2005 年度卒業研究概要

英単語のカタカナ表記生成手法の研究

指導教員
承認印

大谷　紀子　研究室
0232024　入井　歩

1. 背景と目的

近年では、日本語の文章の中で英語をそのままカタカナ表記にした「カタカナ英語」がよく使われるようになった。英単語は難解なものも多く、数多くの英単語を正確にカタカナ表記に直すのは困難である。特に英語の人名は読み方に特徴的なものが多い。さらにカタカナ表記は 1 つとは限らず、1 つの英単語に複数のカタカナ表記がある場合もある。

本研究では、英単語のカタカナ表記を生成することを目的とする。また本手法により導出される結果は英単語のカタカナ表記であり、発音とは異なる。カタカナ表記は一概にスペルから同定できるとはいい難く、また単純なルールに基づいているともいえない。様々な英単語の中から共通のルールを見出し、どのようなものにも対応できるような手法を研究し、未知の英単語でもカタカナ表記を生成できるようにすることが本研究の最大の目的といえる。

2. カタカナ表記生成

カタカナ表記生成手法は、データベース作成とカタカナ表記出力の 2 段階に大別される。

2.1　データベース作成段階

データベース作成段階では「高校入試でる順　英単語ターゲット 1800」[1]に載っているすべての単語を用いて、カタカナ表記生成のためのデータベースを作成する。ローマ字レベルまで分解した単語とカタカナ表記とを比較し、該当する英文字列とカナ文字列を 1 組として、csv 形式のファイルに格納する。表 1 にデータベースの例を示す。

表 1：データベース例

mother	マザー	mo	マ	ther	ザー				
if	イフ	i	イ	f	フ				
school	スクール	s	ス	choo	クー	l	ル		
friend	フレンド	f	フ	rie	レ	n	ン	d	ド
and	アンド	a	ア	n	ン	d	ド		
or	オア	o	オ	r	ア				

左 2 列が英単語とそのカタカナ表記であり、こちらは検索には使用しない。データベースを見やすくするためと、入力された単語がデータベース内の単語と一致するかどうかの判定にだけ使われる。データベースは左から 3 列目以降になる。この場合は mo とマ、ther とザーなどがそれぞれ組となりデータベースに格納される。

2．2　　カタカナ表記出力段階

カタカナ表記出力段階では、仕組みは以下のようになっている。

1-1：データベース中から、入力された英単語の最初の文字で始まるものを全て抽出する。

1-2：(1-1)で抽出されたものの中から、入力された英単語の 2 文字目まで一致するものを抽出する。以下、検索結果が 1 つになるまで検索する文字数を増やして繰り返す。

2：導出される文字が 1 パターンになった場合、次の文字から(1-1)の手順に戻り、入力された英単語の文字が全て検索されるまで繰り返す。

3：導出されたそれぞれの文字のカタカナ表記を組み合わせる。ここで使われるカタカナ表記は、データベース内でそれぞれの文字の中で一番多いカタカナ表記が採用される。同数の場合は両方とも表示する。

3． 評価実験

本手法を実装したシステムを構築し、大谷研究室の 3 年生 8 人を被験者として評価実験を行った。またシステムには、入力された英単語がデータベースにあるかどうかを表示する機能も付けた。手順、質問・評価項目、実験結果の一部を以下に示す。また、表の数字は人数である。

手順： データベース内にある単語とない単語をそれぞれ 3 語ずつ入力してもらい、入力後に出力された結果について評価してもらい、またいくつかの質問に答えてもらった。

評価項目 1：導出されたカタカナ表記について、単語毎に○・△・×の 3 段階で評価してもらう。

評価項目 2：全体としての正確さを 4 段階で評価してもらう。

質問項目：英単語のカタカナ表記を知りたいときに、本システムを使用してみたいと思うかどうか。

表 2：評価項目 2

全体としての正確さ	
正確だと思う	0
まあ正確だと思う	6
あまり正確ではない	2
まったく正確ではない	0

表 3：質問項目

本システムを使用してみたいと思うかどうか	
思う	6
思わない	2

4． 考察

評価実験の結果、表 2 の評価項目 2 で「まあ正確だと思う」と答えた人は、データベース上にない単語でも満足な結果を得られた。逆に「あまり正確ではない」と答えた人は、データベース上にある単語ですら満足な結果を得られなかった。ゆえに一部の単語ではデータベース上にあるものでも不十分な精度ということがわかり、より一層の精度の向上が求められる。また表 3 の質問項目で「思わない」と答えた人は、データベースにない単語の検索で満足な結果を得られず、データベース上の単語しか高い精度を維持できないシステムは不要との意見だった。ゆえにデータベース外の単語でもある程度の精度を維持できるような、新しい手法を考案する必要がある。

参考文献

[1]　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.