原表記とカナ表記の対応判定アルゴリズム

6B-3

大深悦子

日本アイ・ビー・エム株式会社　東京基礎研究所

1．はじめに

日本語のかな等の音標文字を使って外来語を表記する場合、複数の表記のしかたが存在することが多い。[1] 例えば "interview" の表記には、"interview", 「インタビュ」,「インタヴュ」,「インタビュー」,「インタヴュー」等がある。これらの表記結果同士が同一語を表わしているかどうかの判定が求められている場として、文書校正システムの表記ばらつき検出や情報検索システムがある。従来は、これらの表記を辞書に登録する方式がとられたが、辞書の作成・保守が大きな問題であった。ここでは、辞書を使わずに、発音をもとにして表記の対応を判定するアルゴリズムについて、カナ表記ー英語表記の対応判定システム(図1)を例に説明する。

2．表記間の対応判定アルゴリズム

2.1 日本語音素列生成

ローマ字表記結果に、語頭以外の(w→u/y→i)、二重母音の長音化、母音にはさまれた長音削除などの変換規則を適用した統一表記を日本語音素列とする。

例: プレーヤー → pure-ya- → pureia- ..→ purea-

2.2 英語音素列生成

英語のつづりから音素列を得る。(例えば[2]参照)

2.3 英語音素と日本語音素の対応

英・日音素対応のため、音素を以下の形に表わす。

子音音素(C)=R+Pr+J;　母音音素(V)=R+Pr+VL

R :音素に対応する、日本語でのヨミ
(促音,長音,拗音(j音)はヨミに含めない)

Pr:各音素に対応するRの優先順位

J :拗音の有無 (1:有, 0:無)

VL:母音音素の長さ
(-6:促音 ~ 0:促・長音なし ~ +6:長音)

例:キャッ →R(k)+Pr(1)+J(1) + R(a)+Pr(1)+VL(-6)

英語音素をこの形式に表わした例を表1に示す。

2.4 英語音素列の変換結果調整

2.3で得た英語音素列の変換結果を、音韻環境、つづりによって、以下の二点から調整する。

(1) ヨミ(R)の優先順位

例:音素UHのつづり ='a' → Pr(R:a)=0

(2) 母音音素の長さ(調整値をA1とする)

例えば、母音音素のつづりが2文字以上ならカナ表記上、長音になる傾向がある。→ A1=+1。

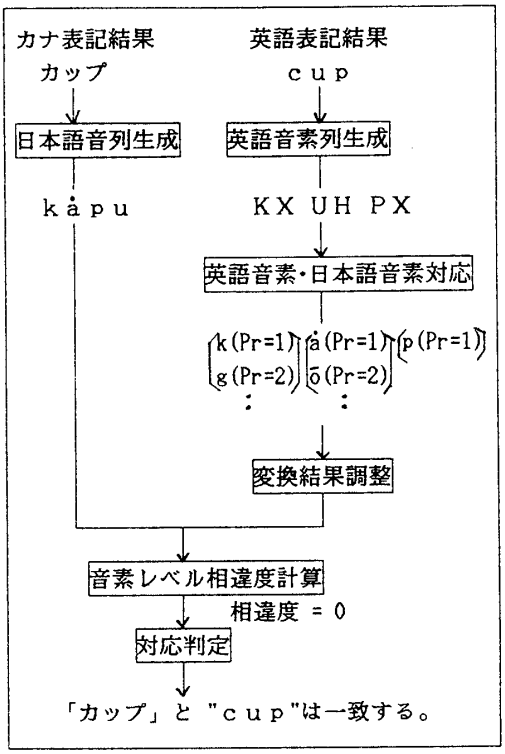

図1　カナ表記ー英語表記の対応判定の一例

表1　英語音素・日本語音素の対応表

```
@:phoneme:Pr:R(yomi):J:VL:+v:
-----------------------------
** KX -> /k/
c:KX      :1:k        :0:  :u :
c:KX      :2:g        :0:  :u :
c:KX      :3:c        :0:  :i :

** UH -> /ə/ or /ʌ/
v:UH      :1:a        :0:-1:
v:UH      :2:o        :0:1 :
v:UH      :2:u        :0:-1:

** PX -> /p/
c:PX      :1:p        :0:  :u :
```

A correspondence finding algorithm between words in original language and kana

Etsuko OFUKA

IBM Research, Tokyo Research Laboratory

2.5 音素レベル相違度計算

2.5.1 チャンク数マッチング

音素列を子音の先頭で区切った固まりを"ﾁｬﾝｸ"とよぶ。(例: ｶｯﾌﾟは{kå}{pu}で2ﾁｬﾝｸ; • は促音)

ﾁｬﾝｸ数が一致しない場合、両表記は一致しないと判断する。以下、相違度は各項目ごとに定められたﾍﾟﾅﾙﾃｨｰの総和とする。

2.5.2 子音部のヨミ(Rc)のマッチング

日本語音素列の子音部のﾖﾐ(JRc)と英語音素列の子音部のﾖﾐ候補(ERci)を第1ﾁｬﾝｸから比較する。ERciは表１のRで与えられる。JRcと一致するﾖﾐ(ERcm)が見つかれば、ERcmの優先順位によって定められたﾍﾟﾅﾙﾃｨｰを相違度に加算していく。

(例:Pr=0,1→ﾍﾟﾅﾙﾃｨｰ0, Pr=2 →+2, Pr>=3 →+4)

JRcと一致するﾖﾐが見つからない場合は、両表記は一致しないと判断する。

2.5.3 拗音と母音部のマッチング

第1ﾁｬﾝｸから順に以下のマッチングを行なう。

(1) 拗音の有無(J)のマッチング

Jが一致しなければ +5のﾍﾟﾅﾙﾃｨｰを加算する。

(2) 母音部のヨミ(Rv)のマッチング

日本語音素列の母音部のﾖﾐを(JRv)、英語音素列の母音部の音素をERv=ERv1 ERv2 .. ERvnとする。各ERviのﾖﾐ候補(表1のR)から任意に１つずつ選んで作ったヨミ列とJRvを、左から順に比較する。英語ﾁｬﾝｸが子音で終わっている場合は、ｶﾅ表記上つける母音(表1の+v)を補って比較する。一致したERviのﾖﾐが見つかるたびに、そのﾖﾐの優先順位によって定められたﾍﾟﾅﾙﾃｨｰを加算する。(例:Pr<=1→ﾍﾟﾅﾙﾃｨｰ0,Pr=2→+1 Pr>=3→+2; Pr=0の別候補があれば ﾍﾟﾅﾙﾃｨｰ→ ﾍﾟﾅﾙﾃｨｰ+1)

JRvの途中までしか一致するﾖﾐがない場合は、(JRvの余った音素数 × 2)のﾍﾟﾅﾙﾃｨｰを、JRvの最後まで一致するﾖﾐがあって、かつ英語音素が余る場合は、(余った英語音素数 × 2)のﾍﾟﾅﾙﾃｨｰを、加算する。

(3) 各ﾁｬﾝｸ最後の母音音素の長さ(VL)のﾏｯﾁﾝｸﾞ

・ 日本語音素列の母音音素の長さ(Kとする)

K=-6(促音), 0(促・長音以外), +6(長音)

・ 英語音素列の母音音素の長さ(Aとする)

A=A1(2.4 (2)の調整値) + VL(表1)

K,Aと表2よりﾍﾟﾅﾙﾃｨｰを求め、相違度に加算する。

3．相違度を使った対応判定

２種類のデータ

(A)同一語を表わす 原表記－カナ表記 300対

例: expert - ｴｷｽﾊﾟｰﾄ

表２ 母音音素の長さﾏｯﾁﾝｸﾞにおけるﾍﾟﾅﾙﾃｨｰ

条件			ﾍﾟﾅﾙﾃｨｰ
k=0	|A|<=1		0
	|A|=2		+1
	|A|>=3	最後のﾁｬﾝｸ	0
		最後以外のﾁｬﾝｸ	+2
k≠0	|K-A|<=4		0
	|K-A|=5,6		+1
	|K-A|=7	最後のﾁｬﾝｸ	+1
		最後以外のﾁｬﾝｸ	+2
	|K-A|>7	最後のﾁｬﾝｸ	+3
		最後以外のﾁｬﾝｸ	+4

表３ ﾃﾞｰﾀ(A),(B)の相違度の分布

	相違度<3	相違度=3	相違度>3
(A)	９５%	3%	2%
(B)	１１%	１４%	７５%

(B)別語を表わし、かつ子音並びの一致する

原表記－カナ表記 300対

例: expert - ｴｸｽﾎﾟｰﾄ

を使って、相違度の分布を求めた。

ﾃﾞｰﾀ(A)については、表３よりエラー率２%で、

相違度 < 3 ならば 一致

〃 = 3 〃 類似

〃 > 3 〃 不一致

といえることが、わかった。

ﾃﾞｰﾀ(B)で相違度<3のものについては、発音を対応判定の基盤とする本方式では、やむをえない。しかし、校正に使う場合には、同じテキスト中に、子音の並びが完全に一致していて別語を表わす表記が問題になるほど多く出現するとは考えられないので、本方式は十分有効だといえる。

４．おわりに

原表記の発音をもとにカナ表記－原表記対の相違度を計算し、その対応判定をするアルゴリズムについて、そのアルゴリズムと有効性を述べた。

参考文献

[1] 後藤他:片仮名表記をとる技術用語における表記の多様性, 三田図書情報学会大会,1985

[2] Elovitz et al.: Letter-to-Sound Rules for Automatic Translation of English Text to Phonetics, IEEE Trans Vol.ASSP-24,No.6,1976