（ 続紙 1 ）

京都大学	博士（情報学）	氏名	宋　海越（Song Haiyue）
論文題目	Studies on Subword-based Low-Resource Neural Machine Translation: Segmentation, Encoding, and Decoding （サブワードに基づく低資源ニューラル機械翻訳に関する研究：分割、符号化、及び復号化）		

（論文内容の要旨）

This thesis investigates issues in subword-based neural machine translation systems in low-resource scenarios during three stages: subword segmentation, encoding, and decoding. To address them, this thesis proposes methods to 1) obtain the optimal subword segmentations, 2) incorporate multiple perspectives of one word during encoding, and 3) perform accurate word generation probability estimation during decoding.

In Chapter 1, we provide an overview of the subword-based low-resource neural machine translation system, including its structure, mechanism, and how subwords are used in the model. We emphasize the advantages together with the challenges brought by using subwords.

In Chapter 2, we propose SelfSeg, a neural subword segmenter that yields linguistically intuitive subword segmentation and is faster during training and decoding compared to previous neural methods. SelfSeg takes a word in the form of a partially masked character sequence as input, optimizes the word generation probability, and generates the segmentation with the maximum posterior probability, which is calculated using a dynamic programming algorithm. Additionally, we propose a regularization mechanism that allows the segmenter to generate k-best segmentations for one word. Moreover, it is trained in a self-supervised way that relies on only monolingual word-level data, making it applicable to low-resource languages without large-scale parallel resources.

In Chapter 3, we propose a BERT-based subword segmenter that generates subword segmentation, utilizing the contextualized semantic embeddings of words from the BERT model. During training, it maximizes the marginal probability from all possible segmentations of one word using a dynamic programming algorithm. During inference, it selects the one with the highest probability. Furthermore, we propose a probability-based regularization method that enables the segmenter to produce multiple segmentations for one word to improve the robustness of neural machine translation systems. Based on a pre-trained BERT encoder, it only requires little training data to achieve reasonable segmentations, making it especially applicable in low-resource scenarios.

In Chapter 4, we propose DiverSeg to exploit diverse segmentations from multiple subword segmenters that capture the various perspectives of each word in the encoder. In DiverSeg, multiple segmentations are encoded using a subword

lattice input, a subword-relation-aware attention mechanism integrates relations among subwords, and a cross-granularity embedding alignment objective enhances the similarity across different segmentations of a word. We found incorporating information from multiple aspects enhances the performance of neural machine translation, especially in low-resource scenarios.

In Chapter 5, we propose SubMerge, a decoding algorithm that merges the probabilities of multiple subword segmentations that form the same word, which we call equivalent segmentations. This is specially designed for the subword regularized neural machine translation model that leverages multiple subword segmentations of one target sentence during training as a data augmentation method, which is effective for low-resource scenarios. SubMerge is a nested search algorithm where the outer beam search treats the word as the minimal unit. The inner beam search provides a list of word candidates and their probabilities, merging subword segmentations that form the equivalent word. SubMerge estimates the probability of the next word more precisely, providing better guidance during inference. We show SubMerge consistently outperforms the beam search algorithm in several low-resource machine translation datasets in terms of BLEU scores.

In Chapter 6, we summarize the thesis and outline the possible directions for future work.

（続紙 ２ ）

（論文審査の結果の要旨）

本論文はサブワードに基づく低資源ニューラル機械翻訳に関する分割，符号化，及び復号化の三つの課題に取り組んだものである．本論文の提案手法を用いて，低資源ニューラル機械翻訳の品質を向上させた．具体的には主に以下の成果が挙げられる．

1. 最適なサブワード分割を得るために，新たなニューラルサブワード分割モデルを提案した．提案モデルは、入力を文字単位でマスクした上で、単語生成確率を最適化するために，動的プログラミングアルゴリズムを使用して最大事後確率でサブワード分割を生成する．また，複数の分割を生成できる正則化メカニズムも提案した．実験では提案手法は効率的に訓練と推論ができ，言語学的に直感的なサブワード分割を生成できることを示した．機械翻訳において，特に低資源の場合に品質の向上を確認した．

2. 多様なサブワード分割を活用するために，複数のサブワード分割を利用できるモデルを提案した．提案手法はラティス入力の符号化，サブワード関係認識のための注意メカニズムとクロスグラニュラリティの埋め込みアライメントという三つのモジュールからなる．実験では提案手法は複数の視点からの情報を組み込むことで，特に低資源のデータセットにおいてニューラル機械翻訳の品質を向上させることが確認された．

3. サブワード化されたニューラル機械翻訳モデルは低資源の場合に有効であるが，単語生成確率は不正確である．正確な単語生成確率を得るために，新たな推論アルゴリズムを提案した．提案アルゴリズムは二つのビームサーチから構成される．外部ビームサーチは単語を最小単位として扱う．内部ビームサーチは同じ単語を構成する複数サブワードの分割の確率を合計し，単語候補とその確率のリストを提供する．実験では複数のサブワード分割をマージすることによって，次の単語の確率をより正確に推定することを確認でき，特に低資源機械翻訳において一貫して従来のビームサーチの性能を上回ることを示した．

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，論文内容とそれに関連した事項について試問を行った結果，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．