カタカナ表記からの英訳推定による専門用語辞書作成

熊野 明
kmn@eel.rdc.toshiba.co.jp
(株)東芝 研究開発センター

1 はじめに

対訳コーパスは，新しい対訳知識を抽出する対象として注目され，多くの研究が行われている．これらは，機械翻訳システムのカスタマイズに必須な辞書作成に応用することで実用性が明らかになってきている[山本93]．われわれは，既存対訳文書から専門用語辞書用対訳データを作成するために，日本語の文書から専門用語表現を抽出し，その訳語を英訳文書から推定することを目指している．日本語の専門用語としては，複合名詞と未知語を対象にして実験を行っている[熊野94]．訳語の推定には，機械翻訳用辞書の知識を中心とした言語情報と対訳文書中の出現頻度を中心とした統計情報を利用している．新しい専門用語が多く現われる特許明細書の対訳文書で実験したところ，言語情報が有効に利用できる複合名詞に関しては，7割から8割の訳語を正しく推定できた．ところが，未知語では言語情報がほとんど利用できないため，低い精度に止まっていた[1]．

推定精度の低い未知語を分析したところ，カタカナ語が多く含まれていることがわかった[2]．技術文書の専門用語は英語から発生するものが多く，その日本語訳語が固定するまでカタカナで表現することが多い．また，日本語訳語を決めるより，カタカナ語が一般化するものも少なくない．

ところで，カタカナ語は通常，英語の発音を日本語で表現したものである．このことから，英語の綴りと日本語のカタカナ表記の対応は，語彙的な言語知識がなくてもある程度推定できると判断した．

本稿では，カタカナ表記と英語の綴りの関連性に注目し，対訳文書中に現れるカタカナの未知語に対する英訳を，英文書から推定して抽出する方法とその評価結果について述べる．

[1]詳しいデータは表5を参照のこと．

[2]用語47種類，出現頻度合計371のうち，21種類(45%)，頻度合計223(60%)がカタカナ語であった．

2 背景

異なる言語間で語句の対応を推定する研究は多く行われている[Kupiec93] [Church93]．しかしその多くは，英語・仏語のような同族言語間の特徴を利用するものであり，日英対訳文書からの辞書作成にはそのまま利用できない．また，言語情報を全く利用せず，統計処理だけで対訳データを抽出する研究[Fung94]もあるが，必ずしも精度の高いものではない．

カタカナと英語の照合に関しては，[宮内93]が，カタカナ表記から英単語の検索を実現している．しかし，英語の発音記号を利用したものである．この方法では，少なくとも発音記号情報をもつ辞書データが必要であり，既存の辞書情報のない未知語には対応できない．

一般に，カタカナ表記からもとの英語綴りを推定することや，逆に英語綴りからカタカナ表記を推定することに可能性はある．しかし，カタカナ表記からもとの英語綴りを1つ推定する場合，英語より子音・母音種類の少ない日本語のカタカナから英語綴りを正確に推定することはかなり困難である．

たとえば，カタカナの“シ”では，“シグナル”(signal)，“システム”(system)，“シフト”(shift)，“シン”(thin)のように，“カ”では，“カン”(can)，“カット”(cut)，“アルカリ”(alkali)，“カップル”(couple)，“ブレーカ”(breaker)のように，いくつもの英語綴りの可能性があり，特定は容易ではない．

同様に，一定の英語綴りに対する標準的なカタカナ表記は1つとはかぎらない．たとえば，子音綴り“ch”では，“change”(チェンジ)(“*CH*”[3]に相当)，“school”(スクール)(“*K*”に相当)，“machine”(マシン)(“*SH*”に相当)のように，母音綴り“a”では，“map”(マップ)(“*A*”に相当)，“phase”(フェイズ)(“*EI*”に相

[3]ここでの斜体大文字は，3章で導入する照合コードである．

当), "stage"(ステージ)("*E*"に相当), "character"(キャラクタ)("*YA*"に相当), "talk"(トーク)("*O*"に相当)のように, いくつものカタカナ表記の可能性があり, やはり単純に特定はできない.

3 照合コード

そこで, カタカナ表記と英語綴りの両者を, 発音の観点から直接比較できるコード(以下, 照合コードという)に変換し, 英語からの変換候補の1つが, カタカナ表記からの変換候補の1つに照合すれば, 両者が照合していると判断することにする. 以下, 照合コードは斜体の大文字[4]で表す.

3.1 英語綴りの変換

一般的には, 英語の音節(syllable)は, 1つの母音とその前後の(0個以上の)子音の連続を意味するが, 今回の実現では, カタカナ表記との照合を考慮して, 1つの子音(に相当する綴り)とそれに続く(0個以上の)母音(に相当する綴り)を音節と考える. 以下では, "音節"をこの意味で使用する. つまり, "character"は, 一般的には"char·ac·ter"の3音節であるが, 本稿では, "cha·ra·c·ter"の4音節で表す. 以下の例では, "·"で英語音節の区切りを, "-"で英語音節中の子音綴り部と母音綴り部の区切りを示す.

まず英語綴りを音節に分離し, さらに各音節を子音綴り部と母音綴り部に分離する. 続いて, それぞれの綴りを照合コードに変換する.

例えば, "character"の第1音節"cha"では, 子音綴り部"ch"と母音綴り部"a"に分離する. さらに子音綴り部"ch"を, *"CH"*, *"K"*, *"SH"*の3種類の照合コードに変換し, 母音綴り部"a"を, *"A"*, *"EI"*, *"E"*, *"YA"*, *"O"*の5種類の照合コードに変換する.

長母音と短母音は, 日本語でしばしば混用される[5]. カタカナ表記での両者の差は長音記号の有無だけで表されるので, 照合を容易にするために, 両者を区別しないで短母音に統一した. また, 一部の二重母音は長母音と混用されることが多い. 英語では二重母音にしかならない場合でも, 長母音に相当する照合コードも候補とする. 例えば, "a"は二重母音になる場合のために*"EI"*に変換するが, "エイ"が"エー"と表記されることを考慮して, *"E"*も変換候補としている. カタカナで区別のできない"r"と"l", "b"と"v", "s"と"th"などもそれぞれ同一の照合コード(*"R"*, *"B"*, *"S"*または*"Z"*)に変換する.

"character"を照合コードに変換した結果を表1に示す.

表1: "character"の照合コード

音節 i	綴り $\mathbf{E}_i$	子音部 $\mathbf{C}_i$	母音部 $\mathbf{V}_i$
1	ch-a	*CH*, *K*, *SH*	*A*, *EI*, *E*, *YA*, *O*
2	r-a	*R*	*A*, *EI*, *E*, *YA*, *O*
3	c	*K*, *S*	-
4	t-er	*T*	*A*

多くの子音綴りには, 無条件で照合コードコードに変換できるが, 特定の条件でのみ特定の照合コードに変換するケースもある. 例えば, 英語綴り"sc"は通常"scale"のように*"S"*+*"K"*の2音節に変換できるが, "e"や"i"の前では, "descent", "science"のように*"S"*の1音節に変換する.

なお, 黙字は一定の関連文字列(音)と連続して現われるため, 関連音と黙字を合わせて単独の子音綴り, あるいは母音綴りとして照合コードに変換する. すなわち, "know"の子音綴り部は黙字"k"を含む"kn"と見なして"n"と同様に*"N"*に変換し, "high"の母音綴り部は黙字"gh"を含む"igh"と見なして"i"と同様に*"AI"*に変換する.

3.2 カタカナ語の変換

カタカナ表記を1音節(拗音を含む)ごとに照合コードに変換する.

例えば, "キャ"は*"KYA"*, "ラ"は*"RA"*に変換する. 多くの音節は, 1種類の照合コードに変換するが, "チ"は, "ticket"(チケット), "chip"(チッ

[4]コードの割当ては, ほぼヘボン式ローマ字に基づいたものである.

[5]例: "return"は"リターン"とも"リタン"とも書かれる.

プ), "match"(マッチ)のように複数種類の英語綴りと照合させる必要がある．このような音節に対しては，複数の照合コード *"TI"*, *"CHI"*, *"CH"*を候補として与える．

長音記号"ー"は省略されることが多く，また長母音・短母音の区別が明確でないので，予め削除する．また，促音は英語に対応する綴りがないため，同様に削除する．カタカナ1音は英語の1音節に対応するのが原則であり，照合コードは子音部と母音部を合わせたものであるが，英単語語尾の子音，例えば"system"＝システムの"m"(ム)はカタカナで1音に表現されるので，単独子音を表す可能性のある音(一般にウ段の音)には，母音部のないコードも与える．

"キャラクター"を照合コードに変換した結果を表2に示す．

表 2: "キャラクター"の照合コード

音節 j	表記 $\mathbf{K}_j$	照合コード $\mathbf{R}_j$
1	キャ	*KYA*
2	ラ	*RA*
3	ク	*KU*, *K*
4	タ	*TA*

3.3 英語綴りとカタカナ語の照合

英語綴り全体の照合コード候補の1つとカタカナ表記の照合コード候補の1つが照合すれば，一致するとみなす．表1によると"character"は150($3\times5\times1\times5\times2\times1\times1$)通りの候補をもつ．一方，"キャラクター"は，表2により2通りの候補をもつ．それぞれ下線を施した照合コードを連結することにより，*"K-YA"*+*"R-A"*+*"K"*+*"T-A"*で両者が一致することがわかる．

また，表3は，5音節の"example"と6音節の"イグザンプル"が，照合することを示している．この例では，英語の第2音節の子音部"x"にカタカナ2音節に相当する照合コード *"GZ"*を与えることにより，*"GZ-A"*がカタカナの第2，3音節(*"G"*+*"ZA"*)と一致させている．

表 3: "example"と"イグザンプル"の照合

i	$\mathbf{E}_i$	$\mathbf{C}_i$	$\mathbf{V}_i$	j	$\mathbf{K}_j$	$\mathbf{R}_j$
1	e	-	*E*,*I*,*A*,ϕ[6]	1	イ	*I*,*YI*,*WI*
2	x-a	*KS*,*GZ*	*A*,*EI*,*E*,*YA*,*O*	2	グ	*GU*, *G*
				3	ザ	*ZA*
3	m	*M*	-	4	ン	*N*, *M*
4	p	*P*	-	5	プ	*PU*, *P*
5	l-e	*R*	*E*,*I*,*A*,ϕ	6	ル	*RU*, *R*

4 評価

[熊野94]と同様の方法で，半導体特許明細書7件(全部で2,148文)と，人手による英訳文書7件を使って，専門用語の抽出と訳語の推定を行い，正解データとの照合を行った．

これまで1位推定できなかった未知語のうち，今回の改良で推定が向上したものを，表4に示す．下線を施した語は，本方法によりカタカナと英語が照合したものである．用語中の / は，合成名詞の構成要素の区切りを示す．また，"改良前"とは[熊野94]での推定順位を示す．

推定訳語が正解データと一致した割合を表5に示す．ここで未知語とは，単独(複合語でない)の未知語を意味し，未知語を構成語に含む複合語は，合成名詞に含めている．

表 5: 訳語推定精度

用語	正解訳語	改良前	今回
合成名詞	1位推定	72.9%	**81.8%**
	3位推定以内	83.3%	**90.1%**
未知語	1位推定	54.0%	**74.1%**
	3位推定以内	65.0%	**74.7%**

[6] "machine"の"e"のように，綴りはあるが無音であることを示す

表 4: 訳語推定の向上した用語

専門用語	出現頻度	正解訳語	改良前	今回
ポリ/シリコン	36	polycrystalline silicon	5位	1位
ワード/線/ハイレベル/電圧	27	word line high level voltage	6位	1位
ドレイン	23	drain	2位	1位
シリサイド	12	silicide	6位	1位
リソグラフィ	12	lithography	9位	1位
キャパシタ/容量	9	capacitance of the capacitor	22位	3位
素子/形成/用/トレンチ	8	element formation trench	4位	1位
P/型/ウェル	7	P type well	2位	1位
ボロン	7	boron	12位	1位

カタカナ語の多い未知語に対する訳語推定精度を大きく向上することができた．合成名詞に対する訳語推定でも，構成要素として未知語を含む場合は今回の改良の効果が大きく現われた．

5 結論

対訳文書からの専門用語辞書作成において，英語の綴りとカタカナ表記の照合処理を利用し精度の向上を実現した．ここでの訳語推定処理は，カタカナ表記と英語綴りの発音の類似性に基づいたものであり，語彙辞書のような大規模知識を必要としない．また，対象とする対訳文書は必ずしも大量である必要はない．

対訳文書からの専門用語辞書作成における従来の処理では，未知語に対して言語情報がほとんど利用できなかったために訳語推定精度を上げることができなかった．今回開発した処理を辞書作成に導入した結果，未知語の1位推定精度が54.0%から74.1%に，20ポイント向上した．

この処理を利用することにより，従来対訳文書から専門用語辞書を自動的に作成でき，機械翻訳システムのカスタマイズに費やす時間を短縮できる見通しが得られた。

参考文献

[山本93] 山本由紀雄, 坂本仁: “対訳コーパスを用いた専門用語対訳辞書の作成”, 情報処理学会研究報告 NL94-12 (1993).

[熊野94] 熊野明, 平川秀樹: “言語情報と統計情報を用いた対訳文書からの機械翻訳辞書作成”, 情報処理学会研究報告 NL100-12 (1994).

[Kupiec93] Julian Kupiec: “An Algorithm for Finding Noun Phrase Correspondences in Bilingual Corpora”, *Proc. the 31st Annual Meeting of the ACL*, pp.23-30 (1993).

[Church93] Kenneth W. Church: “Char_align: A Program for Aligning Parallel Texts at the Character Level”, *Proc. the 31st Annual Meeting of the ACL*, pp.1-8 (1993).

[Fung94] Pascale Fung and Kenneth W. Church: “K-vec: A New Approach for Aligning Parallel Texts”, *Proc. COLING 94*, pp.1096-1102 (1994).

[宮内93] 宮内忠信: “カタカナ表記からの英単語検索システムの実現”, 情報処理学会研究報告 NL97-17 (1993).