日本文訂正支援システム　ＲＥＶＩＳＥ

4J-9

安田恒雄，島崎勝美，高木伸一郎，池原　悟

ＮＴＴ電気通信研究所

１．はじめに

日本文データベースの作成や、新聞記事データ等の作成において、計算機に入力した多量の漢字かな混じり文に対して誤字等の誤りを検出し、訂正する作業は、従来人手で行われている。膨大なデータを処理する場合、この作業の精神的・肉体的負担は大きく、計算機による大幅な省力化が求められている。

本稿では、新聞記事の校正等の日本文訂正作業の省力化を目的として、本格的な日本文解析を行うことにより、計算機による実用的なレベルでの誤り検出を実現したシステム、ＲＥＶＩＳＥについて報告する。

２．従来の問題点と多段解析法の導入

入力された日本文の誤りを検出する方法として、誤りのパターンを持ち、パターン・マッチングによって誤りを検出する方法がある[1][2]。この方法は、限られた文の特定の誤りの検出には有効であるが、新聞記事等の多量の”開かれた”文に対して、多種類の誤りを検出するには処理が膨大になり過ぎる。このため、対象の文を文法的、意味的に解析して誤りを検出する方法の実現が必要であるが、誤りを含む文の「誤り」を正しく認定するためには、解析の基本である形態素解析の精度の高さが問題となる。従来、この形態素解析の方法として最長一致法、総当り法，ＤＰ照合法等が提案されているが、いずれの方法においても正しい文を対象とした解析において、精度が十分でないという問題があった。

この問題に対して、我々は複合語等の解析において、部分的に係り受け解析など深いレベルの解析を行う多段解析法を提案し、実際の日本文音声出力システムに適用して、正しい文に対する十分な有効性（９９．８％の高音韻正解率を実現）を確認している[3]。ＲＥＶＩＳＥでは、この方法をさらに発展させる事により、誤りを含む文に対しても実用的なレベルで「誤り」を認定、検出できる見通しを得た。

３．ＲＥＶＩＳＥの機能

計算機に入力された新聞記事等の訂正は、一般に誤字、脱字等、文のスタイルレベルのチェック、修正を原稿とつき合わせて行う校正作業と、文の内容レベルまでチェックする校閲作業とがあり、２人一組みで読み合わせながらチェック、修正する等の方法が行われる。

ＲＥＶＩＳＥは、人間が行う日本文の訂正を支援するという立場から文法的な検証、スタイル・チェックによる誤りや注意語の自動検出等により、校正、校閲を支援する誤りチェック機能と、従来人間が行っていた読み合わせを合成音声で代行するための読み合わせ機能を実現している。

ＲＥＶＩＳＥを利用した日本文訂正システムの例を図１に示す。

３．１誤りチェック機能

高精度の単語認定を基本として次の２種類の誤りを検出する。

①基本誤り――文法的に検出できる未知語、承接チェックエラー等

②スタイル・チェック誤り――単語の持つ誤り属性で検出する誤用語、禁止語等

ＲＥＶＩＳＥの誤りチェック機能の項目例とチェック対象例を表１に示す。この内、誤りの認定基準に利用者によってゆらぎがあるもの（送り仮名エラー等）や、利用者に特有の誤り易い語、注意したい語等についてはユーザによる辞書の登録を可能として、利用者が登録した情報に基づいて認定する方式とした。

３．２読み合わせ機能

訂正支援のための読み合わせ機能として、連濁化処理等を考慮した自然読みと、校正等の作業を行い易くする約物読みや特殊用語読み等の校正特殊読みの２種類の読

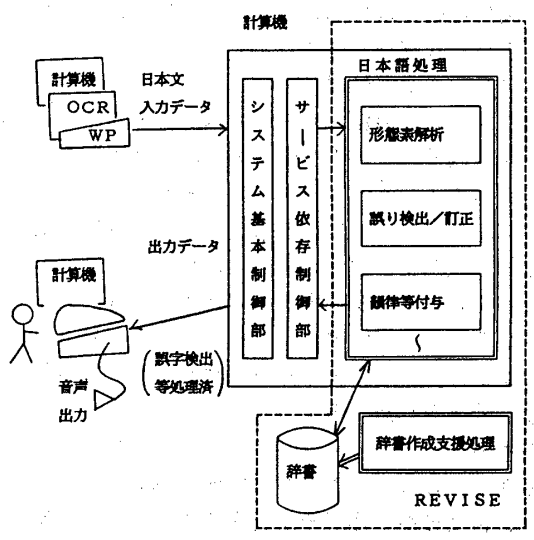

図1．REVISEと日本文訂正システム構成例

Revision of Interactive Support System for Japanese Verbal Error, REVISE

Tsuneo YASUDA, Katsumi SHIMAZAKI, Shinitirou TAKAGI, Satoru IKEHARA

NTT Electrical Communication Laboratory

みをサポートしている。この例を表２に示す。

読み合わせ機能は、誤りチェックを補完するという立場から非常に有効である。例えば、文法的なチェックでは断定出来ない誤字や誤挿入文字（日本後「ニホンウシロ」＝＝日本語、憤情報「ナサケジョーホー」＝＝情報）が、読み合わせによってチェックアウトでき、文意による判定が必要な例（昭和十一年＝＝昭和六十一年、今日は＝＝今回は）においても発見し易くなった。又、ユーザ辞書で指定された校正の特殊読みを行う事によって、同音異義語の識別や記号抜け等のチェックが可能となった。

４．ＲＥＶＩＳＥの構成

ＲＥＶＩＳＥは、以下の処理から構成される。

４．１日本語処理

処理の流れを図２に示す。多段解析法による単語認定を、誤りチェックと日本文音声変換処理の２つの機能の共通機能として取り出し、処理を融合させることにより効率良く訂正支援を実現した。又、誤りの認定や重要度の設定を利用者の判断基準に応じて可能とする柔軟な訂正支援を実現するため、以下の工夫をした。

(a)ＲＥＶＩＳＥとして豊富な誤りチェック機能を持つが、入力された文に対して有効とするチェック種別を指定可能とした。

(b)誤りの重要度の決定はＲＥＶＩＳＥでは行わず、誤り検定の結果をサービス依存制御部へ詳細コードとして通知する方式とした。

誤りの検出の例を図３に示す。校正済みの文には、チェックアウト対象語の直前にチェックアウト種別と誤り文字数からなる校正制御文字を挿入する。日本語処理に処理を依頼したサービス依存制御部では、この制御文字により利用者の要求に応じてディスプレーに表示する色種別等の自由な制御ができる。

４．２辞書作成支援処理

辞書作成の処理の流れを図４に示す。未知語による誤り等、基本誤り（３．１、①）の検出精度は辞書の充実度が重要なポイントとなる。このためＲＥＶＩＳＥでは、約４３万語のシステム辞書を備えた。又、利用者の特定の誤用語や注意語、読み合わせにおける特殊読み等のユーザ辞書の定義を容易にするため、ユーザ辞書には必要最低限の情報だけを定義すれば良いように、統合辞書作成プログラムによって日本語処理に必要な詳細情報を補って統合辞書を作成するよう工夫した。

５．おわりに

今回開発したＲＥＶＩＳＥの機能、構成について報告したが、精度の高い形態素解析を基本としたＲＥＶＩＳＥの実現により、今後は、固有名詞の実在性のチェックや、校閲レベルの機能の実現等、より高度な誤りチェック機能の実現も可能となった。

６．参考文献

(1) James L. Peterson: Lecture Notes in Computer Science, Vol. 96: Computer Programs For Spelling Correction. Springer-Verlag. 1980. P11

(2) 建石ほか: DictionとStyleの日本語化についてP1381-1382、情報処理学会第31回全国大会

(3) 宮崎ほか: 日本文音声出力システムの言語処理、P157-166、通研実報第35巻第2号(1986)

表1．誤りチェック機能項目例

項番	チェック項目 C_0	校正表示属性 分類	C_1	例
1	属性（辞書収録時指定された語）<A>	非正規語	A	ホバークラフト（ホーバークラフト）
		俗語	B	宅急便（宅配便）
		禁止語	C	盲
		注意語	D	解放（開放）
		誤用語	E	墨田川（隅田川）
		常用漢字外表記	F	合鍵
		固有名詞(名前)	G	聖子
2	送り仮名<B>		A	払い込み金（払込金）
3	未知語 <C>	漢字表記	A	漢字のみの未知語
		カタカナ表記	B	カタカナのみの未知語
4	非承接関係語 <D>	固有名詞	A	（複合語中に表記され係り受けがない）
		数詞	B	

表2．読み合わせ機能

項番	選択項目	例	モード 自然読み	モード 特殊読み
1	数詞	￥１００	ヒャク・エン	エン・イチ・ゼロ・ゼロ
2	用語	「太田」	オータ	フトダ
		「追求」	ツイキュー	ツイキュー・モトム
3	約物	？	（無音）	ミミ
		（	（無音）	パーレン

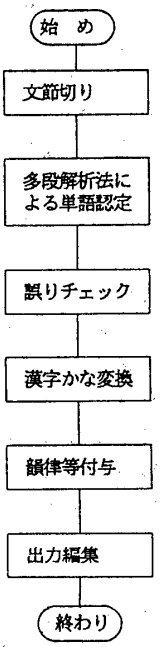

図２．処理の流れ

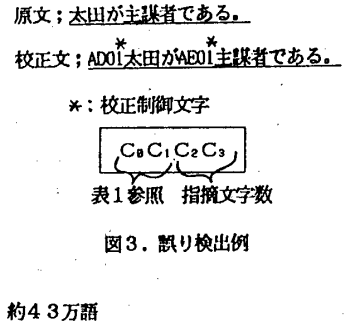

図３．誤り検出例

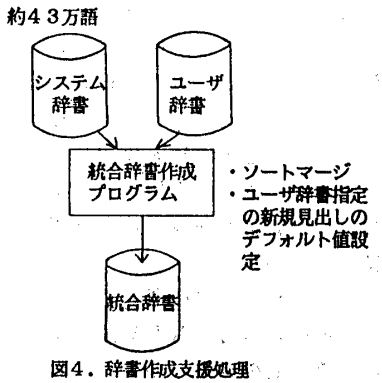

図４．辞書作成支援処理