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.

This is an appendix.

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. We also explored data sampling and data augmentation techniques during the pretraining step to ensure better generalization performance.

Our systems achieved...

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.