2009年度 卒業研究概要

フォニックスを用いた英単語のカタカナ表記生成

大谷　紀子　研究室
0632152　長井　香織

1.研究の背景･目的

今日、カタカナ表記を介して、外来語が日本語文中に頻繁に見受けられる。カタカナの取扱いは、日英方向の翻訳において重要であると考えられる[1]。カタカナ表記を生成する場合、辞書を用いて英単語の発音記号からカタカナ表記を生成する。しかし、新出単語や専門用語は辞書には載っておらず、発音がわからない。辞書に載っていない場合には、辞書に載っている英単語と同じ綴りの箇所を当てはめる方法が適用できるが、英単語は同じ綴りであっても発音が異なる場合がある。そのため、英単語からカタカナ表記を生成すると複数のカタカナ表記ができたり、正確ではないカタカナ表記が生成されたりする可能性がある。

入井[2]は英単語のデータベースを利用したカタカナ表記生成手法を提案している。データベースには「高校入試でる順　英単語ターゲット1800」[3]に載っている単語とそのカタカナ表記が記録してある。入力された英単語と記録されている単語を比べ、綴りが同じカタカナ表記を採用している。しかし、データベース内の単語が生成結果に大きな影響を与え、正解率が低いこと、特徴のある綴りには対応できないことが問題である。

本研究では、英単語のカタカナ表記生成の支援を目的とする。フォニックスを用いて英単語からカタカナ表記を生成するシステムを構築し、評価実験により本システムの有用性を示す。

2.システム

本システムは、入力した英単語からフォニックスを用いてカタカナ表記を生成するシステムである。フォニックスは英語圏で子どもに読み方を教えるための方法として利用されている。英単語の75%はフォニックスを用いて読むことができるといわれている。本システムの処理の流れを以下に示し、“apple”を対象とした処理例を表1に示す。

①入力された英単語を<子音+母音>の形に分解する。
②フォニックスのルールに従い、子音の連続や末尾の“e”などを削除する。
③それぞれの母音の読み方をすべて選択し、子音のみの組には対応する母音を添える。
④すべての<子音+母音>の組に対してローマ字表記を生成する。
⑤ローマ字表記からカタカナ表記に変換する。
⑥カタカナ表記をYahoo!検索APIを用いて検索結果数の1番多いカタカナ表記を出力する。

表 1　処理例

	apple
	↓
①	a / pp / le
	↓
②	a / p / l
	↓
③	a(ア:a,エイ:ei) / p(u) / l(u)
	↓
④	apulu,eipulu
	↓
⑤	アップル，　エイップル
	↓
⑥	アップル

3.実験･結果

先行研究の評価実験にて用いられた英単語のカタカナ表記を本システムで生成し、出力された結果をカタカナ辞書を元に評価した。先行研究では“off”を“オブフ”、“sense”を“スンズエ”と間違った変換をしたが、本システムでは正確に変換することができた。反面、先行研究で正確であった“baseball”を“バセバル”、“July”を“ジュリー”と間違った変換をした例も見られた。先行研究の評価実験に用いられた英単語のカタカナ表記の正解率は先行研究が54%、本システムは58%とさほど差は見られなかった。

また、大谷研究室の3年生9人を対象にアンケートを実施した。辞書に載っている英単語(以下、既存単語)と載っていない英単語(以下、新出単語)それぞれに評価する。既存単語では、 本システムが生成したカタカナ表記と発音記号を基に生成したカタカナ表記から正確だと思うものを選択する。新出単語では本システムで生成したカタカナ表記を“正確である”“まあまあ正確である”“正確ではない”の3段階で評価する。既存単語では本システムのカタカナ表記が61%で正確であると評価された。発音記号から正しく生成したカタカナ表記が選択されるとは限らなかった。新出単語では83%で正確である、まあまあ正確であるという評価が得られた。

4.考察

発音記号からすると正確ではないカタカナ表記でも正確であると判断されている。特に“ultra”のカタカナ表記は全員が“ウルトラ”を選択している。“ultra”の発音記号は[ˈʌltrə]であり最初の音は“ア”である。正しいとされるカタカナ表記は発音だけではなく、ローマ字の要素もある程度含まれていることがわかる。アルファベットの“u”は“ウ”とも“ア”とも読むため、本システムの生成段階で“ウルトラ”“アルトラ”両者とも生成される。Yahoo!検索APIを利用しているので、より多く使用されているカタカナ表記を採用することができるが、正確ではないカタカナ表記を出力することがある。特に短い単語の場合、生成結果が文字列の一部として認識され、検索結果に影響があるため正確なカタカナ表記を生成しているにも関わらず、間違ったカタカナ表記を採用する。

アンケートをとっている際、新出単語の読み方がまったくわからないという声が上がった。今までにない綴りで発音もわからない英単語のカタカナ表記を生成するには本システムを使用することで発音を基にした結果が得られるため、本システムは有効であるといえる。

しかし、本システムでは単語を組み合わせた複成語が認識されないため、複製語の構成単語ごとにフォニックスのルールを適用することができない。また、本システムは母音の発音が複数あることに注目していたが、子音にも複数の発音があり、母音との関係や、位置などが影響していたり、まったく関係がなかったりと複雑な処理ができなかったため精度が上がらなかったと考える。複雑な処理をするためにさらに細かいルールを取り入れることや、Yahoo!検索APIに代わる新たな採用手法を提案することで精度が向上すると考えられる。

参考文献

[1] 黒田純子，松永義文，“日本語文におけるカタカナ英語の研究”，情報処理学会自然言語研究会資料，68－3，1988.

[2] 入井歩，“英単語のカタカナ表記生成手法の研究”，武蔵工業大学卒業論文，2005.

[3] 谷口賢，“高校入試でる順 中学 英単語ターゲット1800”，旺文社，2005.