1P−6

カタカナ異表記を考慮したユーザ辞書システムの拡張

相川 勇之，宮原 浩二，高山 泰博，鈴木 克志，丸山 冬樹

三菱電機(株) 情報電子研究所

1 はじめに

日本語における表記のゆれは、自然言語処理システムの開発において大きな負担となっている。中でもカタカナ語については、その種類、量ともに豊富である。複合語の一部となりやすいこと、日々新しい外来語がカタカナ語として日本語に取り入れられていることを考えると、カタカナ語の表記すべてを登録することは事実上不可能である[2]。

そこで我々は、現在開発中の日英機械翻訳システムに、カタカナ異表記変換アルゴリズムに基づいた処理を組み込み、その有効性を既に確認している[1]。しかし、従来の変換処理は、ユーザ辞書中の単語に対して適用することができなかった。そのため、ユーザ辞書にカタカナ語を登録する時には、異表記をすべて登録する必要があり、非常に効率が悪かった。我々は、翻訳システムのユーザ辞書中の単語にもカタカナ異表記変換を適用できるよう、ユーザ辞書システムを拡張し、翻訳実験によりその有効性を確認できたので報告する。また、カタカナ異表記を考慮したユーザ辞書編集が可能となるよう、ユーザ辞書エディタを改良したので、これについても報告する。

2 カタカナ異表記変換処理

かつて導入したカタカナ異表記変換処理は、ユーザ辞書に対して適用することができなかった。これは、入力文中のカタカナ表記が、規則により正規化され、正規化された表記をキーとして辞書検索を行なうため、ユーザが自由に編集できるユーザ辞書では、正規化表記の衝突の恐れがあったためである(図1参照)。

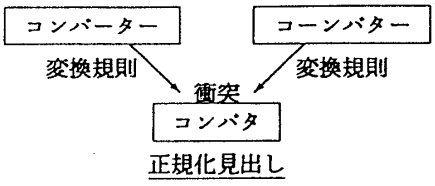

図1: 正規化見出しの衝突

しかし、当社内で試験的に使用している部門のユーザ辞書について、カタカナを含む見出しを調査したところ、このような衝突は存在していなかった。また、ユーザ辞書に登録されるカタカナ語は日本語を含んだ複合語である場合が多く、上記のような衝突が起こる可能性は非常に低い。仮に衝突が起こったとしても、ユーザ辞書エディタの改良によって、ユーザの判断により異表記変換処理の適用を禁止できる。また、カタカナ語の表記のゆれを解消する手段として、原語の辞書を利用する方法が提案されている[2]が、上記ユーザ辞書の調査によると、原語とカタカナ語の訳語とは必ずしも一致していない。そこで、我々はユーザ辞書中のカタカナ語見出しに異表記変換処理を適用した文字列をキーとして、ユーザ辞書を検索することにした。

3 ユーザ辞書システムの拡張

従来のカタカナ異表記処理は、まず正規化された入力文字列をキーとして、正表記の検索を行ない、得られた正表記で辞書内容を検索するものだった。ここで、正規化文字列から正表記を得る検索テーブルは、システムが提供する辞書から静的に作成されたものであるため、ユーザ辞書に登録された単語には対応していない。

ユーザ辞書中の単語にカタカナ異表記処理を適用するために、最初にユーザ辞書に登録された単語を正表記として、検索テーブルを更新する方法も考えられる。しかし、検索テーブルを更新可能な構造にすると、静的に生成する場合と比較して辞書検索速度の低下を招く恐れがある。

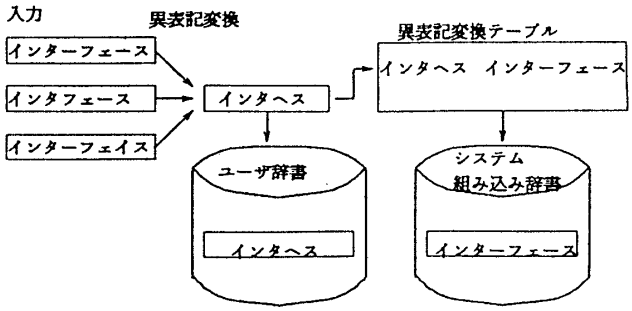

図2: ユーザ辞書からのカタカナ異表記検索

我々は、ユーザ辞書に登録されたすべてのカタカナ語に対して、異表記変換処理を行ない、正規化された見出しを検索キーとして再登録した。さらに、カタカナ語の辞書検索時に、これらの単語が引かれるように辞書シ

An Extension of User Dictionary System in consideration of *katakana* variant notations

Takeyuki AIKAWA, Koji MIYAHARA, Yasuhiro TAKAYAMA, Katsushi SUZUKI, Fuyuki MARUYAMA

Mitsubishi Electric Corp.

ステムを改良した(図2参照)。この方法によれば、検索テーブルを更新する必要はない。従来のユーザ辞書システムをそのまま利用して、単語の登録、削除が可能である。

さらに再登録の際に、元の見出しも辞書情報として格納した。異表記が複数の見出しとして登録されている場合は手作業で辞書内容をマージし、それぞれの元の見出しをリストとして辞書情報に取り込んだ。この元の見出しの情報は後述のユーザ辞書エディタにおいて使用される。

4 翻訳実験による評価

ユーザ辞書システムは、動的な編集を可能とするため、システム組み込みの辞書に比べて検索速度が劣っている。今回の拡張により、ユーザ辞書に対するアクセス回数が増すため、翻訳速度が低下することが予想される。

テスト用の例文により、翻訳速度の比較を行なった。実験結果を表1に示す。数値は従来版の処理時間を1.0としたときの拡張版の処理速度である。

表 1: 翻訳比較実験結果

	実時間比	CPU消費時間比
全文 (796文)	1.07	1.06
カタカナを含む文 (557文)	1.06	1.08
カタカナを含まない文 (239文)	1.01	1.01

カタカナを含む文については6〜7%の速度低下となる。しかし、カタカナ異表記処理がユーザ辞書中の単語にも適用されることにより未知語が減少し、処理できる文が大幅に増加すると考えられるので、実用性は十分にある。

カタカナを含む文557文中28文で、翻訳結果に以下の変化が見られた。

- 訳語が変化する語があった。従来版では、別見出しとして登録された異表記同士で訳語が一致していないものが存在したため訳語の変化が生じた。拡張版では各異表記を一つの正規化見出しでまとめて管理できるため、整合性の維持が容易となる。
- 従来、複合名詞として解析されていた部分が、異表記として検索されることにより、正しい訳語が得られた。

実験に使用した文章中の単語のうち、大部分が既にユーザ辞書に登録されていたため、ユーザ辞書システムの拡張による翻訳結果の変化は少なかった。しかし、より大きな範囲の文書を対象とすれば、ユーザ辞書中のカタカナ語に対して異表記処理が適用できることによる翻訳精度の向上は、本実験より大きなものになる。

5 ユーザ辞書エディタの改良

ユーザ辞書中に含まれるカタカナ語はすべて、正規化見出しに変換されている。正規化見出しは、システムに組み込まれた規則に従って変換された文字列であるため、元の日本語とはかけはなれたものになる場合がある(例:インターフェース ⇒ インタヘス)。そこで、ユーザ辞書エディタを改良して、ユーザには元の見出しのみ見せ、正規化見出しを隠すようにした。

また、元の見出しの情報を、異表記のリストとしてユーザ辞書情報に含んでおくことにより、ユーザ辞書にカタカナ語を登録をするとき、既登録語と正規化見出しが衝突しているかどうかの検査が可能となる。判定は以下のアルゴリズムに従う。

(1) 登録語の正規化見出しが、既に登録されているかどうか調べる。存在すれば(2)へ。存在しなければ(3)へ。

(2) 既登録語の元の見出しのリストと、新規登録語の元の見出しのリストが包含関係にあれば(3)へ。それ以外の場合は(4)へ。

(3) 新規登録語に対して異表記変換を行ない正規化見出しで登録する。

(4) 新規登録語は、既登録語と正規化見出しが一致している。新規登録語が既登録語の異表記であるかどうか、ユーザに入力を要求する。異表記である場合は、(3)へ。異表記でない場合は(5)へ。

(5) 新規登録語は、異表記変換を適用せず、元の見出しで登録する。

このアルゴリズムに従うと、正規化見出しの衝突が生じた場合、既登録語に対しては異表記変換処理が適用されるが、新規登録語については適用されなくなる。しかし、このような場合は非常に稀であると思われること、従来版の解析結果に悪影響はないことを考え、本方式を採用した。

6 おわりに

カタカナ語に関する表記のゆれを吸収する、異表記変換処理が適用可能なユーザ辞書システムについて報告した。実際に、翻訳システムに組み込んで翻訳結果の比較を行ない、その有効性を確認した。さらに、正規化表記の衝突を考慮して、拡張版ユーザ辞書に対応したユーザ辞書エディタの改良を行なった。

参考文献

[1] 伍井、清原、鈴木、太細:カタカナ異表記処理, 情報処理学会第38回全国大会 4E-2, 1989

[2] 野美山:カタカナ外来語の表記の揺れの解消, 情報処理学会第41回全国大会 6S-3, 1990