氏　　名	CHIA ZHENG LIN
授 与 学 位	博士（工学）
学 位 記 番 号	博甲第 217 号
学位授与年月日	令和 6 年 9 月 5 日
学位授与の要件	学位規則第 4 条第 1 項
学位論文題目	Exploring Optimal Settings for Machine Translation of Irony with Application to Multilingual Irony Detection （皮肉文の機械翻訳における最適設定に関する考察及び多言語皮肉検出への応用）
論文審査委員	主査　准教授　ﾌﾟﾀｼﾝｽｷ　ﾐﾊｳｪﾄﾞﾑﾝﾄﾞ 教　授　桝　井　文　人 教　授　奥　村　貴　史 教　授　升　井　洋　志 准教授　澁　谷　隆　俊

学位論文内容の要旨

In this thesis I investigate sarcasm and irony with implementation of machine learning and other related methods. My primary goal of this research is to improve machine learning systems, as well as the natural language processing of figurative language focusing on irony. First I clarify the definitions of irony and sarcasm and present an exhaustive and thorough field review on irony machine translation related works including machine learning, figurative language, evaluation metrics, datasets, and irony and sarcasm detection, from purely linguistic as well as computational linguistic perspective.

Next I construct a novel English-Chinese parallel dataset by manually translating and annotating two datasets originally collected from highly varied reliable sources for the purpose of irony translation. The content of my new dataset is highly polarized with tweets, which characterize in short sentences full of social media slangs, and forum posts, which contain logically structured long paragraph extracted from online debate forums focusing on specific topics.

After construction of the dataset, I propose a new combined evaluation metric for machine translation. The score of my new evaluation metric, COMMET, is calculate by first choosing multiple popular evaluation metrics along with their weights based on other reliable studies. Then I sum the multiplications of the metrics scores and weights and divide with the total number of implemented evaluation metrics. The score is calculated with a variety of evaluation metrics and reliable weights; hence it is potentially more democratic and fairer in the evaluation of machine translations.

With my dataset and evaluation metric ready, I first compare between different types of popular language models for English-Chinese translation. Before conducting the experiments, I review some of the commonly used models and evaluation metrics for machine translation. With enough background information, I start with exploring optimal experiment settings and different combinations of training data. Then after some comparisons, mBART-large-50 becomes my preferred based model for the remaining of experiments.

Next I conduct various types of experiments surrounding our preferred base model training and testing on different combinations of my dataset using my proposed evaluation metric. One of the first experiments is to decide whether the inclusion of hashtags is vital for data originally collected from twitter. With better result on model trained on data without hashtags, I decide to exclude all ironic hashtags such as #sarcasm and #irony from all following experiments. I also compared between model trained on only short tweets of my dataset, model trained on only long forum posts of my dataset, and model trained on both all data from my dataset. With further experiments, I conclude my best model for English-Chinese irony machine translation so far is mBART-large-50 trained on both ironic and non-ironic data only in long forum posts. One of my important discussions here regarding model training for figurative language is that more training data is not always a better approach.

For my new combined dataset, I categorize the ironic data into three different types; self-contained irony which is the normal easy-to-understand irony, contextual irony where the irony is dependent on the context of the data, and ambiguous irony which is the most difficult to comprehend. With the new categories, I perform qualitative analysis of contextual and ambiguous irony by manually process each instance of them due to their low number of samples to find any characteristic in how the lack of sufficient context influence the quality of machine translation of irony. I describe and conclude many of the findings in the experiment results discussion. Next I also compare the results of my model with results from ChatGPT, where my model achieve better in ambiguous data and ChatGPT did better in contextual data due to its large training.

Lastly I conduct another experiment using various combinations of models trained on different data. Some of the trained data worth mentioning includes human translated Chinese data and model translated Chinese data. Improvements are found in various combinations, which proves translation training benefit models in classification, or in more general terms, fine-tuning first on certain tasks can improve latter fine-tuning, if the tasks are related to some extents.

Finally, I discuss some insights and findings of my thesis and my plan to extend this research to the range of other applicable languages. I will also delve further into the potential studies of fine-grained irony types, specifically, ambiguous and contextual, for a better understanding of irony, and figurative languages.

審査結果の要旨

この論文では、機械翻訳という新しい視点から皮肉について調査する。まず、皮肉の定義を明確にし、言語学および計算言語学において皮肉に関する従来研究を徹底的にレビュー調査を行なった。次に、皮肉など比喩的言語表現の翻訳をより公平に評価するために新しい評価指標を提案する。皮肉を含む英語と中国語の並列データセットを作成し、その内容にはツイートとフォーラム投稿という複数種類の書き込みを用い耐用性を保ったデータセットを作成した。さらに、徹底的な実験過程を通じて、性能の高い機械翻訳用のモデルである mBART-50 を特定している。そして、皮肉の翻訳に最適な翻訳設定を実験的に特定し、皮肉及び一般文章両方をより高い性能で機械翻訳するためのモデルの微調整（ファインチューニング）を行う。さらに、近年広く使われている ChatGPT3.5 という大規模言語モデルを使用し、プロンプト工学を用いた機械翻訳結果と、本研究で作成した mBART モデルの翻訳性能を比較した。結果、mBART が ChatGPT を上回ったことを確認した。しかし、両方のモデルの出力の定性的分析をしたら、前者は微調整されていることを見て、皮肉の翻訳に訓練されていない ChatGPT の出力も、十分に理解できると評価ができ、訓練データが存在しない場合には、アドホック手法と見なすことができるという結論に達した。最後に、手動または機械翻訳モデルで翻訳されたデータが皮肉検出というタスクにおいて訓練データとして使用できることを確認した。本研究において明らかにしたことを、中国語と英語以外の言語にも拡張することを期待でき、比喩的言語、特に皮肉の理解を深めることにさらに貢献できると考えられる。

結論として、著者は、今後の機械翻訳とその評価の手法開発への根本的な貢献をした他、自然言語処理分野，特に機械翻訳分野において課題となっている比喩的言語の取り扱いの課題について示唆を与える新知見を得たと判断される。なお、自然言語処理、人工知能など複数の分野に跨る深淵な課題に対して貢献するところ大なるものがある。よって著者は、博士(工学)の学位を授与される資格があるものと認める。