自然言語処理 119－9
（1997. 5. 27）

Using Evidence that is both Strong and Reliable in Japanese Homograph Disambiguation

Hang LI and Jun-ichi TAKEUCHI

C&C Res. Labs., NEC

4-1-1 Miyazaki Miyamae-ku Kawasaki, 216 Japan

{lihang,tak}@sbl.cl.nec.co.jp

Abstract

We address the problem of resolving word sense ambiguities on the basis of corpus data, in particular the problem of resolving Japanese homograph ambiguities. Yarowsky has previously proposed for word sense disambiguation the use of the strongest piece of contextual evidence prepared in a decision list. We argue that in order to improve disambiguation accuracy, as well as to save space for storing pieces of evidence for disambiguation, it is necessary to use not only evidence which is strong but that which is also reliable. We propose a method of testing the reliability of a piece of contextual evidence on the basis of the Minimum Description Length principle. Our experimental results indicate that in word sense disambiguation it is better to use evidence which is both strong and reliable than simply to use that which is only known to be strong.

証拠の強さと信頼度を考慮した日本語同形異音語の読み分け

李 航, 竹内 純一

NEC C&C 研究所

〒216 川崎市宮前区宮崎 4-1-1

本稿では語彙曖昧性解消、特に日本語同形異音語の読み分けの一方式を提案する。Yarowskyの方式では、文脈に現れる読み分けのための証拠をその強さの順に並べて決定リストを構成する。この決定リストを用いると、新しく現れる同形異音語の読みは、もっとも強い証拠で判断される。しかし、証拠の強さだけでは、読みを正しく判断できないことがある。本稿では、証拠の強さだけでなく、その信頼度をも考慮して決定リストを学習する方法を提案する。実験によると、本方式により読み分けの正解率が改善された。

1 Introduction

A number of tasks in natural language processing fall into the category of word sense disambiguation (Yarowsky, 1993). These include homograph disambiguation in speech synthesis, word selection in machine translation, and spelling correction in document processing. Many methods have been proposed to address the difficulties involved in these tasks and a number of them have proved to be quite effective (Black, 1988; Brown et al., 1991; Bruce and Wiebe, 1994; Gale et al., 1992; Golding and Roth, 1996; Golding and Schabes, 1996; Guthrie et al., 1991; Leacock et al., 1993; Lewis and Gale, 1994; McRoy, 1992; Niwa and Nitta, 1994; Ng and Lee, 1996; Schutze, 1993; Voorhees et al., 1995; Yarowsky, 1992; Yarowsky, 1993; Yarowsky, 1994; Yarowsky, 1995).

Among these methods, that proposed by Yarowsky (Yarowsky, 1994), which merely makes use of the strongest piece of contextual evidence found in a decision list, is remarkable for its significant simplicity, ease of implementation, and clarity. There is, however, still room to improve his method. We must note that a strong piece of evidence is not necessarily a statistically reliable one and it is likely that the simple use of strong pieces of evidence will sometimes degrade disambiguation results. In order to improve disambiguation accuracy, as well as to economize on the space necessary for storing pieces of evidence, it is necessary to devise a method that discards those unreliable pieces of evidence. We propose here a method of testing the reliability of a piece of evidence on the basis of the Minimum Description Length (MDL) principle (Rissanen, 1989) used in statistical estimation.

Our method calculates the value of the mutual information between a piece of evidence in question and a target word. If that value exceeds a certain threshold, which depends on the size of the training data, it then considers the piece of evidence unreliable and prunes it from the decision list, after which it uses the remaining pieces of evidence as the basis for its word sense disambiguation.

We have applied our method to the task of homograph disambiguation in Japanese speech synthesis and have found that our method slightly outperforms Yarowsky's method in terms of disambiguation accuracy, and significantly outperforms his method in terms of space for storing knowledge, indicating that for this task it is better to use both strong and reliable pieces of evidence than simply to use strong pieces of evidence alone.

2 Simply Using Strong Pieces of Evidence

Let us begin here by considering the basic problem of homograph disambiguation in Japanese speech synthesis. By 'homograph disambiguation' is meant the process of determining the intended pronunciation of a homograph (a word having multiple pronunciations) appearing in a given sentence. In Japanese, there are many homographs and usually each pronunciation of a homograph represents a different meaning (word sense) . Thus for any speech synthesis system which inputs a plain text and 'reads' it, it is crucial to determine the pronunciation of a homograph. For instance, '今日' is a Japanese homograph having two possible pronunciations [kyou/konnichi], and each of them stands for a different meaning, i.e., 'today/nowadays.' [1]

In this section, we introduce Yarowsky's method through an example of resolving the ambiguities of the homograph '今日.' Yarowsky's method first constructs a decision list[2] based on training data, and then determines the pronunciations of the target homograph in *new* test data on the basis of that constructed decision list.

Suppose that training data like that in Figure 1 are given, where the parts-of-speech of words in sentences are assigned, and the pronunciations of homographs are labeled (by humans). In this paper, we make use as evidence of eight categories of information: the part-of-speech of the homograph, the parts-of-speech immediately to the left and to the right of the homograph, the words immediately to the left and to the right of the homograph, the characters immediatedly to the left and to the right of the homograph, and content words[3] found within $\pm k$ content words around the homograph. We refer to the first seven categories as 'adjacent evidence,' and the last as 'collocational evidence.' For example, the part-of-speech 'PTCL'[4] appearing immediately to the left of the homograph '今 日' in the training data is an example of adjacent evidence. The content words '明日 (tomorrow)' and '変化 (change)' appearing within $\pm k$ content words around the homograph '今日' are examples of collocational evidence. For any piece of evidence, we can obtain from the training data the frequency of co-occurrence between it and the target homograph. Table 1 shows some examples of frequencies of co-occurrence.

[1] English has fewer homographs than Japanese. Examples of homographs in English include *lives* [livz/laivz], *lead* [lid/led], etc.

[2] It is in fact a very limited form of a normal decision list (Rivest, 1987).

[3] Nouns, verbs, adjectives, and adverbs are referred to as 'content words' in this field.

[4] PTCL = a Japanese post-positional particle, similar in function to an English preposition.

Sentence		Pronunciation
明日/NOUN は/PTCL	今日/NOUN より/PTCL も/PTCL ...	[kyou]
	今日/NOUN の/PTCL 変化/NOUN は/PTCL ...	[konnichi]
さらに/ADV 、/MARK	今日/NOUN と/PTCL 明日/NOUN ...	[kyou]

図 1: Example sentences in the training data

表 1: Co-occurrence between pieces of evidence and homograph

'今日'	Pronunciation: [kyou]	Pronunciation: [konnichi]
'変化':presence	0	2
'変化':absence	195	261

'今日'	Pronunciation: [kyou]	Pronunciation: [konnichi]
'明日':presence	5	1
'明日':absence	190	262

Yarowsky's method utilizes the co-occurrence information obtained from the training data and constructs a decision list for each homograph on the basis of that information.[5] More precisely, it first calculates a likelihood ratio for each piece of evidence, where 'likelihood ratio' refers to the likelihood that a decision based on a given piece of evidence will be correct[6]:

$$\log_2(\hat{P}(D=x|E=1)/\hat{P}(D=y|E=1)), \quad \text{if} \quad \hat{P}(D=x|E=1) \geq \hat{P}(D=y|E=1), \tag{1}$$

where random variable D represents the possible pronunciations of the homograph, random variable E assumes a value of 1 or 0. (1 denotes the presence of the corresponding piece of evidence, 0 its absence.) We estimate the probabilities here by using the so-called Expected Likelihood Estimator (Gale and Church, 1990), i.e., by adding 0.5 to actual counts.

In Yarowsky's method, the likelihood ratio of any piece of evidence is considered to represent its strength. A decision list can then be constructed with pieces of evidence sorted in descending order with respect to their likelihood ratios, as seen in Table 2. The final line of a decision list is always defined as 'a default,' where the likelihood ratio is calculated as:

$$\log_2(\hat{P}(D=x)/\hat{P}(D=y)), \quad \text{if} \quad \hat{P}(D=x) \geq \hat{P}(D=y). \tag{2}$$

[5]For simplicity, we consider here only cases in which a homograph has two pronunciations. One can easily extend them, however, to cases with a greater number of pronunciations.

[6]In the case in which a homograph has more than two pronunciations, the likelihood ratio of a piece of evidence is *heuristically* calculated as the ratio of the largest conditional probability of pronunciation (given the presence of that piece of evidence) to the second largest.

Yarowsky's method then uses the constructed decision list to determine the most likely intended pronunciation of a homograph. Suppose that the sentence in Figure 2 is given, and that we are to determine the pronunciation of its homograph '今日.' Yarowsky's method identifies which piece of evidence will be located on the highest line in the decision list and recommends the corresponding pronunciation. Here, the piece of evidence '変化' is located on the highest line in the decision list in Table 2, and thus the pronunciation [konnichi] is indicated. (This is in fact an incorrect decision.)

The advantages of Yarowsky's method are its significant simplicity, ease of implementation, and clarity, but as we have previously noted, strength of evidence is not necessarily an indicator of statistical reliability, and it is likely that the simple use of strong pieces of evidence alone will sometimes degrade disambiguation results.

In the example in Figure 2, the observed frequency of the presence of the piece of evidence '変化' and the pronunciation [konnichi] in the training data is two, and that of its absence and the pronunciation [kyou] is zero (see Table 1). It is very likely that these occurrences are observed only *by chance*, and thus it would be better not to use the piece of evidence '変化.' Moreover, the frequency of the presence of the piece of evidence '明日' and the pronunciation [kyou] is five, and that of its absence and the pronunciation [konnichi] is one (see Table 1). Thus '明日' seems to be more reliable than '変化,' since it has a higher observed frequency of occurrence, and it would be better to use it rather than '変化' in disambiguation, even though the former is not as strong as the latter in terms of its likelihood ratio.

表 2: Example decision list

	Evidence		Decision	Likelihood ratio
†	'変化' within $\pm k$ content words	⇒	[konnichi]	2.322
	'明日' within $\pm k$ content words	⇒	[kyou]	1.874
	Left POS = PTCL	⇒	[konnichi]	1.288
	default	⇒	[konnichi]	0.431

今日 /NOUN 売れた /VERB もの /NOUN が /PTCL 明日 /NOUN
売れなくなって /VERB しまう /VERB ほど /PTCL 消費者 /NOUN の /PTCL
ニーズ /NOUN の /PTCL 変化 /NOUN が /PTCL 激しい /ADJ 。 /MARK
The needs of customers change so rapidly that products selling well today are not guaranteed to sell well tomorrow.

図 2: An example sentence in the test data

3 Using Evidence That Is both Strong and Reliable

We have devised a method of testing the reliability of evidence used in word sense disambiguation.

Assuming that the pieces of evidence prepared in a decision list are all stronger than the default, for any piece of evidence, the inequality

$$\frac{\hat{P}(D=x|E=1)}{\hat{P}(D=y|E=1)} > \frac{\hat{P}(D=x)}{\hat{P}(D=y)} \qquad (3)$$

should hold. Since the probabilities in Inequality (3) are all estimated on the basis of training data, however, this cannot be known for certain, and it becomes necessary to test whether the inequality is statistically significant, in other words, whether or not it is reliable, for any given piece of evidence.

We may note here that if the two random variables D and E are judged to be independent, or approximately independent, i.e.,

$$\hat{P}(D|E) = (\approx)\hat{P}(D), \qquad (4)$$

then we have

$$\frac{\hat{P}(D=x|E=1)}{\hat{P}(D=y|E=1)} = (\approx)\frac{\hat{P}(D=x)}{\hat{P}(D=y)}. \qquad (5)$$

Thus Inequality (3) will not be reliable. [7] Therefore, it is desirable to prune those pieces of evidence from a decision list whose corresponding random variables are judged to be independent of the random variable D.

The question now turns to how to test for such independence. Our method is based on the Minimum Description Length Principle. We first calculate the value of the mutual information between the two random variables E and D. If that value exceeds the following threshold

$$\theta = \beta \cdot \frac{(k_E - 1)\cdot(k_P - 1)\cdot \log_2 N}{2 \cdot N}, \qquad (6)$$

where β is a weighting parameter, $0 \leq \beta \leq 1$, k_E is the number of values E takes on, k_P is the number of values D takes on, and N is the data size, we judge the two random variables to be independent (see Appendix for the derivation of this expression).

In our homograph disambiguation, we first construct a decision list for each homograph, using the method proposed by Yarowsky. We then calculate the value of the mutual information between each piece of evidence in the decision list and the target homograph and prune those pieces of evidence whose mutual information values do not exceed their respective thresholds[8]. We then use only the pieces remaining.

Table 3 shows the mutual information values between pieces of evidence and the target homograph with respect to our previous example, and it also gives relevant threshold values. Since the mutual information value between the piece of evidence '変化' and the homograph '今日' does not exceed the threshold, we prune it from the decision list in Table 2. (i.e., only the piece of evidence marked with † is pruned from the decision list in Table 2.) Using the piece of evidence '明日' remaining on the highest line in the decision list, we determine the pronunciation of the homograph '今日' in the sentence in Figure 2 to be [kyou], which is correct.

One important characteristic of our method is that the threshold in Equation (6) is based on data size. Note that when we do not have enough data (i.e., when N is small), the threshold will be large

[7] Note that the converse proposition is not necessarily valid.

[8] Note that when $\beta = 0$, our method becomes equivalent to Yarowsky's method.

表 3: Mutual information and threshold

Evidence	Mutual information	Threshold ($\beta = 0.2$)
'変化' within $\pm k$ content words	0.001	0.002
Left POS = PTCL	0.077	0.008
'明日' within $\pm k$ content words	0.006	0.002

and few pieces of evidence will remain in the decision list. This is reasonable, since with little data most pieces of evidence cannot judged with any significance to be reliable.

The advantages of our method over Yarowsky's are its reliability and efficiency. By pruning the unreliable pieces of evidence in a decision list, we can improve disambiguation accuracy and save space for storing pieces of evidence as well.

An alternative way of testing the reliability of a piece of evidence would be to employ the χ^2 test to test the dependency between a piece of evidence and the target homograph. We have instead adopted an MDL-based method in part because we wished to investigate how effective it would be to employ MDL in various natural language processing tasks. (Refer to (Li and Abe, 1995; Li and Abe, 1996b; Li and Abe, 1996a) to see how MDL is applied in other natural language tasks.)

Our method of testing the reliability of a piece of evidence is not only applicable to the decision-list approach alone; it can also be applied, for example, to the example-based method used by (Ng and Lee, 1996) for pruning unreliable features.

4 Experimental Results

In this section, we describe the results of experiments we have conducted to compare the performance of our method with that of Yarowsky's.

In this work, we used EDR corpus data[9], which includes 208,156 sentences in its Japanese corpus. Every word appearing in the corpus is tagged with its part of speech (POS) and pronunciation. There are 2,214 (types of) homographs in the corpus. (We noted pronunciation taggings that appeared to be mistaken, but we did not correct them.) We randomly selected 50 homographs that appeared more than 60 times in the corpus, and collected sentences containing them to create a data set referred to here as 'R50.' We then selected another 50 homographs appearing more than 60 times in the corpus, but we did so this time on the basis of their being difficult to disambiguate, and we used sentences containing them to create a data set referred to here as 'S50.'

We employed 'ten-fold cross validation' to evaluate the effectiveness of our method and Yarowsky's. That is, for each homograph, we used nine tenths of the sentences containing that homograph as training data, saving what remained as test data, constructed a decision list based on the training data, and used the constructed decision list to conduct homograph disambiguation on the test data. We repeated this process ten times and calculated average disambiguation accuracy over the ten trials. Here 'accuracy' means the success rate of decisions achieved by the respective methods.

We compared the performance of our method with that of Yarowsky's in six experimental settings, using as collocational evidence: (1) nouns within ± 5 nouns, (2) nouns within ± 10 nouns, (3) nouns within ± 20 nouns, (4) content words within ± 5 content words, (5) content words ± 10 content words, and (6) content words ± 20 content words. In each experimental setting, we used all the types of adjacent evidence. Table 4 shows the disambiguation accuracy averaged over the homographs in each data set for each setting. 'Base' stands for the accuracy which would be achieved simply by always selecting the pronunciation known to appear most frequently in the training data. When using our method, we set the parameter β in Equation (6) to 0.2 because we experimentally found that doing so provided more accurate results.

Our method slightly outperforms Yarowsky's method in terms of accuracy. We evaluated the significance of this difference by using the 'sign test' (c.f. (Guttman and Wilks, 1965)). In this test, we assumed that the disambiguation accuracies of homographs in one data set for one setting achieved by our method x_i and those achieved by Yarowsky's y_i were independently generated according to the distributions of $F(X)$ and $F(Y)$, respectively. The null hypothesis is $F(X) = F(Y)$, and the alternative hypothesis is $F(X) > F(Y)$. Further, we defined a random variable T such that

$$T = \begin{cases} 1 & x_i > y_i \\ 0 & x_i < y_i, \end{cases} \tag{7}$$

letting

$$\begin{aligned} n &= \text{number of } i \text{ such that } \quad x_i \neq y_i \\ r &= \text{number of } i \text{ such that } \quad x_i > y_i. \end{aligned} \tag{8}$$

We computed the probability of observing r occurrences of 1 in n trials of T according to the binomial distribution $(n, 0.5)$. The probabilities for R50 for some settings are small. Therefore we can

[9] EDR homepage: http://www.iijnet.or.jp/edr/

表 4: Average disambiguation accuracy

Data set	Experimental setting	Base	Yarowsky(%)	Our method(%)
S50	±5 nouns	62.6	92.2	92.3
S50	±10 nouns	-	92.2	92.3
S50	±20 nouns	-	92.2	92.2
S50	±5 content words	-	92.2	92.2
S50	±10 content words	-	92.1	92.2
S50	±20 content words	-	92.1	92.2
R50	±5 nouns	83.7	94.0	94.1
R50	±10 nouns	-	94.0	94.1
R50	±20 nouns	-	94.0	94.1
R50	±5 content words	-	93.9	94.1
R50	±10 content words	-	93.9	94.1
R50	±20 content words	-	93.9	94.1
Average		73.2	93.1	93.2

reject the null hypothesises for them at the significance level of 0.1, i.e., we can say that our method improves Yarowsky's method, and that the improvements are sometimes statistically significant. The probabilities for S50 for some settings are not small, however, and therefore we cannot reject the null hypothesises, i.e., although our method outperforms Yarowsky's, we cannot say that the improvements are statistically significant.

We also evaluated the two methods in terms of space for storing knowledge. We counted the number of lines in the decision lists constructed using our method and that of Yarowsky. Table 5 shows the number of lines per decision list for each data set. The decision lists constructed with our method are much shorter than those constructed with Yarowsky's method.

In summary, then, our method slightly outperforms Yarowsky's method in terms of disambiguation accuracy, and significantly outperforms his method in terms of space for storing knowledge, indicating that in word sense disambiguation it is better to use evidence that is both strong and reliable than simply to use strong pieces.

From the experimental results we can also reach the following conclusion for Japanese homograph disambiguation: it is enough to use only nouns within ±5 nouns around a homograph as pieces of collocational evidence.

表 5: Average length of decision lists

Data set	Yarowsky	Our method (Reduction)
S50	1146	553 (51.7%)
R50	858	578 (32.6%)

5 Summary

We have proposed here a method of using evidence that is both strong and reliable in word sense disambiguation. The main contributions of this research are:

1. we have shown the advantage of limiting evidence to this type;
2. we have devised a method of testing the reliability of a piece of evidence on the basis of the MDL principle; and
3. we have demonstrated that it is effective to employ the decision list approach in Japanese homograph disambiguation. Average disambiguation accuracy for a randomly selected data set was 94.1%.

Acknowledgements

We are grateful to Tomoyuki Fujita of NEC for his important and constant encouragement. We also thank Naoki Abe of NEC, Takayoshi Ochiai of NIS, and Mark Petersen of Meiji Univ. for their valuable comments and suggestions, and express deep appreciation to Yuuko Yamaguchi of NIS for her programming efforts.

References

Ezra Black. 1988. An experiment in computational discrimination of english word senses. *IBM J. RES. DEVELOP*, 32(2):185–193.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1991. Word sense disambiguation using statistical methods. *Proceedings of Annual the 29th Meeting of the Association for Computational Linguistics*, pages 264–270.

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

Williams A. Gale and Kenth W. Church. 1990. Poor estimates of context are worse than none. *Proceedings of the DARPA Speech and Natural Language Workshop*, pages 283–287.

William Gale, Kenneth Ward Church, and David Yarowsky. 1992. Estimating upper and lower bounds on the performance of word-sense disambiguation programs. *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, pages 249–256.

Andrew R. Golding and Dan Roth. 1996. Applying winnow to context-sensitive spelling correction. *Proceedings of the 13rd International Conference on Machine Learning.*

Andrew R. Golding and Yves Schabes. 1996. Combining trigram-based and feature-based methods for context-sensitive spelling correction. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.*

Joe A. Guthrie, Louise Guthrie, Yorick Wilks, and Homa Aidinejad. 1991. Subject-dependent co-occurrence and word sense disambiguation. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 146–152.

I. Guttman and S.S. Wilks. 1965. *Introductory Engineering Statistics.* John Wiley and Sons, Inc.

Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. *ARPA Workshop on Human Language Technology.*

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. *Proceedings of the 15th International Conference on Computational Linguistics.*

Hang Li and Naoki Abe. 1995. Generalizing case frames using a thesaurus and the mdl principle. *Proceedings of Recent Advances in Natural Language Processing*, pages 239–248.

Hang Li and Naoki Abe. 1996a. Clustering words with the mdl principle. *Proceedings of the 16th International Conference on Computational Linguistics*, pages 4–9.

Hang Li and Naoki Abe. 1996b. Learning dependencies between case frame slots. *Proceedings of the 16th International Conference on Computational Linguistics*, pages 10–15.

Susan W. McRoy. 1992. Using multiple knowledge sources for word sense discrimination. *Computational Linguistics*, 18(1):1–30.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An examplar-based approach. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics.*

Yoshiki Niwa and Yoshihiko Nitta. 1994. Co-occurrence vectors from corpora vs. distance vectors from dictionaries. *Proceedings of the 14th International Conference on Computational Linguistics*, pages 304–309.

Jorma Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Co.

Ronald L. Rivest. 1987. Learning decision lists. *Machine Learning*, pages 229–246.

Hinrich Schutze. 1993. Word space. *Advances in Neural Information Processing Systems 5, Ed. by S.J.Hanson, J.D.Cowan, and C.L.Giles, Morgan Kaufmann*, pages 895–902.

Ellen M. Voorhees, Claudia Leacock, and Geoffrey Towell. 1995. Learning context to disambiguate word senses. *Computational Learning Theory and Natural Language Learning Systems 3: Selecting Good Models, ed. by T Petsche, S.J.Hanson, J.W. Shavlik, MIT Press*, pages 279–305.

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

David Yarowsky. 1993. One sense per collocation. *Proceedings of ARPA Workshop on Human Language Technology.*

David Yarowsky. 1994. Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french. *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196.