漢字知識を利用した韓中専門用語翻訳

Korean-to-Chinese Technical Term Translation using Chinese Character Knowledge

京都大学大学院情報学研究科 中澤　敏明

PROFILE 2010 年京都大学大学院情報学研究科知能情報学専攻博士課程修了。博士（情報学）。機械翻訳の研究に従事。

nakazawa@nlp.ist.i.kyoto-u.ac.jp　TEL 075-753-5346

楽天株式会社 盧　元梅

PROFILE 2015 年京都大学大学院情報学研究科知能情報学専攻修士課程修了。修士（情報学）。現在は楽天株式会社に勤務。

京都大学大学院情報学研究科教授 黒橋　禎夫

PROFILE 1994 年京都大学大学院工学研究科電気工学第二専攻博士課程修了。博士（工学）。2006 年 4 月より京都大学大学院情報学研究科教授。自然言語処理、知識情報処理の研究に従事。

1 はじめに

統計的機械翻訳の質は使用可能なパラレルコーパスの量に強く依存し、パラレルコーパスがカバーする語彙を拡充することで未知語（out-of-vocabulary, OOV）の数を減らすことができる。しかし、言語における語彙の総数、特に専門用語の総数は増える一方であるため、パラレルコーパスを増強するだけではすべての新語をカバーすることは不可能である。このため、新語に対しては対訳辞書を整備したり、翻字技術を利用して個別に翻訳したりするなどの方法が必要となる。

言語やドメインを絞れば、すでに対訳辞書が整備されている場合もある。また Wikipedia などの資源を利用することで、対訳辞書を自動構築することも可能である。しかしながらこれらの既存の言語資源のカバレージは、特に専門用語に対しては十分とは言えない。さらに、英語を含まない言語対、例えば日中や韓中に関しては、既存の対訳言語資源が英語を含む言語対と比較して非常に少ないか、存在しない場合さえある。

韓国語、中国語、日本語は同じ言語圏に属しており、共通する特徴がいくつかある。その一つが、漢字の使用である。日本語の漢字はもともと中国語の漢字（Hanzi）を起源としており、韓国語では「漢字語（Sino-Korean word）」と呼ばれる語が日常的に使用されている。たとえば国立国語院による標準国語大辞典に収録されている語彙のうち、57% は漢字語であった。漢字語を構成するハングル文字（Hangul）には、対応する中国語漢字（Hanja）があり、漢字で表記することも可能である。【表 1】に Hangul、Hanja、漢字、Hanzi の対応例を示す。

Hanja はその語を強調したいときなどに使われるが、頻度はあまり多くはなく、ほとんどの場合は Hangul で書かれる。しかし Hangul に対応する Hanja は 1 つではないため、同じ漢字語でも文脈によって意味が異なる場合があり、語義曖昧性解消（Word Sense

表 1：各国語での文字の違い

Hangul（韓国）	애정	노동
Hanja（韓国）	愛情	勞動
漢字（日本）	愛情	労働
Hanzi（中国）	爱情	劳动

Disambiguation, WSD）を行う必要がある。

本研究では、漢字情報を利用することで、韓国語の単語、特に専門用語を中国語に翻訳する手法を提案する。提案法ではまず Hangul-Hanzi の文字マッピングテーブルを構築し、これを使って韓国語専門用語の中国語訳候補を生成する。次に生成された候補から、文字の組み合わせ確率および文脈類似度を考慮して、最適な翻訳候補を選択することで翻訳を行う。なお Hanzi と日本語の漢字にも対応関係があるため、生成された中国語訳をさらに日本語に変換することも可能であるが、本稿では対象外とする。

2 関連研究

文字の変換に関する研究はすでにいくつか存在する [1, 2]。これらの手法では対訳辞書が必須であるが、対訳辞書のような言語資源が存在しない言語対もたくさんある。また、多くの手法では、変換の際に文脈情報を利用しない。

韓国語の文字は表音文字であり、ほとんどの中国語漢字には、対応するハングル文字が 1 つ存在する（ただし稀に 1 つ以上存在する場合がある）。Huang らは 436 のハングル文字と 6763 の中国語漢字からなる文字変換テーブルを構築した [3]。しかしながら、KATS（Korean Agency for Technology and Standards）によると、韓国で使われている Hanja は約 5000 文字、日常的に使われている Hanzi は約 18000 文字であり、Huang らの変換テーブルだけでは文字のカバー率が低いことがわかる。

3 提案手法

【図 1】に提案する韓国語から中国語への単語翻訳システムの概要を示す。本研究では韓国語科学技術文書の専門用語（漢字語）の中国語への翻訳を目的とする。

与えられた韓国語の文に対して、まずは形態素解析を行い、韓国語の名詞を抽出する。これは多くの漢字語が名詞だからである。韓国語の形態素解析器として KOMORAN ver 2.3（精度 95%）を使用した。次に韓中対訳辞書を用いて、一部の韓国語単語を中国語

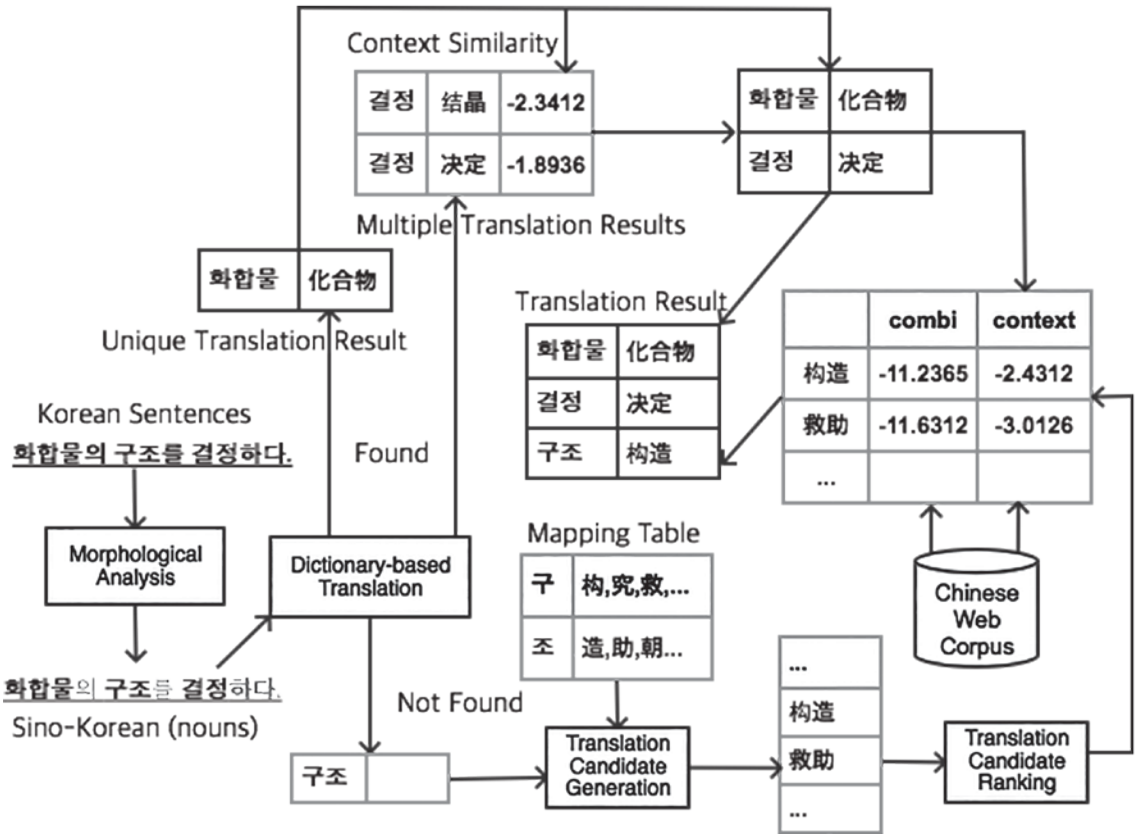

図 1　提案手法の概要

に翻訳する。本研究では対訳辞書として Wikipedia や Wiktionary から自動抽出したものを利用した（約 710 万エントリー）。なお対訳辞書における翻訳候補が複数あり、曖昧性が存在する場合は、それらの翻訳候補から後に説明する文脈類似度スコアが最も高いものを選択する。

対訳辞書に登録がなく翻訳できなかったものは、Hangul-Hanzi マッピングテーブルを用いて中国語の翻訳候補を生成する。Hangul-Hanzi マッピングテーブルは Chu ら [4] による日本語漢字、繁体字、簡体字の中国文字マッピングテーブルを拡張する形で構築した。構築するにあたり、Web 上に存在する様々な情報を統合して利用した。最終的に 5368 の Hanzi と 481 の Hangul を収録した。

最後に文字結合スコアと文脈類似度スコアによってランキングを行い、最終的な翻訳を決定する。文字結合スコアは言語モデルと同様に計算される文字列の生起確率から計算され、大規模な中国語 Web コーパス（47GB）からモデルを構築した。なお文字結合スコアは順方向と逆方向の 2 つの確率の log を足し合わせたものを用いる。

文脈類似度スコアはある単語の入力文における文脈と、大規模中国語 Web コーパスにおける文脈の類似度をスコア化したものである。文脈のウインドウは単語が出現する文内とする。文脈は文字を要素とし、その頻度を値とするベクトルとして表現し、2 つのベクトルの cos 類似度をスコアとする。なお的や了などの 125 文字をストップ文字として除外した。また Web 上での頻度が 100 以下の翻訳候補は候補から除外した（ただし全ての候補が頻度 100 以下の場合は、全ての候補を残した）。入力文における文脈ベクトルの構築には、韓中対訳辞書により中国語に翻訳された単語を利用する。最終的にこれら 2 つのスコアを線形結合し、翻訳候補をランキングする。文字結合スコアを S_{combi}、文脈類似度スコアを $S_{context}$ とすると、翻訳候補のスコア $S(cand)$ は以下のように計算される：

$$S(cand) = \alpha S_{combi} + (1 - \alpha) S_{context}$$

α（$0 \leqq \alpha \leqq 1$）は二つのスコアの結合重みである。

4 実験

提案手法の有効性を示すために、韓中専門用語の翻訳実験を行った。翻訳対象として Web 上の自然科学に関する韓国語の技術文書から 100 文、3281 語を抽出してテストデータとして利用した。なおこのうち名詞は 955 語であった。正解の中国語翻訳は人手で付与した。データを 20 文ずつの 5 セットに分割し、そのうちの 4 セットを使いαの値を調整し、残りの 1 セットで精度を測定する 5 分割交差検定により翻訳精度を測定した。5 分割交差検定により得られた 5 つのαの平均値 $\alpha = 0.39$ を使用して、全てのテストデータの翻訳を再度行った結果を【表 2】に示す。参考として、同じテストデータを Google 翻訳で韓国語から中国語に翻訳した場合の精度は 38.17% であったので、提案手法では十分高精度に翻訳できていることがわかる。

最終的に翻訳が生成できなかった 13 語のうち、11 語は外来語の翻字となっているものであり、これらは提案手法で翻訳することはできない。残りの 2 語は適切な翻訳候補が Web コーパスに出現しておらず、候補から除外されていた。

また生成された翻訳が誤っていたもののうち、24 語は対訳辞書による曖昧性のない翻訳が誤っており、82 語は曖昧性のある対訳辞書エントリーの文脈ベクトルによるランキングが誤っていた。また 12 語は文字マッピ

表 2：韓日専門用語翻訳実験結果

	辞書翻訳		文字結合スコアのみ	文脈スコアのみ	両スコア
	曖昧性なし	曖昧性あり			
正解	549	190	44	47	50
不正解	35	106	18	15	12
翻訳不能	371	75	13	13	13
正解率	94.01% (549/584)	64.19% (190/296)	70.97 (44/62)	75.80 (47/62)	80.65 (50/62)

ングにより翻訳候補生成し、そのランキング結果が誤っていた。

5 結論

本研究では漢字知識を利用した韓国語専門用語の中国語への翻訳手法を提案した。韓国語の形態素解析器を利用して名詞を抽出し、Wikipedia や Wiktionary から抽出した対訳辞書と Hangul-Hanja マッピングテーブルを使用して韓国語専門用語の中国語翻訳候補を生成し、大規模中国語 Web コーパスから構築した文字言語モデルと文脈ベクトルを用いて翻訳結果をランキングすることで、最終的な翻訳を高精度に得ることに成功した。本研究の結果得られた中国語翻訳は、韓日対訳辞書として利用することも可能である。

今後の課題として、より多くの素性を利用してランキングの精度を向上することや、候補の決定しやすいものから決定するなど、翻訳候補の決定順序を工夫することで、より精度を向上することなどが挙げられる。

なお本研究の内容は MT Summit 2015 にて発表予定である [5]。

参考文献

[1] Chen, Z. and Lee, K.-F. (2000). A new statistical approach to Chinese pinyin input. In Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00, pages 241-247, Stroudsburg, PA, USA. Association for Computational Linguistics.

[2] Huang, J.-X., Bae, S.-M., and Choi, K.-S. (2004). A statistical model for hangeul-hanja conversion in terminology domain. Association for Computational Linguistics.

[3] Huang, J.-X. and Choi, K.-S. (2000). Chinese-Korean Word Alignment Based on Linguistic Comparison. In ACL.

[4] Chu, C., Nakazawa, T. and Kurohashi, S.: Chinese Characters Mapping Table of Japanese, Traditional Chinese and Simplified Chinese, Proceedings of the Eighth Conference on International Language Resources and Evaluation (LREC 2012), Istanbul, Turkey, pp. 2149-2152 (2012).

[5] Yuanmei Lu, Toshiaki Nakazawa and Sadao Kurohashi: Korean-to-Chinese Word Translation using Chinese Character Knowledge. In Proceedings of MT Summit 2015, Miami, Florida, USA, to appear.