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 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 based on the number of unique tokens.

From our experiments, **upsampling underrepresented languages** helped reduce our pretraining validation loss. Figure **??** shows that LATM has the most number of unique tokens in the corpora. We upsampled each language by randomly duplicating a document in the training pool until the number of unique tokens is greater than or equal to that of LATM. The same figure also shows the token counts after the augmentation process.

2.2 Model Pretraining

Using the pretraining corpora, we trained two variants of LIBERTUS, a Base model with XXXM parameters and a Large model with XXXM parameters, that serve as foundations for finetuning downstream tasks. LIBERTUS follows RoBERTa's pretraining architecture (Liu et al., 2019) and takes inspiration from Conneau et al. (2020)'s work on scaling BERT models to multiple languages.

Our hyperparameter choices closely resemble that of the original RoBERTa implementation as seen in Table **??**. We also trained the same BPE tokenizer (Sennrich et al., 2016) using the constructed corpora. During model pretraining, we used the AdamW optimizer with β_2=0.98 and a weight decay of 0.01. The Base model underwent training for 100,000 steps with a learning rate of 2e-4 whereas the Large variant trained for 300,000 steps. We used a learning rate scheduler that linearly warms up during the first quarter of the training process, then linearly decays for the rest. Figure **??** shows the training curve for both variants.

2.3 Model Finetuning

3 Results

3.1 Benchmarking results

3.2 Cross-lingual transfer

3.3 Ablations

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco

Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.