論文内容の要旨

博士論文題目 On Language Representation for Low-Resource Neural Machine Translation

氏　名 MARTINEZ Ander

（論文内容の要旨）

In recent years, machine translation systems have taken a qualitative leap. This leap is thanks to the introduction of systems based on neural networks. Neural Machine Translation (NMT) systems have not only yielded unprecedented results, which in certain cases are comparable to those of professional translators, but also come with the promise of being general solutions. However, not all language pairs are the same. If the number of parallel sentences is small, it is considered a low-resource pair. The most popular datasets used as benchmarks in research are large parallel corpora of related languages, such as French-English or English-German.

This thesis explores the problems of machine translation with respect to low-resource language pairs. It focuses on the key elements of a low-resource machine translation system and the critical decisions that have a great impact on the effectiveness of the system. In particular, this research investigates a novel approach that combines subword-level segmentation with character-level information in the form of character n-gram features. The n-gram features are transformed as subword representations in standard encoder-decoder models both for input embedding representation and output prediction representation. A decision tree-based algorithm is employed to select a small number of effective binary character n-gram features.

Experimental results show that the proposed approach significantly improves translation qualities both for low-resource and rich-resource language pairs. Further experiments also reveal the flexibility in designing various features, such as components of ideographic letters, and the benefits of data augmentation setting, in which low-resource training data is augmented by automatically translated data from other sources.

氏　名	MARTINEZ Ander

（論文審査結果の要旨）

In recent years, machine translation systems have taken a qualitative leap. This leap is thanks to the introduction of systems based on neural networks. Neural Machine Translation (NMT) systems have not only yielded unprecedented results, which in certain cases are comparable to those of professional translators, but also come with the promise of being general solutions. However, not all language pairs are the same. If the number of parallel sentences is small, it is considered a low-resource pair. The most popular datasets used as benchmarks in research are large parallel corpora of related languages, such as French-English or English-German.

This thesis explores the problems of machine translation with respect to low-resource language pairs. It focuses on the key elements of a low-resource machine translation system and the critical decisions that have a great impact on the effectiveness of the system. In particular, this research investigates a novel approach that combines subword-level segmentation with character-level information in the form of character n-gram features. The n-gram features are transformed as subword representations in standard encoder-decoder models both for input embedding representation and output prediction representation. A decision tree-based algorithm is employed to select a small number of effective binary character n-gram features.

Experimental results show that the proposed approach significantly improves translation qualities both for low-resource and rich-resource language pairs. Further experiments also reveal the flexibility in designing various features, such as components of ideographic letters, and the benefits of data augmentation setting, in which low-resource training data is augmented by automatically translated data from other sources.

The findings in this thesis are important and would have an influence to other low-resourced natural language processing tasks. The research in this thesis is published as a high-quality peer-reviewed journal paper and a peer-reviewed international conference paper, and would have an impact to spur further researches in this field. As a result, the thesis is sufficiently qualified as a Doctoral thesis of Engineering.