3F−5

日本語校正支援システム *FleCS* －新聞社における実用化報告－

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

新聞製作工程では、組版システムをはじめとして記者ワープロ・自動集配信システムなどにより機械化が推し進められてきた。その中で校正や校閲は専門性の高い分野とされており、ほとんど専門家の手で行われてきた。本稿では校正支援システムＦｌｅＣＳ[1,2]を基に、新聞社で実用に供される入力/校正支援機を構築した報告を行う。

2．校正支援の手法

校正支援の技法には次のようなものが挙げられる。

(1) 文字列のマッチング：予め誤りやすい文字列を登録しておくもの。

(2) 文法解析

(2a)非文の検出：文法解析して日本語として解釈できない部分に誤りがあるものとする手法。形態素解析レベルでの解析不能個所を、未知語という。入力ミスなどを発見する一般的な手法だが、辞書が不備な場合にはこれによる過剰検出がかなりあり得る。構文解析レベルでは、日本語の文生成規則が緩いせいもあってあまり有力ではない。

(2b)誤りを含んだ文法解析：よく起こる誤りはそれらを許容するように文法を拡張し、その文法を用いた部分は誤りであるとする。単語レベルで警告対象として登録されている語を禁止語という。(2a)では別の解釈をして見逃す誤りや、(1)では過剰検出となり得るものに、より的確な検出を成し得る。

(2c)修飾構造：構文解析を行って修飾関係から、曖昧さ・分かりにくさを警告するもの。[3]

(3) 共起：ＡＩかな漢のように共起関係を記述した辞書を持つ。同音語の選択誤りをした場合には、正しい語の方が、周囲の語と共起関係を持つだろうという推論に基づく[4]。

(4) ヒューリスティクス：校正の‘知識’を獲得して、問題ごとに対処する。知識表現の枠組みと、それを実行するメカニズムを持つ。[5,2]ではルールベース・システムをこれに利用している。

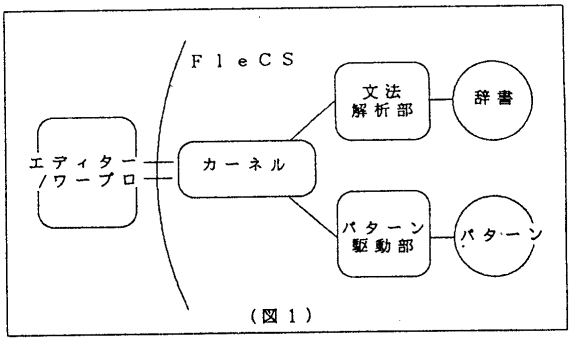
3．ＦｌｅＣＳのシステム構成

実際に使用するユーザは、誤りを全部検出する以上に確実な誤りのみを指摘することを要求するので、今回は比較的正解率が高い手法を中心に用いた。すなわち、(1)、(2a)(2b)の形態素解析レベル、及び(4)である。特に(4)を実現する手段として、校正知識用に特化した‘パターン表記法’を開発したのが本システムの特長である。

ＦｌｅＣＳのシステム構成を図１に挙げる。ユーザがエディターまたはワープロの中で校正支援を要求すると、解析が別プロセスで始められる。カーネルは、文書の種類に応じて空白・改行記号などの位置情報を解釈し、標準フォームに変換する。次に文法解析部に於て各種辞書を参照しながら形態素解析を行う。その結果をパターン駆動部に入力して、誤りパターンとの照合を行う。誤りが発見された時には、カーネルがリクエスターに応じた形でそれを伝える。特殊なハードウェア等は不要である。

これらのシステムは、パーソナル・システム/55＊のOS/2＊のアプリケーションとして稼働している。処理速度は約150文字／秒である。

＊ パーソナル・システム/55、OS/2はＩＢＭの登録商標です。



（図１）

4．校正パターン表記法

校正の知識としては、(1) 誤りの特徴、(2) 警告する部分、(3) 修正のしかた、(4) 警告メッセージを記述する必要がある。(1) として誤りの特徴を、文章の部分に関する評価関数（これをユニットと呼ぶ）の並びで表す。例１のパターン部の[…]がユニットであり、省略可能や有限回の繰り返しも指定できる。また各ユニットに変数を付け、後で参照する事ができる。特徴を記述するには、文字列、品詞等の文法情報、また任意の（Ｃ言語の）関数と、それらの論理結合を用いる。

(2) の警告個所はユニットの変数の並びで指定する。(3) は、書き換えられる個所を変数の並びで、書き換え方を文字列あるいは変数に関する関数で記述する。

例１
```
校正知識: 「例えば～など」は重言。

パターン:[x:" 例えば"|" たとえば"][y:ANY]*
         [z:"等"|" など"]
警告個所:[x],[z]
書換候補:[x]->"";[z]->""
メッセージ:[x]["～"][z][" は重言です"]
```

例２
```
校正知識: 「必ずしも」の後には、否定形が来る。

パターン:[x:" 必ずしも"][!否定()]*[ 文末()]
警告個所:[x]
メッセージ:"呼応: 「必ずしも」は否定で受ける"
```

このパターン記法は正則表現風ではあるが、帰納的集合を識別しうる。何故なら自分より前のユニット全てを変数で参照して関数を書く事が可能だからである。自分より後のユニットはまだ確定していないので参照出来ないものとする。

ルールベース・システムと比べてこのパターン記法は、書きやすくスピードが速い。さらに多くの最適化手法を適応出来る。我々はルールベース・システムから移行して従来の 4.4倍の速度を得た。

5．新聞用カスタマイズ

新聞社用で実際に使用するシステムとするために、以下のカスタマイズが重要であった。

(1) 辞書：辞書が対象とする文書に適合していなければ、正しい言葉が未知語となってしまう。実際、過剰検出のほとんどを未知語が占めるので、これを減らすための辞書修正は必須である。今回は過去の記事を解析して、一定以上の頻度で現れて単語と認めうるものを予め辞書登録することにより、精度向上を図った。また校正中にも、単語の登録や未知語からの単語候補収集を行える。

Japanese Critiquing System FleCS
–Real Use in Newspaper Companies,
Kaoru Okumura, Sakiko Wakita, Hiroshi Kaneko
Tokyo Research Laboratory, IBM Research

(2) 校正対象：一般には誤りではないが、新聞に於ては用いられない記述が数多くある。常用漢字・常用読み以外の使用は固有名詞等を除いて避けられる。また新聞用のスタイルがあり、これらは間違った日本語では無いので気付きにくい。
我々は『記者ハンドブック』[6] と、産経新聞社辞書委員会の作成した誤り/ 使い分けのリスト（約54,000項目）をもとに作業を進めつつある。単語単位で判定出来るものを「誤用語」に登録し、周辺との関係を見て使い方の規則があるものに対しては「校正パターン」を作成するのである。[7]

(3) エディター：ＦｌｅＣＳの外部の問題ではあるが、その現場で用いられているエディター上で校正支援を使えることは、効率上大きな意味がある。本システムは産経新聞社で既に用いられている縦書きエディターから利用できる。通常、エディターはユーザ・インターフェイス関連を追加するだけでＦｌｅＣＳを利用できる。即ち、ＦｌｅＣＳをコールするし、警告があれば誤り個所を表示し、書き換え候補を示していずれかが選択されれば書き換えを実行する等といった機能である。

6. 校閲者による赤字訂正の分析

校正支援システムの性能は、校正の対象文書や誤りの現れ方と独立に評価することは難しく、実感に近い尺度は得られないであろう。ゆえに、まず校閲者が実際に行った赤字訂正を調査・分析する必要がある。
１６４件の赤字を調査した結果を図２に示す。日本語としては正しいが記事には不適切という理由で訂正されたものがかなりを占めているのが分かる。これにより分野毎のカスタマイズとカスタマイズが迅速に出来る枠組みが、実用的精度の校正支援システムを作る上で重要であることが示された。

赤字１６４件中

・用語(1)　45% 一般にはいずれも正しく、新聞ではその一方のみを使うもの。
- 表外字（常用漢字表に含まれない字）
　流暢→流ちょう
- 表外音訓（常用漢字表に含まれない読み）
　全ての→すべての
- 送り仮名（国語審議会答申の本則を用いる）
　悔む→悔やむ

・用語(2)　20%　一般には（単語としては）いずれも正しいが、新聞では使い分けを要するもの。
- 漢字／仮名の使い分け
　欲しい（本動詞）／ほしい（補助動詞）
　丸の内（町名）／丸ノ内（地下鉄駅）
- 類義語
　極める（極限）／究める（探求）
　ボール（球）／ボウル（台所用品）

・同音語　9%　いわゆる変換ミス。
　実践のカン→実戦のカン
　友達同志→友達同士

・入力ミス　13%
　九千万六百万円→九千六百万
　ローマ→ローマ

・字句の修正　10%
日本語の流れとして不自然
　湯川光久八段、結城七段
　→湯川光久八段、結城聡七段
記事のスタイルに違反
　（本社北九州市・＊＊社長）
　→（本社・北九州市、＊＊社長）

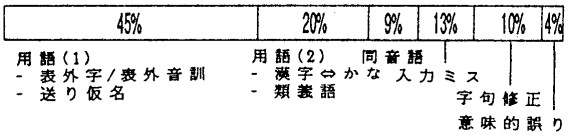
・意味的誤り　4%
　株価は四〇〇円前後で→四五〇円前後で

（図２）

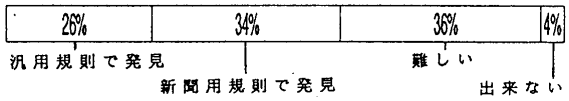
7. ＦｌｅＣＳの評価

前述の赤字１６４件に対してＦｌｅＣＳの発見率を次に記す。

赤字調査　計１６４件中



ＦｌｅＣＳでは、



○汎用の校正規則：
（表外字や送り仮名の辞書無し）
発見：26%(42件)
　禁止語:16 件、表記の統一:17 件、
　入力ミス:9件
未発見：74%

○新聞用カスタマイズ後：（1992年10月時点予測）
発見：60%(94件)
　上記＋表外字/ 送り仮名等の禁止語30件、
　その他辞書:4件、パターン:22 件
発見せず：40%
　共起がある:4件、不自然な流れ:8件、
　記事スタイル:3件、入力ミス:8件、
　使い分け（意味/ 読み方):37件、
　事実誤認:6件

共起の技法を用いれば、検出率がさらに上がると予想されるが、誤りを検出できるほど多くの共起ペアを備えた辞書を用いた場合、正しい単語に対しても「その同音語/ 類義語と共起がある」として過剰な警告をする恐れが大きい。また共起メカニズムで発見できるであろう事例は類義語・同音語のうち 4件(2.4%)と僅かであるので、本システムには組み込まなかった。

8. おわりに

ＦｌｅＣＳは新聞用校正支援として、実際の赤字の過半数を検出する能力があり、校正作業の効率化に役立つとの評価を受けた。本システムは10月中旬より、産経新聞社製作局データー入力部にて入力・校正端末として使用される。今後は校正知識の一層の充実により、さらに正確で使いやすい校正支援システムを目指す予定である。

謝辞
本研究にあたり、校閲事例や新産経辞書、記事データを快く使用させていただき、貴重なコメントをいただいている産経新聞社校閲センター及び製作局の方々に深謝致します。

参考文献
[1] 特集　次世代入力機器の開発状況，新聞技術，1990-2 No.132（1990）
[2] 奥村ほか：日本語校正支援システムFleCS，情報処理学会自然言語処理研究会87-11，（1992）
[3] 箱守ほか：日本語の修飾構造を評価する添削支援システムを実現するための基礎研究，情報処理学会論文誌Vol.33 No.2（1992）
[4] 野崎ほか：かな漢字変換と漢字かな変換を共に用いる同音語誤りの検出方式，情報処理学会第45回全国大会4C-2（1992）
[5] 林ほか：日本文推敲支援システムにおける書換え支援機能の実現方式，情報処理学会論文誌，
Vol.32，No.8（1991）
[6] 記者ハンドブック−用事用語の正しい知識
[7] 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）