising, though only partially successful results can be obtained in the experiments so far. We think we need more thorough experiments with larger test data. Furthermore, we should try to implement 'Better-Way Principle' with other scores than the mutual information, and evaluate its effectiveness. Lastly,

[5]27 sentences change from 'wrong' to 'correct', while 9 change in the opposite.

we will need to devise more elaborate implementation of our 'Better-Way Principle.'

References

R. Bruce and J. Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proc. of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 139–145.

C. B. Callaway and J. C. Lester. 1997. Dynamically improving explanations: a revision-based approach to explanation generation. In *Proc. of the 15th International Joint Conference on Artificial Intelligence*, pages 952–958.

E. Charniak. 1993. *Statistical Language Learning.* MIT Press.

B. E. Cline and J. T. Nutter. 1994. Kalos - a system for natural language generation with revision. In *Proc. of the 12th National Conference on Artificial Intelligence*, pages 767–772.

R. Duda and P. Hart. 1973. *Pattern Classification and Scene Analysis.* Wiley.

EDR. 1996. Edr electronic dictionary version 1.5 technical guide. EDR TR2-007.

W.A. Gale, K.W. Church, and D. Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and Humanities*, 26:415–439.

H. P. Grice. 1975. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press.

G. Hirst. 1987. *Semantic Interpretation and the Resolution of Ambiguity.* Studies in Natural Language Processing. Cambridge University Press.

G. Hirst. 1992. Icky sentences and the better-way principle of ambiguity resolution. draft, July.

K. Inui, T. Tokunaga, and H. Tanaka. 1996. Dependency-directed control of text generation using functional unification grammar. *New Generation Computing*, 14(2).

C. Leacock, G. Towell, and E. Voorhees. 1993. Corpus-based statistical sense resolution. In *Proc. of the Human Language Technology Workshop*, pages 260–265.

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

H.T. Ng and H.B.. Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proc. of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 40–47.

P. Resnik. 1993. *Selection and Information: A Class based Approach to Lexical Relationships.* Ph.D. thesis, University of Pennsylvania.

P. Resnik. 1995. Disambiguating noun groupings with respect to wordnet senses. In *Proceedings of the 3rd Workshop on Very Large Corpora*, pages 54–68.

F. Ribas. 1995. On learning more appropriate selectional restrictions. In *Proc. of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, pages 112–118.

G. Rigau, J. Aserias, and E. Agirre. 1997. Combining unsupervised lexical knowledge methods for word sense disambiguation. In *Proc. of the 35th Annual Meeting of the Association for Computational Linguistics and 18th Conference of the European Chapter of the Association for Computational Linguistics*, pages 48–55.

J. Robin and K. McKeown. 1996. Empirically designing and evaluating a new revision-based model for summary generation. *Artificial Intelligence*, 85:135–179.

D. Sperber and D. Wilson. 1986. *Relevance.* Harvard University Press.

D. Yarowsky. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proc. of the 14th International Conference on Computational Linguistics*, pages 454–460.