テキスト・ベースかな漢字変換

4C−5

鳥原信一・野崎広志

日本アイ・ビー・エム(株)東京基礎研究所

1．はじめに

かな漢字変換用辞書に加えて、蓄積されたテキストにアクセスすることにより、(1)テキストに応じた単語選択(同音異義語選択も含む)することができる。また、(2)辞書に登録されていない未知語も変換することができる。本稿では、(1)字面による処理、(2)フルテキストへの高速アクセスによって、これらの実現を試みる。このテキスト・ベースかな漢字変換を用いれば、例えば、住所録テキストを指定することによる宛名入力、同一分野の論文テキストを指定することにより論文の作成が効率よく行うことができると思われる。

2．関連研究

テキスト・ベースかな漢字変換の研究を進めるに当たり、多くの示唆を与えてくれると思われる研究がある。1つは、実例に基づく機械翻訳であり、もう1つは、フルテキスト・データベース検索である。本稿では、独自の方式「字面による文字列マッチング」をとっている。本方式の利点を生かしながら、さらに高速なテキストへのアクセス、より正確な情報獲得ができるように上記の関連研究を注目して行きたい。

3．テキストの再利用

通常、かな漢字変換によってテキストを入力し、それを文章校正支援システムないしフォーマッタなどにより文章を整えて印刷し保存する。このテキストには、かな漢字変換に有益な情報が多く含まれている。かな漢字変換のためにテキストを再利用するには、(1)テキストから辞書を作成する方法、(2)テキストにアクセスする方法があると思われる。本稿では後者について述べることにする。

4．字面による同一文字種文字列を単位とする処理

あらかじめ形態素解析用辞書を用いて形態素解析をしたファイルにアクセスすることも考えられるが、本方式では、字面によって処理を行っている。また、同一文字種文字列を単位にするとほぼ形態素解析の単位に一致するという性質を利用する。これらの方式により、形態素解析用辞書の単語未登録による解析失敗および形態素解析の誤りから解放されて、テキストからゆがみのない情報が獲得でき、しかも高速に処理できる。図1を参照されたい。

図1　同一文字種文字列を単位とする解析
K=漢字、F=カタカナ、H=ひらがな、N=数字、S=記号

```
/第/9/条/ [/戦争/の/放棄/、/軍備及び/交戦権/の/否認/] /
 KK NN KK  SS KKKK HH KKKK SS  KKKKKK  HH KKKKKK HH KKKK  SS
/日本国民/は/、/正義/と/秩序/を/基調/とする/国際平和/を/誠実 /に/
 KKKKKKKK HH SS  KKKK HH KKKK  HH KKKK  HHHHHH  KKKKKKKK HH KKKK  HH
```

5．同一文字種文字列先頭文字によるインデックシング

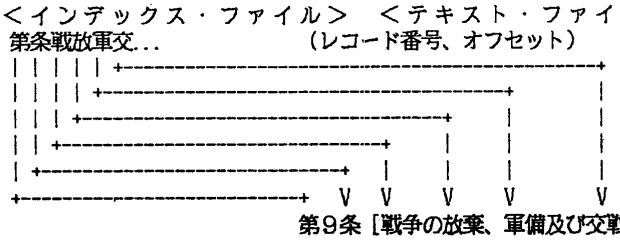
同一文字種文字列によってある程度形態素解析ができるので、同一文字種文字列の先頭文字でインデックシングすることにした。本方式では、半角カタカナ列、全角カタカナ列および漢字列をインデックシングした。ひらがな列は、自立語の一部または全部であったり、付属語であったりするので今後とも特別の考慮をする必要があると思われる。図2に同一文字種文字列先頭文字によるインデックシングを図示したので参照されたい。

Text-based kana-to-kanji conversion

Shinichi TORIHARA, Hiroshi NOZAKI

IBM Japan, Tokyo Research Laboratory

図２　同一文字種文字列先頭文字によるインデックシング



６．テキストから獲得できる情報とかな漢字変換の実際

かな漢字変換をする際に、テキストにある情報を獲得して変換率向上が図れると思われるものを次に示す。

６．１．カタカナ未知語

通常、かな漢字変換では、辞書にある程度カタカナ単語を登録しているがすべてカバーできないのでカタカナ未知語変換のアルゴリズムを有することが多い。それは、辞書引きの際、低頻度でダミー単語を作成しておいて、候補単語が存在しない場合にカタカナ単語を表示する方法などである。この時、テキストにアクセスしてダミーのカタカナ単語と一致すれば、その単語の優先度をあげることができる。すなわち、カタカナ未知語は、テキストにあれば、既知語になれるのである。

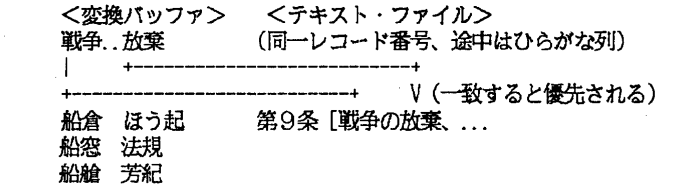
６．２．複合語

複合語は、辞書にある単語の文法・頻度・付加情報などにより、その組合わせで最小の点数をもつパス（文節＋文節）が選択される。意味はほとんど考慮されていないので、かな漢字変換による新造語（誤変換）が表示される。例えば、「けいたいそ」で、「形態素」を変換しようとした際、「系／対ソ」が表示されるような場合である。この時、文節（「形態」＋「素」）が存在し、テキストに、「形態素」が存在したら、文節（「形態」＋「素」）間の優先度をあげることによりユーザの期待した複合語を表示することができるようになる。

６．３．共起

２つの単語が共に生起する関係を用いると同音異義語選択ができる。シソーラス・コードおよび共起単語を辞書に納めて単語選択を行っているかな漢字変換は多い。しかしながら、すべての共起単語を辞書に納めることは不可能である。相補的に、辞書とともにテキストにアクセスして共起単語の選択をするのが本方式である。共起単語には、途中に付属語がない複合語と途中に付属語・修飾語などがあるものとがある。テキスト・ベースかな漢字変換における両者の違いは、前者はテキストを参考にして文節境界を調整するが、後者はしない。前者は漢字列であるが、後者は漢字列の間にひらがな列があるものである。このひらがな列を付属語・修飾語とみなしている。例えば、テキストに「胸がとても躍る」があれば、「がとても」は付属語・修飾語として「胸」「躍」を共起単語かどうかパス（文節＋文節）と比較することになる。テキスト・ベースかな漢字変換の共起単語選択について図３を参照されたい。

図３　共起単語選択　（「せんそうのほうきとは」）



７．まとめ

同一文字種文字列による形態素切り出しおよび字面による高速処理によって、テキストにアクセスしてカタカナ未知語、複合語、共起の変換をテキストに応じて行う方式について述べた。今後、インプリメントを進め変換率および変換スピードのデータをとり実用化に近づけたい。当面の課題は、ひらがな列の扱いである。形態素切り出しおよび文法の観点から研究を進めるつもりである。