英語辞書と英文法を用いたカタカナ表記語の翻訳

4L－7

松尾 義博　畑山 満美子†　池原 悟‡
NTTコミュニケーション科学研究所
北海道大学工学部†　鳥取大学工学部‡

1 はじめに

日本語の文書にはカタカナ表記の外来語が頻出し、これらの外来語をどうやって取り扱うかは、機械翻訳を行なう際の大きな問題である[1]。

日英機械翻訳システムでは、一般に、日本語辞書にこれらのカタカナ外来語を登録する仕組みを取っている。しかし、商品名などで新語が次々に出現してくるため、すべてを登録することは困難である。また、日本語訳が固まるまでの間、暫定的に用いられるカタカナ外来語も多いことを考えると、辞書作成のコストやその後の辞書保守のコストの面からも、すべてを日本語辞書に登録することは現実的ではない。

このカタカナ外来語の問題は機械翻訳に限らず、日本語解析全般で生じる問題である。そのため、日本語辞書未登録のカタカナ表記語を解析するために、カタカナからの英語表記の自動推定[2]や英語表記からカタカナへの自動変換[3][4]などが提案されている。しかし、自動変換の正答率は80%程度に留まっており、これらの手法を機械翻訳に適用するには、精度に問題がある。

本稿では、これらの問題を解決するために、カタカナ表記の英単語辞書と英語文法を日英機械翻訳に適用する手法を提案し、その効果を示す。

2 カタカナ英語辞書

カタカナ表記語を網羅的に日本語辞書に登録することが困難なのは、カタカナ表記語は新語が多くて総数が無数であることに起因する。これは、日本語の複合語をすべて辞書登録できないのと同じ理由である。

したがって、網羅的に収集するためには、日本語複合語と同様に単語の単位で辞書登録し、英熟語(＝カタカナ表記複合語)については、プログラム処理で合成することとすればよい。登録語を単語に限定することにより、英和辞典や英英辞典をベースに作成することができ、網羅的に収集できると考えられる。

また、日本語文書に通常出現するカタカナ表記英単語の総数は、固有名詞を除けば、英和辞典の語数程度と考えられるため、人手による作成も十分可能である。

単語単位の辞書構成で、一般の日本語複合語を扱う場合には、単語間の意味関係の解析や、英訳した時の名詞句構造の選択が問題となる。しかし、カタカナ外来語の場合、熟語になって意味が変わったり、特殊な意味の専門用語であったりしても、次節に述べるように、翻訳にはあまり影響がないと考えられる。

3 カタカナ英語辞書の日英機械翻訳への適用

次に、前節のカタカナ英語辞書を日英機械翻訳へ適用する場合の問題点について考える。

3.1 カタカナ部分の翻訳

カタカナ部分は外来語とは限らず、日本語や和製英語なども含まれる。そのため、まず日本語辞書による解析を試み、次に外来語としてカタカナ英語辞書で解析する(図1)。

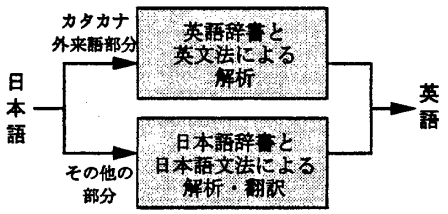

図1: 英語辞書と英文法によるカタカナ語の解析・翻訳

これらのカタカナ語は、外来語(ほとんどが英語)の単語・熟語の発音をそのまま表記したものであり、英語の構造・語順を持っている場合がほとんどを占める。したがって、普通の日本語の場合のように構文解析や意味解析を行なって英語側の適切な構造を選択する必要はなく、カタカナ英語辞書による形態素解析結果をそのまま英単語におきかえればよいと考えられる。

3.2 同音語の解消

カタカナ単語から英単語への置換えで問題になるのは、同音語[1]の問題である。例えば「エアポートイン」というカタカナ複合語を翻訳する場合、"airport in"と"airport inn"などの候補が考えられる。

Translation of 'katakana' words using an English dictionary and grammar

Yoshihiro Matsuo, Mamiko Hatayama† and Satoru Ikehara‡
NTT Communication Science Laboratories
Hokkaido University†, Tottori University‡

[1] ここでは、カタカナ表記が同じであれば、発音が異なっていても同音語と呼ぶ

これらの同音語の解消のために、本手法では、カタカナ語の英文法による解析を提案する。例えば、

```
S -> NP          PP -> P NP
NP -> N NP       N -> airport
NP -> N          N -> inn
NP -> NP PP      P -> in
```

の文法で解析すると、“airport in” は解析に失敗し、“airport inn” は解析に成功するので、曖昧性を解消できる。英語として解析すれば、意味分類による制約などを導入することも容易である。

3.3 辞書の構成・構築

上記の英文法による解析を実現するには、カタカナ英語辞書は英語の品詞体系をもとに構築する必要があり、日本語辞書とは別体系のものとなる。

前節で述べたように、辞書の必要語数は人手構築可能な範囲である。英語解析に用いる英語電子辞書をベースにすれば、追加すべき情報は、各単語のカタカナ表記のみである。カタカナ表記の揺れを吸収するため、辞書には考え得るすべてのカタカナ表記を記載することとする。

4 効果

以上述べた手法の有効性を確かめるため、日本経済新聞3日分の記事中のカタカナ表記語について机上検討を行なった。調査対象は、5983文332210文字である。同記事中のカタカナ表記語は表1の通りである。

	語数
(1) カタカナ表記語の総数	5816語
(2) (1) から日本語を除いたもの	5204語
(3) (2) の ALT-J/E 正解数	4148語 (79.7%)
(4) (2) の ALT-J/E 誤訳数	1056語 (20.3%)

表 1: 調査対象のカタカナ語

記事には平均して1文に1語のカタカナ表記語が含まれていた (表1-(1))。また、カタカナ表記語の約90%は外来語であった (表1-(2))。

調査では、記事中のカタカナ表記語から、本来日本語であるもの (「1カ月」の“カ”や、“ジリジリ”など) を除いたもの (表1-(2)) を、NTTの日英機械翻訳システムALT-J/E[5]で翻訳させ、誤訳となった1056語 (表1-(4)) を対象とした。誤訳のほとんどは未知語で、カタカナ表記がそのまま出力されたものであった。

机上検討では、研究社新英和中辞典第5版を用いた。同辞書を元にカタカナ辞書を作成したと想定して翻訳正解率を求めた。結果は表2の通りである。検討では同音語の曖昧性解消は考慮に入れず、同音語が存在して誤訳となる恐れがあるものについては、正解とは別分類 (表2-(4)) にした。

	語数
(0) 総語数	5204語 (100%)
(1) ALT-J/E 正解数	4148語
(2) カタカナ辞書で正解訳	675語
(3) (1)+(2)	4823語 (92.7%)
(4) 同音語が解消できれば正解	44語 (0.8%)
(5) 訳せないもの	337語 (6.5%)

表 2: カタカナ語正解数

検討によると、本方式によるカタカナ翻訳方式を日英機械翻訳システムALT-J/Eに適用した場合、カタカナ外来語のうち約93%が正しく翻訳されると期待できる。

訳せなかった337語の内、313語は人名等の固有名詞で、その他は英語以外の単語であった。

5 課題

前節で述べたように、翻訳失敗のほとんどは固有名詞であった。固有名詞をどうやって収集するかが次の課題である。

調査対象中には英語以外の単語も含まれていた。英語以外のカタカナ辞書を構築することは可能であるが、出現頻度が低いことや、複数言語となると語数が多いこと、読みを付与できる作業者が限られることなどから、英語辞書の構築に比べコスト高となることが考えられ、より効率的な構築手法が課題となる。

6 おわりに

日英機械翻訳でカタカナ表記外来語を翻訳するために、カタカナ読みを付与した英語辞書を導入し、英文法を用いてカタカナ部分を解析することを提案した。本方式によるカタカナ翻訳方式を日英機械翻訳システムALT-J/Eに適用した場合、カタカナ外来語のうち約93%が正しく翻訳できると期待できる。

現在、5万語程度のカタカナ語辞書の構築を進めており、さらに、固有名詞辞書の収集を進める予定である。

参考文献

[1] 黒田, 松永, “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)

[2] 野美山, “カタカナ外来語の表記の揺れの解消”, 41回情処全大,3-191(1990)

[3] 宮内, “カタカナ表記からの英単語検索システムの実現”, 自然言語処理,97-17(1993)

[4] 堀内, 山崎, “英単語のアルファベット表記から仮名表記への変換”, 自然言語処理,79-1(1990)

[5] Ikehara, S., “Multi-level Machine Translation Method”, *Future Computing Systems*, Vol.2, No.3, (1989)