OCR文書の認識誤り修正支援システムの開発

6J－8

畑田 稔　野里真喜子　遠藤裕英
日立製作所 システム開発研究所

1. はじめに

ディジタル図書館の実現のために、紙でしか存在しないもののディジタル化手段として、OCR（Optical Character Recognition）の需要が高まっている。OCR結果には、文字認識誤りが多く、誤りの訂正には時間がかかるため、入力工数削減が望まれている。本研究では、技術情報サービスで対象としている文書の特徴をとらえて、訂正能力が優れ、操作性のよい認識誤り修正支援システムを開発した。

スペル誤りのチェックおよび自動訂正の研究は古く、英文で30年、和文で10年以上の歴史があるが、今日もなお研究が続いている[1],[2]。それぞれの手法は用途が限られ、また誤り訂正能力に制約があり、OCRなどでは、スペル誤りの訂正に要する工数は今なお、多大なためである。

従来のスペルチェック・訂正の研究では、1つの単語に含まれる文字誤りは少ないこと（せいぜい2、3個）が前提となっていることが多い。しかし、英字を含む和文OCRでは、長さ10文字の英単語に4、5文字のエラーが含まれていることが珍しくない。われわれは、先に、日本語OCR文において認識エラーの多い英字、カタカナを対象を対象としたスペル誤りの訂正方式を開発した[3]。ここで開発したスペルチェッカ（SpellChecker-1）は、誤りのある単語を一つずつダイアログボックスで確認して、修正するものであった。

本稿では、これをさらに発展させ、単語の切り出し精度の改善と、誤りを含む単語を一斉に表示して、正しい単語が第一候補にある場合、ワンタッチで修正できる認識誤り修正支援システム（SpellChecker-2）を提案する。

2. スペルチェッカの概要

2.1. ユーザー・インタフェース

SpellChecker-2の画面例を図1に示す。スペルチェックは表示されている全行について行われ、エラーを含む単語は赤字で表示され、その上の行間に第1候補が表示される（図1では、Computer、コンピュータ、Internetなど）。誤りがないと判定した（この綴りが辞書にあった）単語は青字で表示される（図1では、Network、ミニコン、トースターなど）。

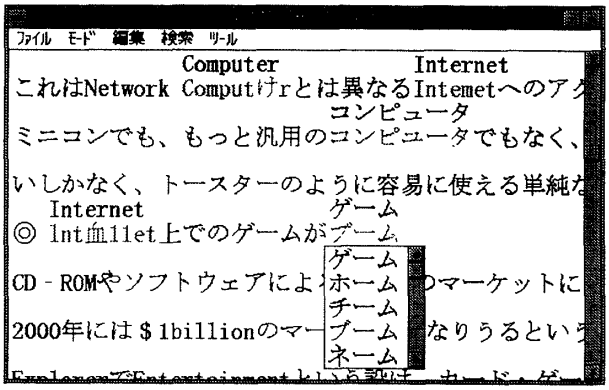

図1. スペルチェッカーの実行例

第1候補への訂正は行間の第1候補をクリックするだけでよい。例えば、図1で第1行の行間にある「Computer」をクリックすると、「Computけr」が「Computer」に置き換わる。

第1候補になければ次のようにする。マウスカーソルを誤りを含む単語の所へ移動すると、その

Development of the Spell Checker for Japanese OCR Text

Minoru Hatada, Makiko Nozato and Hirohide Endoh
Systems Development Laboratory, Hitachi, Ltd.

下にプルダウン式に第1候補から第5候補までが表示される。図1では、マウスポインタを「プーム」に置いた。第4候補の「ブーム」をダブルクリックすると、「プーム」が「ブーム」に置き換わる。

候補単語にない場合は、キー入力で修正する。

2.2. 認識誤り検出・訂正方式[3]

2つの単語間の類似性の評価尺度として、レーベンシュタイン距離を用いた。辞書のすべての単語に対して、レーベンシュタイン距離を計算するには、時間がかかるため、あらかじめ固定長のハッシュ値を求めておき、ハッシュ距離（≦レーベンシュタイン距離）で、候補単語を絞り込むことにより、大幅な時間短縮を図った。誤り率が40%のとき、ヒット率（正しい単語が第一候補）が73～93%、トップ5率（正しい単語が候補の5番目までにある）が89～98%と、実用上十分な訂正能力が得られた。

候補単語選択時間が応答時間の7、8割を占める。クロック180MHzのPentium Pro☆パソコンで実測した候補単語選択時間を表1に示す。

表1　候補単語選択時間

誤り数	候補単語選択時間（単位ms）
2	6～19
4	11～47
8	31～63

認識誤り検出および訂正候補選出は初めて画面に表示するとき行う。図1で一行下にスクロールしたときは、新たな一番下の行について、また、改ページしたときは、新たなページ全体についての誤り検出・訂正候補選出が行われる。
表1に示したように、候補単語の選択は迅速であるため、待ちを感じさせないスピードで画面表示が行われる。

☆ Pentium Proは米国Intel Corp.の商標である。

2.3. 単語の切り出し

SpellChecker-1では、字種の変化を単語切り出しの基本としている。「Info血ation」のように、英単語の中間で英字が漢字に化けたものは英単語として切り出せるが、英単語の先頭、末尾が漢字、ひらがな等に化けたものは検出できないため、予め、手作業で削除していた。

SpellChecker-2は単語の切り出しに、2文字、3文字が隣接して生じる文字の共起関係（これを2-gram、3-gram、一般にn-gramという）[4]を利用した。これによって英単語の切り出し精度が大幅に向上した。

今回、約1,000万字の文書データから共起頻度を求めた。この考え方を、英単語の切り出しだけではなく、漢字、ひらがな等の誤り検出に利用するには、さらに大規模な（少なくとも1億字）文書データが必要である。

3. おわりに

日本語OCR文において認識エラーの多い英字、カタカナに重点を置いたスペル誤り訂正方式を提案した。OCR結果を直接手作業で修正した場合に比較して、修正作業時間をほぼ半減できるという効果を得た。

今後の課題として、スペル誤り率が高いケースでの誤り訂正精度の向上、漢字、ひらがなの認識エラーへの対応などが挙げられる。

参考文献

[1] Webster, R.G., 中川：英語と日本語を対象にした文章誤り検出・訂正の共通点と相違点、情報処理、Vol.37、No.9、pp.865-871 (1996).
[2] Hall, P.A. and Dowling, G.R.: Approximate String Match, *Computing Surveys*, Vol.12, No.4, pp.381-402 (1980)
[3] 畑田、遠藤：日本語OCR文における英字・カタカナのスペル誤り訂正法、情報処理学会論文誌、Vol.38、No.7(1997)（*in press*）
[4] 長尾 真編：自然言語処理、岩波書店（1996）