Deep Learning for NLP research proposal: Multilingual Neural Machine Translation for Low-Resource Languages

Authors: Dominykas Šeputis, Szymon Budziak, Quim Serra Faber, Jozef Ciz

1 Executive Summary

Current **Large Language Models** (LLMs) excel in tasks like reasoning and translation for high-resource languages such as English or German. However, their accuracy declines sharply with low-resource languages like Slovak or Lithuanian due to limited exposure.

The main issue is the lack of data for these languages compared to English. This research aims to study the improvement of model accuracy for low-resource languages without adding new data. Specifically, we explore data quantity versus quality by applying various dataset augmentation strategies in a machine translation task. We will compare model performance when fine-tuned on augmented data generated through different back-translation techniques and sentence rephrasing methods using LLMs, against simply adding previously unseen data.

Our goal is to identify the optimal strategy for enhancing model accuracy for low-resource languages without requiring additional data collection.

2 Survey of Background Literature

Back translation is a technique used for leveraging monolingual data for NMT systems, introduced in the paper ”Improving Neural Machine Translation Models with Monolingual Data”, by Sennrich R., Haddow B., and Birch A. (2016). This technique is used for data augmentation, and it involves translating monolingual data in the target language back into the source language using an MT model and then using these synthetic parallel sentences to retrain the model. Yichuan et. al. explores high quality data augmentation by leveraging LLMs for multiple downstream tasks, where their proposed approach consistently generates augmented data with better quality compared to non-LLM and LLM-based data augmentation methods.

3 Proposed Methodology

3.1 Models

We plan to work with the Helsinki-NLP models, specifically designed for machine translation tasks. For our low-resource languages, Slovak and Lithuanian, we will use the following models as bases for fine-tuning: 'Helsinki-NLP/opus-mt-tc-big-ces_slk-en' for Slovak and 'Helsinki-NLP/opus-mt-tc-big-en-lt' for Lithuanian. Both models will be fine-tuned using the Hugging Face library, which provides a straightforward framework for model learning specific to tasks like machine translation.

3.2 Datasets

We have not yet finalized the exact dataset to use, as different data domains may vary in their compatibility with specific data augmentation techniques (e.g., LLMs). We will select from the openly available datasets provided by the opus.nlpl.eu website based on the data diversity and data quality (source reliability) criteria. Good candidates include OpenSubtitles, TildeMODEL, and TED2020.

3.3 Data Augmentation Techniques

We plan to use back-translation and LLM text generation for data augmentation. For back-translations, we will use either the 'google/t5-base' or 'facebook/nllb-200-distilled-600M' models, as they are popular on the Huggingface platform and yield good results. For LLM augmentation, we will employ Google's Gemini model.

4 Research Plan

Our objective is to determine if different data augmentation techniques can enhance language translation models for low-resource languages. The experimental setup will focus on Slovak and Lithuanian, using datasets not initially included in the training of machine translation models.

We will apply various data augmentation strategies to these new datasets and compare the performance of models trained on high-quality moderated data versus synthetically generated data. Specifically, our synthetic methods will include:

1. **Back Translation**:
   - Translate new data from the original language to English and back to shift word distribution.
   - Alternatively, translate to a similar language (e.g., Slovak to Czech/Polish) and back.
2. **Rephrasing with LLMs**:
   - Use Google's Gemini to rephrase the original text, maintaining meaning but altering word distribution. Experiment with prompts in both languages.
3. **Transfer Learning**:
   - Train a high-resource language pair for the parent model, then transfer parameters to the low-resource pair's child model.

If these augmentation methods are successful, we will combine the original and augmented datasets to assess whether model performance improves or degrades, validating the effectiveness of these methods for low-resource languages.

Model evaluation will be conducted using the BLEU algorithm. The BLEU score will serve as a key metric to quantify translation quality by assessing the correspondence between machine-generated and human-expert translations.

5 Bibliography

1. Improving Neural Machine Translation Models with Monolingual Data

2. Synthetic Data Generation in Low-Resource Settings via Fine-Tuning of Large Language Models

3. Neural Machine Translation For Low Resource Languages

4. Empowering Large Language Models for Textual Data Augmentation