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.

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 data mix provided by the organizers, and then finetuning it 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 data augmentation techniques during the pretraining step to ensure better generalization performance.

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

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

2 Methodology

2.1 Data Preprocessing

2.2 Model pretraining

2.3 Model finetuning

3 Results

3.1 Benchmarking results

3.2 Cross-lingual transfer

3.3 Ablations

	Full	Per-language
CHU		
FRO		
GOT		
GRC		
HBO		
ISL		
LAT		
LATM		
LZH		
OHU		
ORV		
SAN		

Table 1: Comparison between finetuning a model for each language vs. the whole set (dev). [maybe make this into a chart later, then put the table in the appendix]$_{\mathrm{LJ}}$

	LiBERTus	XLM-RoBERTa	mBERT
CHU			
FRO			
GOT			
GRC			
HBO			
ISL			
LAT			
LATM			
LZH			
OHU			
ORV			
SAN			

Table 2: Main results (test)