カタカナ語から英単語への復元の試み

酒井 純
PXL06663@nifty.ne.jp
名古屋大学大学院文学研究科

0. はじめに

日本語の処理システムにおいて、カタカナ外来語や和製英語など（以下カタカナ語と表記）が、処理システムの辞書に記載のない、未知の語となることがある。これは、カタカナ語に新語が多いことやカタカナ語の表記のゆれの問題などによる。このため、すべてのカタカナ語を辞書登録することは不可能であり、従来のシステムでは一定範囲の語を収録の上、これ以外は未知語として例外的に扱う場合が多かった。

この問題に対して、未知カタカナ語を復元変換する方法として、「カタカナ語辞書」を用いる語彙レベルでの変換(松尾 1999)、音韻対照による音素レベルでの変換(Knight1997, 酒井 1998)が考案されてきた。本論では、酒井 1998 で試作したカタカナ語から英単語への復元システムに、新たに形態素・語レベルでの変換を追加することで、復元効率の向上をめざすとともに、これらの方法を組み併せたシステムについて考察している。

1. カタカナ語の変換

1.1 語彙レベルでの変換

松尾 1999 では、「英語電子辞書の各単語にカタカナ語表記を付与した（カタカナ英語辞書）を構築して日英語機械翻訳に適用」することにより、語彙レベルでのカタカナ語から英単語への復元を行っている。

まず電子英語辞書から英語の見出し語を取り出して語彙の一覧を作成する。この一覧の語に、英語の綴りと発音記号からプログラムで生成したカタカナ語をインデックスとして付与する。そして、これらの英単語とカタカナ語のインデックスによる「カタカナ英語辞書」を、カタカナ語から英語への変換に用いている。

この方法では、英語の辞書から「カタカナ英語辞書」の見出し語を作成しているため、英語を借用元とするカタカナ語はほとんど変換が可能であると考えられる。松尾 1999 では、この変換プログラムを日英語機械翻訳に組み込んで検証を行っており、日英語機械翻訳のプログラムと併せて 93% の変換効率を示している。[1]

1.2 音素レベルでの変換

Knight1997、酒井 1998 では、英語とカタカナ語の音韻の対応に着目し、音素単位でのカタカナ語から英単語への復元を行っている。

酒井 1998 においては、まず英単語からカタカナ語への取り入れに用いられる音韻法則を分析することで、英語とカタカナ語の音韻対照をしている。同時に、竹林 1981 のフォニックスのルールを基礎として、英語の綴りと発音の対応をまとめている。この２つをもとにして、カタカナ語から英単語への復元規則を作成し、これに基づいた復元プログラムを試作している。

この方法では、すべてのカタカナ語の音素を網羅的に規則化しているため、どの様な語の入力に対しても何らかの出力を得ることが可能である。しかしながら、音素レベルでのカタカナ語から英単語への復元は、日本語と英語の音韻数・音韻構造の違い、英語の音韻と綴りの不一致、カタカナ語のゆれなどの問題から、変換効率を向上させること容易ではない。

[1] ただしこの数値については、56%が機械翻訳による変換であり、また日本語のカタカナ語、和製英語については変換効率から除いている。このため、語彙レベルでの変換単独では、カタカナ英語正解 176 語÷（カタカナ英語正解 176 語＋翻訳誤り 58 語＋α）で約 75%以下の変換効率であると考えるのが妥当であろう。

1.3 形態素・語レベルでの変換

語彙レベルでの変換と、音素レベルでの変換の中間段階として設定されるのが、形態素・語レベルでの変換である。ここでは、英語の形態素（接辞・語根など）と、英語で複合語を形成するときに用いられる語の形態（以下連結形）を変換の単位とする。まず、音素レベルでの変換との区切り方の違いを次の表 1.に示す。

J	エ	レ	クトロ		ニ		ク	ス
R	e	re	kutor	o	n	i	ku	su
	↓	↓	↓	↓	↓	↓	↓	↓
E	e	le	ctr	o	n	i	c	s

表 1. 音素レベルでの変換

J	エレクトロ	ニクス
R	erekutoro	nikusu
	↓	↓
E	electro	nics

表 2. 形態素・語レベルでの変換

表からも分かるとおり、音素レベルでの変換では 8 つの部分に分割して変換されているのに対して、形態素・語レベルでは 2 つの部分に分かれるのみである。

音素レベル	候補数	形態素・語レベル	候補数
#e-	2	-erekutoro-	1
-r-	4		
-e-	2	-nikusu#	1
-kutor-	1		
-o-	1		
-n-	2		
-i-	2		
-kusu#	5		

表 3. 各レベルにおける候補数

表 3.にあげた候補数は、それぞれの変換目標カタカナ音素列に対しての、候補としてあげられる英語文字列の数である。そして、1 語での候補数を考える場合には、表の候補数の積となり、具体的には音素レベルの変換では 320 候補、形態素レベルの変換では単独候補となる。これは、変換速度の問題、変換効率の問題のどちらの側面からとらえても、形態素・語レベルでの変換が有利な点である。

ただし、形態素・語レベルでの変換では、変換区切りの決定方法が問題となる。本考察の復元システムは酒井 1999 のシステムを基礎とするため、前方最長一致法で変換を行っているが、このために変換に失敗する例を次の表 4.に示す。

前方最長一致法による誤変換			
J	リスト	ラ	クト
R	#risuto	ra	kuto#
	↓	↓	↓
E	list	la	ct

正答		
J	リ	ストラクト
R	#ri	sutorakuto#
	↓	↓
E	re	stract

表 4. 前方最長一致法による誤変換の例

表の通り、前方最長一致法では先頭から最長（この場合には/#risuto-/の部分）が選択される。このため、この後の/-rakuto#/の部分が 2 分割で変換されている。正しくは、最初の部分がまず/#ri-/のみで変換され、残りの/-sutorakuto#/の部分が 1 つままで変換される必要がある。このため、今後は最小コスト法などの変換方法なども検討していく必要があるだろう。

1.4 各レベルの位置づけ

語彙レベルの変換では多くの語が正しい形で復元可能ではあるが、あくまでも固定的な辞書を用いているものであり、辞書に記載のない語は未知語となってしまう。これに対して、音素レベルでの変換は基本的に未知になる語は存在しないシステムであるが、正しい出力を得られる可能性は低い。また、形態素・語レベルでの変換では、語彙レベルの変換に比べて未知語となる可能性は低く、音素レベルでの変換に比べて変換効率が高い。つまり、カタカナ語から英単語への復元変換の正答率としては、

音素レベル<形態素・語レベル<語彙レベル

の順序で高くなると考えられる。

この様な各レベルでの変換の特徴を活かすため、次のような処理が想定される。まず、日本語処理システムから、未知語とし

て出力されたカタカナ語は、語彙レベルでの変換が試みられる。この時、語彙レベルの変換でも未知となってしまったものについては、形態素・語レベルでの変換に送られ、処理される。形態素・語レベルの変換で未知となったもの、また部分的に未知となったものについてのみ、音素レベルでの変換が行われる。そしてこれらの結果が日本語処理システムに返される。これをまとめたのが次の図である。

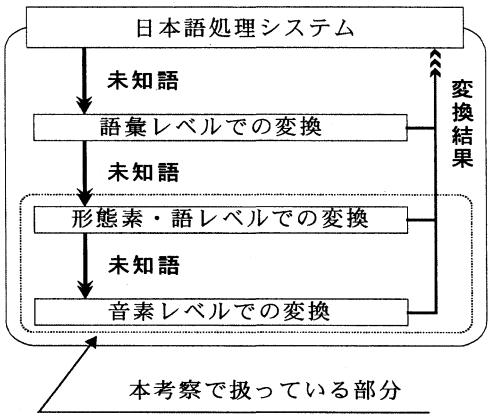

図　未知カタカナ語変換システムを含む日本語処理システム

本考察では、これらのシステムを想定の上、音素レベルと形態素・語レベルでの変換を組み合わせたカタカナ語復元システムを構築している。

2. 使用データと変換規則

形態素・語レベルでの変換規則を作成するために必要となるデータとしては、語根、接辞、連結形が考えられる。

接辞および連結形については、電子版『リーダーズ英和辞典』1996 に、それぞれ接辞または連結形として記載のあるものの中で、「基礎データ[2]」中に用例の存在が確認できるものについて用いている。これは、実例に基づいた変換規則を作成するためである。

一方、語根については、網羅的なデータを収集することができなかったため、基礎データ中から収集している。この時、必ず接辞または連結形が前後両方に接続しているものについてのみ抜き出している。

次の表 5.は、集められたデータ数であり、表 6.はこれを元に作成された語・形態素レベルでの変換規則数である。

	データ数
語根	82
接頭辞	40
接尾辞	112
連結形（前接）	407
連結形（後接）	111

表 5.　語根・接辞・連結形のデータ数

	変換規則数
語根	128
接頭辞	52
接尾辞	232
連結形（前接）	416
連結形（後接）	131

表 6.　変換規則数

一方、音素レベルの変換については、酒井 1998、1999 で試作したシステムを基礎としている。このシステムでは、英単語からカタカナ語への取り入れ法則と、フォニックスの綴り字と音素の対応を組み合わせ、実例を「基礎データ」から収集して復元ルールを作成している。次の表 7.は、音素レベルでの変換ルール数である。

	変換ルール数
母音	155
子音	636
英語語末/e/	3

表 7. 音素レベルの変換ルール数

[2] カタカナ語とその原語の対応、8214 語を一覧としたファイル（酒井 1998、1999）。

3. プログラムの実装と検証

本考察では、音素レベルのみでの変換と、音素レベルと形態素・語レベルを組み合わせた変換のプログラムを試作している。これは、音素レベルの変換と比較することで、形態素・語レベルの変換の有効性を検証するためである。

そして検証方法としては、この２つのプログラムに対して、前出の「基礎データ」8214 語のカタカナ語の部分を入力し、出力された語形を実際の語と比較することで、カタカナ語から英単語への復元効率を算出している。次の表 8.は、それぞれの変換効率を、変換候補１位のみ、１０位まで、１００位までにまとめたものである。

	音素レベル	
	正答数	正答率
1 位のみ	65 語	0.8%
10 位まで	999 語	12.2%
100 位まで	3872 語	47.1%

	形態素・語レベル	
	正答数	正答率
1 位のみ	1593 語	19.4%
10 位まで	4291 語	48.3%
100 位まで	4653 語	56.6%

表 8. 復元変換の正答率 （正答数は 8214 語中）

形態素・語レベルの変換では、音素レベルの変換に比べて上位の正答数が非常にのびている。このことにより、変換候補 10 位までの変換効率で 48.3%と、音素レベルでの変換の 100 位までの変換効率を上回る結果となっている。その一方で変換候補 10 位以上での正答率が低いため、100 位までの変換効率でも 10 位までと比べて 8.3 ポイント差しか付いていない。これは、前方最長一致法による変換区切りの問題、語根データの不足、また音素レベルの変換が行われる部分の効率の悪さなどが原因として考えられる。

4. おわりに

日本語処理システムで未知カタカナ語を処理するための復元方法としては、これまで語彙レベル、音素レベルでの変換が行われてきた。本考察では、この２つの中間段階として、語彙レベルでの変換よりも未知語に対して処理能力が高く、また一方で音素レベルでの変換よりも変換効率を向上できる、形態素・語レベルでの変換を設定し、その有効性を検証してきた。そして、これらを組み合わせた総合的な処理システムについても言及することができた。

プログラムの試作と検証という点からは、語根のデータが網羅的に収集できていないこと、前方最長一致法による問題点、日本語処理の全体のシステムとして検証できていない点など不備が多く、課題となっている。しかしながら、この様な悪条件のなかで、形態素・語レベルでの変換の有効性を証明し、また３つのレベルの特長を活かした未知カタカナ語処理システムを提案することができたと考えている。

参考文献

Knight, K., and Graehl, G. 1997. *Machine Transliteration.* ACL97
http://xxx.lanl.gov/ps/cmp-lg/9704003

竹林滋 1998『英語音声学入門』大修館書店

酒井純 1998「カタカナ語と英単語にみる音韻対照に関する研究」『名古屋大学言語学論集第 14 巻』名古屋大学

酒井純 1998「英語の語中における子音連続について」『名古屋大学言語学論集第 15 巻』名古屋大学

松尾義博 1999『外来語処理のためのカタカナ英語辞書』
http://www.kecl.ntt.co.jp/icl/mtg/members/yosihiro/nlpsym99.html