日中混合テキストの機械翻訳研究

Machine Translation Research of Japanese-Chinese Mixed Text

シュ　ハクエイ 1)
指導教員　亀田　宏之

東京工科大学大学院　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：機械翻訳，コードスイッチング

１．　はじめに

観光や仕事などの目的で、外国人居住者が日本で増加している。国際婚姻の数も約 40 年前と比較して約 3.5 倍に増えている[1]。また、外国人旅行者の数も約 15 年間で約 5 倍に増加している[2]。これらの変化により、国際化が進み、人々のコミュニケーションの仕様にも影響を与えている。

バイリンガルの会話では、言語を混ぜて話す現象が見られることがある。これをコードスイッチング（CS）と呼び、コミュニケーションの変化の１つとして認識されている。実際に、バイリンガルの子供たちが 4 時間に 153 回も CS を使う報告があり[3]、日常生活での CS の使用が明らかになっている。CS には、文の途中で言語を切り替える「文中 CS」[4]と文の切れ目で言語を切り替える「文間 CS」の 2 つのタイプがある。文中 CS 例：「**这**个景色真壮**观**。そう、**让**人心**旷**神怡」。文間 CS 例：「今天真**热**。水分の補給気をつけて」。

CS は、バイリンガル同士が会話をする場合だけでなく、バイリンガルと単言語話者が、会話あるいはテキストによる意思疎通をする場合にも現れる。本研究は、単言語話者が日本語と中国語が混ざっているコードスイッチングされたテキスト（日中 CS）を単言語話者が理解できるように支援することを目的としている。

２．関連研究

テキストからテキストへの CS 機械翻訳の研究がいくつか存在する。Chen Huan[5]らは、トピックモデルにより、英語と中国語が混ざっている CS テキストを中国語への翻訳するシステムを実現した。Sinha[6]らは、各言語を分離することによって、CS テキストから単言語テキストへの翻訳を実現した。また、Johnson[7]らはによって提案された機械翻訳は、翻訳先の言語を入力で指定することにより多言語の翻訳を実現し、CS 翻訳の可能性を示唆した。

BERT（Bidirectiinal Encoder Representations from Transformers）[8]は、単語の意味をより正確に理解することができる文脈に頼る単語埋め込みモデルも適用可能であることを示唆した。

３．研究概要

本研究では、モデル BERT を用いる。

以下の図 1 は、CS 翻訳使用するモデル BERT の構造を示している。

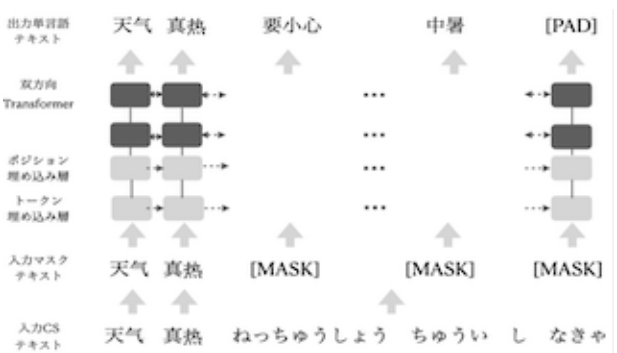

図 1　CS 翻訳使用するモデル BERT の構造

4. データセット

日中 CS データセットは、3 つの手法で収集した。tiktok、weibo、XiaoHongshu、Xiaohongshu third party、kuaishou、Taobao の API により、日中 CS センテンスを収集する[9]。OpenAI の Completion API により、日中 CS センテンスペアを生成する[10]。記の様式に従って下さい。

5. 処理

単言語である中国語のデータセットを使用し、文中 CS あるいは、文間 CS に転換する。収集したセンテンスの質が悪かったので、1 番目の手法を破棄とした。2 番目の手法と 3 番目の手法を使用した[11]。単言語の中国語のデータセットを CS への手法において、ランダムに中国語を単語にする Jieba というワードセグメンテーションを使う。翻訳部分は、Google Translate API により、中国語から日本語へ翻訳する。

6. 今後の実験

対話のデータセットを継続して収集する。BERT の実験を行う。評価指標としては、BLEU スコア[12]を用い、翻訳効果を評価する。

7. 参考文献

[1] 厚生労働省：平成 29 年度人口動態統計, https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/ (2017)

[2] 日本政府観光局(JNTO)：2018 年訪日外客数(総数), https://www.jnto.go.jp/, (2019)

[3] Fotos, S. S. (1990). Japanese-English code switching in bilingual children. JALT Journal, 12(1), 75-98.

[4] 宮原温子. (2011). 日本語英語バイリンガル大学生によるコードスイッチング. 目白大学人文学研究, 7, 239-254.

[5] 陈欢, & 张奇. (2016). 基于话题翻译模型的双语文本纠错. 计算机应用与软件, 33(3), 284-287.

[6] Sinha, R. M. K., & Thakur, A. (2005). Machine translation of bi-lingual hindi-english (hinglish) text. In Proceedings of Machine Translation Summit X: Papers (pp. 149-156).

[7] Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., ... & Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5, 339-351.

[8] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[9] Postman, https://www.postman.com/solar-flare-375895/workspace/spider/collection/3194348-de53a038-17ff-42d9-bc89-f9aa7d71f860, (2022)

[10] GPT models: Documentation of Completion API, https://platform.openai.com/docs/guides/gpt/completions-api, (2023)

[11] Jie: Chinese text segmentation, https://github.com/fxsjy/jieba, (2023)

[12] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).