3Q−1 日本語テキストリーダにおける日本語校正支援機能

山中　紀子　田野崎　康雄　齋藤　裕美　小林　賢一郎*
(株)東芝　情報処理・機器技術研究所　*東芝AVE(株)

1．はじめに

ワープロ等の普及により電子化された文書が大量に作成できるようになった。また、これらの電子化された文書を活用するためのシステムも多く開発されつつある。しかしまだ、印刷された文書の情報量は膨大である。電子化された形での入手が困難な場合、印刷物を見ながらワープロ等で入力するのは、作業的に大変なことである。OCR技術を活用して印刷物から直接電子化された文書に変換する日本語テキストリーダは、このような要求に応えるものとして注目されている。

テキストリーダの最大の課題は、誤認識の解消である。各文字ごとに見た場合、その文字が誤認識でないかどうかの判断は認識の性能そのものに依存し、ある程度の誤認識や識別上の曖昧さは現状では避けられない。しかし、日本語テキストリーダで読み取る文章は日本語であり、認識結果を文字列として見た場合には、基本的に日本語の文章を構成する。

我々は日本語の形態的・構文的な特徴を解析して誤りを指摘する日本語校正支援機能を開発してきた[1][2]。この日本語校正支援機能を日本語テキストリーダに応用し、誤認識の解消および指摘に効果のある方式を開発した。本報告では、日本語テキストリーダにおける日本語校正支援機能の概要と簡単な評価結果について述べる。

2．日本語校正支援機能の概要

日本語校正支援機能は、日本語文を形態的に解析し、その情報などを用いて入力された日本語文の誤りを検出する機能である。誤りを検出して指摘する処理は、大きく分けて以下の3段階で構成されている。

(1) 形態素解析による単語単位の誤り指摘
(2) 文法知識による文単位の誤り指摘
(3) 表記規則による文章単位の誤り指摘

単語単位の誤り指摘では、入力された日本語文に対して形態素解析を行なう。形態素解析は日本語文を単語単位に分割する機能を持つ部分である。日本語の単語辞書を参照しながら、活用語尾変化、単語間の接続検定を行なう。その際に単語辞書に未登録の単語を検出し、送りがなや活用の誤りを指摘する。

文単位の誤り指摘では、簡単な日本語の文法知識を持ち、それを参照しながら文節間のつながりを調べる。この処理では、簡単な構文的な誤りを指摘することができる。

文章単位の誤り指摘では、かっこのバランスが取れているかなど、表記的な誤りを指摘する。

3．校正支援機能のテキストリーダへの応用

日本語テキストリーダの認識結果は、各文字ごとに複数の認識候補文字とその類似度を持っている。類似度とは認識された文字がどれくらい入力イメージに似ているか、すなわちどれくらいの確信度で認識しているかを示す値である。通常日本語テキストリーダは、第一候補の文字を認識結果として提供する。

図1に文字認識結果の一例を示す。1行が1文字に対する認識結果であり、最も左の列が第一候補となる。

日本語校正支援機能を日本語テキストリーダの認識結果に応用して、誤認識の解消および指摘を行なう方法は以下の通りである。

(1) 認識の第一候補よりなる文字列を日本語校正支援機能によって誤り検出する。
(2) 誤り指摘箇所のうち、類似度の近いものを第二候補と置き換える。
(3) 置き換えた文字列を再度日本語校正支援機能によって誤り検出する。

配	V01-1	酎	V01-2	醒	V01-3	酪	V01-4	醇	V01-5
置	V02-1	濆	V02-2	匿	V02-3	清	V02-4	資	V02-5
す	V03-1	ず	V03-2	才	V03-3	オ	V03-4	ォ	V03-5
る	V04-1	ろ	V04-2	乙	V04-3	ゐ	V04-4	不	V04-5
実	V05-1	夫	V05-2	尖	V05-3	美	V05-4	丈	V05-5
川	V06-1	用	V06-2	…	V06-3	巾	V06-4	州	V06-5
的	V07-1	釣	V07-2	曲	V07-3	約	V07-4	向	V07-5
な	V08-1	を	V08-2	々	V08-3				
浮	V09-1	汀	V09-2	汗	V09-3	浬	V09-4	淫	V09-5
上	V10-1	土	V10-2	+	V10-3	士	V10-4	十	V10-5
式	V11-1	武	V11-2	弐	V11-3	戎	V11-4	犬	V11-5

図1　日本語テキストリーダ認識結果例

Text Revision Support Method for Japanese Text Reader
Noriko YAMANAKA, Yasuo TANOSAKI, Hiroyoshi SAITO,
(Information Systems Engineering Lab., TOSHIBA Corp.)
Ken'ichiro KOBAYASHI
(TOSHIBA AVE Co., Ltd.)

(4) 誤りがなくなれば第一候補と第二候補を入れ替える。誤りのままならば第一候補のままでユーザに警告する。

日本語テキストリーダの修正機能では、候補文字の入れ替えが起こった部分および誤り指摘の警告部分が他の部分と色を変えて表示されるので、ユーザは色で指摘された部分を中心に確認すればよいことになる。

4．評価実験

本方式による誤認識の解消および指摘を評価するために、当社で開発した日本語テキストリーダを用いて評価実験を行なった。対象文書は東芝レビュー（技術文書）を中心に、新聞・雑誌・小説などを含む、約４０頁である。

評価の際の日本語校正支援機能が用いた基本語の辞書は１０万語であり、特別に専門用語辞書や固有名詞辞書は用意しなかった。

評価規準としては、誤認識部分の指摘率と置換の成功率に注目した。また逆に、正しい認識候補を誤って置き換えてしまう誤置換も評価規準に加えた。図２は、これらの論理的関係を表わしたものである。

実験の結果を表１に示す。ここでの各数値の単位は文字ではなく箇所である。

5．考察

ここでは漢字・かなの部分と数字・記号の部分に分けて評価した。これは、文字認識の特徴として数字・記号部分には似通った文字が多いだけでなく、通常の漢字・かな混じりの文章部分とは多少性格が異なるためである。

全体としては誤認識全体のうち９４％を指摘することができた。また、誤認識全体の２６％を、置換機能によって正しい候補文字と置き換えることができた。

但し、第一候補が正しいのに候補を置き換えてしまう副作用（誤置換）も少数ながら起こっている。これは主に固有名詞など、辞書に登録されていない単語について多く発生した。

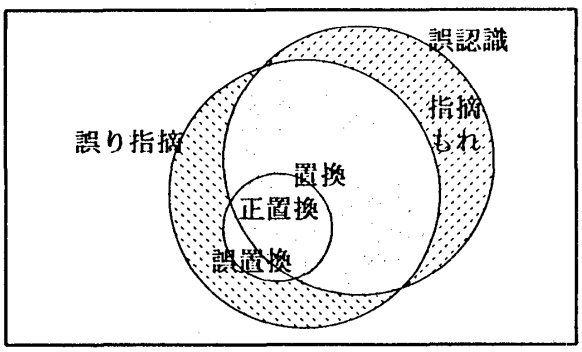

図２　誤り指摘と置換

6．あとがき

日本語テキストリーダに日本語校正支援機能を応用し、誤認識の解消および指摘を行なうための方式について説明した。

本方式は、誤認識ではあるが形が似ていて人間では誤りを見つけにくい文字の指摘、別の候補に正しいと思われる候補がある場合にはそれに置き換えることによる誤認識の解消に効果のある方式である。これにより、ユーザが誤認識を訂正する際の効果的な手助けとなる。

今後、機能をより高めていくためには、指摘もれの解消、正置換率の向上、誤置換の解消が大きな課題となる。

この機能のうち、誤り指摘部分は当社が商品化した日本語テキストリーダ「*ExpressReader 70J*」に搭載されている。

参考文献

[1] 小山他:「文章作成支援システムの機能について」　情報処理学会第35回全国大会4S-4

[2] 小山他:「文章校正支援機能における日本語解析」　情報処理学会自然言語処理研究会 88-NL-69-2

表１　評価結果

	漢字・かな	数字・記号	合　　計
誤認識	２１４	８７	３０１
指　摘	２１０（９８%）	７２（８３%）	２８２（９４%）
指摘もれ	４（　２%）	１５（１７%）	１９（　６%）
正置換	５８（２７%）	１９（２２%）	７７（２６%）
誤置換	１４	１１	２５