外来語の表音的類似性を利用した日韓文アライメント

園尾　聡　　熊野　明
株式会社東芝 研究開発センター
{satoshi.sonoo, akira.kumano}@toshiba.co.jp

1. はじめに

文アライメントは、二言語間で文レベルの対応関係を同定する手法であり、文対応の取れた対訳コーパスを構築する上で非常に重要である。一般的な文アライメントでは、文字数比などの統計情報[1]が利用されるが、より精度を向上させるためには、対訳用語辞書が不可欠となる。しかしながら、豊富な対訳用語辞書を整備するには多大なコストを必要とし、また、未知語になりやすい外来語を全て登録することは難しい。

一方、日本語と韓国語の間には、同一言語から移入した外来語の発音が音韻的に類似しているという特徴がある。この特徴を利用し、ローマ字に変換した日本語の外来語と、ハングル文字の発音記号をローマ字に変換した韓国語とを比較することで、韓国語中の外来語を抽出する研究がなされている[2]。このような比較手法では、表音的な類似性を利用するので、対訳用語辞書を用いずに単語の照合が行うことが可能となる。

本研究では、日本語と韓国語の表層的な類似性を利用した文アライメント手法を提案する。特に、表音的な類似性を基にして外来語の照合を行い、文アライメントの精度向上を試みる。提案手法を日韓対訳特許文書に適用し、その有効性を確認する。

2. 表音的類似性を用いた文アライメント

本研究で提案する文アライメントの処理手順を図 1 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。本研究では、(1)文字数の比、(2)外来語の一致度、(3)格助詞の一致度の 3 つの特徴量を用いて文の類似度を算出する。これらの特徴量はすべて表層的に求められ、対訳用語辞書などを必要としない。文アライメントでは、文の類似度が最大になるような対応関係を動的計画法[3]により決定する。以下、それぞれの特徴量の算出方法について説明する。

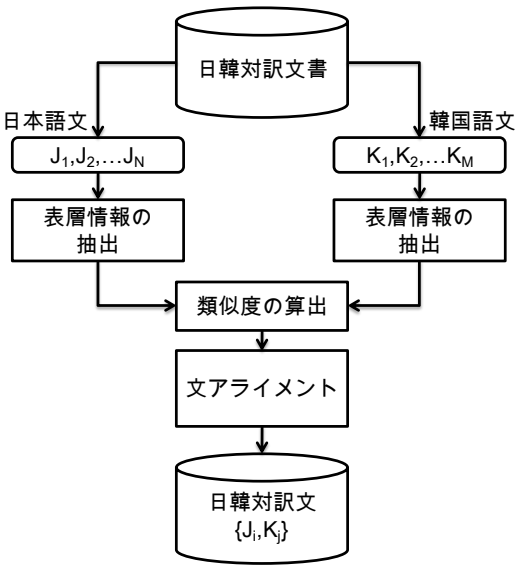

図 1 日韓文アライメントの処理手順

表 1　日本語と韓国語の表音的類似性を用いた外来語の照合例

日本語	ローマ字表記	韓国語	疑似ローマ字表記
ネットワーク	[ne|nei]+[to|t]+wa+[ku|k]	네트워크	ne+[tu|t]+[wo|wa]+[ku|k]
システム	[si|shi]+[su|s]+[te|ti|tei]+[mu|m]	시스템	[si|shi]+[su|s]+tem
コンピュータ	ko+[n|m]+pyu+ta	컴퓨터	kom+pyu+[to|ta]

2.1.　文字数

日本語と韓国語は、文法が類似しており、対訳文間の文字数もほぼ同程度になることが多い。したがって、文字数による類似度(Character Similarity: CS)は、次式によって求められる。

$$\mathrm{CS} = \min\left(\frac{\alpha n_J}{n_K}, \frac{n_K}{\alpha n_J}\right) \qquad (1)$$

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は文字数の比を表し、本研究では$\alpha = 1.1$とした。

2.2.　外来語

日本語と韓国語は、ともに外来語を音韻的に表現する特徴を持つ。本研究では、日本語の外来語（カタカナ表記）をローマ字に変換し、また韓国語の単語をハングル文字の発音を基に擬似的なローマ字に変換し、両者を比較することで外来語の表音的な照合を行った。照合結果の例を表 1 に示す。

ここでは、表音的に一致した外来語の頻度を基にして文の類似度を算出する。日本語文に現れる外来語の出現頻度を要素とするベクトルを$\mathbf{f}_J = (f_{J,1}, f_{J,2}, f_{J,3}, \dots)$、それぞれの外来語と表音的な照合により一致した韓国語文のハングル文字の出現頻度を要素とするベクトルを$\mathbf{f}_K = (f_{K,1}, f_{K,2}, f_{K,3}, \dots)$とすれば、外来語による類似度(Katakana Similarity: KS)は、次式によって求められる。

$$\mathrm{KS} = \frac{\mathbf{f}_J \cdot \mathbf{f}_K}{|\mathbf{f}_J||\mathbf{f}_K|} \qquad (2)$$

2.3.　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、その訳し分けが異なる場合がある。表層的に訳し分けを解析することは困難であるので、表 2 に示す対応表を用いて格助詞のラベル付けを行った。日本語文および韓国文中に現れる格助詞のラベルを文字列$(\mathbf{p}_J, \mathbf{p}_K)$とすれば、格助詞による類似度(Particle Similarity: PS) は、次式によって求められる。

$$\mathrm{PS} = \exp\left(-\frac{\mathrm{LD}(\mathbf{p}_J, \mathbf{p}_K)^2}{2\sigma^2}\right) \qquad (3)$$

ここで、LD は表 2 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本語形態素解析を用いて抽出し、韓国語の格助詞は表に示す格助詞となるハングル文字を機械的に抽出した。

2.4.　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。

$$\mathrm{sim}(J, K) = w_c \mathrm{CS} + w_k \mathrm{KS} + w_p \mathrm{PS} \quad (4)$$

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

表 2　日韓格助詞の対応表（一部）

ラベル	日本語	韓国語
de_ni	で, に, へ	에, 에다, 다, 에게, 에게다, 한테, 에서, 으로, 로
wo	を	을, 를
ha	は, とは	은, 는
ga	が	이, 가, 지만
no	の	의
to_ya	と, や	와, 과, 하고
kara	から	에서부터, 으로부터, 로부터, 부터

3. 実験および結果

3.1. 日韓対訳特許文書

提案する文アライメントを日韓対訳特許文書（PCT 国際出願された日本語特許文書と韓国語特許文書のペア）に適用した。この特許文書ペアは、日本語特許公報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。また構造文書なので、XML タグを利用すれば段落のアライメントが容易に可能となる。

3.2. クローズドテスト

日韓対訳特許文書ペア P_{close}（日本語 413 文、韓国語 486 文、文対応 392 対）に対して文アライメント（クローズドテスト）を実施した。動的計画法では、(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、重み付け係数は、$w_c = 1.0, w_k = 0.4, w_p = 0.2$と経験的に決定した。

文アライメント結果の例を表 3 に示す。ここでは、(J1:K1)および(J2:K2-K3)は正しい文対応となったが、(J3-J4:K4)は、誤対応となった（正しい文対応は(J3:K4)）。このような誤対応となる状況では、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、本研究で用いた対応テーブルの改善が必要であると考えられる。

また、段落アライメントを行った場合と行わなかった場合において、文アライメントを行った結果を図 2 に示す。全ての特徴量を組み合わせた場合(CS/KS/PS)における文アライメントの精度は、35.5%（段落アライメントなし）、81.0%（段落アライメントあり）であった。段落アライメントによって、一部の誤対応の結果が、残りの文対応へ悪影響を及ぼすことを防ぎ、高い精度となった。一方、段落アライ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、段落アライメントが困難の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。

3.3. オープンテスト

クローズドテストで決定したパラメータを用いて別の日韓対訳特許文書ペア P_{open}（日本語 1286 文、韓国語 1602 文、文対応 1206 対）に対してオープンテストを行った。オープンテストでは、CS のみの結果をベースラインとして、CS/KS/PS の結果を提案手法として評価した。結果を図 3 に示す。オープンテストでは、クローズテストに比べ、多対多となる文対応が多く出現し、絶対的な精度が低下する結果となった。しかしながら、段落アライメントを行った場合の提案手法では、ベースラインに比べ、3.6%の精度向上が実現されており、表音的類似性を用いた文アライメントの有効性が確認された。

表 3　提案手法による文アライメント結果の例

J1	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。	K1	이것은 네트워크 프린터에 보안 기능을 제공할 필요성을 일으킨다.
J2	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。	K2	첫 번째로, 대단히 큰 프린트 데이터 크기를 각각 갖는 복수의 작업이 동시에 집중되면, 프린터가 데이터를 수신하는 네트워크 처리부에 너무 많은 부하가 인가된다.
		K3	이것은 아마도 프린터에 접근하기 어렵다는 문제점을 일으킬 수 있다.
J3	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。	K4	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.
J4	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。		

4. まとめ

本研究では、(1)文字数、(2)外来語、(3)格助詞の表層的特徴量を利用した日韓文アライメント手法を提案した。提案手法を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%の文アライメント精度(F 値)を確認した。特に外来語の表音的類似性を利用することにより、対訳用語辞書を用いずに対訳用語の照合が可能となり、文アライメントの精度を向上させることを確認した。

参考文献

[1] Gale, W.A. and Church, K.W., “A Program for Alignment Sentences in Bilingual Corpora”, Computational Linguistics, Vol.19, No.1, pp.75-102, 1993

[2] 金 玉錦、藤井 敦、石川 徹也、“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”、言語処理学会年次大会発表論文集、pp.258-261, 2003

[3] 宇津呂 武仁、松本 裕治、“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

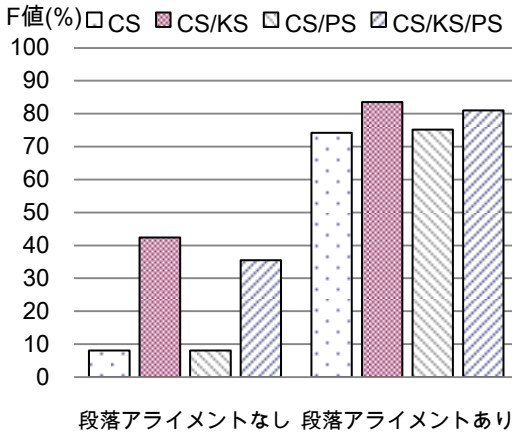

図 2　文アライメント結果
（クローズドテスト）

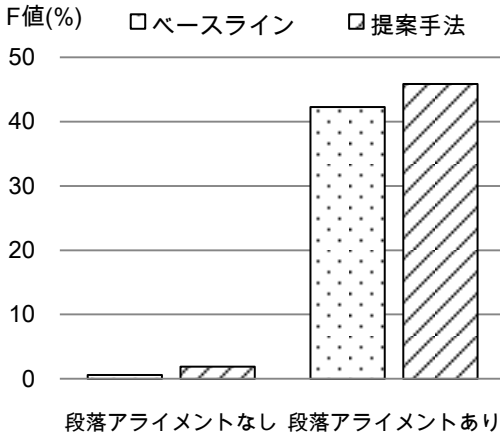

図 3　文アライメント結果
（オープンテスト）