日本語校正支援システム(Joyner)の研究について(1)
－綴り誤り自動訂正について－

2J－4

松井 くにお　中村 直人　伊吹 潤　徐 国偉
富士通研究所

1　はじめに

インターネットなどのネットワークの普及により、電子化された情報を発信する機会が増えてきたが、同音異義語の変換誤りなどのいわゆる”ワープロミス”が散見される。こういった誤りの検出技術として、形態素解析や共起関係を利用する方法が提案[1]されているが、いろいろな問題点を含んでおり[2]、解決策とはなっていない。

本稿では、同音異義語や同音異字語の綴り誤りを検出するだけでなく、自動訂正するシステムを提案し、その有効性を示す。

2　文書情報の誤り

2.1　新聞記事における誤りの種類

新聞社における校正例を分類すると、表１のように分類できる。ここで、割合１は日本語として許容できる範囲で新聞特有の校正処理を含んだ割合を示し、割合２は日本語として許容できないレベルの誤りのみに限定した割合を示す。

表１ 新聞校正の誤りの分類

誤り原因	個数	割合１	割合２
公的表記規則	145	25.0%	-
かな漢字使い分け	131	22.5%	-
用語	16	2.8%	-
同音語使い分け	87	15.0%	30.1%
記号・数字規則	67	11.5%	23.1%
脱落等入力ミス	40	6.9%	13.9%
文章表現	95	16.4%	32.9%

2.2　文書分野別の誤りの種類

文書分野別の誤りの分類では、新聞記事・技術文書・マニュアル別に調査した結果[3]がある。この例でも同音異義（異字）語の誤りは、新聞記事(33%)、技術文書(16%)、マニュアル(8%)となっており、同音異義語の誤りは誤り全体の中で大きな割合を占めている。

3　同音異義語への対処

3.1　従来の方式の問題点

従来の方式は、誤り表記をも含む辞書を用いて形態素解析を行ない、誤り表記にヒットした場合、それとペアの正解表記に置き換える方式である。また、それぞれの形態素の共起関係を利用する方式も提案されている[1]。しかしながらこれらの方式は以下の問題点を持つ。

1. 誤りのバリエーションは無限
   一つの正解に複数個の誤り表記を登録する必要があり、すべての誤りを予想して登録するのは事実上不可能である。

2. 思い込みによる誤りには無能
   文章を作成する際に、思い込みまたは知識の欠落によって誤った表記を用いた場合（例えば「国鉄**精算**事業団」）、形態素解析自体は何の異常も発見できず、誤り指摘には至らない。

3. 過剰指摘を助長
   我々の調査では、現在製品となっているシステムにおいて、誤りの過剰検出（正解なのに誤りの可能性があると指摘）のうち同音異義語の過剰検出が55%を占めている。また、新聞記事22件に27個の誤りを作為的に盛り込んだテストデータにおいて、誤り指摘されたものの正解率（本当に誤りであるもの）は、わずか8%である。

3.2　綴り誤り自動訂正方式

このような問題点を解消するために、文字ラティス型形態素解析の利用した綴り誤り自動訂正方式を提案する（図１）。

これは、文字列を一旦形態素解析した後、同音異義語などの候補語をラティス状に展開（図２）し、最もコストが小さくなるようなパスを選択してそれを出力文字

Japanese spelling corrector(Joyner)(1)
Kunio MATSUI, Naohito NAKAMURA, Jun IBUKI, Guowei XU
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki-shi 211 Japan

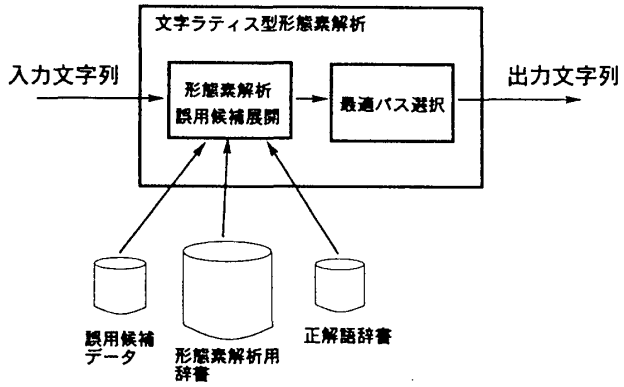

図 1: 綴り誤り自動訂正方式の構成

列とする方式である。ここで、コストとは以下に示す接続コストと展開コストから成る。

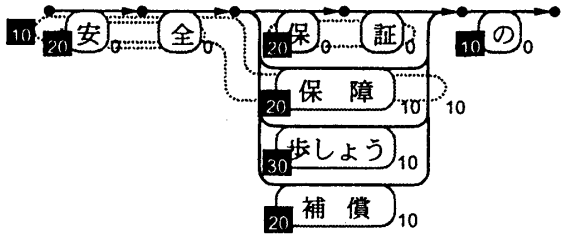

図 2: 文字ラティスとコスト

- 接続コスト ($\mathbf{C_m}$)
  文字列を形態素に分割した場合の形態素自体のコスト及びその連接コストの和で表現する。図２の各形態素の左下に白抜きの数字で示す。
- 展開コスト ($\mathbf{C_e}$)
  同音異義語などの展開を行なった場合のコストで表現する。図２の各形態素の右下の数字で示す。

この方式においては、入力文字列パス (P_0) 以外のパス (P) が出力パスとして判定される時には以下の式が成立するようにコストを調整する。

$$C_m(P_0) > C_m(P) + C_e(P)$$

コスト計算の例を例を表２に示す。「安全保証の」が入力文字列の場合、「安全保障」が辞書になければ「安全保証の」が、辞書にあれば「安全保障の」が出力文字列となる。また、正しい文字列「安全保障の」が入力文字列の場合は、辞書の登録の有無にかかわらず、「安全保障の」が出力文字列となる。

表２ コスト計算の例

文字列	接続コスト	展開コスト	合計
安全｜保証｜の	20+20+10	0	50
安全｜保障｜の	20+20+10	10	60
安全｜歩しょう｜の	20+30+10	10	70
安全｜補償｜の	20+20+10	10	60
安全保障｜の	10+10	10	30

なお、正解語辞書は著名人・組織名・地名を人手で作成し、その他の複合語は新聞記事からの自動抽出を行なった。

4 評価

著名人（約 18,000 語）・組織名（約 28,000 語）を正解語辞書として、ランダムに誤りを含む 150 語をテストした結果を表３に示す。

表３ 評価結果

項目	訂正成功	訂正失敗	変化なし
個数	60(40.0%)	14(9.3%)	76(50.7%)
例	新進塔→新進党	無差別姓→無差別製	行気圧→行気圧
	情報旗艦→情報機関	蛍光党→蛍光等	登山袋→登山袋
	核実検→核実験	最大球→最大九	表面香→表面香
	国務朝刊→国務長官	自動車胴→自動車同	国連期間→国連期
	半生気→半世紀	選挙監視暖→選挙監視弾	回生法案→回生法

5 まとめ

同音異義語の誤りについて、過剰指摘を行なわずに効率的に自動訂正する手法について提案し、実証した。評価結果については原因分析（形態素解析の失敗、誤用候補展開の失敗、正解語の不足）を行ない改良を行なっていく予定である。また、正解文字列を誤り文字列に改悪してしまう場合についても考察を進める予定である。

なお、本研究への貴重なデータの提供及び助言を頂いた福岡克氏に感謝する。

参考文献

[1] 池原他：” 文書校正支援システムにおける自然言語処理”，情報処理 Vol.34 No.10(1993)

[2] 橋本他：” 気付き始めたワープロ 第３部 文書校正支援”，日経バイト 1995 年１月号

[3] 高木：” ワープロ入力における言語統計情報による誤り検出方式の検討”，情報処理学会第 43 回全国大会 6H-5(1992)