日本語学習者のためのカタカナ語理解支援システムの検討

諏訪いずみ† 高橋勇‡ 黒岩丈介‡ 小高知宏‡ 小倉久和‡

† 福井大学大学院工学研究科 ‡ 福井大学工学部

1. はじめに

日本語を母語としない者にとって，カタカナ語の意味を理解するのは困難が伴う．その一因として，カタカナ語の発音が元になった単語の発音と異なっていることがあげられる．一方，専門用語等でのカタカナ語の使用頻度は高く，日常的にも，カタカナ語が使用される機会が多くなっており，外来語辞書にないものも増えている．

現在使用されているカタカナ語の約 80%は英語起源であるといわれることから，カタカナ語から元の英単語を検索するシステムは，カタカナ語理解を支援すると思われる．そこで，カタカナ語理解の支援を目的にとしたシステムの基本部分として，ローマ字表記からカタカナ語の基になった英単語を検索するシステムを試作し，評価をおこなった．

2. ローマ字表記による検索

従来の方法では，検索にカタカナ表記を使用したものが多い[1],[2]．カタカナ表記は日本人にとっては馴染みやすいものであるが，次に述べるような理由により，入力にローマ字表記を採用した．

ローマ字以外の文字で書かれたものをローマ字で表記することは国際的な理解のために一般に行われており，日本語を母語としないものには，カタカナでの入力よりも，馴染みやすいとおもわれる．さらに，日本語を入力する場合ローマ字入力が標準である場合が増えている．

また，ローマ字表記の特徴として，子音と母音が明示的に表記されるということがある．これにより，子音と母音を分けて処理をすることができ，二重母音の処理等の変換を効率よく行うことができる．その結果，カタカナ表記から直接検索する場合よりも，変換のための規則数が少なく済み，効率よく検索をすることができる．

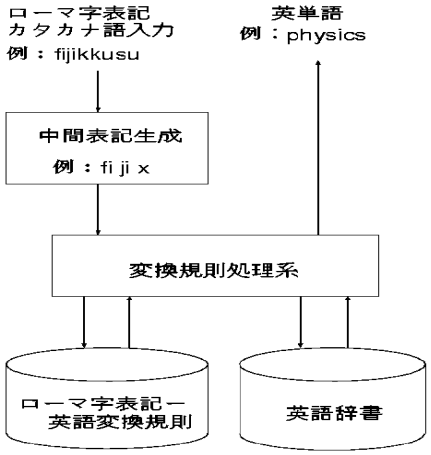

図1 システムの構成

3. システムの構成

本システムの構成を図 1 に示す．入力されたローマ字表記は二重母音，促音，二音音節をひとまとまりとする綴の処理等を行い，ローマ字表記‐英語変換規則を適応するための中間的表記に変換する．ローマ字表記-英語変換規則は，ほぼ日本語の一音節に対応するローマ字表記とそのローマ字表記に対応する英語の文字列が対になったものである．規則表に記述された基本的な規則数は 277 である．この中には，ヘボン式ローマ字表記と訓令式表記に関する変換規則，それぞれの長音表記の変換規則，二重母音などに関する特殊な変換規則が含まれる．

中間的に生成されたローマ字表記に対して先頭から区切りごとにローマ字表記‐英語変換規則を順次適応し，候補となる英単語を英語辞書から検索し，絞ってゆく．完全一致するものがなかった場合，検索に失敗する直前に残ってい

A Support System of Understanding Katakana Loan Words for Learners of Japanese
Izumi Suwa †
Isamu Takahashi ‡
Jousuke Kuroiwa ‡
Tomohiro Odaka ‡
Hisakazu Ogura ‡
† Graduate School of Engineering, Fukui University
‡ Faculty of Engineering, Fukui University

た英単語を候補として出力する．図 2 に “fijikkusu (フィジックス)” の中間表記 “fi ji x” に対する検索アルゴリズムの適用例を示す．

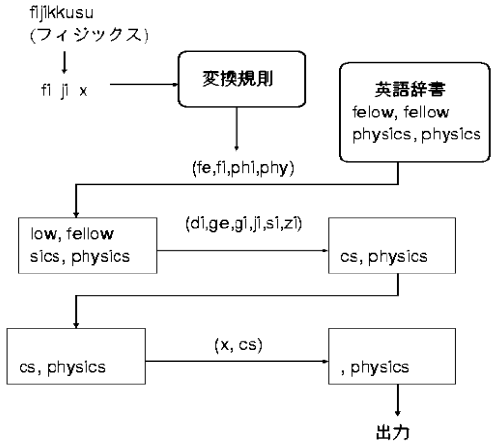

図 2　検索アルゴリズム

4.　評　価

和製英語，短縮語，空白やハイフンで区切られた複合語，固有名詞を除いた英語起源のカタカナ語 7002 語について検索を行った．使用したカタカナ語は，フリーの和英辞書 EDICT に含まれる外来語 12233 語の中から，前記条件を充たす語を抽出したものである．結果の各語数と全体に対する割合を表 1 に示す．

全語数	7002 語	-
正しい候補のみ	3919 語	56.0%
正しい候補を含む	1348 語	19.3%
正しい候補の一部	209 語	3.0%
不適な候補	1526 語	21.7%

表 1　評価結果

正しい候補の一部とは，検索語の単数形や派生語などをいう．不適な候補とは，明らかに間違った候補や，検索が失敗した時点で 20 以上候補が残った場合である．

75.3% の単語について正しい候補がえられた．また，3.0% については，正しい候補ではないが，検索語の単数形や派生語など，候補を類推できるような結果が得られた．

さらに，日本語を母語としない人の評価を得るため，研究室に在籍する中国人留学生に使用してもらった．評価としては，カタカナでの入力よりもローマ字での入力のほうが，使いやすいということであった．これは，日本語の読みを学習する際にローマ字表記を用いるからだそうである．

中国語を母語とする人の場合，判別が難しい音として，促音（例：“ハット” か “ハト” か）がある．促音については，このシステムでは促音なしでも正しい単語を候補としてあげるので，使いやすいという評価を受けた．複数の候補については，各候補に順位づけや使用可能性を表示してもらえるとわかりやすいとのことであった．

5.　考　察

入力表記にローマ字表記を用いた簡潔なシステムで 75.3% の単語について正しい候補が得られた．正しい候補を含むに分類されたものには，短いカタカナ語が多い．これは，外来語辞書にあるものと同時に，類似の発音をする他の単語が出力されることが多いからである．不適な候補の中には，英語の中の外来語，空白やハイフンを含まない複合語が多く含まれる．

本システムの手法は，カタカナ語に対応する英単語がテキストとして存在すれば，高い確率で候補を得ることができる．従って，辞書に登録されていない専門用語などについては，辞書の代わりに対応分野の英論文テキストを直接検索し，候補を得るようにすることが可能である．

現在のところは，候補となる単語の表示のみであるが，英英辞書や英中辞書などとリンクすることによる使用者の母語での意味表示も検討している．

参考文献

[1] 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，3-191，pp.191-192，1990.

[2] 宮内，"カタカナ表記からの英単語検索システムの実現"，情報処理学会・自然言語処理研究報告，93-NL-97，pp.119-126，1993.

[3] 諏訪，西野，小高，小倉，"日本語学習者のためのローマ字表記 に基づいた片仮名語からの英単語検索の試み"，電子情報通信学会論文誌，Vol. J85-D-I，No. 9，pp.927-930，2002.