（ 続紙 1 ）

京都大学	博士（情報学）	氏名	毛 卓遠 (MAO ZHUOYUAN)
論文題目	Breaking Language Barriers: Enhancing Multilingual Representation for Sentence Alignment and Translation (言語の壁を超える:文のアラインメントと翻訳のための多言語表現の改善)		

（論文内容の要旨）

This thesis embarks on a comprehensive exploration of multilingual representation learning, addressing the three identified challenges within this domain: (1) high computational demands, (2) data scarcity, and (3) limitations in Transformer architecture, with a specific focus on sentence alignment and translation tasks. These tasks are essential in the broader context of multilingual NLP, enabling machines to understand and translate across diverse human languages with increased proficiency and efficiency.

Chapter 1 introduces the background of multilingual representation learning and its challenges

Chapter 2 introduces a method for multilingual sentence embedding (MSE) learning that efficiently addresses high computational demands. To balance training efficiency against data and computational needs while maintaining MSE quality, we develop an innovative approach to concurrently train generative token-level reconstruction and sentence-level contrastive objectives. The effectiveness of the proposed method is validated through empirical evaluations on four cross-lingual sentence retrieval tasks and three cross-lingual sentence classification tasks.

Chapter 3 introduces a streamlined model designed to produce compact MSE, addressing the issue of computational intensity during inference. The experimental outcomes indicate that our model achieves robust performance across 109 languages after distilling knowledge from the previous state-of-the-art model, LaBSE.

Chapter 4 introduces Japanese-specific sequence-to-sequence pre-training (JASS) and English-specific sequence-to-sequence pre-training (ENSS), which incorporate syntactic structures of sentences, based on language-agnostic schemes like masked sequence-to-sequence pre-training (MASS), to address data scarcity in low-resource languages for neural machine translation (NMT). Utilizing abundant monolingual data and syntactic analysis, these methods enhance language-specific structure awareness during pre-training. Experiments on various language pairs show that JASS and ENSS surpass MASS and similar methods in low-resource contexts, highlighting the value of language-specific inputs and multi-task pre-training. They significantly improve translation adequacy and fluency, as confirmed by adequacy evaluations, human evaluations, and case studies.

In Chapter 5, we introduce a word-level contrastive learning approach for multilingual NMT to tackle the challenge of data scarcity in low-resource languages. Our experiments demonstrate notable improvements in translation quality across various language pairs, further elucidated by analysis linking BLEU scores to the sentence retrieval capabilities of the NMT encoder.

Chapter 6 presents AlignInstruct, a method that aims at improving the fine-tuning of large language models (LLMs) for NMT in low-resource, previously unseen languages, with a focus on minimizing the need for extra training corpora to address the data scarcity issue. The results from our multilingual and zero-shot experiments highlight AlignInstruct's superiority compared to the baseline method.

In Chapter 7, we unveil an innovative variable-length neural interlingua method, which not only enhances zero-shot translation outcomes but also yields a more reliable model compared to earlier fixed-length interlingua techniques, addressing the issue of suboptimal model architecture for zero-shot NMT. Despite observing a decline in performance in translations to English, our analysis pinpoints the specific model component responsible, setting the stage for targeted improvements in future research.

Chapter 8 thoroughly examines the impact of layer normalization (LayerNorm) on the performance of zero-shot NMT, aiming to overcome issues related to the Transformer model configurations for zero-shot NMT. The findings reveal that post-layer normalization (PostNorm) has a consistent edge over pre-layer normalization (PreNorm) in zero-shot NMT scenarios, independent of language tag and residual connection configurations. By analyzing off-target rates and identifying structural weaknesses in the PreNorm model, we uncover the reasons behind this performance gap. The insights from our study emphasize the importance of careful LayerNorm configuration choices in future zero-shot NMT research.

In Chapter 9, the dissertation is concluded with a summary and a discussion of future work.

（続紙 ２ ）

（論文審査の結果の要旨）

本論文では、多言語表現学習の複雑さに着目し、多言語文の埋め込み学習（MSE）と多言語ニューラル機械翻訳（NMT）の二つの重要なタスクに焦点を当て、以下の三つの主要な課題に取り組む：(1)膨大な計算量、つまり多言語モデルの言語カバレッジを拡大する際に発生する大きな計算オーバーヘッド、(2)データ不足、特に低資源言語における十分で多様な訓練データの欠如、(3)Transformerモデル構造の限界、つまり現在のTransformersモデルが多言語処理の複雑さに適応していない。

１.多言語モデルの言語カバレッジの拡大に伴う膨大な計算量に対処するために、まず、効率的かつ効果的な大規模多言語文埋め込みモデルを提案した。このモデルは、クロスリンガルなトークンレベルの再構成と文レベルの対照学習を訓練目標関数とし、先行研究に比べ極めて少ない訓練データとGPU計算量を使用してMSEモデルを効率的に訓練することができる。次に、MSEモデルの推論プロセスを効率化させるために、新たなMSEモデルの知識蒸留法を提案した。具体的には、軽量モデルを用いてMSEモデルを効率的に学習する方法を体系的に探求し、109の言語に対してロバストな低次元の文埋め込みを構築した。提案した知識蒸留法により、MSEモデルのさらなる改善も達成している。

２.低資源言語におけるデータ不足の課題に対処するために、本論文では、低資源NMTのための新たな事前学習目標関数を提案した。具体的には、訓練データの不足を補うために、フレーズ構造に基づくマスク言語モデリングや並び替えタスクを提案した。また、統計的単語アライメントを活用した低資源NMTのための単語レベルの対照学習を提案し、従来に使われている高品質なバイリンガル辞書を使用せず、翻訳性能の向上を図った。さらに、AlignInstructという新しいアプローチを導入し、統計的単語アライメントを用いてクロスリンガル指示学習を通じて、大規模言語モデルの低資源言語における翻訳精度を向上した。

３.Transformerモデル構造の教師なし多言語NMTにおける性能限界に対処するために、Transformerエンコーダの上にインターリンガル表現を構築する新しいTransformerアーキテクチャを提案し、従来のTransformerアーキテクチャより教師なしNMTの性能を大幅に向上した。さらに、教師なしNMTにおけるレイヤー正規化の影響について包括的に検討した結果、ポストレイヤー正規化がプリレイヤー正規化を一貫して上回ることが分かった。また、モデル設定に関わらず、教師なしNMTにおいてポストレイヤー正規化の設定が重要であることを示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、令和６年２月２２日に実施した論文内容とそれに関連した試問の結果合格と認めた。なお、本論文のインターネットでの全文公表についても支障がないことを確認した。