自然言語処理 91－5
（1992. 9. 18）

日本語校正支援システムＦｌｅＣＳの新聞社における実用化

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

概要：本稿では、日本語校正支援システムＦｌｅＣＳを基に、新聞社に於いて実使用されるシステムを構築した報告を行う。我々は校正知識の表現とその検索に『校正パターン表記法』を開発しており、これによって校正知識を容易に記述し、高速に実行する事ができた。実際の校正パターンを書き進めて行く手順とその効率を紹介する。また、校閲者が赤字訂正を行った事例の分析と、それを基にした本システムの評価についても述べる。

Real Use of the Japanese Critiquing System FleCS in Newspaper Companies

Kaoru Okumura, Sakiko Wakita, Hiroshi Kaneko
Tokyo Research Laboratory, IBM Research

Abstract: This article illustrates our experience on constructing a practical Japanese critiquing system for newspaper company. We have developed a 'pattern description,' which makes easy to write and execute newpaper specific critiquing knowledges. Actual procedure to implememt critiquing knowledges, and its workload, accuracy are described. We also analize the proofreader's check cases, and evaluate FleCS system based on the data.

1．概論

1.1　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校正支援機を構築した報告を行う。

1．2　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で、最も容易に実現できる。しかし登録されている誤りにのみ有効である。また[3]では字面だけから「受け身」や「否定」をかなりの精度で検出している。

(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：よく起こる誤りに対して、それらも許容するように仮想的に文法を拡張し、後で拡張した文法を使用した個所をチェックする。単語レベルで警告対象とな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。[4]

(3) 共起：Ａｌ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。同音異義語の誤りに対する現在唯一のシステマティックな手法である。[5]

(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。[6,2]ではルールベース・システムを利用している。

1．3　本システム構築の指針

現在のところ日本語校正支援システムは確実な自動校正の段階まで到っていないし、また自然言語処理の性格上100%の精度を実現するのは不可能であろう。しかし完ぺきで無くても、使い方によっては現在のレベルで十分強力なツールになり得る。今回は実用に足る校正支援システムを構築することを第一の目的とした。

まず新聞用のシステムに特化する事によって検出率の向上及び過剰検出の低下を図った。実行速度やメモリ使用量等の面でも、通常のパソコンで十分快適に使用できる事を条件とした。また実際に使用するユーザの要求により、誤りを全部検出しようとして過剰検出が出るよりは、確実な誤りのみを指摘する方針を取った。

2．　ＦｌｅＣＳシステム

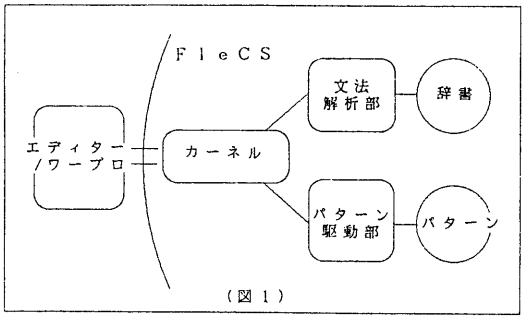
2．1全体構成

今回は1．3節の指針により比較的正解率が高い手法を中心に用いた。すなわち、1．2節の(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。

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から、校正支援を要求すると解析が別プロセスで始められる。カーネルに於ては、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、標準的な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。

---

*パーソナル・システム/55、ＯＳ/2はＩＢＭの登録商標です。



（図１）

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。特殊なハードウェア等は不要である。

2．2　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) 警告をあたえる部分、(3) 修正のしかた、(4) 警告メッセージの４つ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。ユニットには省略可能や有限回の繰り返しを指定できる。また各ユニットに変数を付け、他から参照する事ができる。特徴を記述するには、その文字列、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
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]["～"][y][" は重言です"]

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。実現上の制約として、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。

2．3　カスタマイズ手法

Ｆ l ｅＣＳは様々な目的・環境にカスタマイズしやすいように設計されている。新聞社で実用に足るシステムとするために、以下のカスタマイズが重要となった。

(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。極端な話、囲碁・将棋欄やスポーツ欄は特殊な語や語法で書かれており、これらの単語が一般の辞書にあることは期待できないであろう。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。

(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。それらの表現に対して、「誤用語辞書」または「校正パターン」を作成して対処する。詳しくは３章、４章で述べる。これらは日本語として間違っている訳では無いので、気付

きにくいことも多い。

(3) 入出力（エディター）：ＦｌｃＣＳの範囲外ではあるが、入出力環境も重要である。校正支援システムはその場所で通常用いられているエディター上で使えるのが望ましい。その為エディターはユーザ・インターフェイス関連を追加する必要がある。ユーザの要求にしたがってＦｌｃＣＳをコールする機能、ＦｌｃＣＳの出した警告によって誤り個所を表示し、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。この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ようにした。

3. 新聞記事の校正とは

3．1 記事の表記の基準

新聞記事は、常識的日本語のほかに新聞記事の書き方、さらには各新聞社の決まり事などに従って書かれている。『記者ハンドブック』[7] から、それらの幾つかを列挙する。

●常用漢字表に含まれない字種（表外字）、音訓（表外音訓）を含む語は原則として使わない。ただし固有名詞は例外。
●送り仮名は国語審議会答申の「本則」を用いる。
●漢字/ 仮名の使い分け：代名詞、連体詞、接続詞、感動詞、助詞、助動詞、補助用言、形式名詞、接辞は原則的に平仮名。副詞は語により書き分ける。しかし、例外や意味によるものも多く、単語ごとに決まるのに近い。
●類義語の使い分け：要注意の語については、その使い分けの基準が[7] 中の『用字用語集』の各単語の欄に記されている。一般にも使い分けすべきものも含まれる。

例
基軸[基準・中心] 基軸通貨
機軸[期間・車輪の軸、活動の中心・方式]
新機軸を打ち出す

以上は、＜用字用語＞と呼ばれている。その他、記事のスタイルに関する規則がある。
●日時の書き方：基本的には12時間制で、「零時」は「〇時」とは書かない等。
●数字の書き方：基本的には漢数字で、原則として単位語（十、百、千、万など）を付ける。次の場合は単位語を省略する。(1) 所番地 (2)西暦年 (3)船のトン数…（後略）

3．2 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に大きく左右されるものである。よって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だろう。１６４件の赤字を調査した結果を、訂正理由とその割合をともに図２に記す。

用語１は、一般にはいずれも正しいが、新聞ではその一方を使うという決まりに反したものである。用語２は、「使い分け」と言われるものであり、平仮名／片仮名／送り仮名／類義語／同音語などを使い分ける基準が[7] に定められている。この中には新聞以外でも守るべきものと、さほど厳しくないものがないものがある。使い分け以外の同音語誤りは９％あった。

また、新聞記事で無いとしたら必ずしも間違いとは言えないものが全体の６～７割を占めている。故に分野ごと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確認された。

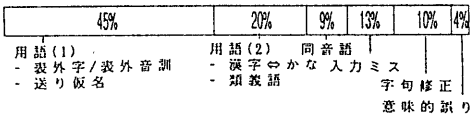
4. 新聞用校正知識の実装

産経新聞社辞書委員会が作成した「用語リスト」と、同社製作局より提供された過去の記事データを用いて、校正知識を作成しつつある。「用語リスト」には、読み・誤用語・品詞・正しい語などが記されている。

読み	誤用語	品詞	変換候補	備考
きべん	詭弁	名詞	詭弁	表外字

（図３：用語リストの形式）

赤字１６４件中



・用語１（統一）　45%
　- 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（使い分け）　20%
　- 漢字/ 仮名の使い分け 14%
　　欲しい（本動詞）／ほしい（補助動詞）
　　丸の内（町名）／丸ノ内（地下鉄路線）
　- 類義語・同音語　6%
　　極める（極限）／究める（探求）
　　ボール（球）／ボウル（台所用品）
・同音語（用語以外の）　9%
　　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%　読点句点の修正等。
　日本語の流れとして不自然 7%
　　湯川光久八段、結城七段
　　→湯川光久八段、結城聡七段
　記事のスタイルに反する 3%
　　（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）

このリストは、校正支援及び仮名漢字変換要の辞書作成の双方に使う意図で作成されたものである。リストの多くはよくある書き誤りなどで、語単位で「誤用語辞書」に登録すればよい。その外に同音語・類義語の使い分けなどがあり、書かれた例から規則性を抽出して『パターン記法』で表現する。

4．1　パターン作成

「嫌い：きらい」の書き分けを例にとって校正パターン作成手法を説明する。

(1) 用語リストから対象語句を定める。

誤用語	品詞	変換候補	備考
独断の嫌い	名詞	独断のきらい	◆きらい
楽観する嫌い	名詞	楽観するきらい	◆きらい

（図４：用語リスト中の使い分け）

このリストを見れば人間はどう使い分けるか大体理解できるであろう。しかし、これを「誤用語辞書」に登録しても、そのままの形で記事中に出てくるとは限らない。

(2) 記事データを調べる。

そこで過去の記事データから問題の単語を抽出し、その前後関係を調べる。（この作業もＦｌｅＣＳのを用いて容易に行える）

…策の一貫性をかく｜きらい｜がある。
…に信用してしまう｜きらい｜がある。
…五歩は巧遅拙速の｜きらい｜があるが、この局…
　好き｜嫌い｜の激しい子がいても…
…の雰囲気は好きか｜嫌い｜か、などの主観を求…
　酒は｜嫌い｜でない主人なのに、…

（図５：用例抽出）

実際の使い方を見ると、「～のキライが」「～するキライも」などは平仮名で書くとして大概よいことが分かる。

(3) 校正パターンを書く。

```
pattern:["の（格助詞）"|(動詞連体形)]
        [@a:" 嫌い"]["が"|" も"|" は" &(助詞)]
replace: @a -> "きらい"
※実際には（…）は品詞番号で示されている。
```

（図6：校正パターン）

これは、［格助詞の『の』または動詞の連体形］と［嫌い］と［助詞の『が』『も』『は』のいずれか］が連続していたら「嫌い」を仮名書きになおす事を意味している。

(4) 他の記事データで正当性を検証する。

4．2　校正パターンの精度

次に校正パターンの正確さを幾つかの例で評価しよう。いずれも、1992年2月の記事(6Mbyte)を参照して校正パターンを作成し、4月分の記事(9Mbyte)を解析してその正確さを試した。結果を図7に示す。図8が使用した校正パターンである。

誤りの数自体が少ないために精度を数値では言えないが、この中で見る限り発見率は十分に高く、誤警告は全く無かった。検出漏れの2件は形態素解析の誤りによるものであった。ただしこれらの校正パターンは、一般的感覚では、厳密性を欠くもので、パターンに反する例はいくらでも創作できるだろう。しかし、新聞記事に限ればこの程度の基準でかなりうまくいくようである。

	2月分		4月分			
	総数	誤り	総数	誤り	正警告	誤警告
極める	59	4	65	2	2	0
究める	0	0	3	0		0
強硬	71	2	210	2	2	0
強行	10	1	26	3	3	0
嫌い	31	1	46	1	1	0
きらい	7	0	8	0		0
く（る）	584	0	21	0		0
来（る）	213	67	371	129	127	0

（図7：使い分けルールの評価実験）

```
pattern:[(動詞の連用形)]["て"| "で"&( 接続助詞)]
        [@a: "来"& (カ変)]
replace: @a -> "く"

pattern:[@a: "強行"| "強攻"]
    ["意見"| "姿勢"| "手段"| "派"| "路線"|
     "論"| "に（断定助動詞，助詞）"|
     "な（断定助動詞）"|
     "で（断定助動詞，助詞）"]
replace: @a -> "強硬"

pattern:[@a: "強硬"]
      ["採決"| "突破"| (サ変活用語尾)| (読点)]
replace: @a -> "強行"

pattern:["真相"| "真実"| "本質"| "道"| "教"|
         "性"| (学問その他)]
        ["を（格助詞）"] [! (述語区切り)]*
        [@a: "極め"& (一段動詞)& (１単語)]
replace: @a -> "究め"

注: 「述語区切り」は語の遠さを判定する。[8]
```

（図8：使い分け校正パターン）

4．3　作業量

一つのパターンを記述するのに、正味の作業時間はおよそ20分かかった。テストを含めて一日に10個程度の校正パターンを作るのは難しくないと思われる。表記リスト全体で約53,000項目の内、「き」で始まるものが1,744 事例ある。そのうち誤用辞書では解決できない［くる/ 来る］の類が30種類ほどあった。ゆえに約1,000 種類がパターン表記の対象となるが、校正パターンでは十分に検出できないものを除くとすると、2～3ヶ月後にはかなりの使い分け校正パターンが出来るようになると思われる。（2人で作業中）。

以上の経験より、まだ作業中ではあるが、

- 対象を新聞記事に限れば
- 指示された同音異義語のうち2/3 について

・誤りの８割りを発見し、
・過剰警告が１割以下になる校正パターンを
・数ヶ月で作成できる

と予測する。これは分野を限定してかなり大胆なヒューリスティクスを用いたためと、校正パターンが書きやすく処理やテストが高速に行えるためと考えられる。

４．４　ＦｌｅＣＳの評価

以上のような作業を進めた後に、ＦｌｅＣＳがどのくらい実際校閲者が直すべき誤りを発見するだろうか。前述の赤字１６４件に対して、新聞用カスタマイズ後には、60%(94件) を発見できると思われる。その内訳は：

●発見　60%(94件)

原因	発見方法	
- 入力ミス	未知語	9件
- 表記１	誤用語辞書	63件
- 表記２	パターン	22件

●発見せず　40%(70件)

- 意味／読みによる使い分け	37件
- 不自然な流れ	8件
- 入力ミス	8件
- 同音語、語用法で共起がある	4件
- 記事スタイル	3件
- 意味的誤り	6件

（図９：発見率）

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ある。また共起メカニズムで発見できるだろう事例は類義語・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。

５　さいごに

本システムは1992年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、より正確で使いやすい校正支援システムを目指していきたい。

謝辞

本研究にあたり、校閲事例や表記リスト、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、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，情処自然言語処理研究会87-11，(1992)

[3] 牛島：日本語文書推敲支援ツール『推敲』，Ｂｉｔ，Vol.23，No.1 (1991)

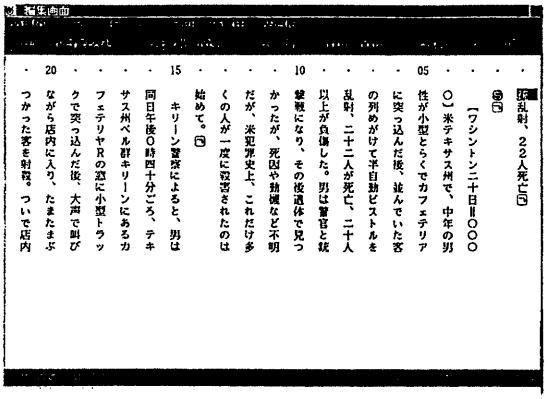
[4] 箱守ほか：日本語の修飾構造を評価する添削支援システムを実現するための基礎研究，情報処理学会論文誌Vol.33 No.2 (1992)

[5] 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)

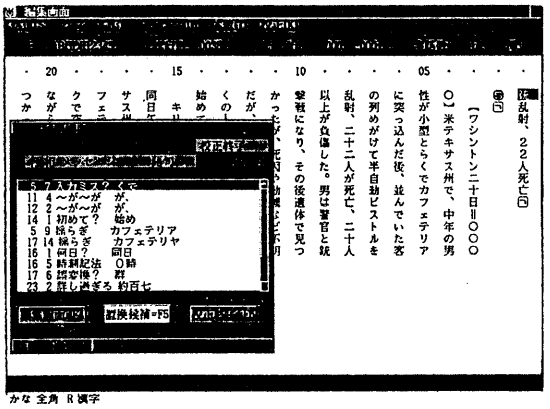
[6] 林ほか：日本文推敲支援システムにおける書換え支援機能の実現方式，情報処理学会論文誌，Vol.32，No.8 (1991)

[7] 記者ハンドブック－用事用語の正しい知識

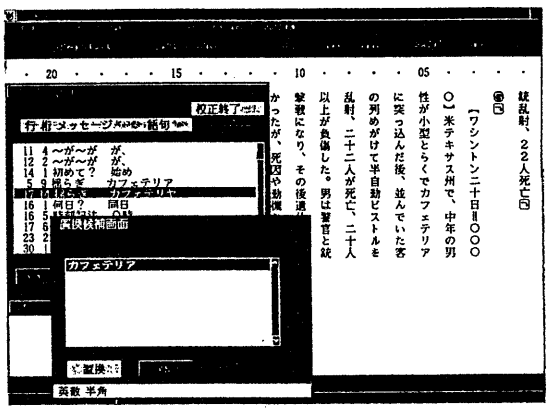
[8] 脇田ほか：文中における語句の『近さ』について，情報処理学会自然言語処理研究会97-NL-90，(1992)



（ａ．新聞用縦書きエディターの画面）



（ｂ．警告メッセージを表示する）



（ｃ．置換候補を表示・選択する）

図１０：入力・校正端末での校正支援システム使用画面