Team 21a's Submission to the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages

Anonymous ACL submission

Abstract

In this paper, we describe Team 21a's submission to the constrained track of the SIGTYP 2024 Shared Task. Using only the data provided by the organizers, we built transformer-based multilingual models finetuned on the Universal Dependencies (UD) annotations of a given language. We also explored the effect of different data mixes, and the cross-lingual capability of our trained models. [Our systems achieved]$_{\text{LJ}}$

1 Introduction

This paper describes Team 21a's submission to the *constrained* track of the SIGTYP 2024 Shared Task on Word Embedding Evaluation for Ancient and Historical Languages. Our general approach involves pretraining a transformer-based multilingual model on the shared task dataset, and then finetuning the pretrained model using the Universal Dependencies (UD) annotations of each language. Throughout this paper, we will refer to the pretrained model as LIBERTUS. We also explored data sampling and augmentation techniques during the pretraining step to ensure better generalization performance.

Our systems achieved...[stuff]$_{\text{LJ}}$

We detail our resource creation, model pretraining, and finetuning methodologies. In addition, we also show the results of our cross-lingual transfer learning set-up.

2 Methodology

2.1 Resource creation

We constructed the pretraining corpora using the annotated tokens of the shared task dataset. Then, we explored several data augmentation techniques to ensure that each language is properly represented (computed by the number of unique tokens).

From our experiments, **upsampling underrepresented languages** helped improve our pretraining validation loss. Figure **??** shows that LATM has the most number of unique tokens in the corpora. We upsampled each language by randomly adding a document from its original pool until the number of unique tokens is greater than or equal to that of LATM.

2.2 Model Pretraining

We then pretrained two transformer models—LIBERTUS$_{\text{base}}$ and LIBERTUS$_{\text{large}}$. Both models follow the RoBERTa pretraining architecture (Liu et al., 2019). The goal of these models is to serve as bases for finetuning downstream tasks.

2.3 Model Finetuning

3 Results

3.1 Benchmarking results

3.2 Cross-lingual transfer

3.3 Ablations

References

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.