4W－7

日本語校正支援システムＦｌｅＣＳによる新聞記事校正

奥村薫　脇田早紀子　金子宏

日本アイ・ビー・エム㈱　東京基礎研究所

１．はじめに

日本語校正・校閲を計算機で支援する研究は、牛島ら［１］に始まり、最近では実用システムの登場を見るようになった［2,3］。その一つである新聞社向けＦｌｅＣＳは１９９２年１２月より産経新聞社にて稼働している。著者らは柔軟な校正支援、とくに保守・拡張の容易さを目指してＦｌｅＣＳ(Flexible Critiquing System)を開発してきたが、この６月よりユーザ自身の手で辞書と校正パターンを保守作成するにいたった。

ここに一応の完結を見た校正機能と、ユーザによるメンテナンスの現状を報告する。

２．校正機能

校正支援の手法としてＦｌｅＣＳでは辞書および校正パターンによる検査を用いている［４］。いずれも校閲記者の経験と大量の赤字データとをもとに作成したが、実際に運用しながら微調整することが精度を上げるためにきわめて重要であった。

２．１．タイプミスの検出

タイプミスや修正ミスは、その個所またはその後で形態素解析ができなくなる可能性が高い。未知語の検出はそれらの誤りの発見に有効である。ただし辞書類が十分に充実していることが必要条件となる。

入力ミスでありながら、本来とは異なる単語の列として何らかの形態素解析ができる場合も多い。これらに対して、発現率の低い品詞列や、信頼性の低い解析結果のパターンを警告する手法が用いられてきた。

一文字語の連鎖は何らかの誤りのことが多いが、過剰な警告のもとにもなる。著者らはその前後の品詞・文字列などに条件を付けることにより、過警告の多くを押さえ込んだ［５］。

辞書サイズ（語数）		校正パターン	
基本辞書	79,760	同音語系	220
新聞辞書	19,337	タイプミス系	40
表外辞書	7,004	誤り語系	15
統一辞書	12,105	その他	124
誤用語辞書	13,639	（記事スタイル、文体など）	
要注意辞書	315	計	399 パターン

表１：校正辞書・校正パターンの内訳

２．２．誤り語の検出

いわゆる用字用語は、辞書レベルで対処できるものが多い。常用漢字表外の字や読み、送り仮名の付け方や仮名書き語などがこれにあたる。また語の単位を広くして、同音語や校閲の分野にも利用できる。単純な手法なのだが、実用上よく誤りを発見し、ユーザに重宝がられた機能である。

例）原点に帰　→ 原点に返(ラ行五段語幹)
　　世田谷区梅ケ丘 → 世田谷区梅丘
　　梅丘駅　→ 梅ケ丘駅

ただし辞書による誤り検出にも、思わぬ副作用があり得る。辞書作成時点では以下の項目があった。

例）曙　→ あけぼの　（表外字）
　　ソ連 → 旧ソ連、ロシア

ところが記事では『曙』はほとんど力士名であり、辞書から省くことになった。『ソ連』はそのまま使用してよい場合が多いことが分かり、より弱いレベルの警告（赤→黄色）として、「歴史的使用は可」とのコメントを付けた。

２．３．同音語誤りの検出

同音語や類語の使い分けに付いては、誤りやすいものを中心に、前後の関係を見て判定する校正パターンを記述している［４］。これは同音注意語にかならずチェックを付けるといった辞書的手法に比べると、誤警告が少なく的確な指摘ができる。通常は誤りやすい語が偏っているためにこの手法は十分に有効である。しかし、すべての同音語誤りを検出するにはいたらないので、将来他の手法との併用も考慮する余地がある。

Proofreading News Items by Japanese Critiquing System "FleCS".
Kaoru Okumura, Sakiko Wakita, Hirishi Kaneko
Tokyo Research Laboratory, IBM Japan

２．４．数字の表記法

新聞記事では、洋数字と漢数字を使い分ける必要がある。漢数字では、単位語（十、万など）を付ける場合と付けない場合、洋数字では連数字（一マスに二文字いれる）とそうでない場合がある。

例）昭和六十三年　　漢数字単位語付き
　　一九九三年　　漢数字単位語なし
　　のぞみ５０１号　洋数字
　　14勝11敗　　洋数字　連数字

これら４種の使い分けを校正パターンで検査するようにした。ただし規則に反していてもスポーツ面や広告なら許すなどあって、過不足ない検出は難しかった。

２．５．カタカナ語・括弧・地名の不整合など

上記のほか、アルゴリズムによるカタカナ語の揺れ、括弧の不整合の検査も行う。また県名と市町村名のデータを保持して、実在しない地名の検出を行う。その他必要に応じて作られた表記法の規則が各種ある。

３．ユーザによるメンテナンス

言語が生き物である以上、校正支援システムも時とともに変化していかなければならない。そのためには校正支援システムの使い手であり、日本語の専門家である新聞社の人々が、システムを育てていくのが理想であろう。

対象：辞書の更新と校正規則(パターン表記法による)を記述し、運用できるようになることを目標とした。今回メンテナンスを担当したのは、校閲センターと製作局から各１名である。ワープロは打てるが、パソコンの操作やプログラムの経験はなかった。

内容：日本語の品詞分類、形態素解析結果の読み方、ユーザ辞書・校正辞書の更新、辞書登録のコツ、校正パターンの過警告の原因解明と修正（多くは辞書の追加）、未警告の誤りから校正パターンを作成すること。

メンテナンス技術の講習会を４月から６月にかけて計１３回（１回２時間程度）行って、ユーザは上記のテクニックをほぼ習得した。彼らの書いたパターンを図１に挙げる。

比較的短期間でコンピュータの初心者がこれらを習得したことにより、校正パターン表記法は直観的で十分使いやすいものであるといえよう。

```
#pattern "勘がいい" PTN_C
pattern: [@a:"感"&isPhrTop()]["が"]
         ["よ"|"い"&WrdPos(16)]
message: "変換ミス?"
warning: [@a]->"勘";
```

図１a：ユーザが記述したパターン(1)

```
#pattern "～選を戦う" PTN_A
pattern: ["参院"|"参議院"|"都議"|"府議"|"県議"
  |"市議"|"町議"|"村議"|"区議"|"知事"|"市長"
  |"町長"|"村長"]["選"][isJoshi()]
  [@a:"闘"&WrdPos(10)]
message: "変換ミス？"
warning: [@a]->"戦";
```

図１b：ユーザが記述したパターン(2)

```
#pattern "～選を戦う" PTN_A
pattern: ["院"|"議"|"知事"|"長"]["選"]
         [isJoshi()][@a:"闘"&WrdPos(10)]
message: "変換ミス？"
warning: [@a]->"戦";
```

図１c：パターン(2)添削後

４．おわりに

当システムは新聞記事校正において、意味的誤りと文体の推敲を除いた誤り全体の約９割を検出している。一方、過剰な検出のほとんどは未知語によるものであり、分野に大きく左右されるが１記事あたり数語である。

ＦｌｅＣＳは330種類のパターンを用いて、誤字・脱字に加えて地名チェックや数字表記法など、さらに踏み込んだ検査を行う。現在はユーザ自身がメンテナンスし、成長させていく校正支援システムになっている。

謝辞：産経新聞社製作局および校閲センターの方々、ことにＦｌｅＣＳのメンテナンスを担当する伊能和美さん、時田昌さんに深謝いたします。

参考文献：

[1] 牛島：日本語文書推敲ツール『推敲』, Bit, Vol.23, No.1 (1991)
[2] 高橋ら：計算機マニュアル推敲・査読支援システムMAPLEの開発と運用，情処論文誌, Vol.331, No.7(1990)
[3] 福島ら：日本語校正支援システム St.WORDS, 情処第45回全国大会2C-1 (1992)
[4] 奥村ら：日本語校正支援システムＦｌｅＣＳの新聞社における実用化, 情処研報 92-NL-91 (1992)
[5] 脇田ら：日本語校正支援システムＦｌｅＣＳ-ミスタイプ検出について, 情処研報 93-NL-97 (1993)