Team 12 EnElPi: Writeup

Domain	Description
Machine Translation	ACL 2019 shared task - Machine Translation of News Articles

What we understood about the project

The WMT conference is one of the premier conferences in NLP advances. They came up with the **shared task on machine translation of news** in 2019**.** This task was attempted by several professional companies and institutions including FAIR (Facebook AI Research), Baidu, etc.

This project is to implement machine translation for a given language pair based on the corresponding datasets, both of which they have provided. Each language pair has its own set of resources, some are resource-rich (like German) while other are resource-poor (like Gujarati).

The participating teams have to submit translations of certain test set files and then manual evaluation is performed by the WMT team along with some automated evaluations.

What we are planning to do

Reading official approaches

We intend to first read official papers to get a feel for how to approach this task. We have outlined some premier papers to read:

1. "Facebook FAIR's WMT19 News Translation Task Submission" by Nathan Ng et. al.
    a. https://aclanthology.org/W19-5333/
2. "Microsoft Translator at WMT 2019: Towards Large-Scale Document-Level Neural Machine Translation" by Dowmnt
    a. https://arxiv.org/abs/1907.06170
3. Microsoft Research Asia's Systems for WMT19 by Xia et. al.
    a. https://aclanthology.org/W19-5348/

Choosing a language pair

We are considering choosing German-English language pair from the tasks, as it is the most resource rich language pair in the entire task. Being resource rich will help us train the model faster as we have smaller compute resources available.

Specifically, we intend to perform **English to German one-way translation**.
Note that just swapping the dataset (from input German to output English) *may not* be sufficient to still achieve decent performance. Hence, we are currently opting for one way translation. If time remains, we will also try to perform bidirectional translation.

Another advantage is that many competing teams have made a submission for this language pair, so we'll be able to compare our results with the official participants.

Dataset used

We looked through all the datasets that are listed on the WMT 2019 task page. We plan to use **Europarl v9 en-de** as our primary and only dataset. This dataset has 1,838,567 DE-EN sentence pairs, which we believe would be sufficient to train our model. The dataset is available here.

The other possible dataset is CommonCrawl corpus, however, we know that it has already been used by Transformers to train themselves in a monolingual manner. As transformers as already pretrained on that, there is no additional benefit for us to train further on that.

Training model

Baseline: seq2seq RNN using PyTorch Lightning. In case the model does not perform well (misses vocabulary or grammar), we will use the Teacher Forcing technique to improve it.

Baseline+: a small-size transformer obtained from HuggingFace repository. We will replace the RNN from our baseline with this transformer.

Advantages of baseline+ over baseline: the advantages are obtained primarily due to replacing the RNN with a transformer, which includes faster training time, larger model size possible in same time, better evaluation results.

Evaluation of inference results

We intend to use SacreBLEU as the evaluation metric as this has been used by many teams and was developed for this specific type of shared tasks. Specifically, we will use `BLEU+case.mixed+lang.en-

`de+numrefs.1+smooth.exp+test.wmt18+tok.13a+version.1.3` as has been used by Microsoft's submission and is given in this GitHub repo.

Timeline of project

1. **First five days (12th Nov-16th November)**
    a. understand the project
    b. prepare a writeup on what we intend to do
2. **Next seven days (17th November-23rd November)**
    a. Prepare Dataset and DataLoaders
    b. Implement model, train baseline and baseline+ models
    c. Resolve bugs
3. **Final seven days (24th November to 30th November)**
    a. Tune hyperparameters to improve performance
    b. Analyze results
    c. Prepare final report