Vol. 41 No. 4 情報処理学会論文誌 Apr. 2000

表記情報をデフォルトの証拠として用いた決定リストによる同音異義語の誤り検出

新納浩幸†

本論文では日本語文章中に生じる同音異義語の誤りを検出する手法を提案する．基本的にはYarowskyの提案した決定リストを利用する．さらに表記されている単語の情報（**表記情報**と呼ぶ）をdefaultの証拠として導入することで，誤り検出のF値を向上させる．

同音異義語の誤りを検出するには，同音異義語問題を解けば良い．そして同音異義語問題は語義選択問題と等価であるために，語義選択問題に対する種々の統計手法を利用して解くことができる．ただし同音異義語問題は語義選択問題とは明確に異なった面も持っている．それはほとんどの場合正解となる判別結果がすでに表記として現れていることである．同音異義語問題では表記されている単語を選択すれば判別の正解率が非常に高くなる．しかし，常に表記されている単語を選択すれば誤り検出を全く行わず，誤り検出システムとしての意味をなさない．

同音異義語問題の手法の評価は判別の正解率ではなく，誤り検出の正解率と再現率を考慮したF値で行うべきである．本論文では，F値を向上させるように，表記情報を利用する．具体的には統計手法として決定リストを利用する．そして表記情報をdefaultの証拠として決定リスト内に導入する．この表記情報の予測力の値は，訓練コーパスにおいてF値を最大にする値から得る．

Detection of Japanese Homophone Errors by a Decision List including a written word as a default evidence

Hiroyuki Shinnou†,?

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. We improve the original decision list by using written words as the default evidence. The improved decision list can raise the F-measure of error detection.

In order to detect homophone errors, we only have to solve the homophone problem for the homophone word. The homophone problem is equivalent to the word sense disambiguation problem. Consequently, we can solve the homophone problem by using various statistical methods proposed for the word sense disambiguation problem. However, the homophone problem has a distinct difference from the word sense disambiguation problem. In the homophone problem, almost all of the answers are given correctly. Therefore, the choice of the written word results in high precision. However, the method to always choose the written word is useless for error detection because it doesn't detect errors at all.

The method for the homophone problem should be evaluated by the F-measure to combine the precision and the recall. In this paper, we use the written word in order to raise the F-measure of error detection. To put it concretely, we use the written word as the default evidence of the decision list. The identifying strength of the written word is obtained by calculating the strength that gives the maximum F-measure in the training corpus.

本論文では，日本語文章中に生じる同音異義語の誤りを検出する手法を提案する．ここでは統計手法の一種である決定リストを用いる．また表記の情報を決定リスト中の証拠として組み込むことを試みる．これにより誤り検出のF値を向上させることができる．

近年，日本語文書もワードプロセッサで作成されることが多くなった．日本語ワードプロセッサにおいて，漢字からなる単語の入力は，一般に，まずその単語の読みに対応する平仮名文字列を入力し，システムが提示するその平仮名文字列の読みをもつ単語の中から目的の単語を選択することで行われる．このときの選択誤りが同音異義語の誤りとなる．同音異義語の誤りは単純な選択誤りの他に，“意志”と“意思”や“直感”と“直観”のように意味や用法の違いが微妙であり，その違いを知らないためにも生じる．これらの点から

† 茨城大学工学部システム工学科
Faculty of Engineering, Ibaraki University Department of Systems Engineering

日本語ワードプロセッサで作成された文書には同音異義語の誤りが生じやすい．そのため，それら誤りを自動検出，修正するシステムが必要となる．

同音異義語とは，同じ平仮名表記をもつ単語の集合とここでは定義する☆．「二重」が「ふたえ」と「にじゅう」という2つの平仮名表記をもつように，複数の平仮名表記をもつ単語もあるが，その場合は，その中のどれか一つと同じであれば，同音異義語の条件を満たしていると考える．また「組合わせ」と「組合せ」のような「表記のゆれ」によるものは，同じ平仮名表記をもつが，ここではこの種の単語の集合を同音異義語とは考えない．ここで対象としてる同音異義語は，例えば { 確率, 確立 } のような集合である．この集合の要素である単語は同じ平仮名表記「かくりつ」をもつので同音異義語となる．

同音異義語の中から正しい単語を選択する問題をここでは**同音異義語問題**と呼ぶことにする．同音異義語の書き誤りを検出するには，同音異義語のリストを予め作成しておき，そのリスト中の単語が出現した場合に，その単語に対する同音異義語問題を解けばよい．同音異義語問題を解くことで誤りの修正も同時に行える．ただし，文書校正支援では誤り検出が核となる処理であるため，本論文では誤り検出だけに注目する．

同音異義語問題に対しては，従来から様々な手法が用いられてきた [1)~5)]．これらはすべて同音異義語問題に特化した手法と言える．一方，同音異義語問題における平仮名表記を単語，漢字表記を語義ととらえれば，同音異義語問題は語義選択の問題と等価である．このため，従来提案されてきた語義選択の問題に対する種々の統計的手法 [6)] を用いることで，同音異義語問題を解くことができる．実際，同音異義語問題は文脈依存のスペルチェックの問題（peace と piece あるいは quiet と quite の書き誤りを検出する問題）と同じであり，文脈依存のスペルチェックの問題では，語義選択の問題に対する統計的手法が適用されている [7),8)]．

同音異義語問題は語義選択の問題と等価であるが，明確に異なった面ももっている．それはほとんどの場合正解となる判別結果がすでに表記として現れていることである．文中の “bank” の語義が「銀行」か「土手」かを判別することは困難だが，文中の “確立” の正しい表記が「確率」ではなく「確立」であることはほとんどの場合正しいはずである．同音異義語問題では表記されている単語を選択すれば判別の正解率は非常に高くなる．この表記の情報を利用すべきである．しかし，常に表記されている単語を選択すれば誤り検出を全く行わず，誤り検出システムとしての意味をなさない．つまり同音異義語問題に対する手法の評価は，誤り検出としての正解率と再現率から行うべきである．ここでは検出に対する正解率と再現率を統合して評価する F 値を用いる．

本論文では表記されている単語の情報（**表記情報**と呼ぶ）を統計手法内に取り入れることを試みる．これによって誤り検出としての F 値の向上を目指す．具体的には統計手法として決定リスト [9),10)] を利用する．決定リストは語義の判別に利用される証拠を予測力の強い順に並べたものである．実際の判別では，上位の証拠から，その証拠が文脈中に存在するかどうかを調べ，最初に見つかった証拠によって語義の判別を行う．リストの最下位の証拠は default の証拠と呼ばれ，リスト中の全ての証拠が存在しない場合に語義をある固定した語義に決めうちする．本論文では表記情報を default の証拠として決定リスト内に導入する．この表記情報の予測力の値を求めることが問題であるが，ここでは決定リストを作成するもとになった訓練コーパスから，F 値を最大にするような表記情報の予測力を求めた．

実験では 12 組の同音異義語に対して，従来の決定リストと表記情報を取り入れた本手法による決定リストを作成し，テストデータに対して誤り検出の F 値が向上することを確認した．

1. 決定リストによる同音異義語判別

本章では，同音異義語問題に対する決定リストの作成手順，およびその決定リストを用いた同音異義語の判別方法を述べる．この判別の枠組みは新納により示されたもの [11)] と同一であり，ここでは概略だけを示す．

1.1 決定リストの作成

step 1 同音異義語を設定する．

本論文では表 1 に示した誤りやすいと思われる 12 組の同音異義語について実験を行う．

step 2 同音異義語問題を解消するための文脈情報（証拠）を設定する．

本論文では，証拠として以下の3つの文脈情報を設定する．

- 直前の単語 w: $w-$ と表記する．
- 直後の単語 w: $w+$ と表記する．
- 前後に現れる自立語 w: 近いものから前後最大3つづつ取り出し，それぞれ $w \pm 3$ と表記する．

☆ 同音異義語には品詞が同じであるという条件をつける場合もあるが，ここではその条件はつけない．

表 1 同音異義語のリスト
Table 1 Homophone sets

平仮名表記	同音異義語
さいけん	{ 債券, 債権 }
かいほう	{ 解放, 開放 }
きょうちょう	{ 協調, 強調 }
じしん	{ 自信, 自身 }
かんしん	{ 感心, 関心 }
たいがい	{ 体外, 対外 }
うんこう	{ 運航, 運行 }
どうし	{ 同志, 同士 }
かてい	{ 過程, 課程 }
じっこう	{ 実効, 実行 }
しょくりょう	{ 食料, 食糧 }
しょうがい	{ 傷害, 障害 }

step 3 同音異義語 $\{w_1, w_2, \cdots, w_n\}$ 内の単語 w_i と証拠 e_j とが共起する頻度 $frq(w_i, e_j)$ を，コーパスから得る．

例えば，同音異義語を { 運航, 運行 } とし，以下の 2 つの例文をみる．

例文 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。」

例文 2 「 早朝深夜の運行時間が短縮された。」

例文 1 からは,「 運航」に対する証拠として，

“に ＋”，“の － ”，“飛行機 ±3”，“三 ±3”，
“風 ±3”，“支障 ±3”，“ない ±3”，

が取り出される．例文 2 からは,「 運行」に対する証拠として，

“時間 ＋ ”，“の － ”，“深夜 ±3”，“早朝 ±3”，
“時間 ±3”，“短縮 ±3”，“する ±3”

が取り出される．

step 4 証拠 e_j が生じている場合に，単語が w_i である予測力 $est(w_i, e_j)$ を対数尤度比を用いて以下のように定義する．

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

ここで $P(w_i|e_j)$ は以下のように近似する．

$$P(w_i|e_j) = \frac{frq(w_i, e_j) + \alpha}{\sum_{k=1} frq(w_k, e_j) + \alpha}$$

上式の α は，$frq(w_i, e_j) = 0$ の場合の不具合を回避するために設定している．本論文では $\alpha = 0.1$ とする☆．また， $default$ という特別な証拠も設定する．$frq(w_i, default)$ は w_i の総頻度とする．

step 5 $est(w_1, e_j), est(w_2, e_j), \cdots, est(w_n, e_j)$ の中で最も値の大きな $est(w_k, e_j)$ を取り出し，この w_k を証拠 e_j が現れたときの解答とする．またこのときの予測力は $est(w_k, e_j)$ である．

例えば，step 3,4 によって表 2 のようなリストが得られる．

☆ この程度の値を加えることが，最も簡単で効果的な対処方法であることは，論文 [9] に示されている．

表 2 証拠に対する解答と予測力の例
Table 2 Answers and identifying strength for evidences

証拠	「 運航」との頻度	「 運行」との頻度	解答	予測力
に+	77	53	運航	0.538
の−	252	282	運行	0.162
飛行機 ±3	4	0	運航	5.358
…	…	…	…	…
時間+	14	11	運航	0.345
深夜 ±3	0	48	運行	8.910
短縮 ±3	0	4	運行	5.358
…	…	…	…	…
$default$	1468	1422	運航	0.046

step 6 各 e_j に対して，e_j が現れたときの解答 w_{k_j} を求め，予測力 $est(w_{k_j}, e_j)$ が高い順のリストを作成する．これが決定リストとなる．ただし，$default$ に対する予測力以下のものはリストから外す．

以上より { 運航, 運行 } に対して表 3 のような決定リストが得られる．

表 3 作成できた決定リストの例
Table 3 Example of decision list

順位	証拠	解答	予測力
1	列車 ±3	運行	9.453
2	船 ±3	運航	9.106
3	深夜 ±3	運行	8.910
…	…	…	…
701	時間−	運行	0.358
…	…	…	…
746	の+	運行	0.162
…	…	…	…
760	$default$	運航	0.046

1.2 決定リストの利用

実際に決定リストを用いて，同音異義語問題を解くためには，まず文中から予め用意してある同音異義語のリスト中の単語 w を見つけ，step 2 で設定した w に対する証拠

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

を取り出す．

次に作成してある決定リストの最上位の証拠から順

に，その証拠が先ほど取り出した証拠の集合 E に属するかどうかを調べる．もし e_j が属していれば，e_j に対する解答 w_{k_j} が判別結果となる．w_{k_j} が w と等しければ，正しい表記であり，等しくなければ，w_{k_j} の書き誤りと判定する．

なお，本論文では「誤り検出」だけに限定し「誤り修正」までは行わない．このため，ここでは「正しい表記」あるいは「書き誤り」という判別結果だけが出力される．

2. 表記の利用

本章では，同音異義語問題に対して，表記情報を利用する背景と，決定リストと統合させて利用する方法を述べる．

2.1 誤り検出システムとしての評価

同音異義語の誤り検出は，同音異義語問題を解くことで行える．また同音異義語問題は語義選択問題と等価であるために，決定リストを用いて解くことができる．

しかし同音異義語問題は一般の語義選択問題とは明確に異なった面ももっている．それはほとんどの場合正しい判別結果がすでに与えられていることである．一般に同音異義語の書き誤る確率は非常に小さい．このため同音異義語問題の評価を

$$\frac{\text{正しい判別結果の数}}{\text{全判別数}}$$

という判別の正解率によって行うのであれば，文中の表記を判別結果とすればよい．この判別の正解率は語義選択問題に対する種々の手法よりも高いだろう．ただし文中の表記を判別結果とする判別手法は，誤りを全く検出しない．このため，本来の誤り検出の目的には何の寄与もしない．

つまり誤り検出システムの評価は，判別の正解率ではなく，誤りを検出するという観点から行なうべきである．ここでは情報検索で用いられる F 値を利用して誤り検出システムの評価を行う．F 値は以下のように定義された正解率 P，再現率 R を用いて，式 1 により定義できる．

$$P = \frac{\text{検出した誤りの中で実際に誤りであった数}}{\text{誤りだと判別した数}}$$

$$R = \frac{\text{検出した誤りの中で実際に誤りであった数}}{\text{存在する誤りの数}}$$

$$F = \frac{(1+b^2)PR}{b^2P+R} \tag{1}$$

式 1 の b は，再現率 R への重みを意味する．$b > 1$ ならば，再現率に重みをおいた評価になり，$0 < b < 1$ ならば，正解率に重みをおいた評価になる．正解率や再現率への重みは，文書校正システムの適用する文書やユーザの目的に依存する．そこで本論文では正解率と再現率の重みを等しく取ることにし，$b = 1$ と設定した．

2.2 表記情報の予測力の導入

文中の表記をそのまま判別結果とする手法では，検出システムとしては意味をなさない．ただし，文中の表記をそのまま判別結果とした場合に，判別の正解率が高くなることを考えれば，判別が難しい場合に限って文中の表記を判別結果とする利用法が考えられる．

問題はどのような条件になった場合に，文脈を利用した判別を放棄し，表記情報を採用するかである．ここでは表記情報を決定リスト中の証拠として導入し，その証拠に対する予測力を与えることを提案する．これにより決定リストの枠組みをそのまま利用して，表記情報を利用することができる．

2.3 表記情報の予測力の算出

表記情報の予測力を算出することを試みる．

まず表記情報の予測力を x であると仮定する．x よりも大きな予測力をもつ証拠を証拠 A と呼び，x 以下の予測力をもつ証拠を証拠 B と呼ぶことにする．

ある同音異義語の問題の数を T とする．この問題を従来の決定リスト（ここでは**原型** DL と呼ぶ）で解決することを考える．T のうち証拠 A で判断される問題の割合を G，証拠 B で判断される問題の割合を H とする．また証拠 A で判別された場合の正解率を g，証拠 B で判別された場合の正解率を h とする．またこの同音異義語が書き誤る確率を p とする．

証拠 A で判別された問題のうち，表記が正しい問題数は

$$G \cdot T \cdot (1-p) \tag{2}$$

と近似でき，表記が誤っている問題数は

$$G \cdot T \cdot p \tag{3}$$

と近似できる．式 2 の中では，判別が誤ったものが誤りとして検出されるので，式 2 の中から誤りとして検出される数は，

$$G \cdot T \cdot (1-p) \cdot (1-g)$$

となる．また式 3 の中では，正しく判別された場合に，誤りとして検出されるので，式 3 の中から誤りとして検出される数は，

$$G \cdot T \cdot p \cdot g$$

となる．つまり証拠 A で判別された問題からは

$$G \cdot T \cdot ((1-p)(1-g) + pg) \tag{4}$$

が誤りとして検出される．同様にして証拠 B で判別された問題からは

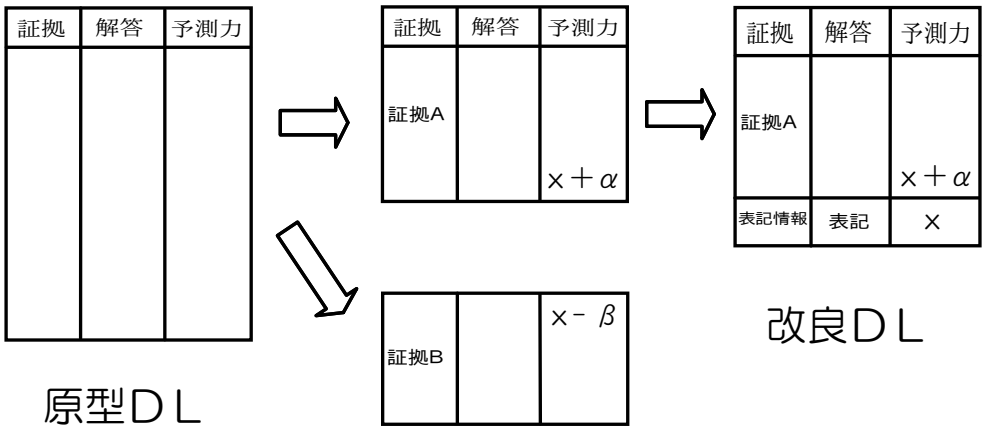


図 1 改良 DL の作成
Fig. 1 Construction of improved decision list

$$H \cdot T \cdot ((1-p)(1-h)+ph) \quad (5)$$

が誤りとして検出される．よって全体として，

$$T(G((1-p)(1-g)+pg)+H((1-p)(1-h)+ph)) \quad (6)$$

が誤りとして検出される．式 6 の中で正解となるのは，

$$T \cdot p(G \cdot g + H \cdot h)$$

なので正解率 P_0 は，

$$P_0 = \frac{p(G \cdot g + H \cdot h)}{G((1-p)(1-g)+pg)+H((1-p)(1-h)+ph)}$$

となる．また T の中の実際の誤りの数は Tp となるので，再現率 R_0 は

$$R_0 = G \cdot g + H \cdot h$$

となる．P_0 と R_0 を用いて，式 1 より，原型 DL を用いた場合の F 値 F_0 が計算できる．

一方，表記情報を原型 DL に埋め込み，証拠 A からは原型 DL の判別を行い，証拠 A からは判別できない場合にのみ，表記情報により判別することを考える．つまり証拠 A に $default$ の証拠として表記情報を加えた決定リスト（ ここでは**改良** DL と呼ぶ）を考える（ 図1 参照）．

証拠 A からの判別部分は同じなので，証拠 A からは 式 4 が誤りとして検出される．また改良 DL の場合，原型 DL において証拠 B で判別される問題は，すべて表記情報によって判別されるので，これらの問題からは誤りを検出しない．

結果として，改良 DL の場合，問題の数 T の中から誤りとして検出される数は 式 4 であり，その中で実際に誤りであるのは，$T \cdot G \cdot p \cdot g$ なので，正解率 P_1 は

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

となる．また，全体の誤りの数は $T \cdot p$ なので，再現率 R_1 は

$$R_1 = G \cdot g$$

となる．以上より表記情報を導入した決定リストを用いた場合の F 値 F_1 が式 1 より計算できる．

次に予測力 x を求める．これは $F_1 > F_0$ という条件を満たす中で，F_1 が最大になるような x を求めればよい．p はある定数であり，G, H, g, h は x に依存した数なので，F_1 は x の関数として表現でき，最大値を求めることができる．ただし，実際は，p は未知であり，G, H, g, h の具体的な式も分からないために，理論的な計算だけで x を求めることはできない．

ここでは，p を 0.05 に設定し，決定リストの作成のもとになった訓練コーパス（ 日経新聞の1 年分の記事）から同音異義語の問題を抽出し，作成できた決定リストにより G, H, g, h の値を求めた．たとえば，“運航” あるいは “運行” の単語を含む文は訓練コーパス中 2,890 文存在する．また訓練コーパスから作成された決定リストの予測力は表 3 に示すように 0.046 から 9.453 までである．今，適当に $x = 2.5$ とすると x 以上の予測力で判別される問題の数は 1,631 であり，この中で正しく判別されたものは 1,593 であった．これから $G = 0.564$，$g = 0.977$ が求まる．同様に x 以下の予測力で判別される問題の数は 1,259 であり，この中で正しく判別されたものは 854 であったため，$H = 0.436$，$h = 0.678$ となる．以上より，$x = 2.50$ の場合，$P_0 = 0.225$，$R_0 = 0.847$，$F_0 = 0.356$ および，$P_1 = 0.688$，$R_1 = 0.551$，$F_1 = 0.612$ が求まる．x を 0.0 から 10.0 まで 0.1 刻みで動かした結果を図 2, 図 3 および図 4 に示す．図 2, 図 3 および図 4 の

横軸は x の値を表す．そして，図 2 の縦軸は原型 DL の正解率 P_0 と改良 DL の正解率 P_1 を示している．また，図 3 の縦軸は原型 DL の再現率 R_0 と改良 DL の再現率 R_1 を示している．また，図 4 の縦軸は原型 DL の F-値 F_0 と改良 DL の F-値 F_1 を示している．また F_1 の最大値をとることで，{ 運航，運航 } の同音異義語に対する表記情報の予測力を 3.0 と設定できる☆．

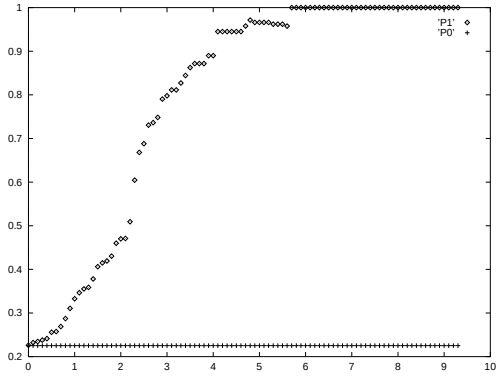


図 2 正解率 P_0 と P_1
Fig. 2 Precisions P_0 and P_1

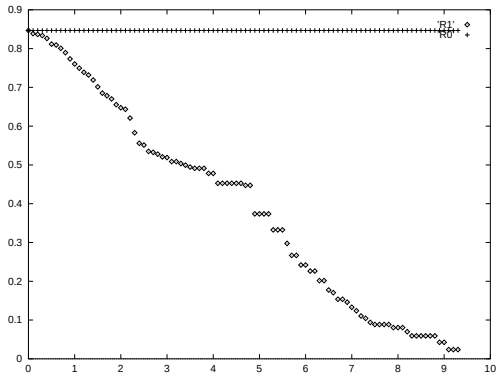


図 3 再現率 R_0 と R_1
Fig. 3 Recalls R_0 and R_1

3. 実　　験

表 1 の 12 組の同音異義語に対して，先の操作により表記情報の予測力を求めた．結果を表 4 に示す．最下位順位0 というのは，原型 DL の最下位の順位のことである．つまり，これは証拠 $default$ の順位を表す．最下位順位1 というのは，改良 DL の最下位の順位の

☆ 最大値を与える点は複数出るが，決定リスト中では順位だけが問題であるので，どれも大きな違いはない．ここでは最も小さな点で代表させた．

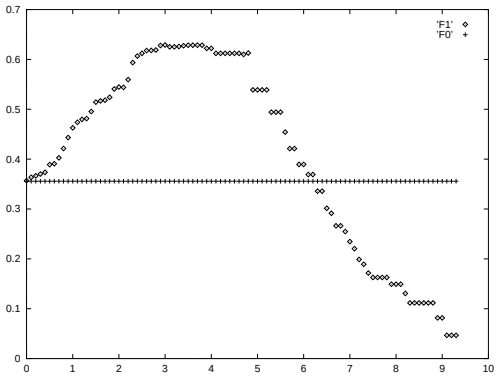


図 4 F 値 F_0 と F_1
Fig. 4 F-measures F_0 and F_1

ことである．つまり，これは表記情報の順位を表す．同音異義語 { 体外, 対外 } の場合，最下位順位0 と最下位順位1 が等しい．これは原型 DL の $default$ の証拠が表記情報に変更されたことを意味する．

表 4 表記情報の予測力
Table 4 Identifying strength of the expression

同音異義語	表記情報の予測力	最下位順位 1	最下位順位 0
{ 債券, 債権 }	4.9	844	1062
{ 解放, 開放 }	4.6	671	1104
{ 協調, 強調 }	4.3	667	1120
{ 自信, 自身 }	4.8	622	1134
{ 感心, 関心 }	5.7	424	1007
{ 体外, 対外 }	3.9	921	921
{ 運航, 運行 }	3.0	319	760
{ 同志, 同士 }	4.5	788	811
{ 過程, 課程 }	5.1	469	799
{ 実効, 実行 }	4.3	665	760
{ 食料, 食糧 }	4.7	255	697
{ 傷害, 障害 }	5.1	397	695

次に訓練コーパスとは別のコーパスである新聞記事 1 年分（毎日新聞'94 年度版）から上記の 12 組の同音異義語が含まれる文を取り出し，これをテスト文とする．このテスト文の同音異義語には誤りは含まれていないとする．

次にこのテスト文の 5% にランダムに誤りを混在させる．この誤りを含むテスト文を対象に，従来の決定リスト（原型 DL）による誤り検出と，表記情報を利用した本手法の決定リスト（改良 DL）による誤り検出の実験を行った．

この実験を 10 回行い，正解率，再現率，F 値の平均を求めた．結果を表 5 に示す．

すべての同音異義語に対して本手法の方が F 値が大きく，本手法の有効性が確認できる．

表 5 実験結果
Table 5 Result of experiments

同音異義語	問題数	原型 DL			改良 DL		
		正解率 P_0	再現率 R_0	F 値	正解率 P_1	再現率 R_1	F 値
{ 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
平均	1,824	0.460	0.906	0.581	0.560	0.794	0.648

表 6 誤り率を変えた実験結果
Table 6 Result of experiments for various error rates

誤り率	原型 DL			改良 DL		
	正解率の平均	再現率の平均	F 値の平均	正解率の平均	再現率の平均	F 値の平均
1%	0.171	0.907	0.271	0.237	0.785	0.347
5%	0.460	0.906	0.581	0.560	0.794	0.648
10%	0.584	0.901	0.690	0.685	0.792	0.731

またこの実験は誤り率 5% のテスト文に対する結果であるが，同様の実験を誤り率 1% のテスト文，及び誤り率 10% のテスト文に対して行った．結果を表 6 に示す．

テスト文の誤り率を変化させてた場合も，本手法の方が F 値が大きく，本手法の有効性が確認できる．

4. 考　　察

改良 DL の再現率は，原型 DL の再現率よりもよい結果になることはない．本手法は再現率を下げる代わりに，正解率を上げることで全体として F 値を向上させることを狙っている．2 章で示した式や 3 章の実際の実験でも，その点は確認できる．そのため再現率の方を重視した評価を行う場合には，本手法の効果は少なくなる．ただし，少なくなるだけであり，オリジナルの決定リストよりも悪くはならないことを注記しておく．

また同音異義語の誤り率をここではすべて等しく 0.05 としている．表記情報の予測力はこの値に依存するので，この値を正確に決める必要があるだろう．ただし実際の文書の誤り率が 0.05 よりも大きい場合でも，小さい場合でも，改良 DL は原型 DL よりも良い結果を出す．これは，再現率が p に依存せずに一定であり，しかも現実的に p に依存せずに P_1 が P_0 よりも高い値を出すからである．再現率が p に依存せずに一定であることや，P_1 が P_0 よりも高い値を出すことは，表 6 からも確認できる．

また今回の実験で用いた同音異義語はすべて 2 つの候補である．実際の同音異義語の判別候補はもっと多いことも考えられる．この場合，本手法の効果は更に上がると予想できる．なぜなら原型 DL の場合，高い予測力の証拠から判断できないときの判別の正解率が更に低くなるからである．このために改良 DL との差は更に大きくなる．

本手法の長所として，決定リストのサイズが小さくなることも挙げられる．決定リストのサイズは判別結果には無関係であるが，小さい方が計算の効率面で有利であろう．また小さい決定リストの方が手作業での保守もやりやすい．

一方，本手法の問題としては，証拠 A からの判別に表記情報を利用していない点があげられる．本論文では，文脈情報と表記情報を統合するために，表記情報に対する予測力を導入し，表記情報を文脈の一種として扱ったととらえられる．しかし文脈上の情報から正しい表記が「運行」であると判別できたとしても，実際に書かれている表記が “運航” であるか “運行”

であるかによって，その予測力は異なるとも考えられる．この点の考察を深め，証拠 A からの判別に表記情報を利用することを検討したい．

5. おわりに

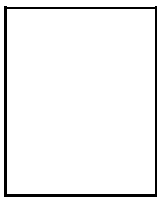
本論文では，同音異義語問題が語義選択問題と等価であることに着目し，同音異義語の判別に決定リストを利用した．さらに同音異義語問題は語義選択問題とは異なり表記情報が有用であることに着目し，表記情報を決定リスト中の証拠として導入した．その際に，表記情報の予測力を，訓練コーパスから計算する方法を示した．12 組の同音異義語を利用した誤り検出実験では，表記情報を組み入れた決定リストの方が，従来の決定リストよりも，高い F 値を出し，本手法の有効性を示すことができた．文脈情報と表記情報をさらに融合させる方法を検討することが今後の課題である．

謝辞 本実験で利用したコーパスは日本経済新聞 CD-ROM '90 版と毎日新聞 CD-ROM '94 版から得ています．利用を許可していただいた日本経済新聞社及び毎日新聞社に深く感謝します．

参考文献

1) 栃内香次, 伊藤太亮, 鈴木康宏: 前後連接文字を利用した同音語選択機能を有するかな漢字変換システム, 情報処理学会論文誌, Vol. 27, No. 3, pp. 313–321 (1986).
2) 伊吹潤, 徐国偉, 斉藤孝広, 松井くにお: 校正支援システム Joyner における表記誤りの訂正方式, 自然言語処理研究会 NL-117-21, 情報処理学会 (1997).
3) 奥雅博, 松岡浩司: 文字連鎖を用いた複合語同音異義語誤りの検出とその評価, 自然言語処理, Vol. 4, No. 3, pp. 83–99 (1997).
4) Oku, M.: Handling Japanese Homophone Errors in Revision Support System; REVISE, *4th Conference on Applied Natural Language Processing (ANLP-94)*, pp. 156–161 (1994).
5) 脇田早紀子, 金子宏: 変換ミスチェッカーのための辞書生成, 自然言語処理研究会 NL-111-5, 情報処理学会 (1996).
6) 藤井敦: コーパスに基づく多義性解消, 人工知能学会誌, Vol. 13, No. 6, pp. 904–911 (1998).
7) Golding, A. and Schabes, Y.: Combining Trigram-based and Feature-based Methods for Context-Sensitive Spelling Correction, *34th Annual Meeting of the Association for Computational Linguistics*, pp. 71–78 (1996).
8) Golding, A.: A Bayesian Hybrid Method for Context-Sensitive Spelling Correction, *Third Workshop on Very Large Corpora (WVLC-95)*, pp. 39–53 (1995).
9) Yarowsky, D.: Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, *32th Annual Meeting of the Association for Computational Linguistics*, pp. 88–95 (1994).
10) Yarowsky, D.: Unsupervised Word Sense Disambiguation Rivaling Supervised Methods, *33th Annual Meeting of the Association for Computational Linguistics*, pp. 189–196 (1995).
11) 新納浩幸: 複合語からの証拠に重みをつけた決定リストによる同音異義語判別, 情報処理学会論文誌, Vol. 39, No. 12, pp. 3200–3206 (1998).

(平成 11 年 8 月 6 日受付)
(平成 12 年 2 月 4 日採録)



新納 浩幸(正会員)

昭和 36 年生．昭和 60 年東京工業大学理学部情報科学科卒業．昭和 62 年同大学大学院理工学研究科情報科学専攻修士課程修了．同年富士ゼロックス，翌年松下電器を経て，平成 5 年 4 月茨城大学工学部システム工学科助手にて着任．平成 9 年 10 月同学科講師，現在に至る．博士(工学)．