A Study on Bidirectional Decoder of Neural Machine Translation

Yang Zhen (1610200)

School of Information Science, JAIST, andy.yang@jaist.ac.jp

Extended Abstract

Machine translation is one of the most active research areas in natural language processing (NLP) field. Although the history of machine translation study can be traced back to the middle of last century, until recently, most of the research is ranged from rule-based direct translation to interlingua method to statistical machine translation (SMT) method. Neural machine translation (NMT) is similar to an idea that appeared firstly in the 1990s, but no further development was undergone because of the constraints of data and computation resources at that time.

As the rise of deep learning, we have witnessed a great number of impressive achievements in various fields, especially computer vision, speech recognition and natural language processing. Of course, it also brings a revolution to machine translation. It is the appearance of NMT, which starts from using the neural network as a component in a phrase-based SMT system, and later developed to a pure neural network machine translation architecture. Although NMT has achieved a lot of impressive results, it still has many challenges, and it needs to make more refinements. Current most successful NMT models include three types: the recurrent neural network (RNN) based, the convolutional neural network (CNN) based and the pure attention based Encoder-Decoder sequence-to-sequence model. In this thesis, we focus on the RNN based one, especially it's decoding process because the current decoding process normally uses only unidirectional information of target sentence leaves the bidirectional information unexploited. Even though there are several works about using bidirectional information of target sentence, you have to train the backward model first, which means longer training time and bigger parameter size. So it is worth to explore training forward and backward decoding in an integrated model.

This thesis presents our research on NMT and our contributions: (1) we implement an NMT model, and make it becomes a strong baseline to compare with state-of-art models in terms of BLEU score through exploring parameters; (2) we implement several multi-task learning models to make bidirectional decoding and compare these models with each other and baseline, and then analyze the real translation result; (3) we also propose a regularization way to build bidirectional decoder and compare with other models, and analyze the translation result; (4) we combine the multi-task learning model and regularization model to make a combined model, which achieve further improvement than baseline model.

At first, we implemented a prevalent NMT model, the RNN based Encoder-Decoder model with attention mechanism. In this model, the source and target sentences are just viewed as sequences of tokens, and our model translates sequences in the source language to sequences in the target language. Based on this basic model, we explore the influence of beam size and auxiliary length and coverage penalty in the beam search decoding process to the translation result. By doing this, we hope to achieve a strong baseline model that can compete with state-of-art models. Our result shows in our model, the optimal beam size ranges from 5 to 10 and if we pass some points, the increase will not help translation and even make the results worse. The reason might be that the big beam size makes a shorter translation, which can lead to worse performance. And also for length and coverage penalty parameters, in terms of BLEU score, it can make a big improvement when compared with baseline model, especially when both of them take an approximate value of 0.33. However, in terms of NIST score, it seems there is no substantial improvement in the translation result. And for both metrics, it shows that both big length penalty and big coverage penalty does not help to improve translation results. Moreover, it can make the performance getting worse. After choosing the parameters, we achieve a strong baseline model, which has better performance than many state-of-arts models.

Our second experiment presented here is focusing on the multi-task learning (MTL) model, which takes forward and backward decoding as two related tasks and trains them together with some components shared between them. The sharable components include attention component, decoder word embedding, and generator, which outputs prediction from hidden states. According to sharing components, we proposed several models, in which some only share one component and some share multiple components at the same time. After implementing and training models, we make comparison among these models and baseline model. The result shows that the proposed models indeed make an improvement when compared with baseline model, especially for the model with sharing generator and the model with sharing both embedding and generator. When we did some deeper analysis of the real translation results, we found that the best-proposed MTL model actually captured more information than the baseline. However, this kind of sharing component model is a quite indirect way to make a connection between the forward and backward decoding. We can see that it will make worse results or no improved results when sharing some components, like sharing attention mechanism, and we do not know the reason clearly.

To make a more direct interaction between the forward and backward decoding, we can make a regularization directly between hidden states of forward and backward. We proposed two models. The first one uses$L2$ regularization loss directly between the forward and backward hidden states. But this method will make less flexibility of our model to produce hidden states. To make more flexible, we proposed the second model which add an affine transformation layer between the forward and backward hidden states. It transforms forward hidden states first before it calculates $L2$ loss. Next, we implemented these two models

and trained them with weight annealing technique, then compared their results with baseline model. We find these two models have better performance than baseline model. And between two proposed model, the model which has affine layer can achieve better results. It can even achieve higher results than the MTL models. When we analyze the translation sentence, it indeed captures more information than the baseline. However, for this model, the disadvantage is that we need to choose a good weight value for the regularization loss term.

After proposed two different ways to exploit bidirectional information of target sentences, we proposed two new models that take the best setting from MTL model, which is sharing embedding and generator model, and apply two kinds of regularization from above. We implement two models and compare them with the former proposed models and baseline model. We find it not only outperform the baseline model, but it also outperforms the former models, especially for the model with *L2* regularization have an affine layer. It becomes the strongest model among all proposed model. For deep analysis, we compare the performance of the best model and the baseline on different sentence length data. And we find the proposed model help the model to translate longer sentence a lot. Besides, for the generality of the proposed model, we also run several selected proposed models in more translation direction. It shows similar improvements. And we also run proposed models on the large dataset to further support the effectiveness of our proposed model. The result demonstrates the effectiveness of our model even training on the big dataset.

Although our work has several limitations, at least it can provide some help for other research work in this field. We wish the method we proposed here can inspire more similar ideas. The proposed methods and finding not only can be applied to NMT field, but also can be applied to many other tasks such as question answering and document summarization, or even image caption that are using RNN based Encoder-Decoder model.

Keywords: natural language processing, neural machine translation, sequence-to-sequence model, bidirectional decoding, multi-task learning, regularization