6C−1

日本語文書校正支援システム St.WORDS

福島 俊一・佐々木 仲太郎†・赤石沢 元博‡・竹元 義美

NEC

C&C システム研究所・C&C 第二公共システム事業部†・C&C 汎用アプリケーション技術本部‡

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ドプロセッサ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を発表し[2][3][4][5]、自然言語処理応用の活発な研究分野の1つとなっている。しかし、その実用化に関しては、日本語ワードプロセッサの拡張機能[6]として一般に普及するには至っておらず、マニュアル作成のための自社内利用[7][8]や、共同開発による新聞社での試験的運用など、対象や使い方を限定したものになっているのが現状である。その背景には、自然言語解析技術の不完全さというシステム側の問題もあるが、日本語の正書法が確立されていない、テクニカルライティング教育が一般に浸透していない、など使う側の文化的な問題もある。そのため、校正支援システムの実用化を進めてゆくには、単に文書検査手法を開発するだけでなく、ユーザとともに運用まで含めた校正支援の方法論を考えてゆくことが必要になる。

筆者らは、上記の考えのもと、設計段階からユーザの要望・意見を取り込んで、出版社向けの日本語文書校正支援システム St.WORDS の開発を進めてきた。従来の試みがマニュアルや新聞記事を対象としているのに対し、St.WORDS では、出版社の扱う一般書籍や週刊誌などの多種多様な文書を校正の対象とする。以下、本稿では St.WORDS の概要・特長を述べる。

2 表記法に重点を置いた文書検査

文書検査を表記法に関するものと内容に関するものとに分けたとき、St.WORDS では前者を充実させる方針をとった。それは以下のような理由による。

校正支援システムにより文書内の様々な種類の誤りを洩れなく検出することは、現状の技術では困難である。おそらく、校正支援システムで文書検査を行なった後、再度、人間自身が検査し直すことになるであろう。したがって、校正支援システムの導入効果は、人手のみの従来の校正作業に完全に置き換わることによる1回の校正作業の時間短縮ではなく、従来と同程度の作業時間で校正支援システムと人間自身との二重の検査が行なえることによる校正の質的向上と考えた方がよい。

このような二重の検査という運用形態を前提としたとき、その作業能率を高めるためには、校正支援システムを用いた検査と人間自身による検査とで役割分担するのがよい。前述の表記法に関する検査と内容に関する検査とでは、作業時の注意の払い方が大きく異なるので、その2つを分けることにより作業能率の向上が期待できる。

役割分担は、表記法に関する検査をシステムで行なうのがよい。内容に関する検査が技術的に難しいこともあるが、表記法に関する検査の方が作業者の心的負担が大きいと考えられるためである。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[9]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。

また、出版社に限っていえば、著者を尊重し、内容に立ち入った文書検査を行なうことは比較的少ない。

以上に述べた運用を含めた検討にもとづき、St.WORDS では、表記法に重点を置いて次のような検査機能を実現した(個々の検査は[4][10]で実現した手法を基本にしている)。

誤字の検出　形態素解析の失敗箇所(例: 開発れた)に加えて、例えば1文字名詞の連続として解析されてしまった箇所(例: 指道力)も信頼性が低いと判断し、未知語として検出する。また、よく誤るパタン(例: 完壁→完璧)を単語辞書に登録しておき、誤用語として検出する。

表記の書き換え推奨　書き換え推奨語(例: 奇蹟→奇跡)やかな書き推奨語(例: 所謂→いわゆる)を、あらかじめ単語辞書にマークしておいて検出する。

同音語表記の注意喚起　使い分けの難しい同音類語(例: 特徴/特長)や、かな漢字変換操作ミスの可能性をもつ同音異義語(例: 構成/校正/更正)を、あらかじめ単語辞書にマークしておいて検出し、注意を促す。

表記のゆれの統一　複数通りの送りがなをもつ語については、規準に合うか否かを単語辞書にマークしておき、規準外のもの(例: 集り→集まり)を検出する。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、文書内で統一されていないもの(例: インタフェース/インターフェース/インターフェイス)をルールにもとづいて検出する。数字表記は、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。

その他　対応のとれていない括弧や、指定した漢字水準(学年別漢字配当にもとづく)に合わない漢字も検出する。

3 大語彙単語辞書を用いた形態素解析

表記法に関する検査の多くは、形態素解析処理を基本にしている。したがって、高い検査精度を得るためには、形態素解析の正確さと単語辞書の充実が要求される。また、出版社で扱う文書は多種多様で、語彙も幅広い。従来の校正支援システムで主に対象とされていたマニュアルや新聞記事と比較すると、口語的表現も多用される(例えば週刊誌)。

そこで、St.WORDS では、高い形態素解析精度を確保するために、複合語(接辞付き語も含む)・固有名詞・口語的表現などを中心に拡充し、約50万語の大語彙単語辞書[11]を構築した。さらに、文書の種類に応じて切り換えて使用できるユーザ辞書も設けた。

大語彙単語辞書は、桁探索法[12]を用いて高速に検索できるように、見出し部を図1のような木構造とした。この辞書は高速検索を優先して直接の追加・削除を行なわず、変更部分は

A Japanese Text Proofreading System St.WORDS

Toshikazu FUKUSHIMA, Shintaro SASAKI, Motohiro AKAISHIZAWA and Yoshikazu TAKEMOTO　NEC Corporation

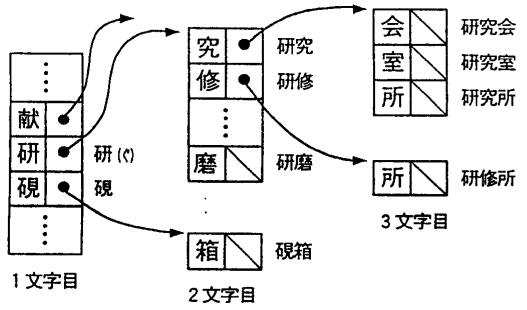

図 1: 単語辞書の見出し部の構造

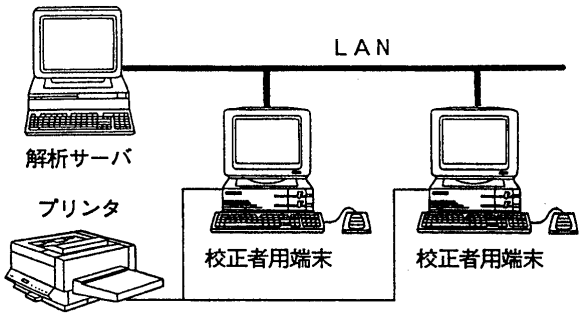

図 2: St.WORDS のシステム構成

別の修正辞書に格納する。修正辞書とユーザ辞書とは B-tree 構造 [12] とした。

さらに、大語彙単語辞書を用いることによる解析候補の組み合わせ的増大、および、それによる処理速度の低下を抑えるために、区間分割文節数限定法を開発した。この方式では、精度の高い形態素解析方式として知られる宮崎らの局所的総当たり法 [13] の考えを踏襲し、字種の変化にもとづいて定めた区間ごとに解析候補を作成する。ただし、全候補作成は行なわず、その区間を最小文節数〜最小文節数 +1 でカバーする候補のみ作成する。定量的な評価はまだ行なっていないが、分割した区間で最小文節数 +2 以上の候補が正解となるケースは極めてまれである。また、このような解析戦略は、複合語を積極的に登録する単語辞書側の方針とも整合がよい。

4 システム構成

校正支援システムの導入・運用では、従来のワードプロセッサ / パーソナルコンピュータなどを用いた作業環境から移行しやすく、かつ、拡張性が高いことが望まれる。そこで、**St.WORDS** では、図 2のように、解析サーバと LAN で接続した複数の校正者用端末 (+ プリンタ) とから成るシステム形態を採用した。解析サーバにはエンジニアリングワークステーション EWS4800 (Unix) を用い、校正者用端末にはパーソナルコンピュータ PC-9800 (WINDOWS) を用いた。校正者用端末では、MS-DOS テキストファイル形式で、既存のワードプロセッサなどと文書の受け渡しができる。解析サーバと校正者用端末は各々、次のような機能を果たす。

解析サーバ　形態素解析プロセス、訂正候補検索プロセス、大語彙単語辞書・ユーザ辞書の更新プロセスが常駐し、校正者用端末からの要求に応じて動作する。未知語と、単語辞書に記述された情報を用いる誤用語・書き換え推奨語・かな書き推奨語・同音語・規準外送りがなは、形態素解析の段階で検出する。

校正者用端末　ユーザに対話的な校正作業環境を提供する。文書編集 (ワードプロセッサ)、文書検査 (解析サーバの形態素解析プロセス・訂正候補検索プロセスを起動、検査該当箇所を色マーク表示)、文書印刷 (検査該当箇所の網掛け印刷)、文書管理 (校正バージョン管理)、辞書編集 (解析サーバの辞書更新プロセスを起動) などの機能をもつ。カタカナ表記・数字表記・括弧・漢字水準などのルールベースの検査は校正者用端末側で実行する。

このようなシステム形態は、昨今の分散処理環境に適合しているのに加えて、単語辞書の共有により、校正規準の統一的管理が行なえる。

5 おわりに

出版社向けに開発した校正支援システム **St.WORDS** の機能・構成・特長などを報告した。**St.WORDS** は、約 50 万語の大語彙単語辞書を搭載した形態素解析部をサーバ化した構成をとり、運用面の検討にもとづいて、表記法に重点を置いた文書検査機能を実現している。**St.WORDS** は現在、ユーザ (出版社の校閲部門) が試験的運用を開始した段階にあり、評価結果については改めて報告したい。

また、**St.WORDS** では、文書検査機能に加えて、キーワード抽出 / 索引作成機能も提供してゆく予定である。

謝辞　本研究は株式会社 講談社との共同開発プロジェクトの一環で進めており、小澤室次長・野村部長・大槻氏・小野寺氏をはじめとするプロジェクト関係者に深謝する。

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] 「次世代ワープロの決め手となるか校正支援 / 可読性評価ツール」、日経バイト、1988 年 3 月。
[7] 高橋 ほか、「計算機マニュアル推敲・査読システム MAPLE の開発と運用」、情処論 31(7)、1990 年。
[8] 大山、「マニュアル検査システム TECS/M − 基本構成 −」、情処 43 全大 6H-3、1991 年。
[9] 講談社校閲局 (編)、「講談社 校正ハンドブック」、講談社、1982 年。
[10] 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。
[11] 福島、「日本語解析マシン − 大語彙言語処理へのハードウェアアプローチ −」、自然言語処理の新しい応用シンポジウム、1992 年。
[12] D.E.Knuth, "The Art of Computer Programming, Vol.3, Sorting and Searching", Addison-Wesley, 1973.
[13] 宮崎 ほか、「日本文音声出力システムの言語処理」、通研実報 35(2)、1986 年。