カタカナ異表記処理

4E-2

伍井啓恭、清原良三、鈴木克志、太細孝

三菱電機㈱ 情報電子研究所

1．はじめに

日本語は、さまざまなキャラクタ（漢字、ひらがな、カタカナ、数字、アルファベット、記号）を吸収することの可能な強力な言語であるが、一方これは機械による自然言語処理にとっては大きな負担となりうる。我々はこのなかでも大きな要因の一つである日本語のカタカナ表記のゆれに対応する処理を当社のＡＩワークステーションＭＥＬＣＯＭ—ＰＳＩⅡ上の日英機械翻訳システムＭＥＬＴＲＡＮ—Ｊ／Ｅ（翻訳速度最大１万語毎時、辞書登録語数約８万語、以下ＭＥＬＴＲＡＮと呼ぶ）に実装実験したので報告する。

2．異表記が及ぼす日本語処理への影響

日本語の処理における解析の失敗の原因として第一に未知語が挙げられる。未知語となるものの中には日本語における表記のゆれ（以下異表記と呼ぶ）によるものが多い〔9〕。日本語処理のなかでも漢字かな変換、文章校正支援、日英機械翻訳などにおいては、異表記処理は特に重要な問題である〔6、8、5〕。異表記には例えば以下のものがある。

カタカナ異表記
漢字異表記
送り仮名異表記
記号異表記
混じり書き異表記

この中で出現頻度の高いものの一つにカタカナ異表記が挙げられる〔5〕。

カタカナは、現在日本語文章中に５～２５％の割合で定着しており〔5〕、特に専門分野において顕著である。カタカナ表記の多くは外来語であり、原語の発音や表記などの日本語にないものをむりやりカタカナ表記にあてはめてしまうために一つの言葉に対して多数の表記が許容されてしまうといったことが起る。例えば

インタフェース
インターフェース
インターフェイス

などである。

これらの表記と日本語処理システムの辞書における表記とが一致していない場合、それは未知語となってしまう。

3．処理方式の検討

表記のゆれを吸収するには大別して次の２通りの方式が考えられる。

① 辞書に全ての異表記を登録

これは最も確実な方式であるといえる。このように辞書を構築している例もある〔10〕。また漢字異表記などは辞書に登録しなければ対応できないであろう。しかし、全ての表記を自然言語処理システムにおいて実際に計算機上にインプリメントしようとした場合そのメモリ効率において問題がある。

② アルゴリズムによる対応

これはアルゴリズムにより表記の曖昧性を吸収する方式である。日英機械翻訳システムにおいてはカタカナ表記から英語の表記を推定し直接変換することにより未知語を減らす実験がなされている〔4、7〕。

しかし機械翻訳にしか使用出来ない点や、必ずしも原語が翻訳目的言語でないこと、省略されたカタカナ表記や和製外来語は推定が難かしい点などが問題点として挙げられる。

4．アルゴリズムの設計

我々は、カタカナ表記自身のもつ曖昧性をアルゴリズミックに吸収することにより辞書検索をする方式によって実験を試みた。

まず、カタカナに対応する音を示す表をもとに各カタカナの異表記となりうる表記を全て算出した。

カ ⇨ カ,クァ,キァ,コァ,キア,クア,コア,キウア
キ ⇨ キ,クィ,コィ,コ,ク,コイ,クイ
ク ⇨ ク,クゥ,キ,キゥ,コゥ,コ,キ,コウ,キウ

この中より実際に異表記として使用される可能性の高いものを抽出し、さらに参考文献〔2、3〕、及び実際の辞書見出しを参照し、カタカナ表記において異表記になりうるルール（約５７）をピックアップした。例を示すと、

☞クに各拗音を付加したものはカ列の各文字になる。
シーケンス，シークェンス

☞イ列、またはエ列にヤが付加された場合ヤはアに置き換えられる。
ダイヤモンド，ダイアモンド

Processing of katakana variant notations

Hiroyasu Itsui, Ryozo Kiyohara, Katsushi Suzuki, Takashi Dasai
MITSUBISHI Electric Corp.

☞促音にクスが付加された場合はその促音を省略できる。

インデックス，インデクス

☞ユームはウムになる。

アルミニューム，アルミニウム

☞長音記号は省略可能である。ただし意味を持った短い文字列単位内では省略出来ない。

インターフェース，インタフェース

これらをアルゴリズム化し異表記変換処理を構成した。

5．異表記変換処理の試験

異表記変換を処理する際に問題となるのは表記の衝突である。我々は、この点を試験するため次の作業を実施した。

ＭＥＬＴＲＡＮにおける辞書のうちの１バージョン（約９万語）においてカタカナを含む語（約３．３万語）を抽出し、さらに同一表記の重複を削除した見出しを抽出した。

これを異表記変換処理にかけ、その出力結果より衝突して重複したものを分析した。

衝突の起きたものは約８００組（１組２語以上）で、そのうち真の衝突（意味が異なる語であるのに衝突が起る）するものが７１組であった。例を示すと、

テラ，テラー，テーラー

ソール，ソウル

ホーム，フォーム

ドラマ，ドラマー

スプリング，スプーリング

などであり、その殆どが語長の比較的短いものであった。

さらに、これらの衝突を解消するためカタカナ異表記変換処理に、これらの見出しが発見された場合処理をパスするように設定した。このように改修し再試験を行ったところ表記の衝突は０組になった。

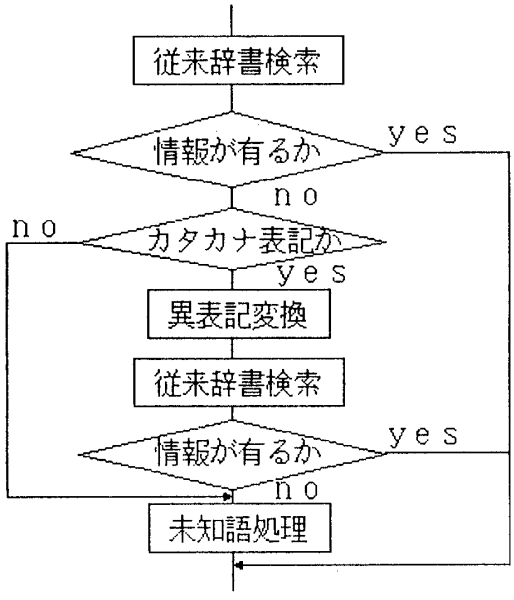

図　1

6．辞書検索部分への導入

このカタカナ異表記変換処理をＭＥＬＴＲＡＮの辞書検索部分へ導入した。これを図１に示す。手順は、まず従来の辞書検索を実行し、これにより辞書情報が検索できた場合は情報を返し終了する。情報が無かった場合にこの文字列が変換処理によって変換される可能性のあるものかチェックし、変換後に変化しなければ未知語処理を実行する。変化するものは変換処理を通し、従来の辞書検索を実行する。これにより辞書情報が検索できなければ未知語処理を実行する。

ここで従来の辞書検索とカタカナ異表記変換処理後の辞書検索との２段の処理に分けた理由は従来辞書にユーザ辞書が含まれており、ユーザが自由に編集できるためユーザ定義の見出しの衝突チェックができないことによる。

7．おわりに

このカタカナ異表記処理を実装することにより今まで未知語となっていたカタカナ表記を検索することが可能となった。この処理の実装後も未知語となるものについては元来辞書に情報が存在していないものが主であった。この処理を導入することにより数万語の見出しを登録するのと同様の効果が得られると思われる。

今後は、アルゴリズム化が困難なもの（曖昧性の比較的大きなもの）にどのように対処するかが課題となる。

参考文献

〔1〕外国人のための日本語　例文・問題シリーズ１１　表記法（荒竹出版）

〔2〕日本語発音アクセント辞典　日本放送協会編

〔3〕第２０回　国語審議会総会術語表記合同部会報告　外来語の表記について　(1954/3/15)

〔4〕大深：原表記とカナ表記の対応判定アルゴリズム　情報処理学会第３７回全国大会　(1988)

〔5〕黒田他：日本語文におけるカタカナ英語の研究　自然言語処理68-3　(1988/9/16)

〔6〕梅田他：漢字カナ変換の一方式　情報処理学会第３２回全国大会　(1986)

〔7〕中村他：カタカナ表記外国語の直接検索　情報処理学会第３５回全国大会　(1987)

〔8〕小林他：文章作成支援システムにおける日本語処理(1)　情報処理学会第３５回全国大会　(1987)

〔9〕亀田他：未知語の分類とその処理規則　情報処理学会第３６回全国大会　(1988)

〔10〕清水他：辞書検証システムの構想　情報処理学会第３６回全国大会　(1988)