Project Proposal NLP

Anna Maria Gnat
agna@itu.dk

Miranda Speyer-Larsen
spey@itu.dk

Josefine Nyeng
jny@itu.dk

1 Introduction

Named Entity Recognition is an important part of Natural Language Processing, however, under-resourced languages often lack dedicated NER models. For such languages, it is common to use either multilingual or English-based models. In this project, we want to investigate whether utilizing a model from a well-established language with a higher degree of similarity, determined through measures such as linguistic distance in the family tree, can improve the NER model performance for under-resourced languages.

2 Current literature

Building natural language processing models to be applied to different languages is a well-researched field, specifically with the focus of developing models that perform well for languages other than English. Several multi-lingual models have been developed such as Google's mBERT (Multilingual BERT). These models train on a large corpus of multilingual data allowing them to capture cross-lingual information (Devlin et al, 2018). Such multilingual language models enable the use of a single Named Entity Recognition model across various languages. Other approaches to multi-lingual Named Entity Recognition include the use of machine translation. A Cross-lingual Entity Projection framework (CROP) was developed which uses a multilingual labeled sequence translation model to translate the target language into the source language allowing an NER model in the source language to be applied. Then a labeled sequence translation model was developed to project the labeled sequence back to the target language (Yang et al., 2022). Unsupervised models have also been proposed, which transfer NER knowledge from one language to another (Bari et al., 2019).

3 Our Project

3.1 New Addition

The use of language similarity in building NER models for under-resourced languages is the new element that this project proposes. This could be useful for languages with very little training data available, allowing NER models to be trained on a well-resourced similar language and then applied to the target language.

3.2 Project set-up

The approach for this project will be to have an under-resourced language such as Serbian as the target language. Then train an NER model in Russian, which is a more similar language in the same family, and see if this model performs better than a model trained in the more common approach, English. The hypothesis is that being trained on a language with a larger degree of language similarity will improve the model's performance on the unseen target language.

A pretrained multilingual BERT language model combined with a linear layer will be used as the NER model. A separate model will be fine-tuned on each training language, Russian and English for the NER task. The performance of each model will then be evaluated on the target language, Serbian.

Further investigation could be done by conducting a similar experiment as described above for a new set of languages to ensure that the hypothesis holds true for other languages as well.

3.3 Current state

Currently, we have a baseline model with the architecture described above, trained on the EWT dataset. We also found multiple datasets with NER annotations in several languages, that we can use to test our hypothesis. Furthermore, we have found resources that allow us to test and quantify the language similarity.

4 References

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2018), BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Hugging Face, https://huggingface.co/google-bert/bert-base-multilingual-cased

Yang, J., Huang, S., Ma, S., Yin, Y., Dong, L., Zhang, D., Guo, H., Li, Z., Wei, F, (2022), Crop: Zero-shot cross-lingual named entity recognition with multilingual labeled sequence translation, arXiv.org, https://arxiv.org/abs/2210.07022

Bari, M. S., Joty, S., Jwalapuram, P. (2019), Zero-resource cross-lingual named entity recognition, arXiv.org, https://arxiv.org/abs/1911.09812