3L−3

校正支援システム St.WORDS の文書検査機能

福島 俊一・佐々木 伸太郎・山田 洋志 (NEC)
小澤 基伸・大槻 道夫・小野寺 裕 (講談社)

1 はじめに

日本語文書を対象とした校正支援システムの研究は、牛島ら[1]に始まり、特に1985年頃からは日本語ワープロの普及を背景にコンピュータメーカ各社が次々に試作システム[2][3][4][5]を発表して活性化してきた。これまでに、校正支援/文書検査のさまざまな手法が考案されている。

日本語文書を扱う手法は概して、辞書や経験則に大きく依存したものになる。そのため、辞書の構築や経験則に関する副作用の調整などの労力が、アイデアと実用との間に大きなギャップを生んでいる。そこで、実際の運用を通して改良を加え、有効性を検証してゆくアプローチが必要になる。

そのようなアプローチをとっている一事例として、校正支援システムの新聞社での運用報告がある[5][6]。一方、筆者らは、St.WORDS を開発し[7]、1992年6月末より出版社における試験的運用を開始した。St.WORDS の文書検査手法は、基本的には、筆者らが過去に COMET [4][8] において実現したものであるが、実用化のために、辞書を大規模化し、校正規則の条件判定などに細かい改良を加えている。本稿では、St.WORDS の文書検査機能とその改良について報告する。

2 校正支援システム St.WORDS

St.WORDS は、一般に普及しているワープロやパソコンの環境からの移行の容易性と今後の拡張性を考慮して、解析サーバ (EWS4800) に校正者用端末 (PC-9801、複数台可) を LAN 接続する構成をとっている[7]。校正者用端末はワープロ機能をもち、作業者は文書検査機能による画面上の色マークを確認しながら対話的な校正操作が行える。

St.WORDS で実行する文書検査は、表1に示すような表記法に関する11項目である。これらは、筆者らが過去に試作したプロトタイプ COMET [4][8] を出版社での校閲作業を想定して試用することで、有効性(欲しい機能であっても現状の技術で十分な性能が得られるか)と必要性を以下のように判断して決定した。現状の技術レベルから考え、文書の内容に関する検査は人間が行うものとした場合、それと同時に表記法に関して気を配ることは、作業者にとって心的負担が大きい[7][9]。日本語では本来、正書法が確立されておらず、同じ語に対して複数通りの表記の仕方(ゆれ)が許容されている。このような背景のもとで細かく定められた一定の規準[10]に適合しているかを判断するのには、かなりの専門的な知識と経験とを要する。知識と経験があっても、文書の異なる箇所を比較して表記のゆれを検出するような作業は労力が大きい。

3 形態素解析にもとづく表記法検査

表1に示した検査項目の1～7は、解析サーバの実行する形態素解析の段階で判断する。

表 1: 文書検査項目

検査項目		例
1	未知語	これあで　指道力
2	誤用語	完璧→完璧
3	送りがな	集り→集まり
4	書き換え推奨語	奇蹟→奇跡
5	かな書き推奨語	所謂→いわゆる
6	同音類語	特長/特徴
7	同音注意語	週間/週刊/習慣
8	カタカナ表記	インタフェース/インターフェース/インターフェイス
9	数字表記	50万(漢数字を規準としたとき)
10	括弧	週刊「モーニング」
11	漢字レベル	絹(3年生レベルを規準としたとき)

3.1 未知語チェック

未知語チェックは、形態素解析に失敗した箇所を誤字・脱字の可能性があるとして検出するものである[4][5]。したがって、単語辞書の語数が少ないと、誤字・脱字ではない単なる未知語が数多く検出されることになり、文書検査を目的とすると煩わしいものになる。出版社では一般書籍や週刊誌など多種多様な文書を検査の対象とするため、その傾向は顕著になる。そこで、St.WORDS では単語辞書を大規模化して50万語とした(表2参照)。週刊誌1冊分(約11.5万字)を形態素解析した際の文節分割誤りを分析したところ、単語辞書の語数を約8万語から約50万語に増加させることにより、未知語による文節分割誤り件数は約1/7 (1818件→254件)に減少し、効果を確認できた。さらに、週刊誌などには口語的な表現(例:混んでる、冗談じゃない、困っちゃう)も頻出するので、未知語として過剰検出しないように対処した[11]。

また、本来は未知語や誤字・脱字であっても、形態素解析が失敗せずに、なんらかの解釈を出してしまうことがある(例:「指道力」を「指」「道」「力」と解釈)。そこで、1文字名詞の連続など、解析結果において信頼性の低いパターン(現在は約30通りを用意)も検出するようにした。

週刊誌の文書(約5.2万字)で未知語チェックの正確性を評価したところ、266箇所が検出され、そのうち未知語や誤字・脱字でないのに検出されてしまった箇所は9件で、検出できなかった未知語は13件であった。したがって、適合率と再現率は次のようになる。

- 適合率 $= \frac{\text{検出件数} - \text{過剰検出件数}}{\text{検出件数}} = 96.6\%$
- 再現率 $= \frac{\text{真の件数} - \text{検出洩れ件数}}{\text{真の件数}} = 95.2\%$

Japanese Text Proofreading Methods Implemented on St.WORDS
Toshikazu Fukushima, Shintaro Sasaki and Hiroshi Yamada (NEC)
Motonobu Ozawa, Michio Otsuki and Yutaka Onodera (Kodansha)

3.2 誤用語チェック・送りがなチェック・書き換え推奨語チェック・かな書き推奨語チェック

第3.1節で述べたような形態素解析失敗による検出だけでは検出洩れが発生してしまう。また、誤りではないが書き換えを促すことも必要になる。そこで、**St.WORDS**では、単語辞書に誤り語も含めて登録しておき、誤用語・書き換え推奨語などの校正用マークを付与しておく手法[5][8]も用意した。現在までに、校正ハンドブック[10]をもとにした辞書登録はほぼ終えたが(表2参照)、さらに増強・見直し中である。

また、ハンドブックの校正知識を辞書に登録する際に、単に一語では正誤を判断できず、文書中の前後の表現と組み合わせて判断すべきものもあった(たとえば「訳(わけ)」は直前が連体形のときはかな書きを推奨する)。そこで、前後の表現に関する条件を分類して単語辞書に登録し、その条件に応じて正誤を判断する機構も用意した。

3.3 同音類語チェック・同音注意語チェック

同音類語チェックでは、読みが同じで意味が類似していて使い分けの難しい語を、第3.2節と同様に辞書にマークを付与しておいて検出する。また、かな漢字変換入力で作成した文書の場合、類語に限らず変換操作ミスによる同音語の使用誤りが発生するので、同音注意語チェックを設け、同音異義語をもつ語についても注意を要するものには辞書にマークを付与しておいて検出するようにした。

この同音注意語チェックでは、誤りでない語が過剰に検出されることになる。しかし、校正支援システムを使って校正作業を行う場合、作業者はマークの付いた箇所にしか気を配らなくなりがちなので(文献[9]でも同様の指摘がある)、同音語については多少過剰でも洩れのないように検出しておきたいというのが、出版社の校閲現場からの要望である。ただし、過剰に検出した場合、単に煩わしいだけでなく、作業者が十分な知識をもっていないと、正誤の判断ができずに修正ミスを犯してしまう可能性が生じる。正誤を判断するための情報が容易に参照できるような仕組みも必要であろう。

3.4 ユーザ辞書機能

50万語の大語彙辞書であっても専門用語や新語はカバーしきれないので、単語の登録や削除、単語情報の変更が行えるユーザ辞書機能を用意した。また、送りがなや書き換え推奨・かな書き推奨の規準は個人や組織によって異なり得るから、前述の校正用マークの変更も行えるようになっている。

4 字面処理による表記法検査

表1に示した検査項目の8～11は、校正者用端末の側で字面処理と規則によって判断する。

4.1 カタカナ表記チェック

カタカナ表記については、固有名詞・新語などが多数存在するので規準表記は定めず、文書内(複数ファイルにまたがっても可)で統一されていないものを変形アルゴリズム[4]により検出する。この変形アルゴリズムでは、「フェイ→フェー」「ヴァ→バ」「ー(長音)→削除」など、複数通りの表記が可能なカタカナ文字列を同形化するような変形を加えて、変形結果は一致するが、もとの表記が異なるカタカナ列を検出する。

この機能により、出版社において従来は手作業で多大な時間と労力を要していたカタカナ表記の統一作業が、本1冊あたり数十分という短時間で行えるようになった。

表2: 辞書の内訳

単語辞書の内訳		校正用マークの内訳	
ワープロ辞書	約8万語	誤用語	1789件
基本語追加	約6万語	規準外送りがな	8003件
異表記追加	約3万語	書き換え推奨語	1114件
派生語追加	約13万語	かな書き推奨語	608件
固有名詞追加	約19万語	同音類語	465件
校正特有語追加	約1万語	同音注意語	66641件
合計	約50万語	合計	78620件

4.2 数字表記チェック

数字表記チェックは、規準として算用数字か漢数字かを指定して、それと異なる箇所を検出する。現状のものは単なる数字テーブルと照合処理であるため、算用数字を規準とした場合に「一任」「一存」のように算用数字書きしない語まで過剰に検出されることが問題になった。そこで、それらを算用数字禁止語として辞書にマークを付与しておくことで過剰検出を防ぐことになった。また、逆に算用数字禁止語を算用数字書きしたもの(例: 1任、1存)も算用数字誤用語として検出する。

4.3 括弧チェック・漢字レベルチェック

括弧チェックでは、スタックを用いて、対応のとれていない括弧を判断して検出する。漢字レベルチェックでは、学年別漢字配当表にもとづいて、指定した漢字水準に合わない漢字を検出する。

5 おわりに

出版社での運用を通して改良を加えてきた校正支援システム**St.WORDS**の文書検査機能について述べた。今後は、さらに辞書の増強や見直しを進めるとともに、同音注意語に関する共起情報を利用した検査なども実現してゆく予定である。

謝辞 NECと講談社の共同開発プロジェクトの関係者、形態素解析部の開発に協力いただいた赤石沢氏・竹元氏、未知語チェックの評価をお願いした宇都宮氏に深謝する。

参考文献

[1] 牛島 ほか、「日本語文章推敲支援ツールの試作とその作成環境」、情処研報84-SW-35-2、1984年。
[2] 空閑 ほか、「文書作成・校正支援用OANERS」、情処32全大1L-6、1986年。
[3] 鈴木 ほか、「日本語文書校正支援システムCRITAC」、情処研報86-JDP-8-5、1986年。
[4] 福島 ほか、「日本語文章作成支援システムCOMET」、信学技報OS86-21、1986年。
[5] 池原 ほか、「日本文訂正支援システム(REVISE)」、通研実報36(9)、1987年。
[6] 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」、情処研報92-NL-91-5、1992年。
[7] 福島 ほか、「日本語文書校正支援システムSt.WORDS」、情処45全大6C-1、1992年。
[8] 福島 ほか、「日本語文章作成支援システムCOMET－文章解析応用の統合化方式を中心に－」、情処研報88-DPHI-20-2、1988年。
[9] 下村 ほか、「人間の文章誤りの検出能力と誤り検出機能の効果に関する実験」、情処論33(12)、1992年。
[10] 講談社校閲局(編)、「講談社 校正ハンドブック」、講談社、1982年。
[11] 竹元 ほか、「口語的表現を含む日本語文の形態素解析」、情処46全大1B-2、1993年。