2022 年 8 月 10 日(水)
思考と言語究室　修士中間審査会資料

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

G2121022　　魏 晟淼

1. はじめに

言語は、人間の心と心をつなぐ大切なツールであり、とても美しく、素晴らしいものです。しかし、太古の昔から言語障壁という問題があり、文明間のコミュニケーションは限られていた。コンピュータの発明により、人類はこの障壁を機械翻訳で取り除こうとしている。

機械翻訳(Machine Translation、MT と略称)は、自動翻訳とも呼ばれ、数学、論理学、統計学、自然言語学、コンピュータ科学、人工知能などの多学科を利用して、書面形式、音声形式の自然言語をプログラミング処理によって同じ自然言語の別の目標言語に変換する多学科交差形成のエッジ学科である(フィンチ アンドリュー, 安田 圭志,2012)。

世界の多くの国では長い間、機械翻訳という研究に従事してきた。実際には、1940 年代に電子計算機が誕生した日から、コンピュータを言語翻訳に応用するための探索が始まっている。50 年代初め、人々は電子計算機が機械翻訳に適していることを発見した。この新しい機械は大量の情報を記憶することができ、月は記憶と入力の情報に対して便利に各種の操作を行うことができるため、そして、当時アメリカはソ連に対抗して、アメリカは大量のロシア語の技術材料と情報を英語に翻訳する必要があった。そのため、当時は多くの大学や会社が機械翻訳の研究や開発に携わっていました。コンピュータの出現と適量の需要は、機械翻訳史上初のクライマックスをもたらした。

1950s、機械翻訳はアメリカで最初の繁栄時代を経験したが、翻訳能力が機械翻訳に対する過度な期待を満たすことができなかったため、発展はしばらくの間停滞を経験した。1970s-1980s、物質発展の客観的な需要のため、機械翻訳はまた再び人々の視線に戻ったが、その時の人々は、機械翻訳が情報時代に人類発展史にどのような深い影響をもたらすか分からない(森藤 淳志, 稲葉 崇,船守 茉美,2012)。

最初、機械翻訳の応用範囲は言葉だけに対して、電子辞書が紙質辞書を引くことに相当して、それから、機械翻訳は文法と文法に対する分析を導入し始めて、今、人々はますます多くの翻訳方法を結合する方式を採用して翻訳の過程の中で現れた曖昧さ、誤りなど人類の言語の特徴に合わない問題を解決する。

しかし、インターネットが発達し、人間のニーズが高まるにつれて、新しい言葉がどんどん生み出されている。 このような言葉は辞書に載せるのが間に合わないので、結果として異なる言語の人たちが理解しあえない状況になる。より多くの言葉が、より速いスピードで発明されていく中で、この問題は迫ってきている。

この問題では、単語の埋め込みを使って、新しい流行語をより速く翻訳したいと思いました。

2. 関連研究

2.1　Lexical Diversity in Statistical and Neural Machine Translation. Information [1]

機械翻訳システムの研究は、コンピュータの黎明期からすでにてがけられており、規則に基づく言語処理や用例に基づく言語処理がかつて多く手掛けられていた。しかしながら、深層学習手法が提案された以後は、ビッグデータに基づく深層学習型の機械翻訳が中心となってきた。例えば、ディープラーニングの分野で有名な Bengio 教授は、2003 年にニューラルネットワークベースの言語モデルを提案し、分散表現に

よってデータの疎密問題を効果的に緩和している（Brglez Mojca & Vintar Špela, 2022）。

2.2 Neural machine translation for Indian language pair using hybrid attention mechanism [2]

Nal らは 2013 年に初めてエンドツーエンドのニューラル機械翻訳を提案し、入力された原語文をエンコーダーで連続的で密なベクトルにマッピングし、それまでの疎密問題を解決した上で、デコーダーで日常言語文に変換することで翻訳の確率を直接モデル化する「エンコード・デコード」のモデル化枠組みを提案した。 彼らは、エンコーダーの構築に畳み込みニューラルネットワーク、デコーダーにリカレントニューラルネットワークを用い、履歴情報の取得と可変長文字列の処理に対応した。 学習中に起こりがちな勾配減衰や勾配爆発の問題に対して、Sutskever らは end-to-end ニューラル機械翻訳に長短記憶 LSTM を導入し、閾値スイッチを設定してリカレントニューラルネットワークを改良した（Nath, Basab et al.、2022 年）。

2.3 Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. [3]

ソース言語の文脈情報をシード辞書によってターゲット言語に翻訳し、2 つ前の文脈ベクトルとの類似度を類似度アルゴリズムで計算し、得られた類似度を降順にランク付けし、上位の単語をソース言語の翻訳候補語のペアとしてターゲット言語で選択する。(Khatri, Jyotsana et al.、2022 年)

以上のことより、機械翻訳に関する研究は数多くあるが、流行語の中日機械翻訳に関する研究は皆無である。本研究の目的は、Twitter から自動的に流行語を抽出し、中国語と日本語の翻訳を行う機械翻訳アプリケーションを提供することである。これは、翻訳が中日両国の文化交流の手段として活用される明るい未来につながる。

3. 本研究の概要

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。具体的な手順は以下の通りです。

1. ウェブクローラーを使って Twitter からデータを収集

Twitter データの詳細は公開されていないため、数値化して分析することができないため、Twitter データを大規模に収集することでインターネット上の流行語の使用状況を近似的に把握した。 ここでは、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、タイムベースラインに応じた関連語の使用率のカウント、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。

2. データの流行語の抽出

(1)テキストの前処理 まず、Twitter から得られた大量のコーパスに前処理を施し、Web 流行語抽出に無意味な非日本語の文字を除去する。

(2) 統計とランキング インターネット上の流行語候補をランキング形式で分析。 これは主に、使い方の動的な特徴

と、判定を補助するための静的な特徴に基づくものである。

(3) 候補のノイズ処理 ランク付けされた候補語をさらにフィルタリングするために、手動で収集したノイズ辞書が使用される。

3. 使用度を量化

Twitter 日本語流行語の抽出は、得られた流行語の時間的特徴に基づいて、その使用度を詳細に量化する。このため、検索エンジン技術でよく用いられる TF－IDF 手法を採用した。

4. 語流行度メトリック

ネット上の流行語をより適切に抽出・測定するために、主にネット上の流行語の人気度を用いて評価を行っている。人気の主な指標は正規化 DF（文書頻度）と正規化 TF*DF（用語頻度*文書頻度）であり、Twitter の投稿は短くノイズが多いため、利用度の評価指標として DF を直接用いている。具体的には、以下の 2 つの特徴を考慮し、候補用語の人気度を測定する。

(1) 動的特徴：Twitter における候補用語の使用率の 2 年ごとの差。 2013 年の候補語句 w、2014 年の Un4 の使用率を U で表すので、w の動的な使用率特性は次のように表すことができる。

$$w_d = \frac{U_{14} - U_{13}}{U_{13}}$$

(2)静的特徴：Twitter におけるインターネット流行語候補の 2020-2021 年の使用状況である。

ランキング基準：静的特徴量と比較して、より流行語の使用頻度を反映できる動的特徴量を主なランキング基準として使用した。さらに、上位にランクされた流行語候補を静的特徴量にしたがって補正し、流行語抽出の精度を確保した。

5. コンテキストベクトルを構築

情報検索の分野では、ベクトル空間モデル（VSM）が広く使われている。 物事をいくつかの基本要素に分解して座標系を構築し、それぞれの基本要素で物事をベクトルとして座標系に表現することができる、というのが主な考え方である。シード辞書を使用して、2 つの言語のコーパスを同じ空間にマッピングする。

6. 類似度測定により候補翻訳を抽出する

テキストをベクトルとして表現し、従来のベクトル空間モデル形式を用いて、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。そして類似度に基づいて、翻訳の出力を与える

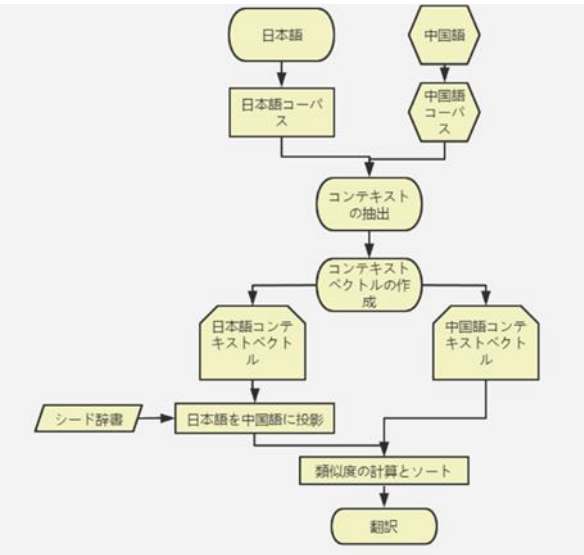

図 1. 本研究で構築するシステムの概要図

4. 研究計画

	7月	8月	9月	10月	11月	12月	1月
データ収集							
アプリケーション化							
実装と改善							
最終レポートの作成							

表1　研究のスケジュール表

5. 進捗状況

- Twitter のデータ収集が完了しました。
- コーパスの前処理が完了しました。 前処理はストップワード削除、emoji 削除、日本語以外の部分削除、数字削除、句読点の削除と単語頻度統計。
- 12 月に開催される八王子の第 14 回学生発表会での発表を予定しています。

6. おわりに

この研究を通じて、日本のネット文化を知らない中国の人たちに、異文化の魅力を感じてもらえたらと思う。 このように、日中間のコミュニケーションを促進し、言葉の壁をなくすことを目標に、これからも努力を続けていきたいと思う。

参考文献

[1] Brglez Mojca & Vintar Špela, Lexical Diversity in Statistical and Neural Machine Translation., Information (2). doi:10.3390/INFO13020093.

[2] Nath, Basab,Sarkar, Sunita,Das, Surajeet & Mukhopadhyay, Somnath, Neural machine translation for Indian language pair using hybrid attention mechanism. Innovations in Systems and Software Engineering(prepublish). doi:10.1007/S11334-021-00429-Z.

[3] Khatri, Jyotsana,Murthy, Rudra,Banerjee, Tamali & Bhattacharyya, Pushpak, Simple measures of bridging lexical divergence help unsupervised neural machine translation for low-resource languages. Machine Translation(prepublish). doi:10.1007/S10590-021-09292-Y.