4J-5

構造化文書上における校正・推こう手法の検討

鈴木恵美子 武田浩一 藤崎哲之助

日本アイ・ビー・エム株式会社サイエンス・インスティチュート

1 はじめに

ワード・プロセッサで文書を作成する際に、(ワード・プロセッサ・ソフトにもよるが)どのような問題点が生じるか、また、そのワード・プロセッサの使い勝手がどのようであるかについては、東京工業大学の木村らがかなり詳しい報告をしている[2,3]。彼等の報告によると[4]、著作作業にワード・プロセッサを使用した場合、出版社の要請による「1 行の文字数の調整」や「校正に備えた手直し」にかかった時間も含めてワード・プロセッサの方が原稿を直接原稿用紙に書いたときに比べて2 倍以上時間がかかったとのことである。

2 CRITACの概要

我々は現在CRITACという文書を自動的に校正するシステムを試作しており[7]、文書を高度に構造化し、文節区切りや、読み等の情報を付加した、構造化文書の上で校正ルールを働かせることを試みている[7] 入力された文書は文節切りされ[5]、自立部と付属部に分離され、漢字複合語は短単位に分割され[8]、付属語は接続検定されて[1]、最終的にPrologの節集合に変換される。

CRITACの校正知識はソース表現[10](通常ワード・プロセッサで作成される文書と同じように見える)上で働くものと、KWIC表現[10](構造化文書の各自立語をキーワードとしてその読み順などによって並べ代え、前後の文脈とともに表示したもの)上で働くものの2 種類あり、どちらもPrologのルールとして表現されている。

3 ワード・プロセッサで作成された文書に現われた誤りの調査

我々はワード・プロセッサで作成された2 種類の比較的短いサンプル文書でそこに現われる誤りを調査してみた。この時は、あまり作成中の画面に注目せず、下書きの原稿を見ながら、無造作に変換キーや文字種キーを押した。なお、ここではローマ字かな変換入力を行ない、文節単位のかな漢字変換を行った。文書の大きさとしては、1 つ(文書A)が、2099 文字、もう一方(文書B)は、2779 文字、2 文書とも情報工学分野の論文である。文書Aも文書Bも筆者の内の1 人が入力したもので、文書Aはしばらく時間をおいてから、同じワード・プロセッサを使用して今度は比較的注意深く入力してみた結果文書A' を得た。誤りを分類した結果を表1 に示す。

	文書A	文書B	文書A
誤変換α	23	21	3
誤変換β	2	1	0
未変換α	8	0	0
未変換β	1	2	0
ミスタイプ	7	5	2
文字種キーの押し忘れ	2	1	0
変換されすぎ	10	3	2
固有名詞	2	0	0
表記のゆれ	0	4	0
スタイルの乱れ	0	1	0
誤修正	0	1	0
誤りの数の合計	55	41	7

表1 . サンプル文書に現われた誤り

ここで、誤変換に2 種類あるのは、誤変換αというのが、いわゆる同音異義語による変換誤りで、誤変換βというのが、ミスタイプによる誤変換である。

未変換についても誤変換と同じく、未変換αは単純に変換キーの押し忘れと思われるもので、未変換βはミスタイプのせいで、該当する漢字がみつからず、変換されずに残ってしまったと思われるものである。

その他のミスタイプには、熟練者には起こらないであろうが、「でけあがった←できあがった」(本当は右手中指の「i」を打とうと思っているのに、左手中指の「e」を打ったことによる誤り)のような、右手と左手の動作誤りが特徴的であった。

文字種キーの押し忘れによる誤りは思いのほか少なく、全体で3 種しか現われなかったが、これらのうちの2 つはカタカナ語の後でひらがなキーを押さずに続けて付属語を入力してしまっており、残る1 つは、カタカナ語(プログラム)がひらがなで書かれていた。

『変換され過ぎ』というのは、長い単位で変換したときや、または助詞がカタカナ語とカタカナ語をはさんで途中にあるときに変換キーによって変換されてしまったような場合を指す。

表記のゆれは、外来語、専門用語に多い。

誤修正についてはもともと正しく書かれていたであろう語の一部が消去されてしまっていた。

On a method to proofread using structured text
Emiko Suzuki, Koichi Takeda, Tetsunosuke Fujisaki
Science Institute, IBM Jpapn, Ltd.

4　構造化文書上での取り扱い

構造化文書では、各自立語には読みがふられ、自立語や付属語の品詞といった情報も付与されたＰｒｏｌｏｇの節が与えられている。この'読み'と段落(章、文書)内の統計的情報から、誤変換の内、70% は検出できる。残り30% の内、前回の発表［9］でも述べた、KWIC表現上で単語を読みの順に並べ、読みが同じで表記の異なる語が現われていたときに警告するようなルールを働かせることで20% 程度はカバーできると考える。これは、文書中ある語を1回しか使用せず、かつそれが誤変換である、あるいは同じ語を何度も同じように誤っている確率は全体の誤変換αの内の10% 程度であろうと予測するためである。誤変換βについては、「救って←作って」や「苦情←向上」など、この文節が本当に誤りかどうかを判定するには品詞や語の使用頻度だけでなく、文の意味まで考えなければならない。ＣＲＩＴＡＣでは現在のところ、構文的な情報や意味は扱わないので、誤変換βに対しては検出・訂正の方法がない。

未変換αについては、変換されるべき正しいかな文字列が入力されているが、構造化文書では正しく自立語と付属語に分離できない可能性がある。今回のサンプルでは8 個の内の1 個については正しく文節切りできない。残り7 個については文節切りは正しいので、文節内の構造をうまくとらえられれば辞書を引くことで辞書の正書を取り出せる。一方未変換βの内文書Ｂに現われた「ひょうげんん←表現」についてはひらがなで「ん」が2 つ続くような語はないので警告できる。しかし文書Ａに現われた「ねべてきた←述べてきた」については、例えばミスタイプは1 つの文節中では1 カ所にしか現われず、かつ文書の入力が常にローマ字かな漢字変換であるとすると、そのローマ字表現「nebetekita 」のうちのどこか1 文字を別のアルファベットで置き換えることで、語を創造できるが、そこから派生する語は非常に多くなる可能性があり、処理も複雑なので、当面はこの種の誤りの訂正は考えない。

ミスタイプについては英語で行なわれているようなスペルチェックを応用することで将来は正しい語を推定できるようにしたい。

文字種キーの押し忘れによる誤りはＣＲＩＴＡＣでは、付属語接続検定の折に接続に失敗し、かつそのような自立語(ワード・プロセッサニオイテ、チェックサエナク)も辞書にないことから、検出し警告することができる。

『変換され過ぎ』には「～賀←～が」「此のため←このため」「Ａ戸Ｂ←ＡとＢ」のようなものがあり、構造化文書の前処理段階で読みがふられている。ＣＲＩＴＡＣではその読みが助詞の読みに等しいような場合に検出し、訂正することができる。

アルファベットで書かれる語はそれが大文字か小文字かで、またカタカナ化する場合はその発音により様々に表記がゆれる可能性があり、辞書等で管理するほかはないと考えている。

誤修正についてはそれが長い語の一部で、自立語が一短単位でも残っていれば、ＫＷＩＣ表現上で見つけることができる。

この検討結果を、現在及び将来どの程度の誤りが検出・訂正できるかで、誤りの種類別にグラフにしてみた(図2 参照)。詳しい校正ルールについては紙面の都合上、文献「6」を参照されたい。

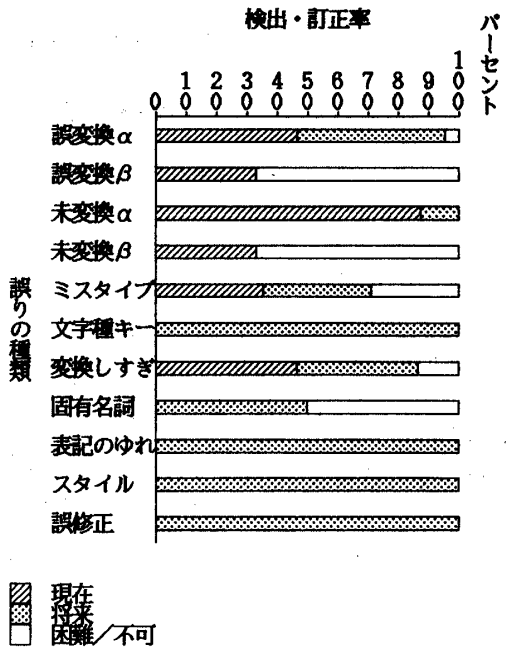

図２．誤りの検出・訂正率

5　おわりに

以上ワード・プロセッサによって作成された文書に現われた誤りの種類とその原因について検討した。今後はこの結果を生かし、それぞれの原因に対する校正・推こう方法を実際に試行してみるつもりである。

［参考文献］

1. 大河内：仮名漢字変換のための形態素接続規則、東京サイエンティフィック・センター・レポート、N:G318-1560-1、19281。
2. 木村ほか：「松」とJWORDの比較論、コンピュータソフトウェア、vol.2、no.2、pp.36-62、1985。
3. 木村ほか：一太郎の「住み込み」評価、日経バイト、pp.185-201、1986。
4. 木村ほか：著作過程の業務分析、pp.1643-1644、情報処理学会第30回全国大会。
5. 鈴木：漢字かな混じり文に現われるひらがな列の文節推定方法について、pp.1383-1384、1985。
6. 鈴木ほか:日本語文書校正支援システムCRITAC、日本語文書処理研究会、1986年9月。
7. 武田、藤崎：統計的手法を用いた漢字複合語の短単位分割、自然言語処理研究会、48-2、1985。
8. 武田ほか：日本語文書校正支援システムCRITAC、pp.1691-1692、情報処理学会第32回全国大会。
9. 武田ほか：日本語文書校正支援システムCRITACの校正知識、pp.1693-1694、情報処理学会第32回全国大会。
10. 武田ほか：CRITAC - A Japanese Text Proofreading System 、ＣＯＬＩＮＧ、1986。