日本語外来語に対する中国語訳語の自動抽出

後藤 功雄　　田中 英輝
NHK 放送技術研究所

1 はじめに

NHK では，中国語を含む 22 の言語でニュースを放送している. 中国語のニュースは日本語のニュースを翻訳者が翻訳して作成している. 翻訳の際に，人名などの固有名詞は対訳辞書に登録がないことが多く, 訳語の調査は時間のかかる作業となっている. この固有名詞の日中対訳辞書を自動的に最新のニュース記事から生成できれば, 翻訳支援に有効である.

そこで, 日本語ニュース記事中の外来固有名詞のカタカナ表現を元にして, 対訳の中国語ニュース記事から対応する中国語訳語を自動的に抽出し, 対訳辞書を作成する手法を提案する. 翻字や音訳により翻訳された外来語は言語が異なっても発音が類似しているという特徴がある. 本手法では，この特徴を利用して，日本語のカタカナ表現と中国語の漢字表現を発音表現に変換して, 発音表現が類似の中国語表現を抽出する.

外来語として他の言語で表現した場合には, 発音は類似したものとなるが, 言語により母音や子音の種類や数などの発音体系が異なる場合が多く，また，表音文字でない場合は，発音の推定が容易ではないことが多い．そのため，２つの言語で表現された語の発音を同一の発音体系の表現に変換しても，完全に一致するとは限らない．そこで，発音の類似度の比較を精度良く行うことが重要である.

外来語が類似した発音を表現しているということを手がかりに, 対訳のコーパスから訳語の抽出する方法として, 日本語／英語[3], 日本語／韓国語[4]などの研究例がある. 松尾ら[3]は，発音表現において，全て一致したもの，または，子音列が全て一致したもののみを抽出している．この手法では, 一部が一致していない場合には抽出できない. 金ら[4]は，発音表現間の類似度を計算して訳語を抽出している．彼らは，二つの発音表現の文字列間の類似度として挿入・置換・削除による最小差異数を正規化した値を用いている.

本手法は，金ら[4]と同様に挿入・置換・削除の編集距離を基にして類似度を定義する．ここで，編集操作の回数以外に，表現の長さに応じた編集コスト，コーパスから学習した挿入と削除のコストの比率, 置換する文字に応じた置換コストを導入して編集距離のコスト値を補正することにより，精度を向上させる．また，本稿では，編集距離が近い上位の表現を抽出する方法も提案する．以下，２章で提案手法について説明し，３章で本手法を用いた実験の結果と今後の課題を示し，４章で結論を述べる.

2 提案手法

2.1 手法概要

本稿の問題設定は, 日本語ニュース中のカタカナで表現された外来固有名詞の訳語を対訳の中国語ニュース記事から抽出することである.

提案手法の概要を図１に示す. はじめに, 両方の言語の表現を文字または部分文字列単位で変換テーブルを用いて発音表現に変換する．さらに，発音表現間の類似度が高い中国語表現を訳語の候補として抽出する. 類似度が等しい場合は，出現記事数の頻度で順位付けする.

本手法は，全て文字単位で処理を行うため，中国語の形態素解析を必要としない. 次に各処理の詳細について説明する.

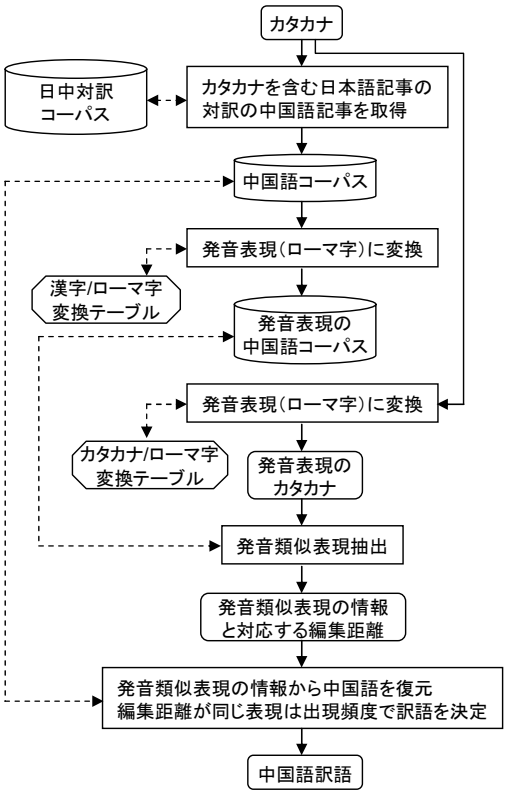

図１ 概要図

2.2 発音表現への変換

発音表現にはローマ字[1]を用いる.

中国語の漢字をローマ字に変換する際には, 漢字１文字に対応するローマ字が登録された変換テーブル[2]を用いて１文字単位で変換する. 漢字１文字に対して変換候補が複数ある場合は，全ての変換結果を用いる.

カタカナをローマ字に変換する際には, 変換テーブルを用いて１文字または部分文字列単位で一意に変換する.

2.3 編集距離を基にした類似度の定義

類似度を編集距離の値により定義する. 編集距離のコス

---

[1] 訓令式を用いた．ただし，一部例外の規則を追加した．変換の際に，長音「ー」，促音「ッ」は削除した.

[2] 日中韓辞典研究所発行の２つのテーブル「漢字１文字／ピンイン」と「ピンイン／カタカナ」から作成した.

トの値が小さい程，類似度が高いとする．

ここで，編集距離は，中国語の発音表現を日本語の発音表現と同じになるように編集する際のコストが最小となるコストとして定義する．本手法で用いる編集の種類と対応するコストは次の３種類を用いる．

- 文字の追加，コスト= *CostInsert*
- 文字の削除，コスト= *CostDelete*
- 文字の置換，コスト= *CostReplace*(*char*1, *char*2)

ここで，*char*1 と *char*2 は，置換する文字を示す．

本手法では，編集距離を補正することにより，類似度判別の精度を向上させる．次に３つの補正手法について述べる．

2.3.1 表現長に応じた編集コスト

長い文字列中での１文字の編集と短い文字列中での１文字の編集では，文字列全体が表現する発音への影響の大きさが異なる．例えば，10 文字中の 1 文字の編集の影響は全体の 1/10，5 文字中の 1 文字の編集は全体の 1/5 の影響があると考えることができる．そこで，この影響の大きさの差を編集コストに反映させる．

編集先である日本語発音表現の文字列の長さを a 文字，編集元の中国語発音表現の文字列の長さを b 文字とする．挿入・削除・置換の各操作による合計の最小コストである編集距離 C を計算した後に，1 式により，文字列の長さを考慮したコストに変換する．

$$\text{表現長に応じた編集距離} = C \times \frac{a}{b} \qquad (1)$$

2.3.2 コーパスから学習した挿入と削除のコストの比率

同一の発音を表現することを目的としている場合でも，例えば，日本語の発音表現では短い表現になり，中国語側の発音表現では長い表現になる傾向があれば，挿入と削除のコストの比率を挿入よりも削除が起こりやすくなるように調整したほうがよいと思われる．

この比率を日中の外来語の対訳のコーパスから求める．比率の設定方法は以下の通りである．ここでは，編集の方向を「中国語の発音表現を日本語の発音表現と同じになるように編集する」としているので，はじめに中国語の発音表現を全て削除し，次に発音表現が 0 文字の状態から日本語の発音表現を生成する操作を考える．中国語の発音表現を全て削除する操作にかかるコストと，0 文字の状態から文字を追加することで日本語の発音表現を生成する操作にかかるコストは，どちらも１つの発音表現全体の編集コストと考えることができる．これらの表現全体の編集コストを等しいとすることで，挿入と削除の比率を調整する．そこで，追加と削除の比率として，2 式を用いる．

$$CostDelete \times b' = CostInsert \times a' \qquad (2)$$

ここで，a'はコーパス中の日本語発音表現の平均文字数，b'はコーパス中の中国語発音表現の平均文字数である．さらに，文字追加のコスト *CostInsert* を 1 とする．この場合，削除コストは 3 式となる．

$$CostDelete = \frac{a'}{b'} \qquad (3)$$

日本語／中国語の外国人名辞書[3]を用いて 3 式の値を求める．日本語と中国語の見出し語[4]を発音表現のローマ字に変換し，平均の発音表現文字数の a'と b'の値[5]を求めた．この結果，$CostDelete = a'/b' = 0.91$ の値を得た．

2.3.3 置換する文字に応じた置換コスト

置換する文字 *char*1 と *char*2 との発音の近さに応じた置換のコストを用いることで，発音の類似度比較をよりきめ細かく計算することができる．ここでは，経験的に置換のコストを定義したテーブルを作成して用いる．

2.4 中国語訳語の抽出

2.3 節で定義した編集距離が近い上位の類似表現を，長い表現から抽出する処理の流れを，図 2 を用いて以下に示す．本手法は，抽出する最大文字数 N[6]を指定し，それ以下の長さの任意の文字列の編集距離を全て得ることができる．ただし，図 2 では，削除，挿入，置換のコストは全て 1 とし，N は 6 としている．

(1) 日本語のローマ字と最大で N 文字の中国語の漢字を対応させた複数のテーブル（MainTable）を作成する．これらのテーブルは，開始位置が中国語の漢字で 1 文字ずつ異なるテーブルとする．ここで，各テーブルの先頭行の前にダミー[7]の行を追加する．ダミーの行頭のみコストを 0 とする．
(2) (3)の処理を MainTable の各テーブルに対して行う．
(3) (4)～(6)の処理を MainTable の各行に対して行頭から順番に行う．
(4) 中国語の漢字 1 文字に対応するローマ字表現と日本語のローマ字表現を対応させたテーブル（SubTable）を作成する．ここで，SubTable にはダミーの行として MainTable 上での 1 つ前の行の値を設定する．
(5) SubTable の各位置における編集距離のコストをダイナミックプログラミングにより計算する．
(6) SubTable の末尾行のコストを MainTable にコピーする．１つの漢字に対して SubTable が複数存在する場合は，最もコストが小さい値を MainTable にコピーする．
(7) MainTable の開始位置が中国語漢字の開始位置となり，MainTable の末尾列の位置と値が，中国語漢字の各終了位置と，対応するローマ字表現の編集距離のコストとなる．このコストを小さい順にソートして，上位の開始位置と終了位置を抽出する．
(8) 取得した中国語漢字の開始位置と終了位置から中国語表現を抽出し，コストと共に出力する．

---

[3] 日本語／英語の外国人名辞書（日外アソシエーツ発行「カタカナから引く外国人名綴り方字典」）と中国語／英語の外国人名辞書（日中韓辞典研究所発行「中英欧米人名辞書」）から作成した．

[4] 日本語側をローマ字に変換した際に 2 文字以上に変換された全ての見出し 92,407 語

[5] 中国語側で変換先に複数の候補が存在する場合は，編集距離が最小の候補を選択した．

[6] 元になるカタカナの文字数に応じて設定することにより文字数の不足を防ぐことができる．

[7] 図中では，”#”記号でダミーを示す．

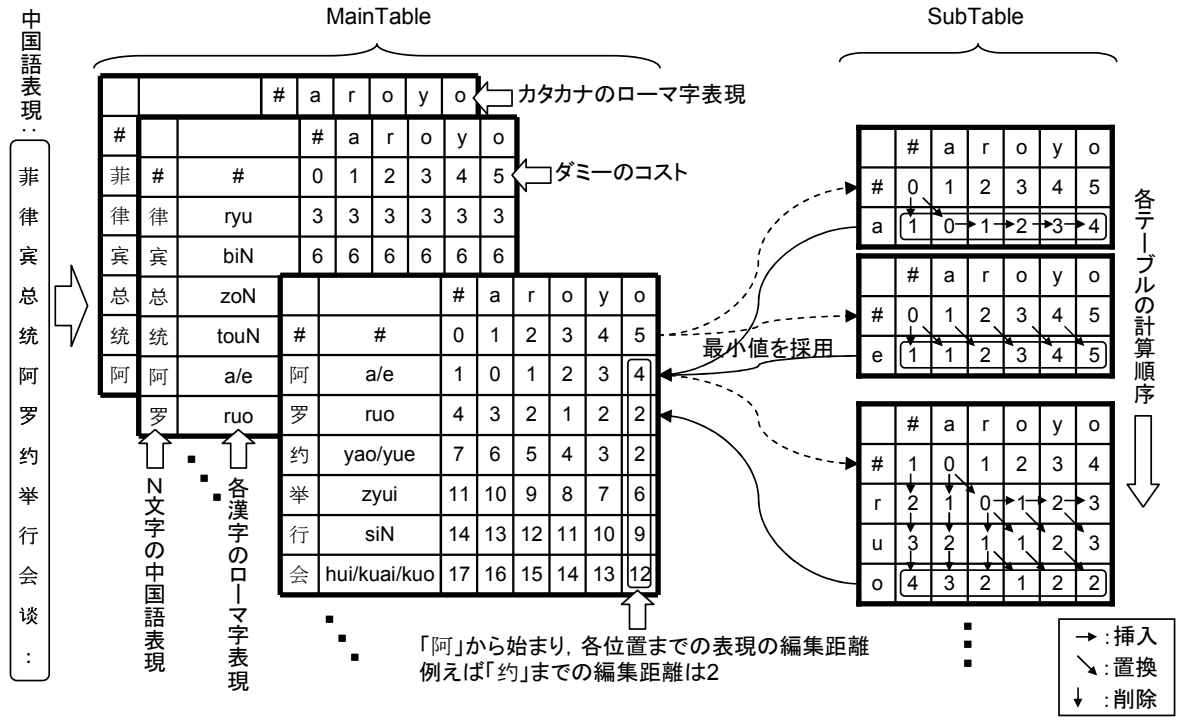

図 2　N 文字以下の任意の中国語表現の編集距離の計算

(9) コストが同じ中国語表現は，抽出した中国語表現を含む記事の数が多い順にソートする．

3 評価実験

本手法を用いて，放送に用いられた NHK の日中対訳ニュースコーパスから中国語訳語を抽出する実験を行った．対訳ニュース記事は，日本語記事を中国語に翻訳[8]して作成されたものである．2003 年 5 月から 9 月までの 5 ヶ月間の記事中で，音訳により中国語に翻訳された日本語記事中のカタカナで表現された 203 の外国人名[9]を日本語側の単語として用いた．実験に用いた置換スコアは付録に示すテーブルを用いた．このテーブルに存在しない文字の置換コストは全て 1 とした．抽出する漢字の最大文字数 N は 10 とした[10]．句読点やアラビア数字などの人名に用いられない文字を区切りとして，中国語表現を分割して処理を行った．実験に用いた手法 A,B,C,D,E の内容を表 1 に，実験の結果を表 2 に示す．なお，手法 A は，挿入・削除・置換の編集コストを全て 1 とした．また，実験で抽出に成功した訳語と日本語表現の例を表 3 に示す．

表 1 各手法の内容

手法名	A	B	C	D	E
表現長に応じた編集コスト導入	×	○	×	×	○
挿入と削除のコスト比率導入	×	×	○	×	○
文字に応じた置換コスト導入	×	×	×	○	○

表 2 中国語訳語抽出の正解含有率(%)

手法名	A	B	C	D	E
1 位	66.2	77.0	75.3	68.6	79.7
2 位以上	79.1	83.0	81.1	79.5	83.9
3 位以上	80.8	85.8	82.7	82.9	87.0
5 位以上	83.9	88.2	86.9	86.7	88.7
10 位以上	86.4	91.1	89.7	90.2	92.6

表 3 中国語訳語抽出成功結果例

日本語	中国語抽出結果
アウン・サン・スー・チー	昂山素季
アロヨ	阿罗约
アーミテージ	阿米蒂奇
ウォルフォウィッツ	沃尔福威茨
エカテリーナ	叶卡特林娜
シャロン	沙龙
ドビルパン	德维尔潘
ハッサン	哈桑
ビンラディン	本拉登

[8] 直訳ではなく意訳されており，情報の削除や追加などがある[1].

[9] 漢字を母国語に持たない言語の人名．中国，日本，朝鮮が出身の人以外のほぼ全て．中国語への音訳の際に，対応するカタカナにおいて「ー」「ッ」を除いて 2 文字以上連続して発音が省略された語は除いた（2 語）．

[10] 正解訳語の最大文字数が 8 であったため，それよりも十分大きな値とした．

本手法(E)の実際のニュース記事からの日本語外来語に対応する中国語訳語抽出結果は, 第１位の抽出結果の正解率は 79.7%であり，上位 10 位以上の正解含有率は 92.6%であった.

第１位の正解率は, 全ての編集コストを１とした手法 A に比べて，表現の長さの情報を考慮した手法 B は 10.8%精度が向上し, 挿入と削除のコスト比率を導入した手法 C は 9.1%精度が向上し，置換する文字に応じた置換コストを導入した手法 D は，2.4%精度が向上し，これらの３つの補正を導入した手法 E は 13.5%精度が向上した．これにより，「表現長に応じた編集コスト」,「挿入と削除のコストの比率」,「文字に応じた置換コスト」はそれぞれ有効であることが確認された.

１位が正解でなかった結果は，主に以下の３種類に分類することができる.

１）抽出対象部分の先頭の文字または末尾の文字の発音表現が一致しない場合

先頭や末尾の１文字の抽出漏れや余分な抽出が起こる

＜例＞エルナンデス　　抽出結果：尔南德斯
　　訳語：埃尔南德斯

テネット　　抽出結果：特纳德发
　　訳語：特纳德

２）記事中の訳語ではない文字列よりも発音表現が類似していなかった場合

元になるカタカナの表現が短い場合は発音が類似した多くの表現が特に抽出されやすい

＜例＞フィッシャー　　抽出結果：会上
　　訳語：菲舍尔

ヨー　　抽出結果：这
　　訳語：杨

３）中国語への音訳の際に，発音の一部が省略された場合

＜例＞クリスターズ　　抽出結果：克利斯特首
　　訳語：克利斯特

（対応する中国語側の発音表現はクリスタのみ）

１），２）で正解が１位にならない原因は，カタカナを発音表現に変換した結果と比べて, 中国語の漢字を発音表現に変換した結果があまり類似していない表現となったことである. 文字に応じた置換のコストを大規模に取り入れることによりある程度対応できると思われる．また，中国語側で, 辞書を用いて固有名詞以外の単語を除いて抽出対象範囲を限定することにり, エラーを削減することができると思われる．３）への対応は，コーパスから音訳時に削除されやすい部分を抽出して, その部分の挿入コストを下げるなどである程度対応できると思われる.

4 おわりに

本研究では, 日本語と中国語の対訳ニュースコーパスから, 日本語記事中のカタカナで表現された外来語に対する中国語の訳語を抽出する手法を提案した．本手法により，毎日作成される日中のニュース記事データから外来固有名詞の対訳関係を自動的に取得する見通しを得た. 本手法の結果は, 辞書の自動構築や, 用例提示による翻訳支援[2]の際に有効である. 実際のニュース記事コーパスを利用して, 本手法による中国語訳語を抽出する実験を行った結果, 正解訳語抽出率 79.7%を得た.

今後は, コーパスから各文字に応じた最適な編集コストを求めて用いることにより, より細かい精度で類似度を計算できるように手法の改良を行う予定である.

参考文献

[1] Isao Goto, Naoto Kato, and Terumasa Ehara, 2001, "*A multilingual news database and its application to a translation memory system*," NLPRS Post-Conference Workshop on Language Resources in Asia, pp.1-6.

[2] Isao Goto, Naoto Kato, Noriyoshi Uratani, Terumasa Ehara, Tadashi Kumano, and Hideki Tanaka, 2003, "*A Multi-language Translation Example Browser*," Machine Translation Summit IX, pp.463-466.

[3] 松尾義博，白井　諭，1996，“発音情報を用いた訳語対の自動抽出，”情報処理学会研究会報告，NL-116-15，pp.101-106.

[4] 金　玉錦，藤井　敦，石川徹也，2003，“韓国語コーパスからの外来語自動抽出と言語解析への応用，”言語処理学会第９回年次大会，pp.258-261.

付録 置換コスト

置換する文字		コスト
p	b	0.6
k	g	0.7
s	z	0.7
t	d	0.7
m	n	0.8
h	p	0.9
h	b	0.9
h	s	0.8
h	z	0.9
t	z	0.8
g	z	0.7
z	d	0.7
g	d	0.8
p	d	0.9
a	i	0.9
a	u	0.8
a	e	0.8
a	o	0.8
i	u	0.8
i	e	0.8
i	o	0.9
u	e	0.8
u	o	0.8
e	o	0.8

置換元と置換先の区別は無し