自然言語処理 124－9
（1998. 3. 13）

複合語判定を優先させた決定リストによる同音異義語判定

新納 浩幸

茨城大学 工学部 システム工学科

本論文では同音異義語の誤り修正を目的に，同音異義語問題に対して Yarowsky の決定リストによる手法を試みる．更に，同音異義語問題において，もし同音異義語が複合語内に存在すれば，前後の単語から判断して問題を解決できる場合が多いというヒューリスティクスから，そのような証拠に対する予測力に重みをつける工夫も採り入れる．新聞記事１年分を利用した実験を通して，決定リストによる手法は同音異義語問題に対して有効であること，更に上記の重みをつける工夫はオリジナルの決定リストよりもよい精度が得らることを確認した．

Japanese Homophone Disambiguation using a Decision List weighted Evidences on Compounds

Hiroyuki Shinnou
shinnou@lily.dse.ibaraki.ac.jp

Ibaraki University.
Dept. of Systems Engineering
4-12-1, Nakanarusawa-cho, Hitachi-shi, Ibaraki-ken 316-8511, Japan

In this paper, in order to disambiguate Japanese homophone, we apply the decision list proposed by Yarowsky. Moreover, we improve the decision list by giving added weight to the identifying strength for evidences corresponding to words in front of and behind the homophone word in a compound. By experiments, we showed that the decision list is effective for the Japanese homophone disambiguation, and the weighted decision list is superior to the original decision list.

1 はじめに

日本語文章を自動的に校正する，あるいはそれを支援するシステムが有益であることは明らかである．しかし，文章中の誤りは様々な種類，様々なレベルがあり，高精度の文書校正システムの実現は難しい．本論文では文章中の誤りのうち特に，同音意義語の書き誤りに対象を絞り，同音意義語の書き誤りを自動発見するための手法について論じる．

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[1]．例えば，{「確率」，「確立」} 中の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．同音意義語の書き誤りを発見するには，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．

同音意義語問題に対しては，従来，前後の1文字から判定する手法 [1] があったが，その判定力には限界がある．また複合語内の同音異義語に対して，前後の文字や単語を調べる手法 [2, 3, 4] は，複合語内にない同音異義語には対処できない．また対象となる単語の回りの語列，品詞列などを手がかりとして，同音意義語問題を解消する試みもあるが [5]，手がかりの質や量の問題の他に，手がかりの組合せ方の問題も残っている．一方，同音意義語問題における平仮名表記を単語，漢字表記を語義と捉えれば，同音意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（例えば [6, 7, 8, 9, 10] など）を用いることで，同音意義語問題を解くことができる．

本論文では Yarowsky の提案した決定リストによる手法 [11, 12] を同音意義語問題に適用することを試みる．決定リストとは，語義選択に影響を与える文脈中の証拠を語義の予測力の順に並べたものである．予測力はその証拠 $evidence_i$ の基で，語義 $sense_a$ が選ばれる確率と，語義 $sense_b$ が選ばれる確率との対数尤度比で表される．

$$log(\frac{P(sense_a|evidence_i)}{P(sense_b|evidence_i)})$$

語義は文脈中の最も予測力の強い証拠から判定される．

ただし決定リストの予測力の順位は，直感的な判断とずれる場合がある．例えば，本論文で行なった実験では，「化学」と「科学」の選択の場合，『「化学」の直前に「積水」が現れる』という証拠よりも『「科学」の回りに「補助金」が現れる』という証拠の方が予測力が高いという結果が生じている．しかし，「積水科学」という表現がない以上，これは我々の直感に反する．これは，異なったタイプの証拠に対して，同一の重みで予測力を測るために，生じている．しかし，従来の研究にもあるように，複合語の一部を構成する単語の同音意義語の判定に関しては，その前後の単語を調べることで解決できる場合が多い．つまり決定リストにおける複合語の一部であることから生じる証拠は，通常の証拠よりも予測力が強いと考えられる．このためそのような証拠に対する予測力には，重みをつける方が，より適切な決定リストを作成できると考えられる．

本論文は，複合語の一部であることから生じる証拠に対する予測力に重みをつけた決定リストを作成し，同音異義語問題に対処した．実験では，予め用意した同音異義語 128 組と新聞記事1年分のコーパスを利用した．結果，決定リストは同音意義語問題で有効であることが判明した．また，通常の重みをつけない決定リストよりも，正解率が上がった．

2 決定リストによる同音異義語の判定とその問題点

本節では，同音異義語問題に対して，決定リストを用いる道筋，そこで生じる問題および本論文での対処方法について述べる．

[1] 同音異義語には品詞が同じであるという条件を含める場合もあるが，ここではその条件は入れない．

2.1 決定リストの作成

決定リストの作成は概略以下の手順に従う.

step 1 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.

step 2 同音異義語 $\{w_1, w_2, \cdots, w_n\}$ 内の単語 w_iと証拠 evd_jとが共起する頻度 $frq(w_i, evd_j)$ を，コーパスから得る.

例えば，同音異義語を {「衛星」,「衛生」} とし，以下の2つの例文をみる.

例文1 「一昼夜にわたった本番は、衛星放送で全国に中継された。」

例文2 「衛生上の問題が見つかった。」

例文1からは,「衛星」に対する証拠として,

"放送 +", "、 −", "本番±3",
"わたる±3", "一昼夜±3", "放送±3",
"全国±3", "中継±3"

が取り出される. 例文2からは,「衛生」に対する証拠として,

"上 +", "問題±3", "見つかる±3"

が取り出される.

step 3 証拠 evd_jが生じている場合に，単語が w_i である予測力 $est(w_i, evd_j)$ を対数尤度比を用いて以下のように定義する.

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

ここで $P(w_i|evd_j)$ は以下のように近似する.

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

上式のαは，$frq(w_i, evd_j) = 0$ 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[2]. また，$defalut$ 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 4
$est(w_1, evd_j), est(w_2, evd_j), \cdots, est(w_n, evd_j)$ の中で最も値の大きな $est(w_k, evd_j)$ を取り出し，この w_k を証拠 evd_j が現れたときの解答とする. またこの時の予測力は $est(w_k, evd_j)$ である.

例えば，step 3,4 によって表1のようなリストが得られる.

表 1: 証拠に対する解答と予測力の例

証拠	「衛星」との頻度	「衛生」との頻度	解答	予測力
公衆−	0	49	衛生	8.940
通信−	549	0	衛星	12.423
...	...	...	...	...
事業+	12	5	衛星	1.426
...	...	...	...	...
環境±3	6	111	衛生	4.187
...	...	...	...	...
default	3396	736	衛星	2.206

step 5 各 evd_j に対して，evd_j が現れた時の解答 w_{k_j} を求め，予測力 $est(w_{k_j}, evd_j)$ 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 $est(w_{k_j}, evd_j)$ 以下のものはリストから外す.

以上より {「衛星」,「衛生」} に対して表2のような決定リストが得られる.

[2] 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，[11] に示されている.

表 2: 作成できた決定リストの例

順位	証拠	解答	予測力
1	通信−	衛星	12.423
2	人工−	衛星	11.584
…	…	…	…
36	公衆−	衛生	8.940
…	…	…	…
1080	環境±3	衛生	4.187
…	…	…	…
1122	defaul	衛星	2.206

表 3: 直感と合わない決定リストの例

順位	証拠	解答	予測力
1	石油−	化学	12.577
…	…	…	…
169	役割±3	科学	7.033
170	補助金±3	科学	7.033
…	…	…	…
233	積水−	化学	6.794
234	空想−	科学	6.794
…	…	…	…

2.2 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 1 で設定した wに対する証拠

$$E = \{evd_1, evd_2, ..., evd_i\}$$

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 evd_jが属していれば，evd_jに対する解答 w_{k_j}が判定となる．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．

2.3 同音異義語問題における決定リストの問題点

決定リストは簡易であり，判定の根拠が直感的にわかりやすいため，語義選択問題に対してかなり有効であると考えられる．しかし証拠の予測力の順位は，言語的な直感に反する場合もかなりある．例えば，{「科学」,「化学」} の同音異義語に対して，我々の実験では表 3 のような決定リストが作成されている．

証拠に対する解答は妥当に見えるが，リストの順位には疑問がある．直感的に“積水−”や“空想−”は「積水科学」や「空想化学」という表現がない以上,「化学」あるいは「科学」を判定する決定的な証拠になるはずであり，“役割±3”や“補助金±3”という証拠よりも下位に位置するのはおかしい．例えば，これでは「積水化学からの補助金を...」という文章では,「化学」に対する同音異義語の判定が誤る．

これは $w \pm 3$ の証拠の予測力と $w-$ あるいは $w+$ の証拠の予測力とを同じ重みで評価していることから生じる．一般に同音異義語問題を解く場合，その単語が複合語の一部になっている場合には，前後の単語の情報が非常に有効である．そこで本論文では名詞であるような $w-$ あるいは $w+$ の証拠の予測に重みをつけて，決定リストの順位を作成することを試みる．具体的には対象となる本来の予測力の値に重みβ を乗じる．

$$est(w_i, evd_j) = \beta \cdot log(\frac{P(w_i|evd_j)}{1 - P(w_i|evd_j)})$$

これによってより適切な決定リストが作成できると考える．

3 実験

本手法の有効性を確認するために実験を行なった．まず同音異義語のリストとして，辞書のなかから特に誤り易いと思われる同音異義語 128 組を取り出した．一部を表 4 に示す．

表 4: 同音異義語のリスト

平仮名表記	同音異義語
いし	{意思，意志}
いぶつ	{異物，遺物}
うんこう	{運航，運行}
えいせい	{衛星，衛生}
おうせい	{王制，王政}
かいてい	{改定，改訂}
かいとう	{回答，解答}
かがく	{化学，科学}
かくりつ	{確率，確立}
かせつ	{仮説，仮設}
かねつ	{加熱，過熱}
かてい	{過程，課程}
...	...
やせい	{野生，野性}
れいじょう	{礼状，令状}
れんけい	{連携，連係}
ろてん	{露店，露天}

次に'90 年度日経新聞 1 年分の記事を 5 等分し，それぞれの 1/5 をテスト用にし，残りの 4/5 をトレーニングデータとして，128 組の同音異義語に対して，本来の決定リスト（List1）と，本手法による決定リスト（List2）を作成した．ここで重みは 2.6 に設定した[3]．テストデータで生じている同音異義語は表記がすべて正しいという仮定で正解率を測った．実験の結果を表 5 に示す．

まず，決定リストが同音意義語の判定に対して有効であることがわかる．また，本手法による決定リスト (List2) の方が，若干ではあるが，本来の決定リスト (List1) よりも正解率が高い．また，128 組の各同音異義語ごとに正解率を出し，その平均を求めると，List1 の場合が 89.93% であるのに対し，List2 の場合は 90.37% であった．この点からも本手法の工夫に効果があることがわかる．

次に正解率が List1 の正解率と比べて高かった同音異義語の 10 組を表 6 に示す[4]．これらの同音異義語問題に対して効果的であることがわかる．

表 6: 効果のあった同音異義語

同音異義語	List1	List2	差
{漂白, 漂泊}	84.69%	91.84%	7.15%
{露店, 露天}	86.13%	90.51%	4.38%
{反攻, 反抗}	73.97%	78.08%	4.11%
{重傷, 重症}	89.74%	93.57%	3.83%
{異物, 遺物}	78.13%	81.88%	3.75%
{学会, 学界}	84.53%	88.05%	3.52%
{帰路, 岐路}	80.65%	83.87%	3.22%
{占有, 専有}	80.71%	83.76%	3.05%
{台地, 大地}	81.78%	84.76%	2.98%
{野生, 野性}	87.50%	90.38%	2.88%

逆に正解率が List1 の正解率と比べて低かった同音異義語の 10 組を表 7 に示す．

表 7: 効果のなかった同音異義語

同音異義語	List1	List2	差
{渦中, 火中, 家中}	81.73%	75.00%	-6.73%
{王制, 王政}	77.33%	74.67%	-2.66%
{投降, 投稿}	81.58%	78.95%	-2.63%
{五感, 語感}	68.52%	66.67%	-1.85%
{断行, 断交}	93.49%	91.78%	-1.71%
{紙面, 誌面}	82.40%	80.80%	-1.60%
{侵食, 浸食, 寝食}	67.16%	65.67%	-1.49%
{原型, 原形}	75.97%	74.67%	-1.30%
{私服, 私腹}	96.10%	94.81%	-1.29%
{仮説, 仮設}	91.34%	90.25%	-1.09%

これらの同音異義語問題に対しては，本来の決定リストの方が正解率が高い．これは「科学研究」「化学研究」のように “研究+” からの判断では予測力が低い場合でも，重みをつけたことで，他のより有力な証拠よりも高い順位の予測力になってしまったために生じている．ただ

[3] 2.6 という値は重みを変化させた実験により，最良のものを選んだ結果である

[4] 本手法の正解率は data 1 ～ data 5 の正解率の平均から算出した．

表 5: 実験結果

テストデータ	データ数	Base	List1	List2
deta 1	49,476	39,702 (80.24 %)	46,723 (94.44 %)	46,982 (94.96 %)
deta 2	49,422	39,681 (80.29 %)	46,708 (94.51 %)	46,992 (95.08 %)
deta 3	49,373	39,664 (80.34 %)	46,690 (94.57 %)	46,896 (94.98 %)
deta 4	49,317	39,631 (80.36 %)	46,633 (94.56 %)	46,856 (95.01 %)
deta 5	49,270	39,610 (80.39 %)	46,585 (94.55 %)	46,843 (95.07 %)
平均	49,372	39,658 (80.32 %)	46,668 (94.53 %)	46,914 (95.02 %)

し，このように本手法の工夫が，逆に正解率を下げてしまった同音異義語は 128 種類中，19 種類あったのに対し，残りの 109 種類は本手法の方が正解率が高く，総合的にみると本手法による効果があることがわかる．

4 考察

4.1 低頻度の複合語表現について

決定リストの手法においては，低頻度の証拠の扱いが問題になる [13]．先の実験では単純に，頻度が 1 の証拠は決定リストには含めなかった．しかし，同音異義語の単語が複合語の一部であるような場合，その前後の単語からの証拠は他の同音異義語の単語を全く支持しないことも多い．この点を考えると，同音異義語の単語が複合語の一部であるような場合，その複合語が低頻度であっても，予測力を計算し，証拠としてリストに加えることも有益であることが予測できる．

追加実験として，頻度 1 の証拠でも，それが複合語の 1 部から生じているような場合には，決定リストに加える実験を行なった．結果的には，平均の正解率は 95.11% となり，更に向上したが，しかし副作用も更に大きくなった．これは低頻度のものを加えると，トレーニングデータ内に存在する誤りからの悪影響を受けてしまう点も原因としてある．例えば「ＩＣ内臓プリペイドカード」という複合語がトレーニングデータ中にあるが,「内臓」は「内蔵」の誤りである．「～内臓プリペイドカード」という複合語の頻度は 1 であるが,「～内蔵プリペイドカード」という語は出現しなかったために，*est*(内臓,“プリペイドカード+”) の値が高くなっている．

4.2 ベイズ分類器との比較

同音異義語問題に対しては決定リストではなく，決定木のようなベイズ分類器を用いることもできる．Golding は，スペル修正に対して，決定リストよりもベイズ分類器の方が優れていることを報告している [14]．本質的にはベイズ分類器は複数の証拠から判断することに対応しているので，決定リストよりも利用できる情報が多い．そのため決定リストよりも悪い結果になることはない．ただし直観的なわかりやすさ，インプリメントの容易さなどから決定リストを使う利点はある．

また決定リストで十分かベイズ分類器を用いるべきかは，設定した証拠に依存すると考える．本論文のように前後の単語と数語のウインドウ内の自立語を証拠とする場合には，大きな差はでないと予想する．

4.3 今後の課題

今後の課題として，以下の 3 点を挙げる．

- 記述された表記の利用

  同音異義語問題は他の語義選択問題とは明確に異なった面を持っている．それはほとんどの場合正解である表記が既に示されているという点である．同音異義語問題を語義選択問題として捉えた場合には，この情報を全く利用していない．この情報を統合して利用していく方法を考えるべきである．

例えば，スペルミスを雑音のある通信路モデル (noisy channel model) として捉える Kernighan らの研究 [15] では，正しいスペル α が文字列 β に書き誤る確率 $P(\beta \mid \alpha)$ を導入している．

- 同音異義語のリストの作成

同じ平仮名表記を持つ単語を同音異義語とここでは定義したが，同音異義語は一般に多数存在する．例えば，本実験では「かいとう」という平仮名表記に対する同音異義語として { 回答, 解答 } の 2 種類の単語をあげているが，実際は少なくとも

{ 垣内, 会党, 会頭, 回答, 灰陶, 快刀, 戒刀, 怪盗, 械闘, 開冬, 解党, 解凍, 解答, 解糖 }

の 14 種類の単語がある．これらを全て列挙することは容易だが，無用に問題を複雑にするだけである．同音異義語中には誤りやすい単語，ほとんど誤らない単語，めったに利用されない単語，区別せずにつかってよい単語などが混在している．同音異義語問題に対しては誤りやすい単語のリストを作成することが重要である．

- 既存知識の統合

コーパスだけから決定リストを作成するには，低頻度あるいは未出現の証拠の扱いが問題となる．この問題に対しては，コーパスを大量にするという戦略やスムージングの手法以外に既存知識を統合させる戦略も考えられる．例えば，辞書の見出しから同音異義語を構成している複合語を取り出し，そこから証拠を集めることも可能である．この場合，その証拠の順位を高くしておけば良い．辞書の例文も同音異義語を判定する非常に有益な情報となっているはずであり，これらも証拠として取り込む工夫も可能である．

5 おわりに

本論文では同音異義語の誤り修正を目的に，同音異義語問題に対して Yarowsky の決定リストによる手法を試みた．本論文では，更に，同音異義語問題において，もし同音異義語が複合語内に存在すれば，前後の単語から問題を解決できる場合が多いというヒューリスティクスから，そのような証拠に対する予測力に重みをつける工夫を採り入れた．実験により，決定リストによる手法と上記の工夫の有効性を示した．記述された表記の利用，同音異義語のリストの作成，及び既存知識の統合などを今後の課題とする．

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"，情報処理学会自然言語処理研究会，NL-117-21,pp.153–160 (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", In *Proceedings of 4th Conference on Applied Natural Language Processing (ANLP-94)* ,Vol.1,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"，情報処理学会自然言語処理研究会，NL-111-5,pp.27–32 (1996).

[6] Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer: "Word sense disambiguation using statistical methods", In *29th Annual Meeting of the Association for Computational Linguistics*, pp.264–270 (1991).

[7] Susan W. McRoy : "Using multiple knowledge source for word sense discrimination", *Computational Linguistics*, 18 (1),pp.1–30 (1992).

[8] Hwee Tou Ng and Hian Beng Lee : "Integrating multiple knowledge source to disambiguate word sense: An examplar-based approach", In *34th Annual Meeting of the Association for Computational Linguistics*, pp.40–47 (1996).

[9] J. Veronis and N. Ide : "Word sense disambiguation with very large neural networks extracted from machine readabledictionaries", In *Proceedings of COLING-90*, Vol.2, pp.389–394 (1992).

[10] David Yarowsky : "Word-sense disambiguation using statistical models of Roget's categories trained on large corpora", In *Proceedings of COLING-92*, Vol.2, pp.454–460 (1992).

[11] David Yarowsky : "Decision lists for lexical ambiguity resolution: Application to accent restoration in spanish and french", In *32th Annual Meeting of the Association for Computational Linguistics*, pp.88–95 (1994).

[12] David Yarowsky : "Unsupervised word sense disambiguation rivaling supervised methods", In *33th Annual Meeting of the Association for Computational Linguistics*, pp.189–196 (1995).

[13] Hang Li and Takeuchi,J : "Using Evidence that is both strong and Reliable Japanese Homograph Disambiguation", 情報処理学会自然言語処理研究会, NL-119-9,pp.53–59 (1997).

[14] Golding A.R. "A Bayesian hybrid method for context-sensitive spelling correction", In *Proceedings of VLC-95*, pp.39–53 (1995).

[15] Kernighan,M.D., Church,K.W. and Gale,W.A. : "A Spelling Correction Program Based on a Noisy Channel Model", In *Proceedings of COLING-90*, Vol.2, pp.205–210 (1990).