Project Topic

Jayveersinh Raj, Makar Shevchenko, Nikolay Pavlenko

February 2022

1 Project Topic

After considering various options presented to us, we have decided to select a topic of our own. The decision was driven by our desire to create a more original work, as all the presented topics were already completed by the students who passed the NLP course in previous years. However, in order to make sure that our topic would be compliant with the general direction of the course and would allow us to explore the NLP domain, we made sure that it partially covers the same areas of interest as some of the topics on the list: toxic comment classification challenge for text classification, and multilingual word sense disambiguation.

Our own topic, named **Cross-Lingual NLP**, will focus our project on developing NLP models that can effectively process and translate multiple languages, with special attention given to low-resource languages.

2 Topic Description and Motivation

Cross-lingual (or multilingual) models' defining characteristic is that they are pre-trained on text belonging to a mix of different languages. Such method of pre-training allows them to perform better in terms of cross-lingual classification than alternative cross-lingual sentence encoders, which are trained on text belonging to one language (such as BERT, that was pre-trained on English Wikipedia and BookCorpus).

One of the aspects of cross-lingual models that interested us the most in the topic is zero-shot transfer. It provides a way for a multilingual model to take the knowledge it learnt in high-resource languages like English, and to apply it on a low-resource language, such as Nepali, achieving much better accuracy than models that were tailored specifically for the latter language and had to contend with significantly smaller datasets.