Twitter を用いた日中機械翻訳精度向上のための流行語埋め込み手法に関する研究

A Study On　Buzzword Embedding Method For Improving Japanese-To-Chinese Machine Translation Accuracy Using Twitter

魏 晟淼
指導教員　亀田 弘之

東京工科大学 バイオ・情報メディア研究科
コンピュータサイエンス専攻 亀田研究室

キーワード：Twitter，機械翻訳，精度向上，流行語，埋め込み手法

1.緒 言

近年、IoT や AI などテクノロジーが急速に進化したことにともない、 機械翻訳技術は劇的に変わっている。「DeepL」、「Google translator」、「みらい翻訳」など超高性能な機械翻訳システムが公開され、やがて文化や言葉のニュアンスまでも読み解けるようになっている[1]。しかし新しい言葉や人名などの「未知語」は辞書に登録されてないため、珍しい単語をうまく処理できないことがしばしば問題になる［2]。「未知語」のひとつとして、「エモい」「タピる」などの「流行語」を含む文の形態素解析およびその語意の推定に関する研究は現在も注目を集めている。日本の文化や慣習に基づいた日本独自の「流行語」を翻訳する時に、どんな対象にどのようなニュアンスで伝えるのかが翻訳業者に対しても考えるべき問題である［3]。

現代社会では、自動化・人工知能の理由から質の高い機械翻訳システムが求められている。例えば、現時点において、ニューラルネットやコーパスなどの手法・システムが提案されたり、実用化されたりしている。

しかしながら、翻訳の質（精度）においては、ニューラルネットやコーパスなどの問題点がなおもある。本研究では、日本語―中国語機械翻訳の処理精度改善を目的として、流行語埋め込み手法を提案し、その有効性の確認を行った。その結果、従来のものよりも翻訳精度の高い日中翻訳システムを構築することができた。

2.方 法

本研究は、この問題を改善することを目的としています。Twitter プラットフォーム上の日本語データを利用して、Twitter プラットフォーム上の流行語を自動的に抽出して中国語に翻訳する処理精度の改善を目指す。流行語の抽出は、流行語の使用度が短期間で急速に向上し、低下する特徴に基づいて行う。実際の Twitter データの分析を通じて、年間をまたぐ期間での語の使用頻度変動の挙動を明らかにし、語の流行の程度を量化する。また、流行語の翻訳は、意味の近い語が通常類似したコンテキストに現れるという特徴を利用し、コーパスという大規模に取得しやすいバイリンガルリソースを利用して、各語のコンテキストベクトルを構築し、類似度測定により候補翻訳を抽出する。

大まかな手順は以下の通りです。

1) Twitter で流行語を取得し、コーパスを構成する
ウェブクローラーを使って大量の実 Twitter データをダウンロードし、前処理としてキーコンテンツの抽出、TF-IDF 手法（式 2.1）を採用して統計とランキングする、最後にノイズをフィルタリングしてインターネット上の流行語を得ている。

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \qquad (2.1)$$

2) コンテキストからベクトルへの変換
Word2vec でコンテキストからベクトルに変換する。
3) 中国語に日本語を投影する
Weblio 辞書を用いて自動的に構築された日中対訳のシード辞書を使用って 2 つの言語間のリンクを確立する。
4) 類似度の計算と出力翻訳
余弦ベクトル法によりテキスト木ベクトル間の角度を計算し、テキスト間の類似度を測定する。類似度の高い単語が翻訳結果として出力される。

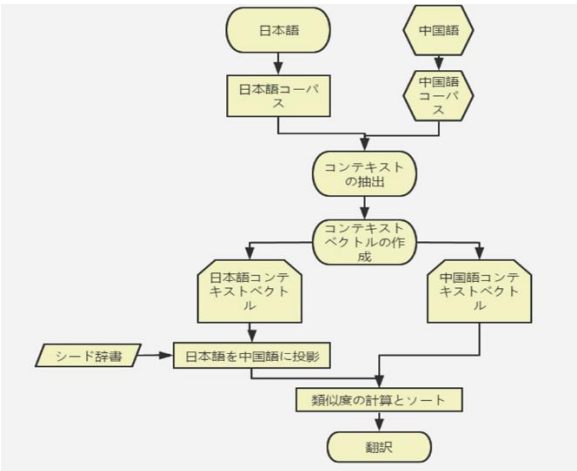

図 1:文脈に基づく日中間の機械翻訳プロセス

3. 評価指標

相関精度率とは、得られた候補語の総数に対する、翻訳類似度の高い上位 n 個の候補語のうち、対象 Web バズワードに関連する語の数の比率を意味する。 相関精度率は、式 3.1 のように定義される。

$$Relevance = \frac{R}{N} \qquad (3.1)$$

ここで、R は類似度ランキング上位 N 位のうち、対象語に関連する流行語の数、N はテスト翻訳候補の数である。

4. 結論

(1)実言語に基づいてデータを用いてネットワーク流行語の自動抽出を行う方法を提案する。この方法は流行語の使用度の特徴を考慮し、動的特徴、静的特徴などの指標を設計することによって真実のネットワークフォーラムの使用データを分析し、流行語の正確な抽出が有望であると考えられる。

(2)比較可能コーパスに基づくネット流行語自動翻訳(和訳中)の戦略を設計した。このポリシーは、流行語を含む比較可能なコーパスを自動的に収集することによって、語のコンテキストを取得し、コンテキスト類似度の比較によって候補翻訳語を取得する。上記の仕事は日本語訳中の機械翻訳の精度を高めるのに有利で、一定の創造性を持っている。

参考文献

[1] 一般社団法人情報処理学会情報システムと社会環境研究会,IS ディジタル辞典: https://ipsj-is.jp/isdic/（参照：2020-6-15） 牛久保佑樹, 藤田茂:Twitter 上の未知語の意味推定方式, 平成 23 年度情報処理学会関西支部支部大会論文集

[2] 井上 博之　特許情報における機械翻訳の活用について:特許庁の取り組みを中心に（機械翻訳技術の向上）　Japio year book 2011 p.216-219

[3] 井佐原 均　科学技術文献を対象とした日中機械翻訳システムの開発 ： 日中・中日言語処理技術の開発研究 （機械翻訳技術の向上）　Japio year book 2011 2011 p.220-223