自然言語処理 116－15
（1996. 11. 19）

発音情報を用いた訳語対の自動抽出

松尾 義博　　白井 諭
ＮＴＴコミュニケーション科学研究所

概要

実用的な機械翻訳システムには辞書情報の充実が欠かせないが、対象分野ごとに辞書を収集するのは容易ではなく、特に、通常未知語となってしまう固有名詞は収集が難しかった。しかし、固有名詞は一般に、ある言語から他の言語への輸入語であるため、その発音はどの言語でもよく似ているという特徴がある。

そこで、本論文では両言語の発音を推定し、その発音を比較することによって対訳を抽出するシステムを提案する。本方式は厳密に対応づけられた対訳コーパスを必要としないにもかかわらず、再現率 62%、適合率 98% の正確さで、未知語固有名詞対を抽出できた。

Using Pronunciation to Automatically Extract Bilingual Word Pairs

Yoshihiro Matsuo and Satoshi Shirai
NTT Communication Science Laboratories

Abstract

In spite of the importance of building domain specific bilingual dictionaries for large scale machine translation systems, it is difficult to gather word pairs for each domain. Proper nouns are especially hard to collect since they often do not appear in dictionaries. Fortunately, the pronunciation of proper nouns does not vary much between languages as they are normally realized as loan words.

This paper describes a method of automatic extraction of bilingual word pairs using an estimation of their pronunciation. This method can extract 62% of unknown proper nouns pairs with a precision of 98% from untagged bilingual corpora.

1 はじめに

実用的な機械翻訳システムを構築する場合には辞書情報の充実が欠かせない。現段階では多くの分野をカバーした汎用的な辞書をシステム側で提供することは望めず、分野毎にユーザーによる辞書構築が必要となっている。しかし、新たな分野に適用とする場合に用語集を収集することは容易ではなく、機械翻訳導入の大きな障害となっていた。

この辞書構築作業を軽減するために、統計情報や辞書方法を用いて対訳コーパスから翻訳情報を抽出する手法が、近年、いくつか提案されている。

辞書情報を用いた手法では、あらかじめ構築された対訳辞書を用いて、訳語候補を抽出している [1]。したがって、専門用語の訳語選択の絞り込みには使えるが、未知語を抽出することはできなかった。

また、統計情報を用いた手法では、まず、コーパス間で文対文程度の対応を取り、それぞれのペア中での、語の出現頻度などを手がかりにして、対訳候補を抽出している [2][3]。この方式は、両言語のコーパスで内容が一致していることを前提としており、内容に差異があるコーパスには適用できない。また、内容が完全に一致している場合でも、文対文程度の対応が取れることが必要であり、大きく意訳されたコーパスには適用できない。さらに、文対文の対応づけの精度によっても訳語抽出能力が制限される。

上記のような、内容が完全に一致し、かつ、大きく意訳されていない「質のいい」対訳コーパスは、どちらかの言語の文章を原本として、翻訳されたものに限られる。しかし、内容の完全一致を求めなければ、例えば、日英の新聞記事などで大量に入手可能である [4]。これらの新聞記事では、例えば、紙面の制約や説明すべき背景事情の差異などで、ある情報が片方の言語にはあるが、もう一方には存在しない、といった場合もあり得る。

そこで、本稿では、上記のように内容が完全に一致しているかどうかの保証のない対訳コーパスから、未知語訳語対を自動的に抽出するために、発音情報を用いる手法を提案する。

2 発音情報を用いた対訳抽出

日本語と英語は、それぞれの語彙や構文が大きく異なっており、英語と仏語のような双方の直接の比較 [5] は困難であった。しかし、どちらかの言語から他方へ輸入された言葉は、その語を表音文字で表記することが多く、発音は両言語で良く似たものとなっている。たとえば、日本語から英語に輸入された言葉は、ローマ字綴りで表記され、また、逆に英語から日本語に輸入された言葉は、カタカナで表記されることが多い。これら輸入語の発音は、原言語のものとは多少異なっているものの、よく似ている。したがって、対訳コーパスの作成・利用にあたって、発音は重要なキーとなることが期待できる。

そこで、本論文では両言語テキストの発音を推定し、その発音を比較することによって訳語対を抽出するシステムを提案する。

日本語にとっての輸入語は、カタカナ表記外来語と固有名詞に大きく分けられる。このうち、カタカナ表記外来語については、英語辞書と英文法を日本語解析に採り入れることで、日英翻訳できることが示されており [6]、同解析で用いた英語辞書を利用すれば、対訳抽出は容易であると考えられる。逆に、日本語から英語に輸入された語は、固有名詞以外にはあまり多くない。そこで、本論文では、固有名詞対訳を抽出することを目標に設定する。

基本的な構成を図 1 に示す。

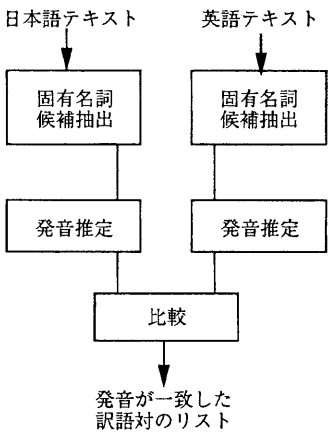

図 1: 発音情報を用いた固有名詞対訳抽出システム

処理の概要は以下の通りである。

1. 両コーパスから形態素解析等の手段によって、固有名詞候補を抽出する。
2. 抽出された固有名詞候補の発音を推定する。
3. 両言語から得られた発音を比較して、一致するものを出力する。

本方式で抽出されるのは、発音が一致する語であるから、元々どちらかの言語の言葉であったものを、他方に輸入した語である。しかし、語源がどちらの言語にあるかによって発音が異なる場合がある。例えば、“make” は英語の発音では [meik] であるが、この語が日本語からの輸入語であれば、[マケ] とな

る。そのために、日本語語源の対訳と英語語源の対訳を別々に抽出し、双方から得られた対訳リストを後に連結することとする (図2)。

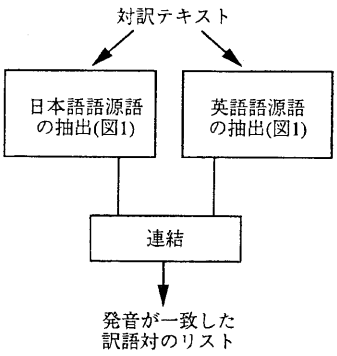

図 2: 輸入語の対訳抽出

ここまでの段階で抽出されるのは、発音が一致した部分である。しかし、固有名詞対訳辞書を作成することを考えると、発音が一致しない部分も含めて辞書登録されている方が好ましい。例えば、“三菱自動車”の英語名は“Mitsubishi Motors”である。この場合、発音一致の手法で抽出されるのは“三菱”と“Mitsubishi”であるが、固有名詞辞書には、“三菱自動車”と“Mitsubishi Motors”で登録された方が望ましいと考えられる。そのために、発音が一致した部分をキーにして、もう一度固有名詞候補を検索し、そのキーを含む最長の固有名詞候補を最終的な訳語対として出力することとする。したがって、システム全体の構成は図3の通りとなる。

次章から、

1. 日本語語源語の発音一致訳語対リストの抽出
2. 英語語源語の発音一致訳語対リストの抽出
3. 抽出された発音一致訳語対リストから、固有名詞対訳辞書の作成

の順でシステムを説明し、その後、本システムを新聞記事に適用した場合の効果を述べる。

3 日本語語源の語の対訳抽出

3.1 固有名詞候補の抽出

3.1.1 日本語テキストからの抽出

日本語の場合には形態素解析にかけることにより、固有名詞を推定する。しかし、固有名詞は形態素解析にとって未知語のものも多いため、固有名詞周辺の解析が完全であることは期待できない。また、形態素解析で既知語であっても、ある分野で特定の訳語になる場合もある。例えば「明星」は[みょうじょう]と既知語一般名詞に解析されるが、ある分野では“Meisei”(企業名)となる必要がある。そのため、以下の基準で抽出し、候補になり得るものを広く得られるようにする。

- 名詞か未知語。 … (a)
- (a) の前後に自立語か未知語がつながった単語列。 … (b)
- (b) の前後に接辞がつながった単語列。 … (c)

3.1.2 英語テキストからの抽出

英語の場合には、固有名詞は大文字で始まっていると期待できるので、以下の基準で単語及び単語列を抽出する。

- 大文字で始まっている単語。 … (a)
- (a) の連続。 … (b)

ただし、文頭の大文字語を除去するために、ピリオドの直後に現れた候補を英語辞書で引き、それが既知語であれば、候補から取り除く。

3.2 発音の推定

3.2.1 日本語固有名詞候補の発音推定

日本語の発音の推定には、Ngramを用いる方法[7]などが提案されているが、本システムでは、可能性のあるすべての発音を生成するために、漢字1文字の読みの辞書を用いて、読みを合成することとする。読みの辞書としては、NTTの日英機械翻訳システムALT-J/E[8]の日本語辞書(読み情報を含んでいる)から、漢字1文字の語を抽出し、読みの辞書とした。残念ながら、ALT-J/Eの辞書はすべての漢字を含んでいるわけではないが、漢和辞典を用いることにより、より網羅的な辞書を構築することが可能である。

前節で抽出された日本語の訳語候補を1文字ごとに分割し、それぞれの文字の読み候補を辞書から取り出す。一般に複数の読みが文字ごとに得られるが、それぞれを連結したすべての組合せを生成し読み候補とする。また、2文字め以降では濁り得る語(カサタハ行)は濁らせた候補も生成する。その他、“「づ」は「ず」に統一する”、“長音記号は取り除く”などの処理を行なう。

上記で生成された読み候補に加え、形態素解析が成功した語については、形態素解析結果の読みも候補に加える。

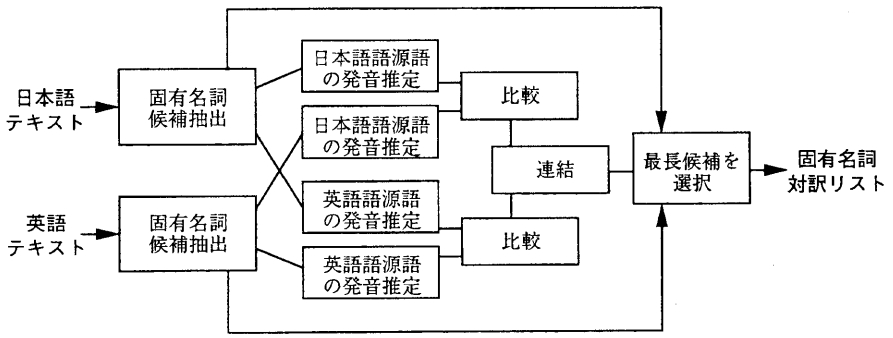

図 3: システム構成図

3.2.2 英語固有名詞候補の発音推定

もともと日本語の固有名詞が英語に翻訳された場合、英語ではローマ字で表記されることが多い。したがって、抽出された候補がローマ字表記の規則に従っていれば、日本語から来た固有名詞である可能性が強い。

本システムでは、ローマ字表記に従っている語はそれをひらがなに変換し、発音候補とする。なお、"ou" のような母音については「お」と「おう」の双方を生成し候補とする。

3.3 発音の比較

抽出されたひらがな列を比較し、一致したものを対訳リストとして出力する。

4 英語語源の語の対訳抽出

4.1 固有名詞候補の抽出

4.1.1 英語テキストからの抽出

英語の訳語候補は、3.1.2節で抽出された候補を用いる。

4.1.2 日本語テキストからの抽出

日本語の場合、輸入語はカタカナで表記されることが多い。したがって、日本語テキストからカタカナ表記語を抽出し、固有名詞候補とする。

4.2 発音の推定

4.2.1 英語固有名詞候補の発音推定

ローマ字表記に従っていない固有名詞候補は、日本語以外が語源の固有名詞、特に英語語源の固有名詞が多いと考えられる。また、ローマ字表記に偶然従っていても、それが英語語源の固有名詞である場合も考えられるので、全固有名詞候補に対して英語の発音推定を行なう。

英語の発音は綴字からある程度推測可能である [9][10] が、完全な推定は困難である。特に発音をカタカナ表記にする場合には、慣習的に表記が決まる場合もあり、決定することは不可能である。

そこで、本システムでは、子音部分のみを発音推定し、日本語の子音との比較を行なうこととする。子音部分の変換規則は、ALT-J/E で用いている、カタカナ発音辞書 [6] の 4000 語から人手で抽出した以下のようなルールを用いる。

1. 子音の前にはモーラ区切り
2. 発音する母音を抽出し、区切りを追加 (8 ルール)
3. 発音しない子音を抽出し、区切りを削除 (26 ルール)
4. 例外的な子音発音を決定 (17 ルール)
5. アルファベットをデフォルトの発音に変換

4.2.2 日本語固有名詞候補の発音推定

前節の子音列と比較できるように、抽出されたカタカナ語を子音列に変換し発音候補とする。

4.3 発音比較

抽出された子音列を比較し、一致したものを対訳リストとして出力する。

5 固有名詞対訳の抽出

前記の方法で発音が一致する単語が抽出される。しかし、固有名詞には発音が一致する部分としない

部分がある。例えば、“三菱自動車”の英語名は“Mitsubishi Motors”である。この場合、“三菱”と“Mitsubishi”は発音一致で抽出可能であるが、固有名詞辞書には、“三菱自動車”と“Mitsubishi Motors”で登録された方が望ましい。そのために、3,4章で抽出された語を含む、最長の固有名詞候補(3.1.1節, 3.1.2節で抽出されたもの)を固有名詞対訳リストとして出力する。

6 実験

本方式の有効性を確かめるために、プロトタイプシステムを作成し、日本経済新聞の市況速報文からの固有名詞対訳抽出を試みた。それぞれの記事は、記事単位で1対1に対応づけられている。

実験では、それぞれの記事のペアに対して本方式を適用し、固有名詞対訳リストを得た。したがって、複数の記事に出現する固有名詞対は複数回抽出されることになる。もし、同一の日本語に対して異なる英訳語が抽出された場合、頻度の高い方を優先する。頻度が同じであれば、長い英訳語を優先することにする。逆に、同一の英語に対して異なる日本語訳語が抽出された場合も同様である。

6.1 再現率の測定

まず、再現率を測定するために、18記事114文(日本語の文数)に対して適用した。それぞれの記事は、記事単位で1対1に対応づけられているが、それぞれの内容は、表1に示すように、完全には一致していない。

本コーパスに適用した結果を表2に示す。

記事番号	1	2	3	4	..	計
日本語記事中の数	17	18	16	14	..	144
英語記事中の数	10	8	10	11	..	97
一致する数	8	7	8	9	..	87

表 1: 対象記事中の固有名詞数

	固有名詞数
記事中の固有名詞総数	87 (異なり 61)
日本語語源法(3章)で抽出	35 (異なり 25)
英語語源法(4章)で抽出	22 (異なり 16)
抽出された総数	54 (異なり 38)

表 2: 抽出された固有名詞数

(1) 日本語でもアルファベットのもの (SANKYO=SANKYO)	2
(2) アルファベットの読み (アイジーエス=IGS)	1
(3) 日本語読み推定失敗	2
(4) 英語発音推定失敗	5
(5) 形態素解析失敗	3
(6) 文頭にあったため固有名詞抽出失敗	1
(7) 英語の略語 (東証=TSE)	2
(8) 日本語の略語 (邦チタ=Toho Titanium)	3
(9) 発音でない対訳 (朝鮮半島=Korean Peninsula)	4
計	23

表 3: 抽出されなかったもの

抽出された総数が両方法の和になっていないのは、カタカナ表記の日本企業名が両手法で抽出されているためである。

したがって、本手法の再現率は、出現総数で見ると、

$$54 / 87 = 62\%$$

異なり語数で見ると

$$38 / 61 = 62\%$$

である。

抽出できなかった33語(異なり23語)の内訳は、表3の通りである。

6.2 適合率の測定

上記の実験の範囲では、誤って抽出された例はなく、適合率100%であった。適合率を測定するために、対象文を増やし、253記事1638文(日本語の文数)で実験した。本実験で抽出された語の例を表4に示す。

抽出された固有名詞の数は、表5の通りである。異なり語数で適合率を計算すると、

$$167 / 170 = 98.2\%$$

であった。

誤って抽出された3例は、表6の通りである。(a)例の場合“東京外為市場”に対する訳語はなく、“Tokyo Electric Power”の訳語は“東電”であった。また、“Nippon Steel Semiconductor”の訳語は“日鉄セミコン”であった。

日本語	英語
アイコー	Aiko
アコム	Acom
アドヴァン	Advan
アムウェイ	Amway Japan
アルプス	Alps Electric
沖電気	Oki Electric Industry
三菱石	Mitsubishi Oil
紀州紙	Kishu Paper
森田ポ	Morita Fire Pump
日興証券	Nikko Securities

表 4: 抽出された固有名詞の例

	固有名詞数
抽出総数	607 (異なり 170)
(うち日本語語源法)	441 (異なり 131)
(うち英語語源法)	211 (異なり 61)
正解	604 (異なり 167)

表 5: 抽出された固有名詞数

7 議論

再現率については、表3のうち、(1) と (2) は抽出は容易と考えられる。また、(3) と (4) については漢和辞典の導入などにより改良の可能性がある。(5) は固有名詞の途中で文節が切れてしまったために抽出に失敗したものである。固有名詞候補抽出範囲を改良することにより救えると考える。したがって再現率は、

$$51 / 61 = 84\%$$

まで、改善の可能性がある。

(7)-(9) は本手法では本質的に抽出不可能なものである。辞書情報を用いる手法などと併用する必要があるだろう。

適合率については、かなりの高率が得られている。記事対記事の対応コーパスは自動的に得られることが示されている [4] ので、用途によっては、固有名詞対訳辞書を完全自動で作成させることも可能であると考える。

8 おわりに

本稿では、内容が完全に一致しているかどうかの保証のない対訳コーパスから、発音情報を手がかりにして、固有名詞対訳を抽出する手法を提案した。

	日本語	英語
(a)	東京外為市場	Tokyo Electric Power
(b)	日経記者	Nikkei Thursday
(c)	日本 MIC	Nippon Steel Semiconductor

表 6: 誤抽出例

本手法を用いると、再現率 62%、適合率 98% でコーパス中の固有名詞対訳が抽出できることがわかった。再現率については、84% まで改善の可能性がある。本手法により、適合率 98% が達成できたことから、対訳辞書の完全自動作成への見通しが得られた。

今後は発音推定のルール等を強化し、発音推定の精度を高めていく予定である。

参考文献

[1] 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).

[2] 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).

[3] 高尾哲康, 富士秀: 「対訳テキストコーパスからの対訳語の自動抽出」, 情報処理学会研究報告 NL115-8 (1996).

[4] 高橋大和, 白井諭, 藤波進, 池原悟, 上田洋美, 松島英之: 「DB から抽出した日英新聞記事の自動対応付け」, 言語処理学会第 2 回大会 201-204 (1996).

[5] Kenneth Church: "Char_align: A Program for Aligning Parrallel Texts at the Character Level", ACL93, p.1-8, (1993)

[6] 松尾義博, 畑山満美子, 池原悟: 「英語辞書と英文法を用いたカタカナ表記語の翻訳」, 情報処理学会 53 回大会 2-65 (1996).

[7] 鳥原信一: 「漢字 N-Gram による日本語テキストの読み付与」, 情報処理学会 53 回大会 2-37 (1996).

[8] Satoru Ikehara: "Multi-level Machine Translation Method" In *Future Computing Systems, Vol.2, No.3* (1989)

[9] 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)

[10] 堀内雄一, 山崎一生: 「英単語のアルファベット表記から仮名表記への変換」, 情報処理学会研究報告 NL79-1 (1990)