Detection of Japanese Homophone Errors by a Decision List Including a Written Word as a Default Evidence

Hiroyuki Shinnou
Ibaraki University
Dept. of Systems Engineering
4-12-1 Nakanarusawa
Hitachi, Ibaraki, 316-8511, JAPAN
shinnou@lily.dse.ibaraki.ac.jp

Abstract

In this paper, we propose a practical method to detect Japanese homophone errors in Japanese texts. It is very important to detect homophone errors in Japanese revision systems because Japanese texts suffer from homophone errors frequently. In order to detect homophone errors, we have only to solve the homophone problem. We can use the decision list to do it because the homophone problem is equivalent to the word sense disambiguation problem. However, the homophone problem is different from the word sense disambiguation problem because the former can use the written word but the latter cannot. In this paper, we incorporate the written word into the original decision list by obtaining the identifying strength of the written word. The improved decision list can raise the F-measure of error detection.

1 Introduction

In this paper, we propose a method of detecting Japanese homophone errors in Japanese texts. Our method is based on a decision list proposed by Yarowsky (Yarowsky, 1994; Yarowsky, 1995). We improve the original decision list by using written words in the default evidence. The improved decision list can raise the F-measure of error detection.

Most Japanese texts are written using Japanese word processors. To input a word composed of kanji characters, we first input the phonetic hiragana sequence for the word, and then convert it to the desired kanji sequence. However, multiple converted kanji sequences are generally produced, and we must then choose the correct kanji sequence. Therefore, Japanese texts suffer from homophone errors caused by incorrect choices. Carelessness of choice alone is not the cause of homophone errors; Ignorance of the difference among homophone words is serious. For example, many Japanese are not aware of the difference between '意志' and '意思', or between '直感' and '直観'[1].

In this paper, we define the term **homophone set** as a set of words consisting of kanji characters that have the same phone [2]. Then, we define the term **homophone word** as a word in a homophone set. For example, the set { 確率 (probability), 確立 (establishment)} is a homophone set because words in the set are composed of kanji characters that have the same phone 'ka-ku-ri-tu'. Thus, '確率' and '確立' are homophone words. In this paper, we name the problem of choosing the correct word from the homophone set the **homophone problem.** In order to detect homophone errors, we make a list of homophone sets in advance, find a homophone word in the text, and then solve the homophone problem for the homophone word.

Many methods of solving the homophone problem have been proposed (Tochinai et al., 1986; Ibuki et al., 1997; Oku and Matsuoka, 1997; Oku, 1994; Wakita and Kaneko, 1996). However, they are restricted to the homophone problem, that is, they are heuristic methods. On the other hand, the homophone problem is equivalent to the word sense disambiguation problem if the phone of the homophone word is regarded as the word, and the homophone word as the sense. Therefore, we can solve the homophone problem by using various

[1] '意志' and '意思' have a same phone 'i-shi'. The meaning of '意思' is a general will, and the meaning of '意志' is a strong positive will. '直感' and '直観' have a same phone 'cho-kkan'. The meaning of '直感' is an intuition through a feeling, and the meaning of '直観' is an intuition through a latent knowledge.

[2] We ignore the difference of accents, stresses and parts of speech. That is, the homophone set is the set of words having the same expression in hiragana characters.

statistical methods proposed for the word sense disambiguation problem(Fujii, 1998). Take the case of context-sensitive spelling error detection [3], which is equivalent to the homophone problem. For that problem, some statistical methods have been applied and succeeded(Golding, 1995; Golding and Schabes, 1996). Hence, statistical methods are certainly valid for the homophone problem. In particular, the decision list is valid for the homophone problem(Shinnou, 1998). The decision list arranges evidences to identify the word sense in the order of strength of identifying the sense. The word sense is judged by the evidence, with the highest identifying strength, in the context.

Although the homophone problem is equivalent to the word sense disambiguation problem, the former has a distinct difference from the latter. In the homophone problem, almost all of the answers are given correctly, because almost all of the expressions written in the given text are correct. It is difficult to decide which is the meaning of 'crane', 'crane of animal' or 'crane of tool'. However, it is almost right that the correct expression of '確立' in a text is not '確率' but '確立'. In the homophone problem, the choice of the written word results in high precision. We should use this information. However, the method to always choose the written word is useless for error detection because it doesn't detect errors at all. The method used for the homophone problem should be evaluated from the precision and the recall of the error detection. In this paper, we evaluate it by the F-measure to combine the precision and the recall, and use the written word to raise the F-measure of the original decision list.

We use the written word as an evidence of the decision list. The problem is how much strength to give to that evidence. If the strength is high, the precision rises but the recall drops. On the other hand, if the strength is low, the decision list is not improved. In this paper, we calculate the strength that gives the maximum F-measure in a training corpus. As a result, our decision list can raise the F-measure of error detection.

2 Homophone disambiguation by a decision list

In this section, we describe how to construct the decision list and to apply it to the homophone problem.

2.1 Construction of the decision list

The decision list is constructed by the following steps.

step 1 Prepare homophone sets.

In this paper, we use the 12 homophone sets shown in Table 1, which consist of homophone words that tend to be mis-chosen.

Table 1: Homophone sets

Phone	Homophone set
sa-i-ken	{ 債券, 債権 }
ka-i-hou	{ 解放, 開放 }
kyo-u-cho-u	{ 協調, 強調 }
ji-shi-n	{ 自信, 自身 }
ka-n-shi-n	{ 感心, 関心 }
ta-i-ga-i	{ 体外, 対外 }
u-n-ko-u	{ 運航, 運行 }
do-u-shi	{ 同志, 同士 }
ka-te-i	{ 過程, 課程 }
ji-kko-u	{ 実効, 実行 }
syo-ku-ryo-u	{ 食料, 食糧 }
syo-u-ga-i	{ 傷害, 障害 }

step 2 Set context information, i.e. evidences, to identify the homophone word.

We use the following three kinds of evidence.

- word (w) in front of H: Expressed as $w-$
- word (w) behind H: Expressed as $w+$
- *jiritu* words [4] surrounding H: We pick up the nearest three *jiritu* words in front of and behind H respectively. We express them as $w \pm 3$.

step 3 Derive the frequency $frq(w_i, e_j)$ of the collocation between the homophone word w_i in the homophone set $\{w_1, w_2, \cdots, w_n\}$ and the evidence e_j, by using a training corpus.

For example, let us consider the homophone set { 運航 (running (of a ship, etc.)), 運行 (running (of a train, etc.))} and the following two Japanese sentences.

Sentence 1 「西の風が三メートルで飛行機の運航に支障はなかった。」
(A west wind of 3 m/s did not prevent the plane from flying.)

[3]For example, confusion between 'peace' and 'piece', or between 'quiet' and 'quite' is the context-sensitive spelling error.

[4]The *jiritu* word is defined as an independent word which can form one bun-setu by itself. Nouns, verbs and adjectives are examples.

Table 2: Answers and identifying strength for evidences

Evid.	Freq. of '運航'	Freq. of '運行'	Ans.	Identifying Strength
に+ (to+)	77	53	運航	0.538
の− (of-)	252	282	運行	0.162
飛行機 ±3 (plane±3)	4	0	運航	5.358
...	...	...	...	...
時間+ (hour+)	14	11	運航	0.345
深夜 ±3 (midnight±3)	0	48	運行	8.910
短縮 ±3 (shorten±3)	0	4	運行	5.358
...	...	...	...	...
default	1468	1422	運航	0.046

Sentence 2 「早朝深夜の運行時間が短縮された。」 (Running hours in the early morning and during the night were shortened.)

From sentence 1, we can extract the following evidences for the word '運航':

"に +", "の −", "飛行機 ±3", "三 ±3", "風 ±3", "支障 ±3", "ない ±3",

and from sentence 2, we can extract the following evidences for the word '運行':

"時間 +", "の −", "深夜 ±3", "早朝 ±3", "時間 ±3", "短縮 ±3", "する ±3".

step 4 Define the strength $est(w_i, e_j)$ of estimating that the homophone word w_i is correct given the evidence e_j:

$$est(w_i, e_j) = log(\frac{P(w_i|e_j)}{\sum_{k \neq i} P(w_k|e_j)})$$

where $P(w_i|e_j)$ is approximately calculated by:

$$P(w_i|e_j) = \frac{frq(w_i, e_j) + \alpha}{\sum_k frq(w_k, e_j) + \alpha}.$$

α in the above expression is included to avoid the unsatisfactory case of $frq(w_i, e_j) = 0$. In this paper, we set $\alpha = 0.1$[5]. We also use the special evidence *default*. $frq(w_i, default)$ is defined as the frequency of w_i.

step 5 Pick the highest strength $est(w_k, e_j)$ among

$\{est(w_1, e_j), est(w_2, e_j), \cdots, est(w_n, e_j)\}$,

and set the word w_k as the *answer* for the evidence e_j. In this case, the identifying strength is $est(w_k, e_j)$.

For example, by steps 4 and 5 we can construct the list shown in Table 2.

step 6 Fix the answer w_{k_j} for each e_j and sort identifying strengths $est(w_{k_j}, e_j)$ in order of dimension, but remove the evidence whose identifying strength is less than the identifying strength $est(w_{k_j}, default)$ for the evidence *default* from the list. This is the decision list.

After step 6, we obtain the decision list for the homophone set { 運航, 運行 } as shown in Table 3.

Table 3: Example of decision list

Rank	Evid.	Ans.	Strength
1	列車 ±3 (train±3)	運行	9.453
2	船 ±3 (ship±3)	運航	9.106
3	深夜 ±3 (midnight±3)	運行	8.910
...	...	...	...
701	時間− (hour-)	運行	0.358
...	...	...	...
746	の+ (of+)	運行	0.162
...	...	...	...
760	*default*	運航	0.046

2.2 Solving by a decision list

In order to solve the homophone problem by the decision list, we first find the homophone word w in the given text, and then extract evidences E for the word w from the text:

$$E = \{e_1, e_2, ..., e_i\}.$$

[5] As in this paper, the addition of a small value is an easy and effective way to avoid the unsatisfactory case, as shown in (Yarowsky, 1994).

Next, picking up the evidence from the decision list for the homophone set for the homophone word w in order of rank, we check whether the evidence is in the set E. If the evidence e_j is in the set E, the answer w_{k_j} for e_j is judged to be the correct expression for the homophone word w. If w_{k_j} is equal to w, w is judged to be correct, and if it is not equal, then it is shown that w may be the error for w_{k_j}.

3 Use of the written word

In this section, we describe the use of the written word in the homophone problem and how to incorporate it into the decision list.

3.1 Evaluation of error detection systems

As described in the Introduction, the written word cannot be used in the word sense disambiguation problem, but it is useful for solving homophone problems. The method used for the homophone problem is trivial if the method is evaluated by the precision of distinction using the following formula:

$$\frac{number\ \ of\ \ correct\ \ discriminations}{number\ \ of\ \ all\ \ discriminations}.$$

That is, if the expression is '運航' (or '運行'), then we should clearly choose the word '運航' (or the word '運行') from the homophone set { 運航, 運行 }. This distinction method probably has better precision than any other methods for the word sense disambiguation problem. However, this method is useless because it does not detect errors at all.

The method for the homophone problem should be evaluated from the standpoint of not error discrimination but error detection. In this paper, we use the F-measure (Eq.1) to combine the precision P and the recall R defined as follows:

$$P = \frac{number\ \ of\ \ real\ \ errors\ \ in\ \ detected\ \ errors}{number\ \ of\ \ detected\ \ errors},$$

$$R = \frac{number\ \ of\ \ real\ \ errors\ \ in\ \ detected\ \ errors}{number\ \ of\ \ errors\ \ in\ \ the\ \ text}.$$

$$F = \frac{2PR}{P+R} \tag{1}$$

3.2 Use of the identifying strength of the written word

The distinction method to choose the written word is useless, but it has a very high precision of error discrimination. Thus, it is valid to use this method where it is difficult to use context to solve the homophone problem.

The question is when to stop using the decision from context and use the written word. In this paper, we regard the written word as a kind of evidence on context, and give it an identifying strength. Consequently we can use the written word in the decision list.

3.3 Calculation of the identifying strength of the written word

First, let x be the identifying strength of the written word. We name the set of evidences with higher identifying strength than x the set α, and the set of evidences with lower identifying strength than x the set β,

Let T be the number of homophone problems for a homophone set. We solve them by the original decision list *DL0*. Let G (or H) be the ratio of the number of homophone problems by judged by α (or β) to T. Let g (or h) be the precision of α (or β), and p be the occurrence probability of the homophone error.

The number of problems correctly solved by α is as follows:

$$GT(1-p), \tag{2}$$

and the number of problems incorrectly solved by α is as follows:

$$GTp. \tag{3}$$

The number of problems detected as errors in Eq.2 and Eq.3 are $GT(1-p)(1-g)$ and $GTpg$ respectively. Thus, the number of problems detected as errors by α is as follows:

$$GT((1-p)(1-g)+pg). \tag{4}$$

In the same way, the number of problems detected as errors by β is as follows:

$$HT((1-p)(1-h)+ph). \tag{5}$$

Consequently the total number of problems detected as errors is as follows:

$$T(G((1-p)(1-g)+pg)+H((1-p)(1-h)+ph)). \tag{6}$$

The number of correct detections in Eq.6 is $Tp(Gg+Hh)$. Therefore the precision P_0 is as follows:

$$\begin{aligned} P_0 &= p(Gg+Hh)/\{G((1-p)(1-g)+pg) \\ &\quad + H((1-p)(1-h)+ph)\} \end{aligned}$$

Because the number of real errors in T is Tp, the recall R_0 is $Gg+Hh$. By using P_0 and R_0, we can get the F-measure F_0 of DL0 by Eq.1.

Next, we construct the decision list incorporating the written word into DL0. We name this decision list *DL1*. In DL1, we use the written word to solve problems which we cannot judge by α. That

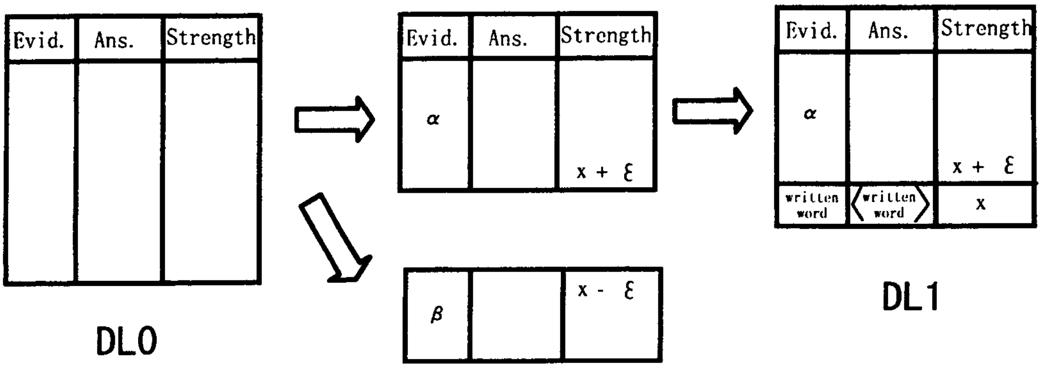

Figure 1: Construction of DL1

is, DL1 is the decision list to attach the written word as the *default* evidence to α (see Fig.1).

Next, we calculate the precision and the recall of DL1. Because α of DL1 is the same as that of DL0, the number of problems detected as errors by α is given by Eq.4. In the case of DL1, problems judged by β of DL0 are judged by the written word. Therefore, we detect no error from these problems.

As a result, the number of problems detected as errors by DL1 is given by Eq.4, and the number of real errors in these detections is $TGpg$. Therefore, the precision P_1 of DL1 is as follows:

$$P_1 = \frac{pg}{(1-p)(1-g)+pg}.$$

Because the number of whole errors is Tp, the recall R_1 of DL1 is Gg. By using P_1 and R_1, we can get the F-measure F_1 of DL1 by Eq.1.

Finally, we try to define the identifying strength x. x is the value that yields the maximum F_1 under the condition $F_1 > F_0$. However, theoretical calculation alone cannot give x, because p is unknown, and functions of G, H, g, and h are also unknown.

In this paper, we set $p = 0.05$, and get values of G, H, g, and h by using the training corpus which is the resource used to construct the original decision list DL0. Take the case of the homophone set {'運航', '運行'}. For this homophone set, we try to get values of G, H, g, and h. The training corpus has 2,890 sentences which include the word '運航' or the word '運行'. These 2,890 sentences are homophone problems for that homophone set. The identifying strength of DL0 for this homophone set covers from 0.046 to 9.453 as shown in Table 3. Next we give x a value. For example, we set $x = 2.5$. In this case, the number of problems judged by α is 1,631, and the number of correct judgments in them is 1,593. Thus, $G = 1631/2890 = 0.564$ and $g = 1593/1631 = 0.977$. In the same way, under this assumption $x = 2.5$, the number of problems judged by β is 1,259, and the number of correct judgments in them is 854. Thus, $H = 1259/2890 = 0.436$ and $h = 854/1259 = 0.678$. As a result, if $x = 2.5$, then $P_0 = 0.225$, $R_0 = 0.847$, $F_0 = 0.356$, $P_1 = 0.688$, $R_1 = 0.551$ and $F_1 = 0.612$. In Fig.2, Fig.3 and Fig.4, we show the experiment result when x varies from 0.0 to 10.0 in units of 0.1. By choosing the maximum value of F_1 in Fig.4, we can get the desired x. In this homophone set, we obtain $x = 3.0$.

4 Experiments

First, we obtain each identifying strength of the written word for the 12 homophone sets shown in Table 1, by the above method. We show this result in Table 4. *LR0* in this table means the lowest rank of DL0. That is, LR0 is the rank of the *default* evidence. *LR1* means the lowest rank of DL1. That is, LR1 is the rank of the evidence of the written word. Moreover, LR0 and LR1 mean the sizes of each decision list DL0 and DL1.

Second, we extract sentences which include a word in the 12 homophone sets from a corpus. We note that this corpus is different from the training corpus; the corpus is one year's worth of Mainichi newspaper articles, and the training corpus is one year's worth of Nikkei newspaper articles. The extracted sentences are the test sentences of the experiment. We assume that these sentences have no homophone errors.

Last, we randomly select 5% of the test sentences, and forcibly put homophone errors into these selected sentences by changing the written

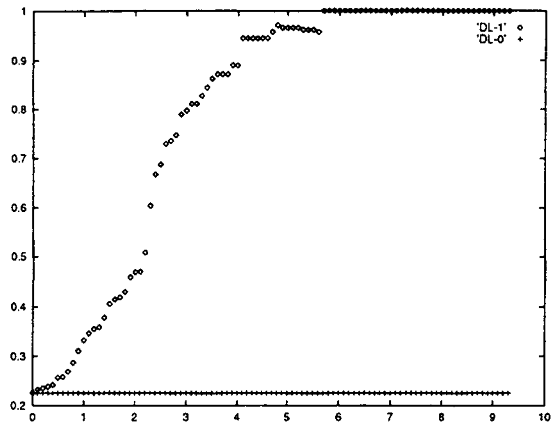

Figure 2: Precisions P_0 and P_1

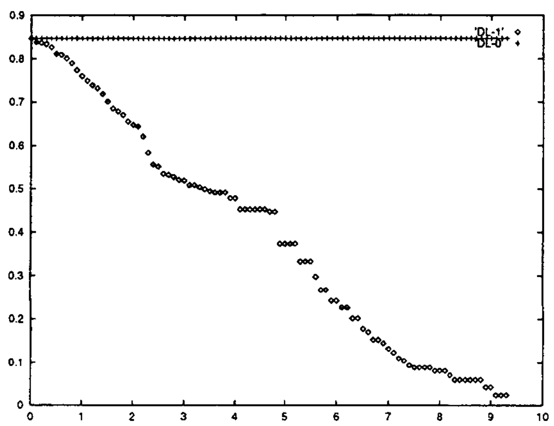

Figure 3: Recalls R_0 and R_1

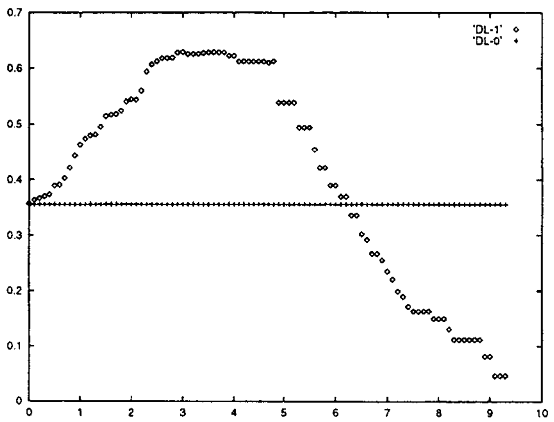

Figure 4: F-measures F_0 and F_1

Table 4: Identifying strength of the expression

homophone set	Identifying strength of expression	LR0	LR1
{ 債券, 債権 }	4.9	1062	844
{ 解放, 開放 }	4.6	1104	671
{ 協調, 強調 }	4.3	1120	667
{ 自信, 自身 }	4.8	1134	622
{ 感心, 関心 }	5.7	1007	424
{ 体外, 対外 }	3.9	921	921
{ 運航, 運行 }	3.0	760	319
{ 同志, 同士 }	4.5	811	788
{ 過程, 課程 }	5.1	799	469
{ 実効, 実行 }	4.3	760	665
{ 食料, 食糧 }	4.7	697	255
{ 傷害, 障害 }	5.1	695	397

homophone word to another homophone word. As a result, the test sentences include 5% errors. From these test sentences, we detect homophone errors by DL0 and DL1 respectively.

We conducted this experiment ten times, and got the mean of the precision, the recall and the F-measure. The result is shown in Table 5.

For all homophone sets, the F-measure of our proposed DL1 is higher than the F-measure of the original decision list DL0. Therefore, it is concluded that our proposed method is effective.

5 Remarks

The recall of DL1 is no more than the recall of DL0. Our method aims to raise the F-measure by raising the precision instead of sacrificing the recall. We confirmed the validity of the method by experiments in sections 3 and 4. Thus our method has only a little effect if the recall is evaluated with importance. However, we should note that the F-measure of DL1 is always not worse than the F-measure of DL0.

We set the occurrence probability of the homophone error at $p = 0.05$. However, each homophone set has its own p. We need decide p exactly because the identifying strength of the written word depends on p. However, DL1 will produce better results than DL0 if p is smaller than 0.05, because the precision of judgment by the written word improves without lowering the recall. The recall does not fall due to smaller p because R_0 and R_1 are independent of p. Moreover, from the definitions of P_0 and P_1, we can confirm that the precision of judgments by the written word improves with smaller p.

Table 5: Result of experiments

homophone set	Number of problems	DL0			DL1		
		P_0	R_0	F_0	P_1	R_1	F_1
{ 債券, 債権 }	1,254	0.190	0.824	0.309	0.310	0.774	0.443
{ 解放, 開放 }	1,938	0.295	0.899	0.443	0.573	0.835	0.680
{ 協調, 強調 }	4,845	0.583	0.957	0.724	0.616	0.934	0.742
{ 自信, 自身 }	3,682	0.343	0.911	0.499	0.470	0.725	0.571
{ 感心, 関心 }	2,032	0.773	0.987	0.867	0.804	0.981	0.884
{ 体外, 対外 }	618	0.708	0.980	0.822	0.806	0.980	0.885
{ 運航, 運行 }	588	0.127	0.745	0.217	0.289	0.420	0.342
{ 同志, 同士 }	1,436	0.391	0.939	0.552	0.440	0.913	0.594
{ 過程, 課程 }	1,220	0.789	0.990	0.879	0.903	0.910	0.906
{ 実効, 実行 }	1,563	0.548	0.966	0.700	0.617	0.911	0.736
{ 食料, 食糧 }	1,074	0.091	0.692	0.161	0.135	0.287	0.183
{ 傷害, 障害 }	1,636	0.681	0.976	0.802	0.760	0.858	0.806
mean	1,824	0.460	0.906	0.581	0.560	0.794	0.648

The number of elements of all homophone sets used in this paper was two, but the number of elements of real homophone sets may be more. However, the bigger this number is, the better the result produced by our method, because the precision of judgments by the *default* evidence of DL0 drops in this case, but that of DL1 does not. Therefore, our method is better than the original one even if the number of elements of the homophone set increases.

Our method has an advantage that the size of DL1 is smaller. The size of the decision list has no relation to the precision and the recall, but a small decision list has advantages of efficiency of calculation and maintenance.

On the other hand, our method has a problem in that it does not use the written word in the judgment from α; Even the identifying strength of the evidence in α must depend on the written word. We intend to study the use of the written word in the judgment from α. Moreover, homophone errors in our experiments are artificial. We must confirm the effectiveness of the proposed method for actual homophone errors.

6 Conclusions

In this paper, we used the decision list to solve the homophone problem. This strategy was based on the fact that the homophone problem is equivalent to the word sense disambiguation problem. However, the homophone problem is different from the word sense disambiguation problem because the former can use the written word but the latter cannot. In this paper, we incorporated the written word into the original decision list by obtaining the identifying strength of the written word. We used 12 homophone sets in experiments. In these experiments, our proposed decision list had a higher F-measure than the original one. A future task is to further integrate context and the written word in the decision list.

Acknowledgments

We used Nikkei Shibun CD-ROM '90 and Mainichi Shibun CD-ROM '94 as the corpus. The Nihon Keizai Shinbun company and the Mainichi Shinbun company gave us permission of their collections. We appreciate the assistance granted by both companies.

References

Atsushi Fujii. 1998. Corpus-Based Word Sence Disambiguation (in Japanese). *Journal of Japanese Society for Artificial Intelligence*, 13(6):904–911.

Andrew R. Golding and Yves Schabes. 1996. Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction. In *34th Annual Meeting of the Association for Computational Linguistics*, pages 71–78.

Andrew R. Golding. 1995. A Bayesian Hybrid Method for Context-Sensitive Spelling Correction. In *Third Workshop on Very Large Corpora (WVLC-95)*, pages 39–53.

Jun Ibuki, Guowei Xu, Takahiro Saitoh, and Kunio Matsui. 1997. A new approach for Japanese Spelling Correction (in Japanese). SIG Notes NL-117-21, IPSJ.

Masahiro Oku and Koji Matsuoka. 1997. A Method for Detecting Japanese Homophone Errors in Compound Nouns based on Character Cooccurrence and Its Evaluation (in Japanese). *Journal of Natural Language Processing*, 4(3):83–99.

Masahiro Oku. 1994. Handling Japanese Homophone Errors in Revision Support System; REVISE. In *4th Conference on Applied Natural Language Processing (ANLP-94)*, pages 156–161.

Hiroyuki Shinnou. 1998. Japanese Homohone Disambiguation Using a Decision List Given Added Weight to Evidences on Compounds (in Japanese). *Journal of Information Processing*, 39(12):3200–3206.

Koji Tochinai, Taisuke Itoh, and Yasuhiro Suzuki. 1986. Kana-Kanji Translation System with Automatic Homonym Selection Using Character Chain Matching (in Japanese). *Journal of Information Processing*, 27(3):313–321.

Sakiko Wakita and Hiroshi Kaneko. 1996. Extraction of Keywords for "Homonym Error Checker" (in Japanese). SIG Notes NL-111-5, IPSJ.

David Yarowsky. 1994. Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French. In *32th Annual Meeting of the Association for Computational Linguistics*, pages 88–95.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *33th Annual Meeting of the Association for Computational Linguistics*, pages 189–196.