CS224U: Experimental procedure: QA and low resource language

October 2022

1 Introduction

Most recent progress in open QA is made for English. Extending this approach to multilingual open QA poses new challenges. Answering multilingual questions requires retrieving evidence from knowledge sources of other languages than the original question since many languages have limited reference documents or the question sometimes inquires about culture-country-specific topic. Previous work in multilingual open QA translates questions into English, applies an English open QA system to answer in English, and then translates answers back to the target language. Such approaches are affected by error propagation of the machine translation performed twice, especially for low-resource languages.

The project is devoted to Polish language in the context of (open)QA design. During the last several years several multilingual models have been made available. On the other hand it is not surprising that models targeted for specific language perform better than multilingual ones [1].

Moreover, specialized architectures are typically smaller due to significantly reduced vocabulary size and they can be trained efficiently [2].

Recently such dedicated model: HerBERT was designed, trained and made available [3]. Also an important milestone is publication of the GLUE analog, KLEJ [4] that allows for bench-marking the models.

2 Hypotheses

Assumptions and hypotheses:
We assume that we can use multilingual language models treating Polish as low resource language and following the methodology designed for such languages. One hypothesis is that continual pre-training of mBART should improve the results.

Longformer architecture and other alternatives for long document are an interesting alternative to BERT. My hypothesis is that such architecture should perform better due to the typical long length of the phrases and sentences in

Polish language [1].

Another hypothesis is that the dedicated language model can perform better than multiligual ones.

3 Data: A description of the dataset(s) that the project will use for evaluation.

Polish language is low resources language in the context of NLP. For that reason there are not many datasets.

For the context of of QA I will use the following:
- Did You Know (DYK) Dataset contains 4,721 question-answer pairs obtained from the Czy wiesz (Do you know) Wikipedia project
- Common crawler CC100-Polish Dataset
- Wikipedia in Polish

There are several product reviews datasets that also might be utilized to augment the pretraining:
- Clarin—Polish entries on the social platform Tweeter
- PolEmo 2.0—Multidomain product review
- AllegroReviews—Multidomain product review
- PolEmo2.0-IN OUT Dataset

Recently the Polish analog of GLUE has been developed. It is a comprehensive multi-task benchmark for the Polish language understanding - KLEJ (eng. GLUE, also abbreviation for "Kompleksowa Lista Ewaluacji Językowych", eng. Comprehensive List of Language Evaluations). KLEJ consists of nine tasks and, similarly to GLUE, is constructed out of existing datasets. The tasks were carefully selected to cover a wide range of genres and different aspects of language understanding.

4 Metrics

I plan to use f1 as metrics for question answering and KLEJ (GLUE) benchmark.

For the multilingual model and machine translation I will use separate metrics: SQuAD for QA and BLEU for the machine translation stages of the project.

5 Models

Several models will be used. Some comparison have been made for other languages with conclusions that some standard models can perform worse than when task is limited to English. Below are the models that are available of HuggingFace or github.

Byte-Pair Encoding tokenizer has been used in all publications devoted to application related to Polish language (BERT uses WordPiece).

5.1 Multilingual models

Pre-trained multilingual models have been extensively used in cross-lingual information processing tasks. The following models were successfully used: mBERT, XLM, XLM-R. In the article [5] it was found that the order of ability in preserving language identity of whole model is: mBERT > XLM-R > XLM and that all three models capture morphological, lexical, word order and syntactic features well, but perform poorly on nominal and verbal features.

I plan to use these 3 models.

5.2 BERT, XLM, ELECTRA

I plan to use BERT, XLM and ELECTRA with machine translation. (Note: In one of the articles XLM has been reported to perform subpar the Polish-English task).

5.3 HerBERT

HerBERT is a BERT-based Language Model (available ofn HiggingFace) trained on Polish corpora using Masked Language Modelling (MLM) and Sentence Structural Objective (SSO) with dynamic masking of whole words [3].

5.4 Longformer

Polish language has often much longer sentences than English. Transformer-based models are unable to process long sequences due to their self-attention operation, which scales quadratically with the sequence length. Because of this I plan to use and evaluate the performance of Longformer - model [6]. Longformer has an attention mechanism that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer. Longformer's attention mechanism is 'a drop-in replacement for the standard self-attention and combines a local windowed attention with a task motivated global attention'. Pretrained Longformer consistently outperformed RoBERTa on long document tasks on WikiHop and TriviaQA [6].

6 General reasoning

Polish language is one of the languages with very complex grammar, vocabulary that has complex origin and special letters in the alphabet. For the purposes of NLP it is low resource language. Since a lot of data can be accesses through Common Crawl I will use it. One possible outcome is production of the good quality data set from the multilingual Common Crawl.

The main part of the project is dedicated to collecting different resources, pretraining the existing models and applying it to OpenQA task.

There are several open questions that one can explore.At this point the effort should be focused on getting best model with limited datasets available for pre-training.

7 References

[1]See e.g.: Louis Martin et al. CamemBERT: a tasty French language model, In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 7203; Hang Le et al., FlauBERT: Unsupervised language model pre-training for French. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 2479

[2] Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin, 2019, Tuning multilingual transformers for language-specific named entity recognition, In Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pages 89–93, Florence, Italy. Association for Computational

[3] HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, Ireneusz Gawlik ACL/arxive

[4] KLEJ: Comprehensive Benchmark for Polish Language Understanding Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik, ACL/arxive

[5]Probing language identity encoded in pre-trained multilingual models: a typological view Jianyu Zheng and Ying Liu, arxive,ACL

[6]Longformer: The Long-Document Transformer Iz Beltagy Matthew E. Peters Arman Cohan, arxive