Project Proposal: Lost in Translation and Transfer

Kelly Patel

20 October 2022

1 Introduction

Sentiment analysis has been a prominent topic of research in numerous areas of application like journalism, marketing, finance, etc. In recent years, data sources rich with opinions have become widely available due to the prominence of social networks. However, some data reflects highly-resourced languages like English, but is found to be lacking in other low-resource languages. As the world becomes more interconnected through global communications, it is becoming more and more imperative to analyze data from all over the world. Lexicons are not always available in other languages (marking them "low-resource languages") and it remains an expensive task to construct and label them.

This motivates me to build upon some existing approaches and run a sentiment analysis for different languages using relatively small datasets. For the sake of this project, the focus will be on Slavic languages (further discussed in the Data Section).

Previous efforts in sentiment analysis for multilingual corpora entailed the translation of the corpora into English to then be analyzed for opinions. Researchers have used language-specific pre-trained models to vectorize English sentences for analysis (using models such as Word2Vec). These embeddings are then passed through supervised machine learning classifiers such as Random Forest, Support Vector Machines, and Gradient Boosting Trees–which provide decent accuracies on out-of-sample data. In one study by Galshchuk, Jourdan, and Qiu in 2019, an accuracy of about 86% was achieved using this method.

I would like to explore whether there are any semantic details lost in the translation from a source language to English. It is often hard to capture the full connotative meanings of words between languages even in person-to-person communication. Therefore, I will be experimenting with transfer learning. Instead of translating from a source to English and risking the loss of some semantic detail in that translation, I propose using BERT to transfer knowledge from a pre-trained English paradigm and then fine-tuning the embeddings for a source language to then be analyzed for sentiment. This will be a domain-adaptation task, where I take a pre-trained BERT and freeze several of the first layers (trained on an English corpus) and re-train the final few layers on a low-resource language. This differs from multilingual BERT since that model was completely trained in other source languages, using high amounts of data from Wikipedia articles in those languages, whereas I would like to see the effects of transfer learning between two different languages. I would also like to perform several ablation studies, perhaps narrowing down which layers are more imperative for each language, whether there are some correlations between English and other languages that can be exploited, etc. This method is often referred to as cross-lingual transfer.

While this style of transfer-learning is not a novel idea, the application on sentiment analysis (especially in other languages) is not widely covered and I think worth looking into. Mostly, these multilingual models are tested on translation tasks (which is valid), but I do wonder if the sentiment is preserved between languages.

2 Research Plan

I propose to first more thoroughly investigate previous efforts in fine-tuning BERT as well as replicating efforts that do not make use of foundational models (like the Word2Vec method mentioned in the introduction). I expect to gather a baseline of performances and methods through this to further guide my own experiments.

As a base goal, I would like to implement a BERT-based model trained on an English corpus with varying numbers of frozen layers trained on another source language and study any correlations between which layers provide which information and whether certain languages have certain sensitivities. I will then use these trained embeddings on Twitter sentiment analysis tasks for different low-resource languages and evaluate the model's performance.

As a stretch goal, I would like to replicate these efforts with ELECTRA, which in fact has a higher GLUE score than BERT, but has less documentation in the realm of sentiment analysis and transfer learning. As I am still exploring machine learning as a field, I thought it might be more manageable for me to stick with the highly documented BERT.

2.1 Data

I will be using datasets in Slavic languages from CLARIN to perform my studies. These are standard .csv files with the tweet and an ID (-1, 0, 1) to indicate negative, neutral, and positive tweets. To compare to Galeshcuk's findings, I will be using corpora in Polish, Slovenian, and Croatian. These languages belong to the Indo-European family, but are members of the Slavic branch, which make them share fewer ties to English than the Latin based languages like Spanish or French.

The dataset comprises 2794 tweets in Polish (1397 positive and 1397 negative), 4272 tweets in Slovenian (2312 positive and 1950 negative) and 3554 tweets in Croatian (2129 positive and 1425 negative).

2.2 Evaluation

I will measure the quality of this model's performance by measuring the proportion of correctly classified observations (accuracy). I hope to also find the false-positive and true-positive rates to further analyze performance perhaps through a confusion matrix.

Expectations

Since multilingual-BERT and ELECTRA do achieve very good performance, I do not expect my own transfer efforts to beat them, but I do want to see if perhaps they are comparable in some regard, considering how much data is required for mBERT to train. If a transfer-learning model can perform at a similar level, then the need for larger resources reduces.

2.3 Timeline

1. **Week 1** (10/30/2022) Process Data (clean text) and implement tokenization.

2. **Week 2** (11/06/2022) Implement BERT without frozen layers. Acquire baseline performance.

3. **Week 3** (11/13/2022) Implement BERT with varying levels of frozen layers.
4. **Week 4** (11/20/2022) Evaluate performance on different numbers of unfrozen layers for Tweet sentiment analysis with low-resource languages.
5. **Week 5** (11/27/2022) Debugging and/or further ablation studies space.
6. **Week 6** (12/04/2022) Test mBERT on low-resource language Tweet Sentiment Analysis. Test ELECTRA on low-resource language Tweet Sentiment.
7. **Week 7** (12/11/2022) Finish Evaluations and prepare final report.

3 Background

One of the most limiting factors in machine/deep learning is the need for large amounts of data. We typically see state-of-the-art results in a fully supervised setting, where a large set of labeled data is mandatory. Unfortunately, labeling training data is expensive and time-consuming. It also often requires specialized knowledge. With these factors in mind, many researchers have devoted their time and efforts into finding ways to train ML models without a large vat of labeled data. One of these methods is called Transfer Learning, which is where I ground my project idea.

Transfer Learning is a machine learning method where a model developed for a task is reused as the starting point for a model on a second task. For this project, I will be using a pre-trained model approach, which involves selecting a source model, reusing/retraining a portion of the pre-trained model, and then tuning it for the task of sentiment analysis on a low-resource language.

I will start by using a pretrained-BERT-base, untrained on low-resource language data, but tested on the low-resource language data to acquire a baseline (12-frozen layers). I will create a fine-tuned custom classifier which extracts the last hidden layer of the [CLS] (Classification) Token and passes it through a single feed-forward neural network. This has been shown to improve accuracy for sentiment analysis by nearly 10 accuracy points compared to the baseline (Tran).

Once the baseline is acquired, I will be testing the effects of layer freezing by evaluating performance on the following number of frozen layers: [0, 2, 4, 8, 10, 12].

Following this, I will hopefully conclude whether or not Transfer Learning is a viable approach for low-resource language sentiment analysis with BERT by comparing to other efforts with mBERT and ELECTRA and non-foundational models.

4 Related Work

Deep learning has revolutionized NLP with powerful models. Efforts in multi-lingual models have been made in many facets by many groups. Among the most popular are the state-of-the-art mBERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2018), ELECTRA, XLM (Lample and Conneau, 2019) and XLM-R (Conneau et al., 2019). While XLM and XLM-R use BERT-large architectures, mBERT uses a smaller architecture (BERT-base). Of these models, XLM-R performs slightly better than the other models on the Cross-lingual Natural Language Inference (XNLI), which is a natural language inference task: given a premise and a hypothesis, does the premise entail or contradict the hypothesis (or is it neutral?).

Transfer Learning is also a broadly explored topic especially in the realm of NLP tasks. As it helps extract knowledge from a source text form and applies it with relative ease to a different use case, it is often used to shorten the time spent on textual analysis. Named Entity Recognition,

Model	D	#M	#lg	en	fr	es	de	el	bg	ru	tr	ar	vi	th	zh	hi	sw	ur	Avg
Fine-tune multilingual model on English training set (Cross-lingual Transfer)																			
mBERT	Wiki	N	102	82.1	73.8	74.3	71.1	66.4	68.9	69.0	61.6	64.9	69.5	55.8	69.3	60.0	50.4	58.0	66.3
XLM (MLM+TLM)	Wiki+MT	N	15	85.0	78.7	78.9	77.8	76.6	77.4	75.3	72.5	73.1	76.1	73.2	76.5	69.6	68.4	67.3	75.1
XLM-R	CC	1	100	**88.8**	**83.6**	**84.2**	**82.7**	**82.3**	**83.1**	**80.1**	**79.0**	**78.8**	**79.7**	**78.6**	**80.2**	**75.8**	**72.0**	**71.7**	**80.1**
Translate everything to English and use English-only model (TRANSLATE-TEST)																			
BERT-en	Wiki	1	1	88.8	81.4	82.3	80.1	80.3	80.9	76.2	76.0	75.4	72.0	71.9	75.6	70.0	65.8	65.8	76.2
RoBERTa	CC	1	1	**91.3**	82.9	84.3	81.2	81.7	83.1	78.3	76.8	76.6	74.2	74.1	77.5	70.9	66.7	66.8	77.8
Fine-tune multilingual model on each training set (TRANSLATE-TRAIN)																			
XLM (MLM)	Wiki	N	100	82.9	77.6	77.9	77.9	77.1	75.7	75.5	72.6	71.2	75.8	73.1	76.2	70.4	66.5	62.4	74.2
Fine-tune multilingual model on all training sets (TRANSLATE-TRAIN-ALL)																			
XLM (MLM+TLM)	Wiki+MT	1	15	85.0	80.8	81.3	80.3	79.1	80.9	78.3	75.6	77.6	78.5	76.0	79.5	72.9	72.8	68.5	77.8
XLM (MLM)	Wiki	1	100	84.5	80.1	81.3	79.3	78.6	79.4	77.5	75.2	75.6	78.3	75.7	78.3	72.1	69.2	67.7	76.9
XLM-R	CC	1	100	**88.7**	**85.2**	**85.6**	**84.6**	**83.6**	**85.5**	**82.4**	**81.6**	**80.9**	**83.4**	**80.9**	**83.3**	**79.8**	**75.9**	**74.3**	**82.4**

Figure 1: Multilingual Model Evaluation on XNLI

Intent Classification, Cross-lingual learning, Sequence Labeling and Sentiment Analysis are all realms where transfer learning has been applied.

As for the combination of Transfer learning and multilingual modeling, there have been several attempts made. Kanclerz et al. (2020) attempt a novel technique for the use of language agnostic sentence representations to adapt the model trained on texts in Polish to recognize polarity in texts in other (high-resource) languages. They focus on creating a language-agnostic representation of each sentence and then predicting the sentiment of the text based on these representations.

Other groups focus on vectorization techniques like Word2Vec in tandem with machine learning methods like random forest, gradient boosting trees, and support vector machines (Galeshchuk et al., 2019). These efforts translate the low-resource language into English to perform the sentiment analysis and achieve decent results, but are overshadowed by the aforementioned large language models.

References

[1] Galeshchuk, Jourdan, Qiu. *Sentiment Analysis for Multilingual Corpora*, Aclanthology. code 2019

[2] Moberg, John. *A Deep Dive into Multilingual NLP Models*, Peltarion. 23 Feb. 2020

[3] Kanclerz, Milkowski, Kocon. *Cross-lingual deep neural transfer learning in sentiment analysis*, ScienceDirect. 2020.

[4] Devlin, et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, arxiv. 11 Oct 2018.

[5] Clark et al. *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*, arxiv. 23 Mar 2020.

[6] Lample, Conneau. *Cross-lingual Languaage Model Pretraining*, arxiv. 22 Jan 2019.

[7] Pires, Schlinger, Garette. *How multilingual is Multilingal BERT?*, arxiv. 4 Jun 2019.

[8] B. Raphael *How many layers of my BERT model should I freeze?* Raphael B.'s Blog 10 May 2021.

[9] Tran, Chris. *Fine-Tuning BERT for Sentiment Analysis* SkimAI