変換ミスチェッカーのための辞書生成

脇田早紀子、金子宏
日本アイ・ビー・エム東京基礎研究所

日本語校正支援システムに対して、変換ミスを漏れなく指摘してほしいという要請は以前からあった。我々が作っていたシステムでは、パターンという校正知識記述法と誤用語辞書を利用して変換ミスを検出していたが、余計な警告を出さない方針に徹していたので検出率は変換ミスの50%程度であった。

漏れなく警告するため同音異義語を持つ語をすべて警告することも考えられるが、警告数が多くなりすぎて実用的でない。

そこで、回りの語・語列・品詞・品詞列などを手がかりに、正しい変換らしいものを除いて警告数を抑える仕組みを作った。本発表では、その辞書を過去の文書の蓄積から自動作成する試みについて述べる。

Extraction of Keywords for "Homonym Error Checker"

Sakiko Wakita,Hiroshi Kaneko
IBM Research, Tokyo Research Laboratory

Users have required Japanese Critiquing System to find out all of homonym errors. Our system 'FleCS' had detected only about 50% of homonym errors, because we thought that to minimize wrong messages was the first priority.

There are so many words which have a possibility of homonym error (=have the same pronunciation with other words) that we can't afford to set warnings to all of them.

We made 'homonym error checker' which can tell each word is 'probably right' or 'maybe wrong' from 'keyword(s)' near it. This paper describes a examination to extract keywords from text automatically.

1.はじめに

変換ミスの可能性がある語に警告を出すにあたって、同音異義語を持つ語が出てくる度に警告するのではなく、回りの語・語列・品詞・品詞列などを「手がかり」にして警告数を抑える仕組み「変換ミスチェッカー」を作った。例えば「容疑者らを追求している。」は警告して「利潤の追求」は警告しないというように（「容疑者」「利潤」が「手がかり」になる）。

この「変換ミスチェッカー」の精度を決めるのは同音語辞書に登録する「手がかり」の質と量である。本発表では、この「手がかり」を過去の文書の蓄積から自動作成する試みを紹介する。

2.同音語チェッカーについて

2.1.目的

文中から、同音異義語を持つ単語を「正しく変換されているらしい（白判定）」「誤って変換されているらしい（黒判定）」「正誤不明（灰判定）」に分けて指摘する。実際には黒判定のみまたは黒＋灰判定のもののみを表示して使用する。同音異義語を持つ単語をすべて指摘するのに比べて警告数を激減させて見やすくするのを目標としている。[i]

2.2.同音異義語辞書の内容

形態素解析を行うための通常の辞書の他に、同音語を持つ語の「危険度チェック」のために「手がかり」を登録しておく「同音異義語辞書」を用意する。以下にその例を掲げる。

表 1　同音異義語辞書の例

[よういん]
***要員**　名詞
直後：の　を　数　枠　面
近く：安全　医療　介護　援助　必要　不足　任務
　　派遣　参加　装備　展開　撤回　内訳　配置
　　代替　保安　足り(る)　送(る)
前後：[サ変名詞]～[読点]
***要因**　名詞
直前：主　諸　一
直後：が　と　に
近く：悪化　圧迫　安定　円安　円高　価格　回復
　　外部　外的　考慮　構造　減少　分析　複雑
　　輸出　抑制　気象　阻害　一つ　除(く)
　　増(す)　多(い)　大き(い)
前後　[が]～[と(引用)]

例に示したように、一つの読みに対して複数の見出し語が登録されていて、さらにそれぞれの見出し語に対して「手がかり」が列挙されている。「手がかり」とは、その見出し語の前後または近くによく出現する語・語列（字面と品詞のいずれかまたは両方で表現する）のことである。

2.3.しくみ

単語それぞれに登録されている「手がかり」が近くにあるかどうかを調べる。例えば、「～は円高の要因になった。」は「要因」の「手がかり」「円高」があるので白判定、「～要員を配置した。」は「要員」の「手がかり」「配置」があるので黒判定。同じ読みで登録されているどの単語の「手がかり」も見つからなければ灰判定になる。「手がかり」が競合する場合はそれぞれの「手がかり」の「ポイント」合計で比較して決める。「ポイント」は「手がかり」の種類によって決めている。[ii]

3.「手がかり」の自動生成実験

3.1.実験の目的

いうまでもなく重要なのは「手がかり」の質と量である。人手で「手がかり」を作成すれば「ちょうどこれが特徴」といえるような質のよいものができる場合もあるが（そううまく思い付かない

ことが多い）、あまりに膨大な作業になってしまうので、やはり大体のところは機械的に生成しておきたい。本実験の目的は、生成の材料とする文章の量・生成の方針・結果として得られる精度の関係について調べることである。

今回は以下のような3種類の抽出方法で、過去の文書の蓄積から「手がかり」を拾い上げることにした。

A. 直前の品詞＋直後の品詞
B. 直前または直後の自立語（字面＋品詞）
C. 近くの名詞（字面＋品詞）

上記の条件にあてはまるものであれば、文中に一度でも出現した語はすべて登録した。結果として、同音異義語の両方と現れた語は相殺される。

この実験で、

- どのくらいの量の文書を材料とすれば十分か
- 「手がかり」抽出方法3種類それぞれの特徴は
- どのくらいの白判定を出せるか（＝警告数をどこまで減らせるか）
- 黒判定がどのどの程度でおさまるか（＝検出率が損なわれずにすむか）
- どんなものを同音語チェッカーに使ったらよいか

を調べたい。

3.2.材料

産経新聞の記事（校正済み）1994年6月分から9月分まで1932万文字。便宜上2万5千文字程度毎に77個の記事に分割し（以後このこの一つ分を文章の量の単位とする）、記事番号20,40,60をテスト用に、それ以外を手がかり語生成用に用いた。

記事一つ分から生成した「手がかり」ごとに辞書に追加しながら、テスト用の記事に対する白判定・灰判定・黒判定の数を調べる作業を「手がかり」の種類別に行った。

3.3.結果

「手がかり」抽出に用いた記事の量と白／灰／黒判定の数の関係を以下に示す。

「手がかり」が登録されていない段階では当然すべてが灰判定（今回のテスト用文書では25,268個）である。材料とするテキストの量を増やしていけばだんだん白判定が増えるが、あるところまでくるとあまり増えなくなってしまう。一方、黒判定も少しは出てくる。

各抽出方法の特徴を見てみよう。

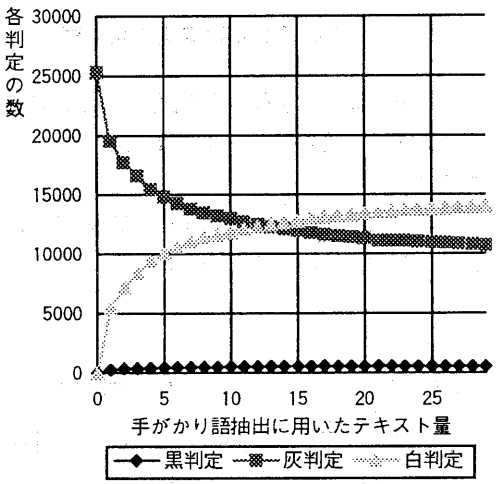

図 1 A.直前の品詞＋直後の品詞

「A.直前の品詞＋直後の品詞」では比較的少ない量のテキストで多くの白判定を出すようになるが、後の伸びはそれほどでもなく、結局6割程度が限界である。

黒判定は2%程度で安定している。

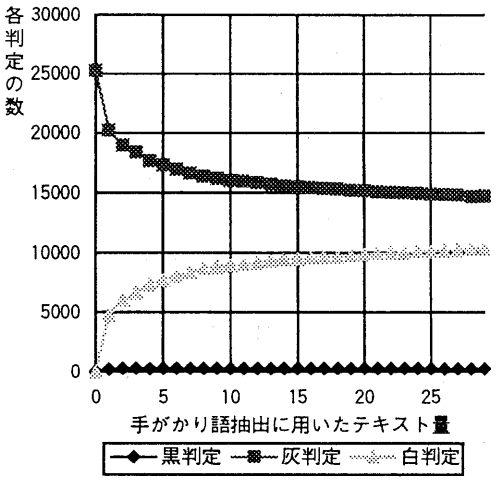

図 2 B.直前または直後の自立語

「B.直前または直後の自立語」では白判定があまり増えず、結局 4 割程度の警告を減らすにすぎない。

しかし黒判定は非常に少ない（1%程度）まま安定しており、増えていかない。

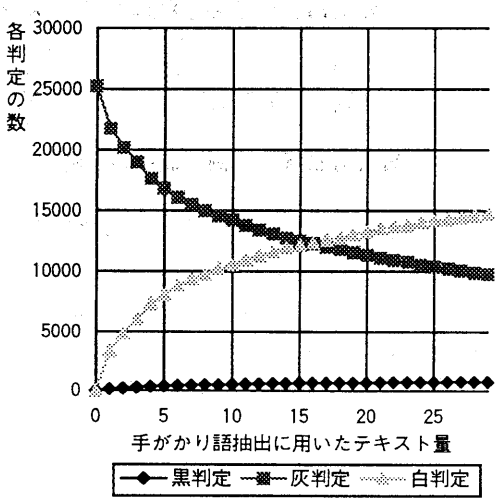

図 3 C.近くの名詞

「C.近くの名詞」の場合、立上りはあまり鋭くないものの、白判定がジリジリと伸び続けている。テキスト 29 個分で 6 割強を白判定にできたが、もう少しは伸びそうである。

ただし、黒判定は比較的多く、3%程度あり、こちらもジリジリ増える傾向にある。

一方、各段階での「手がかり」の数も見ておこう。

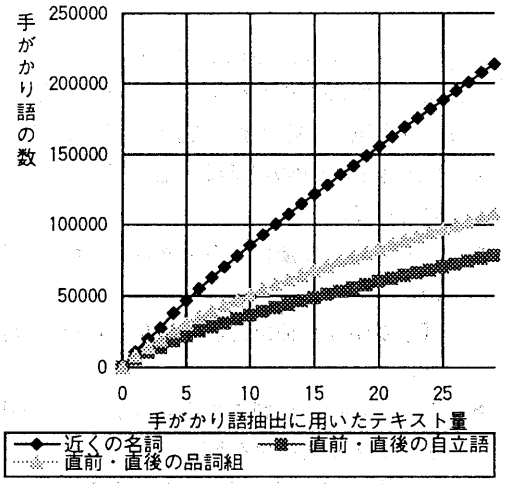

図 4 「手がかり」の数

材料のテキストが増えるのに対してほとんど直線的に「手がかり」数が増えている（辞書が大きくなる）。効果の方は頭打ちになってきているので、ここから先あまりむやみに材料を増やしても効率が悪いことがわかる。

さて、抽出方法はそれぞれに特徴があって一長一短であることがわかったので、この 3 通りを組み合わせることでどうなるかを調べてみた。

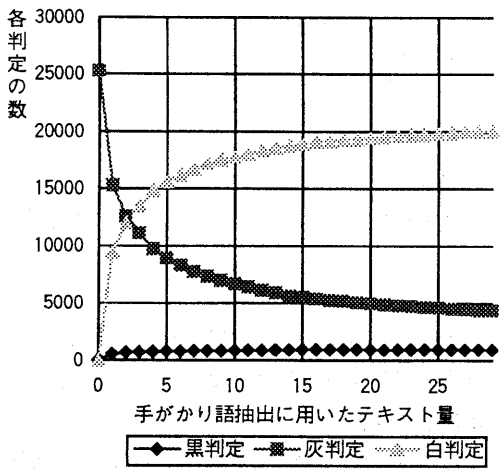

図 5 A+B+C.すべての手がかり語

白判定は8割まで増える。つまり、警告の量を5分の1に減らせる。一方で黒判定はやや多い(3.6%)。

3.4.結果のまとめ

材料となる文章が多くなってくると、

- 白判定は始めのうちは順調に増えるが、テキスト20～30個分（1～1.5Mbyte）でほぼ頭打ちになる。
- 「手がかり」の数の方は、そのあたりでもまだ直線的に増え続ける。
- 黒判定は 1～3%程度で、ほとんど落ち着いている。

そして、

- 3種類の「手がかり」を合わせれば、8割程度を白判定にできる。

4.発展

4.1.黒判定を減らす工夫

正しい判定に対して黒判定が出るということは、裏を返すと、この分くらいは変換ミスを白判定にする（＝見逃す）ということなので、この率は非常に低くするのが望ましい。

前章の実験では、一度登場したものをすぐ「手がかり」として採用してしまったが、採用の基準を少し厳しくすることを考えてみる。

例えば、同音語であるAとBについては、Aの近くにxが2回以上出てきてしかもBの近くには1回も出てきていないとき、またはAの近くにはBの近くに出てきた回数の10倍以上出てきたとき、初めて「手がかり」として採用する。

この基準では当然のことながら、より多くの文書を使わないと「手がかり」が集められない。今回用意した文書すべて(74個)を使って、前章と同様のテストをして比較した。

表 2 「手がかり」採用基準による差

	前章の基準	厳しい基準
白判定	20,033	20,665
灰判定	4,324	4,081
黒判定	918	530
「手がかり」数	399,613	125,090
テキスト量	29	74

すると、前章の場合を上回る白判定(82%)を出していて、黒判定は 4 割以上減り、辞書サイズは激減した。テキスト量を余裕をもって(1800万文字以上)集められるなら、「手がかり」採用基準を厳しくした方がよい。

この基準でするなら、もっと材料テキストを増やしてもよいかもしれない。

4.2.灰判定で残るもの

灰判定で残っているものを見ると、

[左党もＯＫ]「喜易まんじゅう」の最大のウリは、穏やかな甘さ。

の「さとう（左党／砂糖）」のように、人間が見てもちょっとどれを「手がかり」にしたらよいか迷うものもあるが、なんとかなりそうなものがほとんどである。

就職したくて苦闘の最中の女子学生には、

は、「くとう（苦闘／句読）」だが、「句読」はほとんど「句読点」しか使わないので「くとう」はまるごと同音異義語辞書からはずして従来の校正支援辞書の枠組みで扱えばよいだろう。その他、片方の頻度が極端に少ないものや、片方の使い方が限定されているものなどははずしてしまったほうが全体として見やすくなる。

人を見る目には自信がございます。

「じしん（自信／地震／自身／磁針）」では、「近くの名詞」単独では『手がかり』として登録されたものがなかったわけだが、「見る目」とか「～がございます」とあれば「自信」の可能性が濃いことがわかるので、ちょっと違った『手がかり』収集をするとうまくいきそうである。

「そんなことはお安い御用です」

の「ごよう（御用／誤用）」や、

最近受けがいい。

の「さいきん（細菌／最近）」も、『手がかり』の取り方で工夫できそうな例である。

5.まとめ

『変換ミスチェッカー』のための『手がかり』を、過去の文書の蓄積から抽出する実験を行った。抽出方法は3種類を試した。それぞれ一長一短だったが、組み合わせることによってかなりの精度(82%減)を出すことができた。

また、『手がかり』の取り方の工夫などでもう少し向上する余地もありそうである。

現在、今回の実験で作成した辞書（採用基準を厳しくした方）を用いて産経新聞社で『変換ミスチェッカー』のテスト使用をしている。警告の数を 17%程度に抑えたとはいっても、『変換ミスチェッカー』の警告数は新聞記事編集用ワープロ一画面に対して 2.3 個。これは既存の FleCS（我々の開発した校正支援システム）の出す警告の合計に迫るものである。それなのに実際の変換ミスは大変少ない（数画面に一個程度）。

今までの警告（赤、黄、緑）とは色を変えるなどの工夫をしている（灰）ものの、『変換ミスチェッカー』が現場で受け入れられるかどうかは今後の運用にかかっている。

謝辞：本研究に多大なご協力をいただいている産経新聞社校閲センターの方々に感謝いたします。

---

[i] 奥村ほか：日本語校正支援における同音語誤り検出－警告レベル分けの提案，情処 49 全国大会 3K-6,(1994)

[ii] 脇田ほか：日本語校正支援における同音語誤り検出－警告レベル分けの判定基準，情処 50 全国大会 5R-3,(1995)