論 文 の 内 容 の 要 旨

論文題目　Structure-Aware Latent-Variable Models for Neural Machine Translation
（ニューラル機械翻訳のための文構造を考慮した潜在変数モデル）

氏　　名　　朱　中元

Recent advance for training neural networks has allowed end-to-end neural-based machine translation models. Though achieving considerable high translation quality, a regular neural machine translation model still faces multiple folds of challenges. These challenges include the exponentially increasing model size, high translation latency and difficulties of obtaining translation results with diversity. In this thesis, we explore a novel approach that leverage latent-variable models to improve neural machine translation.

Our first contribution directly learns continuous latent variables to capture the information about target tokens, which enables non-autoregressive neural machine translation. By updating the posterior on the latent variables, we found the evidence lower bound of the log-likelihood can be improved significantly. We also show that the proposed latent-variable approach coverages rapidly during updating. The resultant model translates at a speed around 8x-12x faster than the baseline autoregressive models. The translation quality of the proposed model outperforms baselines on ASPEC Japanese-English translation task. On WMT'14 English-German translation task, it narrows the performance gap between baselines down to 1.0 BLEU point.

As our second contribution, we apply the discretization bottleneck in an auto-encoder to learning discrete representation of words. The motivation is to compress neural models by replacing the giant word embedding matrix with discrete codes. Natural language processing (NLP) models often require a massive number of parameters for word embeddings, resulting in a large storage or memory footprint. In our approach, we assign each word a discrete code. To recover the word embeddings, we learn code vectors and composing them according to the codes. To maximize the compression rate, we adopt the multi-codebook quantization approach instead of binary coding scheme. Each code has multiple discrete numbers, such as (3, 2, 1, 8). We directly learn the discrete codes in an end-to-end neural network by applying the Gumbel-Softmax trick. Experiments show the compression rate achieves 98% in a sentiment analysis task and 94%-99% in machine translation tasks without performance loss.

The third contribution is to learn a sentence-level discrete representation. In this work, we add a planning phase in neural machine translation to control the sentence structure. Our approach learns discrete structural representations to encode syntactic information of target sentences. During translation, we can either let beam search to choose the structural codes automatically or specify the codes manually. The word generation is then conditioned on the selected discrete codes. Experiments show that the translation performance remains intact by learning the codes to capture pure structural variations. Through structural planning, we are able to control the global sentence structure by manipulating the codes. By evaluating with a proposed structural diversity metric, we found that the sentences sampled using different codes have much higher diversity scores. To summarize, our approach learns discrete representation for both word-level and sentence-level units with discrete bottlenecks. Through experiments, we are able to compress the neural model with a high compression rate with learned discrete codes. We also found a well trained discrete representation has high interpretability. Finally, we show that the sentence-level discrete representation is useful for controlling the sentence generation with additional constraints. By extending the training data with the discrete codes, no modification to the neural model is required.