論 文 の 内 容 の 要 旨

論文題目　Incorporating Syntactic Structure into Neural Machine Translation
（構文構造に基づくニューラル機械翻訳）

氏　名　江里口　瑛子

Machine translation is one of the most difficult and complicated tasks in Natural Language Processing (NLP) because it is required to let a computer represent semantics of each language and convert the semantics from one language to another language so that humans can interpret the translations. Representation learning based on neural networks treats every language data such as word, phrase, and sentence as vectors. The vector representations allow us to easily compute the objective information and to apply the learned representations to the other NLP tasks.

Neural machine translation (NMT) has been developed in the trends of representation learning. The NMT model is an end-to-end neural network-based model to directly learn translations statistically from a large data set, and the NMT models have already achieved the state-of-the-art performance in several European languages and Asian languages. In comparison to the conventional statistical machine translation models, the NMT has the simpler architecture and powerful performance when there is a lot of training data.

Most of the existing NMT models are modeled as a sequence-to-sequence learning which learns the relations between the input sequences and the output sequences. They do not utilize any syntactic information inherited the languages. It is, however ,well known that syntactic structure helps to improve the machine translation models in a statistical machine translation areas when treating the translation between Japanese and English which are syntactically different languages. In this thesis, we focus on the syntactic structures inherited in languages and extended the existing NMT model to incorporate the syntactic structures. We have studied phrase structures in a source language and dependency tree structures in a target language.

Firstly, we employ the phrase structure in a source language and build a tree-based encoder to explicitly construct a phrase vector following the phrase structure. We call our proposed model "Tree-to-Sequence Neural Machine Translation model". Our proposed model also has an attention mechanism which softly aligns a target output with source word-based units as well as with source phrase-based units. Experimenting on an English-to-Japanese translation task, we evaluated the models by automatic evaluation metrics. We have confirmed that our proposed tree-to-sequence neural machine translation achieved better accuracy than the existing sequence-to-sequence NMT models and achieved the state-of-the-art performance in the RIBES score.

Secondly, we applied the character-based decoding method to the above described tree-to-sequence neural machine translation model. Although most of the NMT models have been developed as a word-based model, the models have the vocabulary coverage problem because of low-frequent words and unknown words in the corpus. We also explored the effectiveness to utilize the phrase label category in the tree-based encoder. We conducted the experiments on the English-to-Japanese translation tasks, and we report the different trends between the word-based decoder model and the character-based decoder model.

Lastly, we have proposed a target-side syntax-based model called "Sequence-to-Tree Neural Machine Translation". We focus on the dependency relations between words in a target sentence as syntactic structure. Our proposed model generates a translation while parsing the translated sentence simultaneously. Experimenting on translation tasks for four language pairs, we have confirmed that our proposed model improved the model performance better than the existing NMT model in every four language pairs.