6B-4 日本文訂正支援システムにおける未知語訂正候補抽出方式

高木伸一郎　安田恒雄　島崎勝美　松岡浩司
ＮＴＴ情報通信処理研究所

1．はじめに

日本文データベースや新聞記事データ等の作成においては、計算機に入力した多量の漢字かな混じり文に含まれる誤字等の誤りを検出、訂正する作業の大幅な省力化が求められている。このため、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計算機による実用的なレベルでの誤り検出を行う日本文訂正支援システムＲＥＶＩＳＥを開発した。(参考文献１参照)本稿では、さらに高位機能として実現した検出誤りのレベルに応じた訂正候補抽出方式の概要、特に未知語により検出した誤りの訂正候補抽出方式について述べる。

2．誤りの分類とＲＥＶＩＳＥにおける訂正候補抽出方式

新聞記事で検出された誤りのうち誤字、脱字、誤挿、送り仮名等の文脈に依存しない校正レベルの誤りについて分類すると、次のように大別される。（図１参照）

(1) 表記の誤り（文字レベルの誤りや送り仮名、表記のゆらぎ）
(2) 表現の誤り（複合語内承接誤りや同音異義語等）
(3) 言語情報以外の一般常識に関する誤り（固有名詞誤り等）

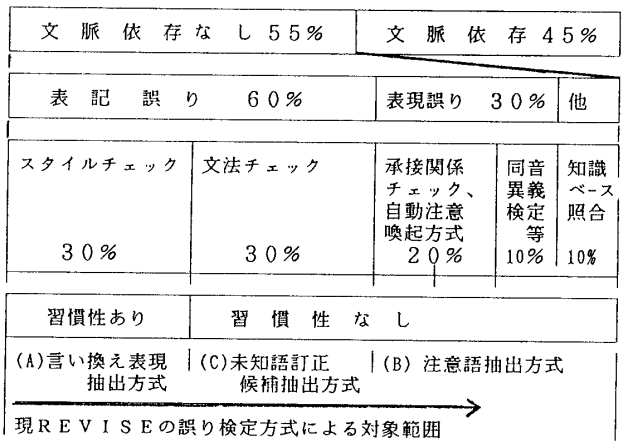

図１　誤り分類と対応する訂正候補抽出方式

ここで、対応する誤り検定方式は、(1) では、習慣性のある誤りを検定するスタイルチェックと習慣性のない誤りを検定する文法チェックである。また(2) では、固有名詞・数詞に関する意味的な接続関係の有無を検定する承接関係チェックや表１に示すヒューリスチックなルールによるシステム自動注意喚起方式である。ここでは、複合語漢字列内の誤字のように文法的に検出されにくくかつ習慣性のない誤りに対応するため、複合語内の単語係り受け解析強化を行っている。このほか、同音異義語誤り検定、固有名詞誤り検定の開発を進めている。

The Correct Candidacy Extract Method on REVISE (Revision Support System for Japanese Verbal Error)
Shinichiro TAKAGI,Tsuneo YASUDA,Katsumi SHIMAZAKI,Kouji Matsuoka
NTT Communications and Information Processing Laboratory

表１　自動注意喚起方式および検定精度

項目	FA	FB	FC	FD	FE	計
検出数	26	0	0	124	126	276
ヒット数	10	0	0	22	32	64
ヒット率	38	0	0	18	25	23%

表記以外の誤り：912中

FA：係り受け無し１文字接辞　　例：参連（参画）
FB：係り受け無しひらがな接辞　例：白げ（白さ）
FC：連続接辞(接頭辞＋接尾辞)　例：不荷（不可）
FD：連続接辞(接頭辞＋接尾辞以外)例：講買層（購買層）
FE：係り受け無し１文字一般名詞・固有名詞　例：日木経済（日本経済）

これらの誤り検定方式によって検出された誤りの特徴に応じた訂正処理を行うために、ＲＥＶＩＳＥでは、次の３種類の訂正候補抽出方式を実現した。

(A) 言い換え表現抽出方式（言い換え表現導入過程:図２参照）

習慣性のある表記誤りについては利用者辞書に予め誤り表記と対となる正解表記を登録し、誤り検定処理で使用する校正辞書にペアで組み込むことによって、誤り表記の単語が認定された場合に訂正候補を抽出する。現在の評価では、ほぼ100％の訂正候補抽出精度を実現している。

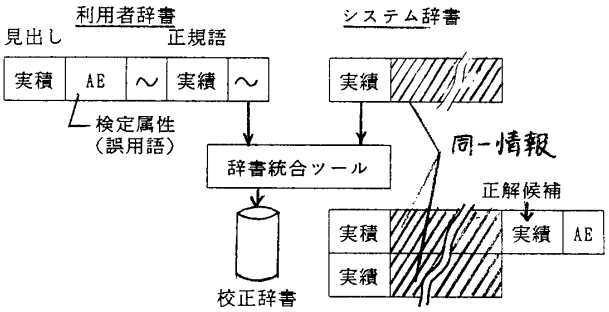

図２　言い換え表現訂正候補の導入法

(B) 注意語抽出方式

同音異義語や使用を誤り易い類義語・慣用語を注意語として予め注意語辞書に登録する事によりこれらの語が出現した場合には、利用者に対する注意喚起を行うとともに、注意語辞書の検索結果からの複数の訂正候補を抽出し提示する。

(C) 未知語訂正候補抽出方式

文法チェックにより検定される誤りは習慣性がなく、文法的に接続が不良な未知語を生成する特徴がある。このため未知語出現状況を分析して６種類の個別訂正候補抽出方式を考案し、その組み合わせによって有効な訂正処理機能の実現を図った。

3．未知語訂正候補抽出方式

新聞記事１日分（約13万文字）に含まれる誤りで未知語によって検出された２５１件について未知語の出現特性を分類し次の６種類の個別訂正候補抽出方式を実現した。

（A）高頻度類形文字・重複文字未知語訂正候補抽出

・高頻度で誤りやすい文字の対を予めテーブルに登録し、該当の文字が、未知語の前後に出現した際に置き換えて訂正候補を抽出する。㊀：未知語

（例） ムルダ 二国軍 ――― ムルダニ（二---- ニ）

彼にとつて ――― とって （つ---- っ）

・同一文字の誤挿により未知語化した際に重複文字の一方を削除する訂正候補を抽出する。

（例） 通じて て 得た ――― 通じて 得た

（B）名詞＋用言未知語訂正候補抽出

助詞抜け誤りにより見掛け上、「名詞＋用言」化した場合に、文法的接続検定の結果発生する未知語を検出し、後方の用言が有する必須の格助詞を訂正候補として抽出する。必須格が複数存在する場合の訂正候補選択に関しては、前方の名詞句の意味カテゴリーについて照合する必要があるので、今後の機能としている。

（例） 決断 迫る ――― 決断 を 迫る（迫る--- を、に）

（C）複合単語訂正候補抽出（図３参照）

漢字列複合語は、一般に２文字単語の連続として構成される頻度が高いことを利用し、予め前後１文字のキーで検索できる高頻度漢字２文字候補テーブルを作成する。そして漢字列複合語内未知語の前後の１文字をキーとしてテーブルを索引し訂正候補を抽出する。この際、候補数が多数となるため、前後との文法的、意味的な接続条件を用いて候補の絞り込みを行う。

（例） 行 販 政 ――― 行 財 政（財政）

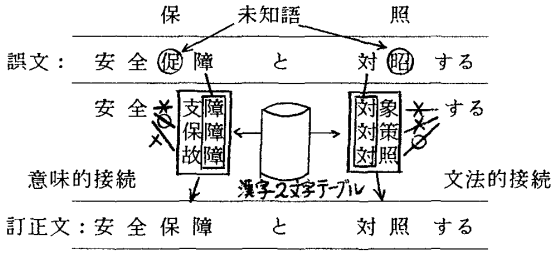

図３　複合語内漢字候補抽出方式

（D）付属語尾文字訂正候補抽出

用言、用言性名詞での付属語尾が未知語となる場合に、語幹の活用型より付属語尾を抽出し訂正候補とする。さらに後方の単語の品詞により文法的に接続可能な活用形を選択する。

（例） 埋込 む ――― 埋め込む（埋 ---- め）

（E）カタカナ・英字語訂正候補抽出

未知語を含むカタカナあるいは英字の文字列は見出し長も長く類似の単語も多いので誤りが混入しやすい。従ってカタカナあるいは英字辞書と照合をとることが有効であるが、対象単語数が多いので、２段階の照合過程による抽出方式を実現した。第１は、予め作成したカタカナ語訂正候補テーブルを２文字をキーに、少なくとも２文字以上マッチするカタカナあるいは英字語候補群を抽出する。つぎに高速に照合を行う連想統合型照合法を用いて照合値を算出し脱字・誤挿補完処理を行って訂正候補順位を付与する。（図４参照）

本方式によって、48事例のカタカナ誤りに対し、正解候補含有率は１位85%、５位以内92%と非常に高い精度が得られた。

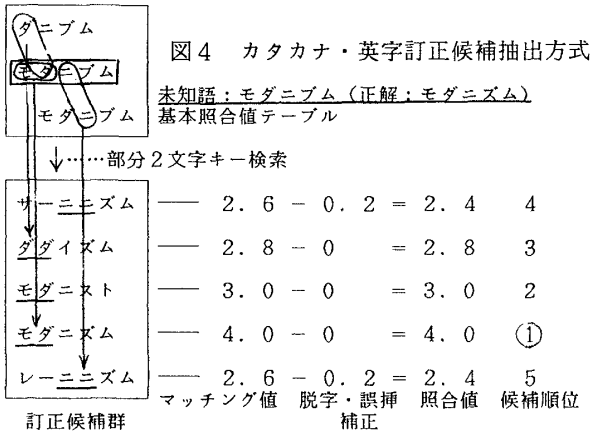

図４　カタカナ・英字訂正候補抽出方式

未知語：モダニブム（正解：モダニズム）
基本照合値テーブル

（F）類音・類形文字（入力装置特性）訂正候補抽出

ペンタッチ・ＯＣＲ等の入力特性に応じた誤りやすい類似文字を予めテーブル化し、未知語文字を用いて検索し候補抽出を行う。

（例）諸国間 わ 関係 ――― の （わ---- の ペンタッチ）

懇案事項 ――― 懸案（懇---- 懸 ＯＣＲ）

4．訂正候補抽出方式の評価

検討対象とした251 件の未知語誤りに対する訂正候補抽出精度の評価結果を表２に示す。この結果、候補抽出処理の起動可能割合が82%、このうち候補数を考慮せず正解候補を抽出可能な割合は53%、さらに上位約10位以内に正解候補を含有する割合が49%となり、ほぼ未知語により検出した誤りの２件中１件は自動的に訂正候補が抽出できることがわかった。また起動されない誤りの６割はひらがな１文字誤字であることから、現在、文字の連接確率等の統計情報を用いる手法を検討している。

表２　未知語訂正候補抽出方式　評価結果

対象未知語数：２５１件（誤りを含む未知語件数）約１３万文字

訂正候補抽出方式	A	B	C	D	E	F	総　計
候補抽出件数	２０	4	４７	２２	９２	２１	２０６ (82.1%)
正解候補抽出件数	１９	3	３１	5	５４	２１	１３３ (53.0%)
上位候補選択件数※	１９	3	２２	5	５２	１１	１２２ (48.6%)

※訂正候補絞り込みの結果、正解が上位に残存するもの（上位約１０候補以内）

5．おわりに

概述した訂正候補抽出方式の導入により、校正者は表記誤りについては計算機が示す候補の中から適当な候補を選択するというわずかの処理で校正できるので、省力化・文書作成の効率化を推進できる。今後は訂正精度を高めるとともに、同音異義語検定や知識ベース照合を用いた地名固有名詞実在性検定等の誤り検定・訂正機能の検討を進める。

6．参考文献

(1)安田ほか:「日本文訂正支援システム ＲＥＶＩＳＥ」情処第33回全国大会 4J-9 1986,10

(2)高木ほか:「日本文訂正支援システムＲＥＶＩＳＥにおける誤り検定方式の検討」 情処第34回全国大会 6X-4 1987,3