Julio Soldevilla

Project Proposal

Proposed Topic:

Unsupervised Neural Machine Translation for low resource domains

Problem importance

In recent years, supervised Neural Machine Translation (NMT) has seen incredible success when used on high resource languages. Survey [3] did a bunch of translation benchmark tasks from english to several low-resource languages and recognized that in the current state, unsupervised NMT on low resource languages is not practically useful since the translation results are not good enough. The survey recognized the two following factors affected the results:

1) The linguistic similarity of the source and target language
2) The domain similarity of training data between source and target language

In papers [1] and [2], the authors devised architectures that offer a solution for factor 1. In particular, paper [1] suggested using a particular pretrained Language Model for initializing the embeddings in the encoder and decoder section. Similarly, [2] suggested a multi-lingual neural machine translation architecture (translating from low-resource to high-resource language and viceversa) using a specific routine for pre-training (generating initializing embeddings for encoder-decoder section) where the architecture used the high-resource language as well as other high-resource languages that are semantically close to the low-resource language.

The questions we want to solve is if the suggested architectures (solutions) also solve factor 2) identified in survey [3]. Additionally, given that there might not be many high-resource domains close to low-resource domains, can we simplify the unsupervised multi-language NMT architecture suggested in [2] (borrowing the pre-training idea from paper [1]) for multi-domain translation?

If the above problems are solved, this would suggest we are close to solve some of the most important factors preventing unsupervised NMT from being practically useful.

Related papers

[1] “Reusing a Pretrained Language Model on Languages with Limited Corpora for Unsupervised NMT”, Chronopoulou et. al., https://arxiv.org/pdf/2009.07610.pdf
[2] “Harnessing Multilinguality in Unsupervised Machine Translation for Rare Languages”, Garcia, et. al. https://aclanthology.org/2021.naacl-main.89.pdf
[3] “When and Why is Unsupervised Neural Machine Translation Useless”, Kim et. al., https://arxiv.org/pdf/2004.10581.pdf
[4] “Neural Machine translation for Low-Resource Languages: A survey”, Ranathunga et. al. ,https://arxiv.org/pdf/2106.15115.pdf
[5] “A Simple Baseline to Semi-Supervised Domain Adaptation for Machine Translation”, Jin, et. al., https://arxiv.org/pdf/2001.08140.pdf

Goal

The main goal for the experimentation phases will be to achieve the highest possible BLUE score and surpass the baseline results which will be computed either taking them from previous papers that establish them as SOTA or computing them following the implementations/processes from papers [1] and [2].

Datasets to use

Julio Soldevilla

For this task we will use the datasets from WMT, in particular monolingual training data for English, French, German (for high resource languages), Kazakh, Turkish, the OSCAR dataset with languages like afrikaans, albanian and several low-resource languages.

Ideas to solve the problem

First, to test if the proposed architectures solve the problem for the mismatch in domain between source language and target language, we can apply the model in [1] to the domain data and see the results. My guess is that as it is, the proposed solution might not achieve high BLEU scores. Perhaps we could try, in the pretraining step, try to do some clustering of the terms and keep the terms in the high-resource language that are in the cluster that has higher similarity with the largest cluster of the low-resource language clustering.

For solving the problem of simplifying the multi-lingual NMT, we can try to replace the pretraining of the architecture proposed in [2] and replace it with some modification of the pretraining from [1]. I suspect this could simplify the architecture and still give relevant results.