氏　　名	Longtu Zhang（ロントゥ チャン）
所　　属	システムデザイン研究科 システムデザイン専攻
学位の種類	博士（工学）
学位記番号	シス博　第 133 号
学位授与の日付	令和 3 年 3 月 25 日
課程・論文の別	学位規則第 4 条第 1 項該当
学位論文題名	Neural Machine Translation Using Sub-Character Level Information(サブ文字レベルの情報を使用したニューラル機械翻訳)
論文審査委員	主査　准教授　小町　守
	委員　教　授　山口　亨
	委員　教　授　高間　康史
	委員　准教授　褚晨翚（京都大学大学院）

【論文の内容の要旨】

Neural machine translation (NMT) is a machine learning task that predicts the target language translation text based on the source-language text. Most NMT models exhibit encoder-decoder architecture and rely on supervised training to learn. A particular type of self-supervised NMT model called "unsupervised NMT" (UNMT) pairs up two NMT models of different translation directions between two languages. The two NMT models can serve as back-translators for each other to create pseudo parallel data for training. The encoders and decoders of the same languages can also form sequence-to-sequence language models to facilitate training.

NMT training usually requires a significant amount of parallel text. The text data are usually in long-tail distribution with large vocabulary size (number of unique words) and always poses a significant challenge for computation. Researchers typically reduce the vocabulary size by grouping the words in the long-tail into one "<unk>" type, and the vocabulary will drop significantly from millions (e.g., the well-received public benchmark dataset wmt14.en-de) to thousands. However, the "<unk>" type is harmful to the model because, from the linguistic point of view, the long-tail types are usually the most informative. Loosing too many of them might lower the translation accuracy and quality.

Different sub-word algorithms, such as BPE (byte-pair encoding), word-piece, etc., were introduced to solve this problem. The general idea is to break up words into smaller "sub-word" pieces and use sub-word vocabulary to train NMT models. During testing, the model's sub-word hypothesis will be composed back to word sequences for evaluation. It turns out that BPE significantly alleviates the long-tail distribution in training data and boosts more shared information within and between source and target text so that the model can learn better. BPE has been widely used not only in NMT but also in other natural language modeling tasks.

However, BPE is most suitable for alphabetic languages but not logographic languages, and researchers have long overlooked the difference between them. Alphabetic languages use a limited set of alphabets (usually dozens) to compose words and use spaces as natural word boundaries. On the other hand, logographic languages use a much larger character set (usually tens of thousands) without any word boundaries. Fortunately, these characters are decomposable. By decomposing characters into smaller sub-character units, we can significantly reduce the number of types and make BPE as effective as in alphabetic languages. Specifically, introducing sub-character level data in NMT training can further help alleviate the long-tail distribution in training data and boost the information sharing within a logographic text and across logographic language pairs.
Our research hypothesis is that sub-character level data can increase the performance of NMT models both in logographic language pairs and alphabetic-logographic language pairs, and the finer granularity data, the better the performance.

The dissertation is organized as follows:

Chapter 1 provides a general introduction to the sub-character NMT problem. The organizations of the thesis and contributions are also highlighted.

Chapter 2 reviews the background of NMT and UNMT tasks in general, including the task formulation, the decomposability of logographic characters, and the machine translation models. It reveals that although most NMT models and popular methods are widely examined, they overlook the difference between alphabetic languages and logographic languages, which can further facilitate their performance if the logographic information can be used wisely. It also discusses the significance of the current study.

Chapter 3 introduces the research method in detail. It first details the character decomposition method for logographic languages and then introduces a method to transform the decomposition sequence back to character sequences safely. Next, it

introduces the (U)NMT models used in this dissertation, namely, LSTM and Transformer based NMT models, and the Transformer based UNMT model. Further, for Transformer models, it describes an extra positional encoding for the sub-character level sequences.

Chapter 4 details the experimental settings. We tested our hypothesis on Chinese-Japanese, Chinese-English, and Japanese-English language pairs, representing NMT models between logographic languages and between alphabetic and logographic languages. We use a character decomposition dictionary (cjkvi-ids) to decompose logographic characters into three granularities: ideograph level data, finest ideograph level data, and stroke level data. We modified the dictionary with special markers to make sure the sub-character sequence can be converted back safely. Lastly, we added an extra positional encoding layer to alleviate the problem of long sub-character level sequences.

Chapter 5 shows the results. Our results suggested that almost from all dimensions, using sub-character level information in (U)NMT tasks outperformed the character level baselines:

- Under the same vocabulary size, models trained on sub-character level data outperformed those on character level data.
- Finer granularity data usually have better performance.
- Sub-character level data was able to increase the performance under both logographic-logographic and alphabetic-logographic language pairs.
- Similar tendencies were found in UNMT models trained on logographic language pairs.

Chapter 6 discusses the results. The small vocabulary size and more shared information might be reasons for better performance. Extra control experiments have confirmed the benefits of both, which is only possible if sub-character level data is used. The finer the granularity of sub-character level data, the better performance the model tends to have. However, the performance can drop when the training sequence is too long in "stroke" level data. Additionally, we compared our results with many other baselines, which also tried to increase the shared information between source and target text to boost NMT performance, such as "character mapping" and "kana decomposition." The results suggested that our approach performs steadily better than these baselines. We also discuss the potential reasons.

Chapter 7 concludes the dissertation and gives out future research directions. This

dissertation first notices the critical difference between alphabetic languages and logographic languages in (U)NMT training and uses a simple character decomposition method to transform character-level data to sub-character level data, and demonstrated significant and steady improvement in using sub-character level information in NMT tasks and shed light on other NLP tasks. In the future, we will try to apply more NMT techniques to sub-character level data, such as curriculum learning. Also, we will try to achieve better results in UNMT settings.