Proposal
Sacred-text Intelligent Coding and Embedding for Multilingual(SICEM) Neural Machine Translation

Nishan Dhoj Karki

09/21/21

Project Background

Johnson et al. in 2017 proposed a relatively simple model for neural machine translation compared to earlier efforts. More recently, in 2019, an intern at Google proposed a smaller model that was able to achieve comparable results on smaller corpus and many more languages. In this project, we will implement and improve these or the latest neural machine translation models by tailoring them to the particular case of a sacred text for which there are many translations available.

Project Objective

The SICEM Neural Machine Translation project will use the Bible text in many languages to train a model to translate to other languages. The input will be a token for the desired target language followed by to be translated. The output will be the sentence translated into that target language. The source language does not need to be specified, but the target language is. These will allow us to train many sentence language pairs and use the model to translate new sentences for which some languages have no known context. When the model is trained, the input should be a source sentence and the desired target language, and the output of this project will be a translated sentence.

Project Approach

We propose a simple solution to use a single Neural Machine Translation (NMT) model to translate between multiple languages. We will 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. New data is simply added, possibly with over or under-sampling such that all languages are appropriately represented, and used with a new token if the target language changes. Using a shared word piece vocabulary, our approach enables Multilingual NMT systems using a single model. Our models will try to adopt zero- shot translation so it can also learn to perform implicit bridging between language pairs.

Data Pre-processing

We will add an artificial token to the input sequence to indicate the required target language, a simple amendment to the data only. The data sets for the project will be the JW300 multilingual corpus from the OPUS webpage. Additional bible corpus will be provided by Dr. Rivas. We will prepare the training set and mini-batch retrieval functions that will enable the learning algorithm to call for mini-batches on-demand from the corpus of language pairs.

Model Selection

The multilingual model architecture is identical to Google's Neural Machine Translation (GNMT) sys- tem (Wu et al., 2016). We will apply our proposed method to train multilingual models in several different

configurations.
Zero-shot translation: A surprising benefit of modeling several language pairs in a single model is that the model can learn to translate between language pairs it has never seen. It has also been observed that when language pairs with little available data and language pairs with abundant data are mixed into a single model, translation quality on the low resource language pair is significantly improved.
After adding the token to the input data, we will train the model with all multilingual data consisting of multiple language pairs, possibly after over- or under-sampling some of the data to adjust for the relative ratio of the language data available. We will use larger batch sizes with a slightly higher initial learning rate to speed up the convergence of these models. We will over-sample the examples from the smaller language pairs to balance the data.
Many to Many: The model reduces the total complexity involved in training and productionization.

Testing and tuning

We will evaluate our models using the standard BLEU score metric and to make our results comparable to previous work (Sutskever et al., 2014; Luong et al., 2015c; Zhou et al., 2016; Wu et al., 2016), we report tokenized BLEU score as computed by the multi-bleu.pl script, which can be downloaded from the public implementation of Moses.

Work Schedule/ Division of work

Task	Timeframe	People
Data Collection and preprocessing	Sep - Oct1st	Nishan Karki, Swapnil Shah, Tanvir Razwan
Data Modelling	Sep - Oct 1st	Agm Islam
Data processing	Oct 1- Oct 15	Nishan Karki, Swapnil Shah, TanvirRazwan, Agm Islam
Data Modelling	Oct 1- Oct 15	Nishan Karki, Swapnil Shah, Tanvir Razwan, Agm Islam
Model Training	Oct 15- Nov 15	Nishan Karki, Swapnil Shah, Tanvir Razwan, Agm Islam
Model Testing and scoring	Oct 30 –	Nishan Karki, Swapnil Shah, Tanvir Razwan, Agm Islam
Improving model	Oct 30 –	Nishan Karki, Swapnil Shah, Tanvir Razwan, Agm Islam

References

[1] Johnson, Melvin, et al. ”Google's multilingual neural machine translation system: Enabling zero-shot translation.” Trans-actions of the Association for Computational Linguistics 5 (2017): 339-351.

[2] Aharoni, Roee, Melvin Johnson, and Orhan Firat. ”Massively multilingual neural machine translation.” arXiv preprintarXiv:1903.00089 (2019).