日韓対訳辞書作成のための表音的類似性に基づく表記照合

東芝ソリューション株式会社　プラットフォームソリューション事業部参事　熊野　明

PROFILE

1982 年東京工業大学卒業。同年東京芝浦電気（株）（現、（株）東芝）入社。2010 年から東芝ソリューション（株）プラットフォームソリューション事業部参事。自然言語処理システムの研究開発に従事。AAMT/Japio 特許翻訳研究会委員。2007 年度から特許版・産業日本語委員会委員。

kumano.akira@toshiba-sol.co.jp　TEL

1 はじめに

日韓機械翻訳用対訳辞書の開発には、日本語見出し語に対して韓国語訳語を調べて付与する方法と、対訳文書から半自動的に対訳知識を抽出する方法 [1] がある。

対訳文書からの対訳知識抽出には、日本語文と韓国語文の単語単位での対応関係の推定が必要である。一般的には、両言語における語順の類似性や格助詞の対応などを利用することができるが、技術文書に頻出する外来語の名詞の場合、日本語、韓国語ともに音訳で表現しているものが多く、表音的類似性があることが分かっている。

これは、韓国語の外来語の表現方法の特性によるものである。ハングルでは日本語のカタカナ語のように、原語である英語などの発音をハングルの発音に置き換え、つまり音訳して外来語表記としているものが多い [2]。原語が同じ場合、日本語のカタカナ表記と韓国語のハングル表記は、自ずと発音が類似している。たとえば、「ネットワーク (network)」に対するハングル表記「네트워크」の発音記号は [ne:tʰɯwɔ:kʰɯ] で、日本語と非常に類似している。

これまで、英語のスペルと日本語のカタカナ表記の照合プログラムを開発して、日英対訳辞書開発に使ってきた [3]。今回これを拡張して、ハングル表記とカタカナ表記との照合プログラムを作成した。また、この機能を日本語と韓国語の特許文書の文アラインメントに利用することで、有効性を検証した。

2 ハングル表記とカタカナ表記の照合

ハングル表記とカタカナ表記を、その表音的類似性に基づいて文字単位に照合する。英語のスペルと日本語のカタカナ表記の照合プログラムを拡張して開発したものである。

2.1. 処理方法

処理 1：ハングル文字を字母に分解し、疑似英語スペルに変換して単語単位で結合する

処理 2：処理 1 の出力を英語スペルとみなし、ハングル・カタカナ照合用の疑似ローマ字に変換する

処理 3：カタカナ (促音、長音を除く) を 1 文字ごとに疑似ローマ字に置き換える。

処理 4：ハングルの疑似ローマ字の先頭から 1 文字ごとに、カタカナの疑似ローマ字と照合する。

2.1.1. ハングル文字から疑似英語スペルへの変換

ハングル文字を字母に分解し、疑似英語スペルに変換して単語単位に結合する。

(例)

① 레인지 (= レンジ)

(1) 字母への分解：

ㄹ + ㅖ , ㅇ + ㅣ + ㄴ , ㅈ + ㅣ

(2) 疑似英語スペルに変換： le + in + ji
(3) 単語に結合： leinji

② 가솔린 (= ガソリン)
(1) 字母への分解：
ㄱ + ㅏ , ㅅ + ㅗ + ㄹ , ㄹ + ㅣ + ㄴ
(2) 疑似英語スペルに変換： ga + sor + rin
(3) 単語に結合： gasorrin

③ 필름 (= フィルム)
(1) 字母への分解：
ㅍ + ㅣ + ㄹ , ㄹ + ㅡ + ㅁ
(2) 疑似英語スペルに変換： pir + reum
(3) 単語に結合： pirreum

④ 쇼핑 (= ショッピング)
(1) 字母への分解：
ㅅ + ㅛ , ㅍ + ㅣ + ㅇ
(2) 疑似英語スペルに変換： syo + ping
(3) 単語に結合： syoping

ハングル文字から疑似英語スペルへの変換テーブルの一部を図１に示す。

" 가 " = "g"+"a"
" 각 " = "g"+"a"+"g"
" 간 " = "g"+"a"+"n"
" 갇 " = "g"+"a"+"d"
" 갈 " = "g"+"a"+"l"
" 갉 " = "g"+"a"+"lg"
" 갊 " = "g"+"a"+"lm"
" 감 " = "g"+"a"+"m"
" 갑 " = "g"+"a"+"b"
" 값 " = "g"+"a"+"bs"

図１　ハングル文字から疑似英語スペルへの変換テーブル（一部）

2.1.2. 疑似英語スペルから疑似ローマ字への変換

処理１の出力を英語スペルとみなして疑似ローマ字に変換する。英語のスペルと日本語のカタカナ表記との照合プログラムと同様の方式である。疑似英語スペルを、音節[1]単位に分解し、さらに子音部分と母音部分に疑似ローマ字を割り当てる。ただし、ハングルの発音と英語の発音の対応には曖昧性があるので、既存の英語の対応表では不十分である。例えば、p は P だけでなく F にもなる。

(例)
① leinji → lei/n/ji
→ R[EI¦AI¦E¦I] + N + J[I¦AI]
② gasorrin → ga/so/rri/n
→ [G¦J][A¦EI¦E¦YA¦O] + [S¦Z][ORU¦O¦A]
+ R[I¦AI] + N
③ pirreum → pi/rreu/m
→ [P¦F]I + R[U¦O¦E] + M
④ syoping → syo/pi/ng
→ [S¦SH¦Z][YOU¦YO] + [P¦F¦B¦H][II¦I¦E¦U] + [NG¦N]

疑似英語スペルから疑似ローマ字への変換テーブルの一部を図２に示す。

"b" : ("B", "P", "F", "W")
"bb" : ("B", "P", "F")
"ch" : ("CH", "TS")
"d" : ("D", "T", "Z", "S")
"dd" : ("D", "T", "Z")
:
"a" : ("A", "O")
"ae" : ("A", "YA")
"e" : ("EE", "E")
"ei" : ("EI", "E")
"eo" : ("A", "O", "EO", "E")
"eu" : ("U", "O", "E", "I")

図２. 疑似英語スペルから疑似ローマ字への変換テーブル（一部）

1 入力が英語のスペルではないので、ここでは厳密な意味での音節ではなく、疑似ローマ字を割り当てるための処理単位である。

2.1.3. カタカナから疑似ローマ字への変換

カタカナ文字を 1 文字ずつ疑似ローマ字に変換する。この変換は、英語スペルとカタカナ表記との照合で行っている処理と同じである。各文字に可能性のある疑似ローマ字を定義した変換テーブルを利用する。促音や長音は省略し、拗音は全体で 1 音と扱う。

カタカナから疑似ローマ字への変換テーブルの一部を図3に示す

" サ " : ("SA")
" ザ " : ("ZA")
" シ " : ("SI", "SHI")
" シェ " : ("SHE", "SYE")
" シャ " : ("SHA", "SYA", "TYA", "SIA")
" シュ " : ("SHU", "SH", "SYU")
" ショ " : ("SHO", "SYO", "TYO")
" ジ " : ("ZI", "JI", "J")
" ジェ " : ("JE")
" ジャ " : ("JA", "DYUA")
" ジュ " : ("JU", "ZYU", "J", "DYU")
" ジョ " : ("JO", "ZIO", "TIO")

図3. カタカナから疑似ローマ字への変換テーブル（一部）

2.1.4. 疑似ローマ字の照合

ハングル表記の疑似ローマ字と、カタカナ表記の疑似ローマ字を、音節ごとに照合する。ハングル疑似ローマ字 1 音節に対してカタカナ疑似ローマ字 1 音節が照合しない場合は、カタカナ 2 ～ 4 音節分で照合を試みる。

(例)

① **레인지** : レンジ

[ハングル表記の疑似ローマ字]

R[EI¦AI¦E¦I] + N + J[I¦AI]

[カタカナ表記の疑似ローマ字]

[RE¦REI] + N + JI

照合： [RE=RE], [N=N], [JI=JI]

② **가솔린** : ガソリン

[ハングル表記の疑似ローマ字]

[G¦K][A¦O] + [S¦SH¦Z][OU¦O¦A] + [R¦RR][I¦I¦E¦U] + [N¦NN]

[カタカナ表記の疑似ローマ字]

GA + SO + RI + [N¦M]

照合： [GA=GA], [SO=SO], [RI=RI], [N=N]

③ **필름** : フィルム

[ハングル表記の疑似ローマ字]

[P¦F¦B¦H][I¦I¦E¦U] + [R¦RR][U¦O¦E¦I] + [M¦MM]

[カタカナ表記の疑似ローマ字]

FI + [RU¦R] + [MU¦M]

照合： [FI=FI], [RU=RU], [M=M]

④ **쇼핑** : ショッピング

[ハングル表記の疑似ローマ字]

[S¦SH¦Z][YOU¦YO] + [P¦F¦B¦H][I¦I¦E¦U] + [NG¦N]

[カタカナ表記の疑似ローマ字]

[SHO¦SYO¦TYO] + PI + [N¦M] + [GU¦G]

照合： [SYO=SYO], [PI=PI], [NG=N+G]

2.2. 評価

インターネットで公開されていたハングル単語とカタカナ表記の対応データ [4] を利用して、今回開発した関数が正しく照合を判断できるか確認した。データの例を以下に示す。

가든파티 ; ガーデンパーティー
가솔린 ; ガソリン
가스레인지 ; ガスレンジ
가이드 ; ガイド
가제 ; ガーゼ
가톨릭 ; カトリック
간다라 ; ガンダーラ
갈라파고스 ; ガラパゴス
갈리마르 ; ガリマール
개그 ; ギャグ
개스트 ; ゲスト
개임 ; ゲーム

2.2.1. 評価結果

評価に使った対訳データは、1,483 対である。そのうち、照合できたものは 1,370 対であった。照合精度は約 92% である。

3 文アラインメントへの応用

ハングル表記とカタカナ表記の照合の効果を実証するため、日韓対訳特許文書を利用して文アラインメントの実験を行った [5]。

3.1. 表音的類似性を用いた文アラインメント

本実験で行った文アラインメントの処理手順を図４に示す。日韓対訳文書から日本語文および韓国語文を抽出し、文の類似度を算出する。類似度の算出には、(1) 文字数の比、(2) 外来語の一致度、(3) 格助詞の一致度の３つの特徴量を用いる。

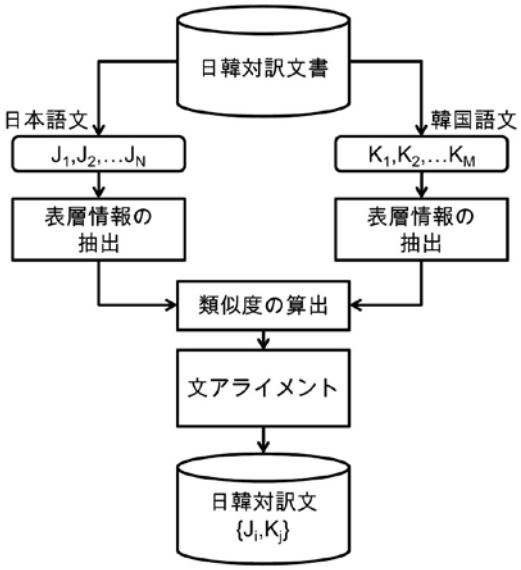

図４．　日韓文アラインメントの処理手順

これら３種類の特徴量はすべて表層的に求めることができ、対訳辞書などの言語資源を必要としない。文アラインメントでは、文の類似度が最大になるような対応関係を動的計画法 [6] により決定する。以下、それぞれの特徴量について説明する。

3.1.1. 文字数

日本語と韓国語は、文法が類似しており、対訳文間の文字数の比は変化が少ない。文字数による類似度 (Character Similarity: CS) は、次式によって求める。

$$CS = \min\left(\frac{\alpha\, n_J}{n_K}, \frac{n_K}{\alpha\, n_J}\right)$$

ここでn_J, n_Kはそれぞれ、日本語文と韓国語文に含まれる文字数を表す。 α は２言語間の文字数の比を表す係数である。事前のデータ分析により、 $\alpha = 1.1$ とした。

3.1.2. 外来語

第２章で紹介した方法に基づいて、韓国語の単語（ハングル表記）と日本語の外来語（カタカナ表記）の照合を行った。

ここでは、照合が成功した外来語の頻度を基にして文の類似度を算出する。日本語文に現れるカタカナ表記の出現頻度を要素とするベクトルを $\mathbf{f}_J = (f_{J,1}, f_{J,2}, f_{J,3}, \cdots)$ 、それぞれの外来語と照合した韓国語文のハングル表記の出現頻度を要素とするベクトルを $\mathbf{f}_K = (f_{K,1}, f_{K,2}, f_{K,3}, \cdots)$ とし、外来語による類似度 (Katakana Similarity: KS) は、次式によって求める。

$$KS = \frac{\mathbf{f}_J \cdot \mathbf{f}_K}{\|\mathbf{f}_J\|\|\mathbf{f}_K\|}$$

3.1.3. 格助詞

韓国語には、日本語と同様に格助詞が存在するが、その対応関係は一対一とはならない。特に、「に」や「で」に対応する韓国語の格助詞は、文節間の修飾関係や係る名詞の意味分類（人、場所、時間など）によって、その訳が異なる。表層的に訳し分けを解析することは困難であるので、表１に示す対応表を用いて格助詞のラベル付けを行った。

日本語	韓国語
が	이 , 가 , 지만
は , とは	은 , 는
を	을 , 를
で , に , へ	에 , 에다 , 다 , 에게 , 에게다 , 한테 , 에서 , 으로 , 로
の	의
から	에서부터 , 으로부터 , 로부터 , 부터

表１　日韓格助詞の対応（一部）

日本語文および韓国語文中に現れる格助詞のラベルを文字列$(\mathbf{p}_J, \mathbf{p}_K)$とし、格助詞による類似度 (Particle Similarity: PS) は、次式によって求める。

$$PS = \exp\left(-\frac{LD(\mathbf{p}_J, \mathbf{p}_K)^2}{2\sigma^2}\right)$$

ここで、LD は表 1 に示すラベルの挿入、削除、置換を 1 回の編集作業とした編集距離 (Levenshtein Distance) を表し、σ は正規化係数である。

なお、日本語の格助詞は日英機械翻訳システムの形態素解析エンジンを用いて抽出したが、韓国語の格助詞は表 1 に示す格助詞と一致するハングル文字を機械的に抽出したため、精度は十分ではない。

3.1.4. 文の類似度

以上の表層的な特徴量を統合し、日本語文と韓国語文の類似度を次式で定義した。

sim(J,K) = wc · CS + wk · KS + wp · PS

ここで、wc, wk, wp は、それぞれの特徴量に対する重み付け係数である。

3.2. アラインメント実験とその結果

3.2.1. 日韓対訳特許文書

提案する文の類似度を日韓対訳特許文書（PCT 出願された日本語特許文書と韓国語特許文書）の文アラインメントに適用した。この対訳特許文書は、日本語特許公報の内容を韓国語に翻訳したものであり、文中に専門用語などの外来語が多く出現する。また、それぞれの特許文書は、分野・背景・課題・解決手段などの段落を明示した XML 形式なので、タグの対応を利用することにより段落単位でのアラインメントが可能である。

3.2.2. クローズドデータでの実験

日韓対訳特許文書 P_{close}（クローズドデータ：日本語 413 文、韓国語 486 文、文対応 392 対）に対して文アラインメントを実施した。動的計画法では、(日本語文 : 韓国語文) = (1:1), (1:2), (1:3), (2:1), (3:1), (2:2) の文対応を制約条件として、最適化を行った。また、事前の実験により重み付け係数を、wc = 1.0, wk = 0.4, wp = 0.2 とした。

文アラインメント結果の例を図 5 に示す。(J1):(K1) および (J2):(K2+K3) は正しい対応が得られたが、(J3+J4):(K4) は誤っている。正しい対応は (J3:K4) である。このような誤対応となる例では、文字数による類似度に比べ、格助詞による類似度の値が小さくなることが多く、格助詞対応テーブルの改善が必要である。

また、XML タグを利用した段落アラインメントを行った場合と行わなかった場合において、文アラインメントを行った結果を図 6 に示す。3 種類全ての特徴量

<table>
<tr><td>J1</td><td>そのため、ネットワークプリンタには、セキュリティ機能を設けることが要望されている。</td><td>K1</td><td>이것은 네트워크 프린터에 보안 기능을 제공할 필요성을 일으킨다 .</td></tr>
<tr><td rowspan="2">J2</td><td rowspan="2">第 1 に、サイズが膨大な印刷データが一度に複数集中すると、プリンタがデータを受信するネットワーク処理部に負荷がかかり、プリンタにつながりにくくなる等の問題が発生する可能性がある。</td><td>K2</td><td>첫 번째로 , 대단히 큰 프린트 데이터 크기를 각각 갖는 복수의 작업이 동시에 집중되면 , 프린터가 데이터를 수신하는 네트워크 처리부에 너무 많은 부하가 인가된다 .</td></tr>
<tr><td>K3</td><td>이것은 아마도 프린터에 접근하기 어렵다는 문제점을 일으킬 수 있다</td></tr>
<tr><td>J3</td><td>不一致の場合、例えば、照合情報に国コードを含め、国コードとして日本を指定した場合は、日本以外の国では出力を行うことができる構成となる。</td><td rowspan="2">K4</td><td rowspan="2">불일치의 경우에 , 예를 들어 , 상기 검증 정보에 국가 코드가 포함되고 , 상기 국가 코드에 일본이 특정되면 , 일본 이외의 국가에서 출력이 수행될 수 있는 방식으로 출력 시스템이 구성될 수 있다 .</td></tr>
<tr><td>J4</td><td>また、所定演算式としては、公開暗号方式やハッシュ関数を用いることができる。</td></tr>
</table>

図 5　提案手法による文アラインメント結果（例）

を組み合わせた場合 (CS/KS/PS) における精度 (F 値) は、段落アラインメントを行わない場合の 35.5% に対して、行う場合に 81.0% であった。段落アラインメントを行わなかった場合では、外来語による類似度によって精度が大きく向上しており、段落アラインメントが困難な状況においても外来語による類似度が効果的であることが分かる。

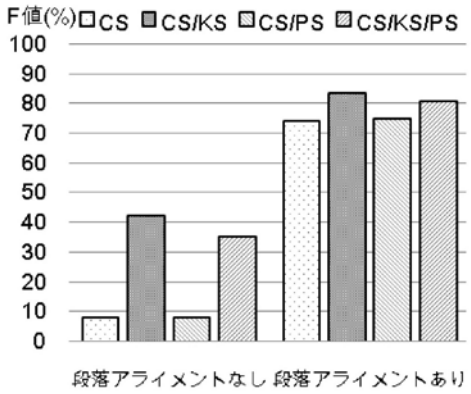

図 6. クローズドデータでの文アラインメント精度

3.2.3. オープンデータでの実験

クローズドデータでの実験で決定したパラメータを用いて新たな日韓対訳特許文書 P_{open}（日本語 1,286 文、韓国語 1,602 文、文対応 1,206 対）に対して文アラインメントを実施した。オープンデータでの実験では、CS のみの精度をベースラインとし、CS/KS/PS の精度を評価した。段落アラインメントのない場合は十分な精度が認められなかったが、段落アラインメントを行った場合は、ベースラインの 42.3% に比べ 3.6 ポイント向上し、45.9% の精度が得られた。表音的類似性を用いた日韓表記照合の文アラインメントに対する有効性が確認できた。

4 まとめ

表音的類似性に基づいて、ハングル表記とカタカナ表記の照合機構を開発した。この照合機構を利用して、日韓文アラインメント手法を提案した。日韓対訳特許文書に適用し、クローズテストにおいて 81.0%、オープンテストにおいて 45.9% の文アラインメント精度を確認した。既存の対訳知識を利用しなくても、日韓の用語を対応付けることが可能になり、文アラインメントの精度を向上させることができた。

[参考文献]

[1] 熊野 , 平川 : 対訳文書からの機械翻訳専門用語辞書作成 , 情報処理学会論文誌 , Vol.35, No.11, pp.2283-2290, (1994)

[2] 蓮池 : 蓮池流韓国語入門 , 文芸春秋 (2008)

[3] 熊野 : カタカナ表記からの英訳推定による専門用語辞書作成 , 言語処理学会 第 1 回年次大会 , pp.221-224, (1995)

[4] 韓国語外来語辞典 , http://www.geocities.jp/mfutatsugi/katakana.htm

[5] 園尾 , 熊野 : 外来語の表音的類似性を利用した日韓文アライメント , 言語処理学会第 16 回年次大会 , pp.482-485, (2010)

[6] 宇津呂 , 松本 : 対訳辞書および統計情報を用いた二言語対訳テキスト照合 , コンピュータソフトウェア , Vol.12, No.5, pp.12-21, (1995)