5 J − 1

カタカナ表記述語の日英機械翻訳

松尾 義博　　小林 正裕 †　　白井 諭
NTT コミュニケーション科学研究所
電気通信大学 †

1 はじめに

日英機械翻訳において、カタカナ表記の外来語をどうやって取り扱うかは大きな問題である。カタカナ外来語の問題点は、新語が多いために辞書登録が事実上不可能なことである。そのため、カタカナ表記からの英語表記の自動推定 [1][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。また、英語辞書と英語解析を用いることにより、高い精度で英語へ翻訳する手法 [5] も提案されている。

しかし、これらの手法では、カタカナ語そのものを英語にすることを目的にしており、翻訳されたカタカナ語を如何に全体訳文中に合成するかについてはあまり考慮されていなかった。カタカナ語が名詞であれば、そのまま英語名詞句として埋め込むことにより、ある程度の訳が得られることが多いが、述語として使われた場合には、何らかの構造変換が必要となる。

本稿では、カタカナ表記述語の構造を分析し、英語への変換ルールを検討する。また、日英機械翻訳への適用について報告する。

2 カタカナ表記述語の特徴

実際にカタカナ述語がどのような構造を持っているかを分類するために、日本経済新聞 90 日分 (147,907 文) 中のカタカナ表記語について調査を行なった。結果を表 1に示す。調査の結果、カタカナ語の総数は 144,309 で、約 1 文に 1 つのカタカナ語が含まれている。カタカナ述語 (動詞と形容動詞) の総数は、2,091 で全体の約 1.5% に相当する。その他の大部分は名詞である。

日本語も一部カタカナ表記されている (例: ヤケドする) が、以下の調査では、外来語のもの (動詞 164 種、形容動詞 88 種) を対象にする。外来語のうち、出現頻度の上位語を表 2に示す。調査では、学研電子英和辞書 (語数: 約 5 万語) を用い、同辞書の品詞分類に従って調査した。

分類は、カタカナ語を英語並びとして見た場合の品詞列によって行ない、I〜VI の 6 分類を得た (表 3)。

表 1: 新聞中のカタカナ述語数

	語数	異なり
カタカナ＋する	1,758	204
(日本語)	350	40
(外来語)	1,408	164
カタカナ＋だ	333	110
(日本語)	58	22
(外来語)	275	88
カタカナ述語の合計	2,091	314
カタカナ語総数	144,309	15,155

表 2: 頻度上位語

	出現回数
スタートする	280
チェックする	80
ストップする	69
リードする	68
カバーする	66
ユニークだ	49
エスカレートする	34

2.1 タイプ I: 単純動詞 / 形容詞

「キャッチする」や「ダーティーだ」のように対応する英語が単語であり、品詞が動詞または形容詞のものである。調査対象の 252 語のうち 187 語 (74%) が、この種であった。

2.2 タイプ II: ing 型

「シールディングする」や「コーティングする」などのように、カタカナ表記が「-i ング」であり、英語の原型が動詞であるものである。コーパス中に 10 語出現した。

2.3 タイプ III: 名詞

「スピーチする」や「シミュレーションする」などのように、名詞に「する」がついたものである。そのままでは英語にすることはできず、"make a speech" のように何らかの動詞を付加するか、"simulate" のように動詞派生語を用いる必要がある。

Japanese-to-English Translation of 'katakana' Predicates
Yoshihiro Matsuo, Masahiro Kobayashi† and Satoshi Shirai
NTT Communication Science Laboratories
University of Electro-Communications†

2.4 タイプIV: 動詞＋副詞

「フォローアップする」や「チェックインする」などのように、動詞と副詞の連続である。英語にするには、"I follow it up" のように群動詞として扱う方法が考えられる。また "follow-up" や "check-in" は一語の名詞としても辞書記載されているため、タイプIIIと同様に何らかの動詞を補う方法も考えられる。

2.5 タイプV: 名詞＋動詞

「バトンタッチする」「モデルチェンジする」などのように、名詞と動詞が連続したものである。前の名詞が後ろの動詞の格要素になっているものが多いが、そのまま英語にするのは困難である。

2.6 タイプVI: その他

「アンケートする」などの非英語や、「ベースアップする」などのI~Vの分類に含まれない品詞並びの語である。

表 3: カタカナ述語の型分類

		動詞	形容動詞	合計
I.	単純動詞 / 形容詞	102	85	187
II.	ing 型	10	-	10
III.	名詞	14	-	14
IV.	動詞＋副詞	18	-	18
V.	名詞＋動詞	6	-	6
VI.	その他	14	3	17
	合計	164	88	252

3 変換ルールとその効果

前節の分類のうち、比較的簡単な規則で英語に変換できると思われる「タイプI: 単純動詞 / 形容詞」と「タイプII: ing 型」について、表4の変換規則を作成し、翻訳正解率を調査した。調査では、カタカナ述語を含む単文を作成し、本規則で翻訳した時に正しい意味に訳せる割合を求めた。作成した単文は新聞記事から抽出した文から修飾句を除いて簡単にした文である。英語辞書を用いればカタカナ表記から英語綴りをほぼ正確に得られることが示されているので [5]、調査では正しい英語綴りが得られているものとして検討する。

表 4: 変換ルール

カタカナの条件	変換処理
(自動詞)＋する	[ガ格]→SUBJ とする
(他動詞)＋する	[ガ格]→SUBJ、[ヲ格]→OBJ とする
(ing)＋する	原型動詞と同じ扱いとする
(形容詞)＋だ	[ガ格]→SUBJ とする

結果は、タイプIとIIの197語中、170語(86%)が正しく訳せることがわかった。この語数は、カタカナ外来語述語の総数252語のうちの67%にあたる。翻訳に失敗した27語の分類は、以下のとおりである(件数は表5)。

他動詞なのにヲ格なし 英語では他動詞であるが、日本語にヲ格がなく、翻訳できなかったもの。正しく翻訳するには何らかのゼロ代名詞補完処理が必要である。

前置詞選択の誤り 「～をアピールする」→「appeal for ～」などのように、決まった前置詞を訳出する必要があるもの。品詞のみの情報を用いての機械的な翻訳は困難であり、正しく翻訳するには格パタン記述が必要である。

意味が異なる 原文とは異なる意味の訳文になるか、全く意味が通じないもの。

表 5: 翻訳失敗の分類

翻訳失敗の原因	件数
他動詞なのにヲ格なし	3
前置詞選択の誤り	4
意味が異なる	20
合計	27

4 おわりに

従来未知語となり、翻訳不能となることが多かったカタカナ述語の構造の分類を行ない、そのうち67%が簡単な変換ルールと英語辞書の情報のみで翻訳できることを示した。本翻訳規則はNTTの日英機械翻訳実験システムALT-J/E[6]に組み込まれている。

参考文献

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

[2] Knight, K. and Graehl, J., "Machine Transliteration", *ACL97*, (1997)

[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] 松尾, 畑山, 池原, "英語辞書と英文法を用いたカタカナ表記語の翻訳", 53回情処全大,2-65(1996)

[6] Ikehara, S., Shirai, S., Yokoo, A. and Nakaiwa, H., "Toward an MT System without Pre-Editing – Effects of New Methods in ALT-J/E –", *MT Summit III*, (1991)