自 然 言 語 処 理 97－19
（１９９３．９．１７）

# 日本語校正支援システムＦｌｅＣＳ
## －　ミスタイプ検出について

脇田早紀子、奥村薫
日本アイ・ビー・エム(株)東京基礎研究所

日本語の場合、単純に見えるタイプミスでも機械に発見させようとするとけっこう難しい。タイプミスを検出するには、形態素解析を行いその失敗箇所を警告する方法が一般的だが、誤り文でも形態素解析ができてしまうことが多いので、この方法だけでは不十分である。本研究は、「形態素解析失敗」では発見できないミスタイプ文を発見することを目的としている。ミスタイプ文の特徴を記述した検出ルールの原型と、それを実用レベルにまで高めるために必要な修正作業について述べる。

# Japanese Critiquing System FleCS
## - How to detect typographical error

Sakiko Wakita, Kaoru Okumura
IBM Research,
Tokyo Research Laboratory

It is hard to make useful Japanese spell-checker. General way to detect typographical error is to regard failure of morphological analysis as the indicator of error position, but it's not perfect. In this paper, we propose rules to detect typographical error which can't be detected by the general method.

## 1　はじめに

「日本語スペルチェッカー」の必要性は誰もが認めるところである。英語のスペルチェッカーはかなり実用的な道具としてすでに広く使用されているが、日本語では気軽に使えて信頼に足るものがまだできていない。日本語の場合、英語などと異なり、単語の区切りが明白でないこと、表記の方法にいく通りもあることなどの条件により、いわゆる「単純な」ミスタイプであっても完全な検出は難しい。

日本語の場合、「スペルチェッカー」というより「校正支援システム」と呼ぶべきものの方がつくりやすい。つまり、人間にとってはめんどうな送り仮名・表外字・カタカナ表記、さらにはもう少し複雑に見える文体チェックなどはかなり機械でできるところなのに、単純きわまりなく見えるミスタイプ検出は意外と難しいのである。我がＦｌｅＣＳも「校正支援システム」と称しているが現在のところ「スペルチェッカー」とはいいにくい。人間にとっては（気付きさえすれば）あたりまえでくだらない間違いをかえって見逃すかもしれないからである。

変換ミスの中でも、よく間違える「越える／超える」「務める／勤める／努める」などについては、対象をしぼりやすい分だけ対策のたてようもあろう。しかしちょっと見ればだれでもわかるはずのうっかり変換ミス・打ち間違いほど毎回違った現れ方をするので、予測がつきにくい。予測できないミスタイプを発見するには、形態素解析を行いその失敗個所を警告するという手法が一般的である。この方法だけでも、例えば「検出するとば、」「警告をを、」のような文なら発見できる。しかし「物価は西側の水準の近づいている。」「関心できません。」「言葉と音間」などのような、辞書にある単語と接続でいちおうつながってしまうものは発見できない。

とはいえ、ミスタイプ文を無理やり形態素解析した結果を見ると、それなりに「不自然な」感じのすることも多いものである。本研究では、形態素列や文字種・単語長に現れる「不自然さ」を利用してミスタイプを発見するルールを考え、さらにそのルールの精度を実用レベルにまで高める工夫を行う。

## 2　本研究の目的

本研究の目的は「形態素解析失敗」とはならない種類のミスタイプ文を発見するルールを作成し、さらにそれを実用レベルまで高めることである。すなわち有用な警告を出し、かつ余計な警告を出し過ぎないようにすること。

## 3　概要

本論文は以下のような構成になっている。

・「形態素解析失敗」にならなかったミスタイプの例文を収集・分類し、原則ルールにまとめる。

・原則ルールをそのまま用いたときに出る過検出（正しい文に当てはまってしまい余計な警告がでること）を洗い出し、対策をたてる。

・例外処理を付け足して修正ルールを作成する。

・原則／修正ルールそれぞれを評価し、考察する。

## 4　例文と原則ルール

校正前の新聞原稿から、形態素解析をしただけでは「接続不良」として警告を出せなかったミスタイプの例文を収集した。これら本研究の出発点として用いたミスタイプの例文は百余りであった。形態素解析の結果と合わせてそれらを検討し、同じ特徴を持つものをまとめてグループを作り、それぞれを「パターン」で表現して「このパターンにあてはまるものはミスタイプとして警告する」という原則ルールを作成する。以下に各グループの例文とＦｌｅＣＳが出す形態素解析結果、およびその特徴を記述したパターン[3)]を示す。'/' は（本システムでいうところの）文節の切れ目を表す。

・グループ１ ： 不自然な文節末

文節の終わり方が唐突に見えるものがかなりあった。例えば名詞があるとき、助詞がついて文節が終わるのなら自然だが、いきなり動詞がつなが

っているような場合である。名詞のほか、助詞・助動詞で通常は読点・句点なしに文節が終わらないようなものも含めた。

```
そのことを指し手いる。
 →指し手(名詞)/い(一段動詞)る
関心できません。
 →関心(名詞)/でき(可能助動詞)ません
自分の足出歩けて、
 →足(名詞)/出歩(カ行五段)けて
刺激され願っている。
 →され(助動詞連用形)/願(ワ行五段動詞)って
認められいてないため、
 →られ(助動詞連用形)/い(一段動詞)て
それを考えるの自分の仕事ではない。
 →考えるの(終助詞)/自分(名詞)
```

pattern: [@a:名詞|使役助動詞連用形
|受身助動詞連用形|断定助動終止形
|終助詞&文節末]
(ただし|はor、&はand、!はnotの意味の論理演算子。優先順位は !,|,& の順)

・グループ２ ： 連体形に用言

動詞や助動詞の連体形・連体詞など、体言が来ることを予想させる品詞がありながら体言が続いていない、というものも多かった。体言でなく動詞・形容詞・形容動詞・副詞・接続詞が来るものはミスタイプの可能性があると考えた。

```
読んだわけた。
 →読んだ(過去助動詞連体形)
   /わけ(一段動詞)た。
導く努力しよう。
 →導く(動詞連体形)/努力(サ変)しよう。
手荷物の入って袋を抱えた。
 →手荷物の(格助詞「の」)/入(ラ行五段)って
考えていたただきたいことがあります。
 →いた(過去助動詞連体形)/ただ(副詞)/
   き(カ変)たい
はっきりつかるない。
 →つかる(動詞連体形)/な(形容詞)い
辞めるのと辞めないとのでは、
 →辞めないとの(格助詞「の」)では(接続詞)
```

pattern: [@a:動詞連体形|助動詞連体形
|格助詞「の」|連体詞&文節末]
[動詞|形容詞|形容動詞|副詞|接続詞]

・グループ３ ： 不自然な文節頭

文頭でない文節頭に形式名詞・サ変語尾が来ているものも多かった。

```
異様で危険やものだった。
 →危険や(助詞)/もの(形式名詞)だった。
上にいくにためには、
 →いくに(格助詞「に」)/ため(形式名詞)
取引を会ししたが、
 →会し(サ変連用形)/し(サ変連用形)たが、
こうした行動をとるとことはないだろう。
 →とると(接続助詞)/こと(形式名詞)は
```

pattern: [@a:形式名詞|サ変語尾&文節頭]

・グループ４ ： 短い単語列

ミスタイプの文を無理やり形態素解析したものは、短い単語の列となることが多い。

```
特色豊だった。
 →特色(名詞)豊(固有名詞・名)だった。
名人に勝事もあるが、
 →勝(固有名詞・名)事(名詞)も
```

すなわち、漢字一文字の固有名詞が他の固有名詞や固有名詞接頭・接尾を伴わずに現れるときはミスタイプ（この場合送り仮名抜け）の可能性がある。

pattern: [!固有名詞(姓)
&!固有名詞(名・姓)接頭]
[@a:固有名詞(姓)|固有名詞(名)&漢字
&(単語長==1)]
[!固有名詞(名・姓)接尾&!固有名詞(姓)]

```
言葉と音間。→音(名詞)間(接尾語)
論義を呼ぶ。→論(名詞)義(名詞)
```

漢字一文字の名詞・接頭・接尾がつながっている

ときはミスタイプの可能性がある。

pattern: [@a:名詞|接頭語|固有名詞|サ変名詞
&漢字&(単語長==1)]
[@a:名詞|接尾語|固有名詞|サ変名詞
&漢字&(単語長==1)]

長い目で見みれば、
→目で/見(一段動詞連用形)/み(一段動詞)れば

pattern: [@a:漢字&(単語長==1)&文節頭&文節末]

同様に、平仮名の短い単語に切られている文も疑わしい。

pattern: [@a:平仮名&(単語長==1)&文節頭
&文節末]

pattern: [@a:平仮名&(単語長==2)&文節頭
&文節末]
[@b:平仮名&(単語長==2)&文節頭
&文節末]

・グループ５ ： 存在しない複合動詞

よく使われる複合動詞はもともと一単語で登録されているので、切って解釈しているということは、本当はつながらないのに複合動詞として解釈されている可能性がある。

電話していもす。
→い(一段動詞)も(サ行五段)す。

pattern: [@a:動詞語幹]
[@b:五段動詞連用形語尾]*[@c:動詞]

・グループ６ ： 接続詞「が」「して」・感動詞

なぜが、まだ下を向いている。
→なぜ(副詞)/が(接続詞)、
入院者ははいなかった。
→入院者は/はい(感動詞)/な(形容詞)かった

「が」や「して」が接続詞として使われるのも、感動詞が出てくるのも文頭ぐらいのはず。

pattern: [@a:"が"|"して"&接続詞&!文頭]

pattern: [@a:感動詞&!文頭]

・グループ７ ： 助詞を漢字変換してしまった

修学旅行画始まるのに、
→旅行(名詞)画(名詞)/始ま(動詞)るのに、
彼の発明出ある。
→発明(名詞)出(接尾語)あ(動詞)る。

pattern: [@a:"名"|"出"|"画"&単語&文節末]

pattern: [@a:"出"&単語&!文節頭]["あ"|"、"]

・グループ８ ： なさそうな接続

昨年実六万円下回る見通しだ。
→昨年(副詞的名詞)/実(接頭語)六万(数詞)円
ポスターやＰＲ紙がなど一般印刷部門。
→ＰＲ紙が(格助詞)など(副助詞)

pattern: [@a:"が"&格助詞][@b:副助詞]

pattern: [@a:一般接頭語][@b:数詞]

・グループ９ ： 助詞三連鎖

立ててのは、
→立て(一段動詞)て(接続助詞)の(格助詞)
は(係助詞)、

pattern: [@a:助詞][@b:助詞][@c:助詞]

## 5 原則ルールの生む過検出とその対策

前項で述べた原則ルールをＦｌｅＣＳシステムに実装し、産經新聞社に実験的に導入した。これまで使われていたルールとは警告レベルを変えて表示し、区別できるようにしておいたが、ともかく警告がたくさん出過ぎるので当初は評判がひどく悪かった。

過検出の原因はいくつかあるが、まずルール自体の欠陥とは違うものから述べる。

原因その１：未登録単語

未登録単語なのに、いままで警告がでていなかったものが一気に現れてしまった。例えば漢字一文字の名詞連鎖として解釈されていた未登録の人名（グループ４のパターン）である。これらはむしろ積極的に警告を出し、発見し次第登録していくべきものではあるが、始めのうちはわずらわしい。

原因その２：形態素解析の誤り

辞書のキズ・コストの付け間違いなどの原因により、形態素解析結果が誤る文があっても、これまで警告がでないため気づかずに済んでいたものが表に出てきた。発見するたびに直していくしかない（ちょっと恥ずかしいのだが）。

原因その３：形態素解析のチューニング不足

形態素解析結果として間違いではなくても、ミスタイプ検出ルールの都合に合う結果ではない場合。主に文節の切れ目などが問題となる。例えば「解決可能な」を「解決(名詞)/可能(形容動詞)な」と切ってしまうとグループ１の「不自然な文節末」にひっかかってしまう。これは「可能な」が直前の名詞とつながりやすい単語なので、特に接続を記述して文節が切れないようにしてしまえば解決する。一般の形容動詞は名詞とつながらないので「机/きれいな」ならば警告できる。同様に、「注目され始めている」を「注目され(助動詞連用形)/始め(一段動詞)ている」と切らずに、「始める」は「～し始める」「～され始める」と接続するようにすることにより不必要な警告を抑えられる。

そして、最大の原因は

原因その４：原則ルールにあてはまるが正しい文

予想されたことではあるが、原則ルールにはひっかかってしまうのにミスタイプでない文は大量に存在する。この対策として行うルールの修正については次章で述べる。

## ６　修正ルール

前章の「原因その４」を取り除くため、それぞれのルールについての過検出例を収集し、例外としてひとつずつ記述していく。

・グループ１：不自然な文節末
にあてはまってしまうもの

| |
|---|
| だれからも愛され慕われるように。<br>→愛され(助動詞連用形)/慕われるように。 |

次の文節の中で同じ助動詞が繰り返されていれば警告しないようにする。

・グループ２：連体形に用言
にあてはまってしまうもの

| |
|---|
| 予算案に関する突っ込んだ議論。<br>→関する(動詞連体形)/突っ込(動詞)んだ<br>つながりの深い事例である。<br>→つながりの(格助詞)/深(形容詞)い |

次の文節も体言に続く格好をしていれば警告しないようにする。つまり「予算案に関する突っ込んだ。」や「予算案に関する突っ込み始めて、」などは警告するが「予算案に関する突っ込んだ＊＊」なら警告しない。

・グループ４：短い単語列

| |
|---|
| 同国は…→同(接頭語)国(名詞)は |

同～、各～、～側、～内などの場合は、一文字漢字の連鎖でもおかしくないことが多い。

| |
|---|
| 雅子さんの母、→雅子さんの/母(名詞)、 |

一文字の漢字が単独で文節をつくっていても、直前が平仮名で直後が読点のような場合は無理やり切った結果ではないと推定できる。

・グループ５：存在しない複合動詞

よく使われる複合動詞は登録しておくとはいっても、「投げかける」「呼びかける」「やりかける」「言いかける」…と登録していたらきりがない。「…かける」をまとめて例外すると都合がよい。いろいろな動詞につきやすい動詞はすでに、特に接続を指定できる辞書に登録してあるのでそ

れを利用し、２つ目の動詞がその辞書に登録してあるときは警告しないことにする。

このように、前後の形態素列や字面を調べて例外記述する。ひとつひとつは気づいてみればなるほどと思うことばかりだが、一度に思い付くことはできないもので、実際に出てきたのを見てひとつず書き加えてゆく。しばらく使っては修正する作業を続けているうちにできた「修正ルール」を資料１に掲げる。

## 7　評価

原因その１～その４はいずれも、一朝一夕になくすというわけにはいかないが、実際に使いながら不都合をひとつひとつつぶしていき、一カ月ほどたつとなんとか落ち着いてきた。原則ルールをそのまま用いると、余計な警告が多過ぎて「いったい機械が何を言おうととしているのかさっぱりわからない」状態だったのがやっと「警告がでたあたりはミスタイプを疑ってかかる」ことができる状態になった。最低限の実用レベルに達したというところだろう。

「実用レベル」に達するために、主に過検出を消すことばかりを考えてきた。警告が多過ぎれば使い物にならないのは明らかなのでしかたがないことではあるが、せっかく原則ルールでは発見できていたミスタイプを逃してしまっていないだろうか。以下はこの疑問に答えるための実験である。

実験：
原則ルールと修正ルールの警告個所を、実際の文書について比較する。

使用したテキスト：
産經新聞社政治面記事原稿（校正前とは限らない）
約130万文字

辞書：
ミスタイプ警告ルールを導入後、実使用の中で１カ月ほどチューニングしたものを共通に用いる。

結果：

表１　警告の個数変化

| | 原則ルール | | 修正ルール | |
|---|---|---|---|---|
| | 過検出 | 正検出 | 過検出 | 正検出 |
| グループ１ | 191 | 2 | 0 | 2 |
| グループ２ | 78 | 8 | 0 | 8 |
| グループ３ | 19 | 4 | 8 | 4 |
| グループ４ | 79 | 20 | 1 | 20 |
| グループ５ | 21 | 2 | 1 | 2 |
| グループ６ | 0 | 2 | 0 | 2 |
| グループ７ | 0 | 0 | 0 | 0 |
| グループ８ | 0 | 0 | 0 | 0 |
| グループ９ | 29 | 0 | 0 | 0 |
| total | 417 | 38 | 10 | 38 |

過検出は417から10まで劇的に減少した。一方、このテストにおいては、ルールを修正したことによって減った正検出はまったくなかった。たまたまこの文章では正検出の減少がなかっただけで、ありえないということではない（資料１の修正ルール参照）。しかし、実際に起こる悪影響がかなり少ないということはわかる。

ミスタイプを発見できた例
（括弧内はルールのグループ番号）

名誉会長就任の報告したり、（２）
辞めるのと辞めないとのでは（２）
五月に予定されいる渡辺外相の再訪露（１）
…と述べるともに、（２）
「…をねらった」とようだ。（３）
価値を生むというもののの解釈が（４）
誤解であるとを強調した。（４）
「…必要がある」して、（６）

ミスタイプのうちどれだけを警告できたのか（検出率）については今回調べなかった。

修正ルールでも残った過検出11のうち8までが同じ原因によるものだった。

額に汗してものを創造していくという、
　→汗して/もの(形式名詞)を
マネーマーケットはものをつくるということではない。
　→マーケットは/もの(形式名詞)を

つまりグループ３の不自然な文節頭（形式名詞「もの」で文節が始まっている）にあてはまったものである。これらは「ものをつくる」「ものを想像する」をひとまとまりの言い方として例外とすることで抑えこめる。このような修正を積み重ねていくことでさらに過検出を減らしていくことができる。

## ８　まとめ

「形態素解析失敗」とはならないミスタイプを発見するため、品詞列・字種・単語長を用いて経験上信頼性が低いと思われるパターンを利用して検出する方法について考察した。

まずミスタイプ例文から抽出した疑わしいパターンを記述した原則ルールをそのまま用いると、正しい文にあてはまってしまう場合が多過ぎて実用にならなかった。そこで、運用の中で実例をもとにして過検出例を収集し、例外をひとつずつぶしてゆく作業を行うと、一カ月ほどで収束し、実用レベルに達した。その間、過検出率は劇的に減少し、正検出率はほとんど変わっていない。

本研究を進めながら実感したのは、実例による「磨き込み」作業の重要性である。この地道な作業なしでは、「磨けば光る」はずのアイディアであっても見向きもされないのである。

なお、今回のような作業
- 品詞、字面などのまざった条件記述
- その実装と実験
- 修正

に、パターン記法はたいへん適しているので、スムーズに進めることができた。

**謝辞**

ＦｌｅＣＳの使い込んでくださった産經新聞社の方々に感謝いたします。

**参考文献**

1) 脇田ほか：日本語校正支援システムＦｌｅＣＳ－新聞用校正ルールの獲得と表現、情報処理学会第45回全国大会3F-4(1992)
2) 奥村ほか：日本語校正支援システムＦｌｅＣＳ－新聞社における実用化報告、情報処理学会第45回全国大会3F-5(1992)
3) 奥村ほか：日本語校正支援システムＦｌｅＣＳの新聞社における実用化、情報処理学会自然言語処理研究会92-NL-91(1992)
4) 奥村ほか：日本語校正支援システムＦｌｅＣＳ－校正規則の衝突について、情報処理学会第46回全国大会3L-4(1993)

**資料１**　修正ルール（下線部分が修正個所）

・グループ１ ： 不自然な文節末

pattern: [@a:断定助動終止形|終助詞&文節末]

pattern: [@a:名詞]
[!記号&!括弧&!名詞&!数詞&!接続詞&!体言文節&!スペース]

pattern: [@a:使役助動詞連用形|受身助動詞連用形&文節末]
[!使役・受身助動詞のある文節]

・グループ２ ： 連体形に用言

pattern: [@a:動詞連体形|助動詞連体形|格助詞「の」|連体詞&文節末]
[副詞]*
[通常動詞(注)|形容詞|形容動詞&!連体文節!体言文節]

pattern: [@a:動詞連体形|助動詞連体形|格助詞「の」|連体詞&文節末]
[副詞]*[補助的動詞(注)]

pattern: [@a:動詞連体形|助動詞連体形|格助詞「の」|連体詞&文節末]
[接続詞]

(注) 補助的動詞とは平仮名書きの「いる・いう・ある・なる」のこと。
通常動詞とはそれ以外。

・グループ３ ： 不自然な文節頭

pattern: [!閉じ括弧&!読点]
[@a:形式名詞|サ変語尾&文節頭]

・グループ４ ： 短い単語列

pattern: [!固有名詞(姓)
&!固有名詞(名・姓)接頭]
[@a:固有名詞(姓)|固有名詞(名)&漢字
&(単語長==1)]
[!固有名詞(名・姓)接尾&!固有名詞(姓)]

pattern: [@a:名詞|接頭語|固有名詞|サ変名詞
&漢字&(単語長==1)
&!"同"&!"各"&!"右"&!"左"&!"前"
&!"後"&!"本"&!"自"&!"他"]
[@a:名詞|接尾語|固有名詞|サ変名詞
&漢字&(単語長==1)
&!"側"&!"状"&!"内"]

pattern: [!読点&!平仮名]
[@a:漢字&(単語長==1)&文節頭&文節末]

pattern: [@a:平仮名&(単語長==1)&文節頭
&文節末]

pattern: [@a:平仮名&(単語長==2)&文節頭
&文節末&!副詞]
[@b:平仮名&(単語長==2)&文節頭
&文節末]

ただし例外パターン[4)]

pattern: [@a:数詞&漢字&(単語長==1)][読点]
[数詞]

pattern: ["「"|"("][@a:平仮名]["」"|")"]

括弧内がすべて平仮名の場合は読み仮名とみなし、形態素解析結果がどうあれ警告しないことにする。

・グループ５ ： 存在しない複合動詞

pattern: [@a:動詞語幹]
[@b:五段動詞連用形語尾]*
[@c:動詞&!接続指定単語]

・グループ６ ： 接続詞「が」「して」・感動詞

pattern: [@a:"が"|"して"&接続詞&!文頭]

pattern: [@a:感動詞&!文頭]

・グループ７ ： 助詞を漢字変換してしまった

pattern: [@a:"名"|"出"|"画"&単語&文節末]

pattern: [@a:"出"&単語&!文節頭]["あ"|"、"]

・グループ８ ： なさそうな接続

pattern: [@a:"が"&格助詞][@b:副助詞]

pattern: [@a:一般接頭語][@b:数詞]

・グループ９ ： 助詞三連鎖

pattern: [@a:助詞][@b:助詞][@c:助詞]

ただし例外パターン

pattern: [@a:"のとは"|"までには"|"などとの"|
"のほどは"|"などに"|"にだけは"|
"などへの"|"とでは"|"かとの"|
"かなどの"|"のかと"|"だけに"]

pattern: [@a:"と"&引用格助詞][係助詞]

助詞が３つ続いてもおかしくない場合を例外として列挙する。