Rahul Thapa

Paper: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Summary

In this paper, the authors introduce a simple method to translate between multiple languages using a single model, taking advantage of multilingual data to improve NMT for all languages involved. Their method requires no changes to the traditional NMT model architecture. In fact, it is almost the same as the model presented in paper Wu et. al. However, the authors in this paper add an artificial token to the input sequence to indicate the required target language. Since they used only a single model to translate between various languages instead of training a lot of individual models for each language pair, it makes the model very simple and drastically reduces the number of free parameters. However, the most important aspect of this model is something the authors call "zero-shot translation". Basically, the model can learn to translate between language pairs it has never seen in the desired combination during training. This is a working example of how transfer learning works within the neural translation.

The model architecture used in this experiment is identical to Google's Neural Machine Translation (GNMT) with a simple modification to the input data, which is to introduce an artificial token at the beginning of the input sentence to indicate the target language. After adding the token to the input data, the authors train the model with all multilingual data consisting of multiple language pairs at once, possibly after over or under-sampling some of the data to adjust for the relative ratio of the language data available. To address the issue of translation of unknown words and to limit the vocabulary for computational efficiency, they use a shared wordpiece model across all the source and target data used for training.

The authors tested three different configurations of their multilingual model: many to one, one to many, and many to many. For example, in many to one, they train the model on many sources' languages and one target language. They also trained a single language pair model by oversampling, if necessary, and without oversampling. For another comparison, they also trained a single language pair model with the same total number of parameters as the baseline NMT models trained on a single language pair. For all the experiments, the multilingual models outperform the baseline single systems despite the disadvantage with respect to the number of parameters available per language pair. They explain that it might be because the model has been shown more English data on the target side and that the source languages belong to the same language families, so the model has learned to generalize well. In One to Many, they found out that the multilingual model is comparable to, and in some cases outperform baselines, but not always. Unlike the previous set of results, there are less significant gains. The authors hypothesize that this might be because the decoder has a more difficult time translating into multiple targets which may even have different scripts, which are combined into a single shared wordpiece vocabulary. They also observed that the oversampling helps the smaller language pair at the cost of lower quality for the larger language pair. In Many to Many model too, the average relative loss in BLEU score across all experiments is about 2.5%. They once again explore the impact of oversampling the smaller language pairs and found a similar trend to the previous experiment in which oversampling helps smaller language pairs at the expense of the larger ones. Although there are some significant losses in the quality of training many languages jointly using a single model, the authors claim that these models reduce the total complexity involved in training and production.

The authors explain that an important benefit of their approach is that it allows them to perform directly implicit bridging (zero-shot translation) between language pair for which no explicit parallel training data has been seen without any modification to the model. Of course, the model is not going to perform as good as other baseline models, however, it gives a very promising result. They even found out that when they trained the zero-shot multilingual model with some parallel data i.e. the direct translated data between the source and target language, they found out that they can easily improve the quality of their translation. The authors finally looked at the activations of the network during translation. Several trained networks indeed show strong visual evidence of a shared representation. Inspection of the visual representation of context vectors showed that the clusters of strands in the figure generally represent a set of translations of the same underlying sentence, but with different source and target languages. However, not all models show such clean semantic clustering. Sometimes, they observed joint embedding in some regions of space coexisting with separate large clusters which contained many context vectors from just one language pair. Therefore, an interesting question that the authors pose is what is the relationship between translation quality and distance between embeddings of the same semantic sentence?

Strengths

- The minor examples they give throughout the paper, such as the example of bridging between English to French and French to German helps explain the concept really well.
- Their experiments and result section are comprehensive. They test their model in various ways compared with the other standard models such as PBMT and a model trained to translate between just a single language pair.
- The visual representations are helpful to understand why their model work.

Weaknesses

- The authors repeat some information in between sections and sometimes even in the same section. For example, in section 4.5 when they are talking comparing the number of parameters in a single vs. multi-model.

Confusions

- I did not fully understand what kind of mechanism they were using on oversampling. I want to learn more about it.
- I also did not fully understand the hypotheses they made towards the end of the paper by using the visualizations.

Discussions

- What is a zero-shot translation and why is it called so?
- In what context a single multilingual model can be better than multiple production language pair models, even though the latter seems to outperform in most cases?
- What can we tell about the internal representation of the model? In other words, how is the model being able to do multilingual translation with comparable performance to that of a model trained on single language pair?