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

Nirajan Koirala

CSC 9010_001

Summary

NMT system's architecture have been mainly built for single language pair translations nd there has not been proposed a sufficiently simple and efficient way to handle multiple language pairs using this single model. In this paper authors introduce a method to translate between multiple languages using single model by taking advantage of multilingual data to improve NMT for all languages involved. This method requires no changes in the current architecture of NMTs and only an artificial token is added to the input sequence to indicate the required target language. It doesn't add any more complexities to the existing system and since no apriori decisions about allocating parameters for different languages need to be made, the system adapts automatically to use the total number of parameters efficiently to minimize global loss. Additionally, model built using these methods also simplifies production deployment significantly. All the parameters are implicitly shared by all the languages being modeled and this helps increase translation quality on the low-resource language pairs. One of the most important benefits of these models is zero shot translation using which model can learn to translate between language pairs it has never seen during training. Authors have went through different ways of merging languages on the source and target side and discuss their results on WMT benchmarks as well as on Google's large-scale production datasets.

Authors discuss the history of multilingual translation systems and their approach to this problem is related to the multitask learning framework. In section 3, they discuss their system's architecture and the only modification required in the input data, which is to introduce an artificial token at the beginning of the input sentence. Although this approach may have some disadvantages like ambiguity in words having same spelling, one of its main advantages is that it is simpler and can handle input with code switching. During the experiments, they apply the proposed method to train multilingual models in several different configurations. The models are then trained on WMT'14 datasets and they have also evaluated the multilingual approach on some Google-internal large-scale production datasets. Evaluation of their models is done using the standard BLEU score metric and they have tested the influence of training data per language pair by either using oversampling or skipping it.

During testing, the many to one models are found to outperform the baseline single systems despite the disadvantage with respect to the number of parameters. On production experiments as well they can see their models outperform the baseline systems by significant BLEU scores. The one to many model is not able to enjoy large gains in BLEU scores like the many to one but still it is found to perform reasonably well. After oversampling they are able to help the smaller language pairs at the expense of lower quality translation for larger language pair but they have found that this effect is less prominent on much larger production datasets. The many to many models are found to have average loss in BLEU scores across all experiment. Although the models built by training many languages have some significant losses in quality, these models help to reduce the complexities involved in their

training and productionization. Later in the section, they discuss the results of combining 12 production language pairs into a single multilingual model. This model has about 5 times less parameters than combined single models and also only requires on about $1/12^{th}$ of the training time to converge compared to the combined single models. The multilingual model is able to perform reasonably well and it enables them to group languages with little or no loss in quality while enjoying the benefits of training efficiency, smaller number of models and easier productionization. More importantly, their models allow zero-shot translation and it halves the decoding speed as no explicit bridging through a third language is necessary. To obtain better results, they introduce the model with small amount of true parallel data while still maintaining the half amount of decoding time when compared to explicit bridging. However, the translation quality is found to drop significantly when zero shot translation is used for unrelated languages like Spanish and Japanese.

In the later sections, they explore the ways of leveraging available parallel data to improve zero-shot translation quality. They found that their shared architecture models the zero-shot language pairs quite well and enables to easily improve their quality with small amount of additional parallel data. Visual analysis is done in section 5 where it is found that several trained networks show a strong visual evidence of a shared representation. They discuss the representations in further sub-sections and provide speculations and reasonings behind their formation. In section 6, authors try a very interesting approach to show how these models deal with mixed languages. They found that the translation for the mixed-language input differs slightly from both of the single source language translations while using source code language code-switching. When they try to mix the target language, they found that most of the time the output just switches from one language to another either partially or fully.

Finally in conclusion, they list down all the major benefits of multilingual NMT models like simplicity, zero-shot translation without explicit bridging and the details of their visual speculation of results. Additionally, their approach has been shown to work reliably in a Google-scale production setting and it enables them to scale to a large number languages quickly.

Strengths

- Authors have provided a very simple and elegant solution for building a multilingual NMT.
- The experiments are very detailed and wide ranging.
- Their methods are practical and it is confirmed by the reliable results in production setting.

Weaknesses

- I think the captioning the figures were a little longer than necessary.
- Some information is discussed multiple times throughout the paper and omitting them may help shorten the size.

Points of Confusion

- I am a little confused about section 5 where they discuss the visual analysis of their models.
- In section 6.2, I couldn't grasp the explanation provided by the authors behind the behavior of network while code-switching target languages.

Discussion Questions

- Authors have about discussed zero-shot translation for unrelated languages. Would the other models (one to many, many to one etc.) be able to produce respectable results when unrelated languages like German and Japanese are used?
- How would changing other factors like the depth of networks used to build these model affect their performance?
- The 4 discussion questions provided by the authors in section 5.