Jenish Maharjan **Week 5**

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

Summary:

In this paper, the authors present a system for multilingual Neural Machine Translation (NMT) using a single model with shared parameters. They show how having all the parameters shared has a positive side-effect on the translation quality of languages for which the resources are limited. They also demonstrate zero shot translation which requires no explicit bridging which they claim to perceive as the first instance of true transfer learning shown to work for machine translation. Their system is able to translate from a source language to a target language without having seen explicit examples from that specific source-target pair.

The model architecture they use is identical to the GNMT system with some optional modifications for some experiments. They introduce a simple modification to the input data in order to make use of multilingual data within the single system. The idea is to add a token to the input data that indicates the target language. Interestingly, they do not specify the source language which they claim is automatically learned by the model. They also implement a shared wordpiece model across all the source and target data used for training in order to address the issue of translation of unknown words and to limit the vocabulary for computational efficiency. The authors claim that their approach is the simplest among alternatives they were aware of.

The models are trained for three different scenarios – many source languages to one target language, one source language to many target languages and many source languages to many target languages. The dataset they use consist of English-German and English-French language pairs. They also use Google's internal large-scale production dataset for multiple language pairs. For the many to one case, the models are trained using a combined dataset for German to English and French to English pairs and compare with models trained in the one to one scenario. And similarly, the datasets are combined to train the models in the one to many and many to many scenarios. They present results for each of these scenarios and the models perform as good as the ones trained with one to one language pairs. They also present results for the large-scale experiments where 12 production language pairs were used to train a single multilingual model. Multilingual models perform very close to single models but the gap decreases as the models become larger.

While using explicit bridging is the most straight-forward approach of translating between languages with no or little parallel data is available, the method can have increased translation times and may potentially lose quality of translating when translating from the intermediate language. The approach proposed by the authors uses implicit bridging between a language pair to address this issue. They make a comparison between different models with different bridging techniques to verify the zero-shot translation approach. The authors also explore ways of parallel data to improve zero-shot translation quality by considering incremental training of the multilingual model on additional parallel data for the zero-shot directions and training of new multilingual model with all available parallel data mixed equally. They show a comparison of zero-shot, from-scratch and incremental training approaches. The authors also use methods of visual analysis to evaluate their models' performances. Another key feature they introduce is to be able to translate to a target language from a random source language for which they implement the source language code-switching technique and weighted target language selection.

Strengths:

- The authors make comparisons of their models with the baselines which gives a very good idea about where the models stand. The results shown in the tables increase the readability.
- The content of the paper does a great job of explaining the additional features and experiments with respect to their previously published work on GNMT. This allows the paper to focus on the novel ideas and concepts.

Weaknesses:

- Some information in the paper seem to be verbose and could have been made more concise.
- The paper requires the reader to have read the previous paper which might be seen as a drawback.

Confusions:

- How would you define direct Parallel Data?
- Why is that they don't mention the GLEU score in this paper?

Discussion Questions:

- A discussion on the visualization technique they use would be very useful.
- Are there any production level data available publicly for testing similar systems?
- Sometimes, even though the literal translation is correct, languages can have a deeper meaning that is different from what the words read. How can machine translation tackle this? Is extensive human evaluation the only way to go?