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 pretrained a transformer-based multilingual model, then finetuned it on the Universal Dependencies (UD) annotations of a given language for a downstream task. Our systems achieved [on the test set...]$_{LJ}$. On the validation set, we obtained $\geq$70% F1-score on most language-task pairs. We also explored the cross-lingual capability of our trained models. This paper highlights our pretraining and finetuning process, and our findings from our internal evaluations.

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]$_{LJ}$. Table 1 shows our systems' performance on the shared task test set. On the validation set, we obtained $\geq$70% F1-score for the majority of language-task pairs.

We detail our resource creation, model pretraining, and finetuning methodologies. The source code for all experiments can be found on GitHub: https://anonymous.4open.science/r/sigtyp-1F05/.

	POS tag.	Morph. annot.	Lemma.
CHU	0.95	0.92	0.92
COP	0.43	-0.91	0.46
FRO	0.85	0.83	0.84
GOT	0.93	0.90	0.91
GRC	0.93	0.87	0.88
HBO	0.27	-1.38	0.62
ISL	0.94	0.89	0.95
LAT	0.92	0.90	0.92
LATM	0.94	0.91	0.97
LZH	0.82	0.79	1.00
OHU	0.94	0.92	0.70
ORV	0.91	0.87	0.78
SAN	0.87	0.85	0.83

Table 1: SIGTYP 2024 Shared Task final leaderboard results as evaluated on the test set.

2 Methodology

2.1 Model Pretraining

The main purpose of pretraining is to obtain context-sensitive word embeddings that we will finetune further for each downstream task. We approach this by training a multilingual language model akin to the XLM-RoBERTa (Conneau et al., 2020) and multilingual BERT (Devlin et al., 2019) architectures.

Preparing the pretraining corpora. We constructed the pretraining corpora using the annotated tokens of the shared task dataset. Initially, we explored several data augmentation techniques to ensure that each language is properly represented based on the number of unique tokens. However, we found pretraining to be unstable when we upsampled tokens to achieve the same count as LATM, the most overrepresented language. In the end, we found that leaving the token distribution as-is leads to more stable pretraining and lower valida-

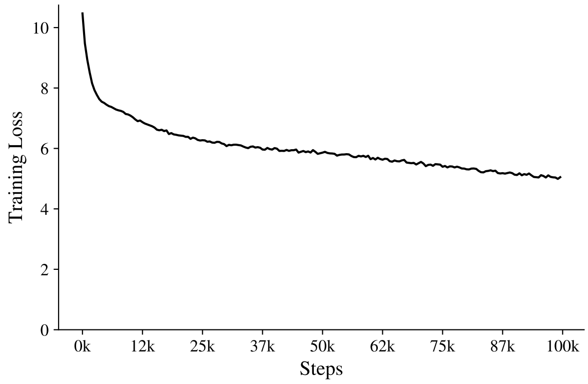

Figure 1: Training loss curve for the 126M-parameter model after 100k steps.

Hyperparameters	Value
Hidden size	768
Intermediate size	3072
Max position embed.	512
Num. attention heads	12
Hidden layers	12
Dropout	0.1

Table 2: Hyperparameter configuration for the LIBERTUS pretrained model.

tion scores.

Pretraining the base model. Using the pretraining corpora, we trained a model with 126M parameters that will serve as a base 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 2. 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 100k steps with a learning rate of 2e-4. We used a learning rate scheduler that linearly warms up during the first 12k steps of the training process, then linearly decays for the rest. Figure 1 shows the training curve.

2.2 Model Finetuning

For each language, we finetuned a multitask model using spaCy (Honnibal et al., 2020). We used spaCy's tokenization rules for the majority of languages except for LZH, where we segmented on characters. The final system consists of a parts-of-speech (POS) tagger, morphological analyzer, and lemmatizer.

Parts-of-speech (POS) tagger. We employed a standard classifier that predicts a vector of tag probabilities for each token. Each POS tag is a unique class that we assign exclusively to a token. We trained a model by taking the context-sensitive vectors from our pretrained embeddings, and passing it to a linear layer with a softmax activation. The network is then optimized using a categorical cross-entropy loss. For languages with subtokens such as COP and HBO, we merged each subtoken and used the full multi-word expression (MWE) during training.

Morphological analyzer. Similar to the POS tagger, we treat morphological annotation as a token classification task. Instead of directly modeling each feature, we made every unique combination of morphological features as a class. The limitation of this approach is that it can only predict combinations that were present in the training corpora. Similar to the POS tagger, we merged each subtoken for every multi-word expression (MWE) during training.

Lemmatizer. We trained a neural-based edit tree lemmatizer (Müller et al., 2015) by first extracting an edit tree for each token-lemma pair. Because this process can result to hundreds of edit trees, we treat the problem of picking the correct tree as a classification task. Here, each unique tree serves as a class and we compute a probability distribution over all trees for a given token. To obtain the most probable tree, we passed the context-sensitive embeddings from our pretrained model to a softmax layer and trained the network with a cross-entropy loss objective. We set the minimum frequency of an edit tree to 3, and used the surface form of the token as a backoff when no applicable edit tree is found. Finally, we ensured that the lemmatizer checks at least a single tree before resorting to backoff.

Finetuning the pipelines. We trained each component of the system in parallel, although the final "pipeline" assembles them together using the spaCy framework. For all components, the pretrained embeddings are passed on to linear layer with softmax activation. Sometimes, the tokenization from the multilingual model does not align one-to-one with

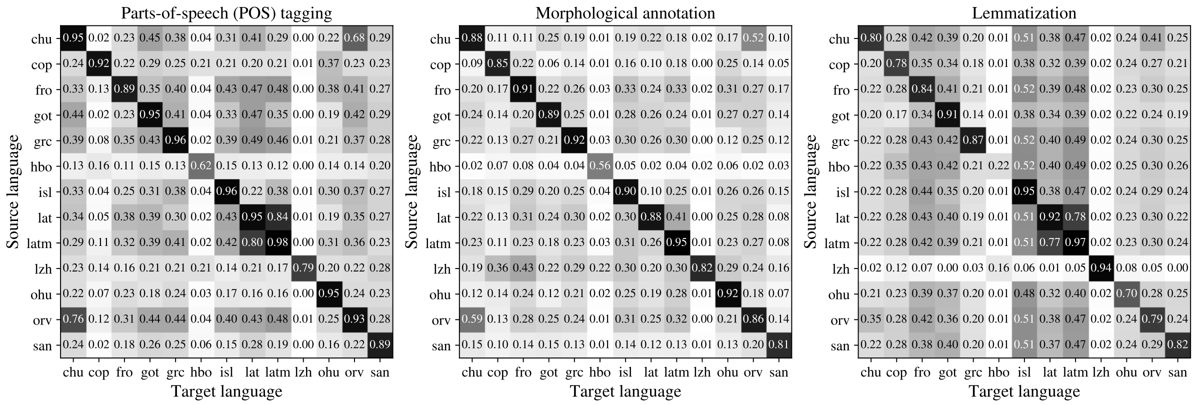

Figure 2: Cross-lingual evaluation given a monolingual model from one language and a validation set in another.

	POS tag.	Morph. annot.	Lemma.
CHU	0.947	0.876	0.803
COP	0.924	0.846	0.776
FRO	0.890	0.912	0.844
GOT	0.951	0.886	0.914
GRC	0.956	0.915	0.873
HBO	0.624	0.561	0.219
ISL	0.963	0.901	0.949
LAT	0.949	0.882	0.922
LATM	0.984	0.951	0.968
LZH	0.795	0.824	0.942
OHU	0.953	0.919	0.697
ORV	0.933	0.859	0.787
SAN	0.888	0.811	0.817

Table 3: F1-score results on the validation set.

spaCy's tokenization. In such case, we use a pooling layer that computes the average of each feature to obtain a single vector per spaCy token.

During finetuning, we used the Adam optimizer with β_1=0.9, β_2=0.999 and a learning rate of 0.001. The learning rate warms up linearly for the first 250 steps, and then decays afterwards.

3 Results

Table 1 shows the test scores for the shared task. [Talk about how you ranked against other teams?]$_{\text{LJ}}$ In this section, we will outline our internal evaluations and benchmarking experiments.

3.1 Performance on the validation set

Table 3 shows the validation scores of our finetuned models. We achieved $\geq$70% performance in most language-task pairs. The top performers, calculated by taking the average across all tasks, are LATM (0.968), ISL (0.938), and LAT (0.918), whereas the bottom performers are COP (0.849), SAN (0.839), and HBO (0.468).

Compared to our validation scores, our leaderboard scores on COP and HBO are poor. This performance is due to our models being unable to accurately predict subtoken information as it has only seen the full MWE during training. In order to align our tokenization with the shared task's validation script in Codalab, we substituted each MWE with its subtokens resulting to potentially incomprehensible text. Finally, for empty tokens such as those found in ORV, we added a rule in our system to produce empty predictions.

3.2 Evaluating cross-lingual capabilities

To test the cross-lingual capability of a language, we evaluated its finetuned model to the validation set of another. Figure 2 shows the results.

We found that it is possible to adapt a language onto another for morphological annotation and lemmatization. However, this does not extend to its morphology, as the validation set performs best only in the language it was trained on.

Some target languages tend to be cross-lingually receptive on lemmatization, i.e., many source languages can perform decently when applied to them. This observation is true for FRO, GOT, ISL, LAT, and LATM. Finally, there is also good cross-lingual compatibility between LAT and LATM—which is expected because they came from similar roots.

4 Conclusion

This paper describes Team 21a's system: a pretrained multilingual model (LIBERTUS) finetuned on different languages for each downstream task. Our system obtained [describe your rank, how you stack, etc. etc.]$_{\mathrm{LJ}}$

Our system's main strength is its downstream performance. The validation scores are high ($\geq$70% F1-score) for the majority of language-task pairs. However, performance on languages such as COP and HBO are low because of our downstream tokenization process. We highly recommend exploring alternative approaches to tokenization.

We also evaluated each language's cross-lingual capability and showed that transfer learning is possible especially on lemmatization. This approach can be a viable alternative on limited corpora.

Our training and benchmarking source code is on GitHub: `https://anonymous.4open.science/r/sigtyp-1F05/`. We will also release the pretrained multilingual model and finetuned pipelines in public shortly after the review period.

Limitations

Pretrained LM size. Due to compute constraints, we were only able to pretrain a model akin to the size of RoBERTa$_{\text{base}}$. We highly recommend pretraining a large LIBERTUS model to obtain performance gains if the resource allows.

Pretraining data mix. In the end, we didn't employ any sampling strategy to balance the token distribution of different languages during pretraining. We only tested simple up-/downsampling strategies and our experiments are limited to repeating available data.

Label combination as individual classes When training the morphologizer and POS tagger, we treated each feature and parts-of-speech combination as its own class instead of modeling them individually. This limits our text classifier to only predicting combinations it has seen during the training process.

Subtoken performance for multiword expressions Our systems performed poorly on COP and HBO in the leaderboard due to how we trained our model. Instead of showing subtokens, we used the full multi-word expression during training.

References

Burton H. Bloom. 1970. Space/time trade-offs in hash coding with allowable errors. *Commun. ACM*, 13(7):422–426.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

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*.

Lester James Miranda, Ákos Kádár, Adriane Boyd, Sofie Van Landeghem, Anders Søgaard, and Matthew Honnibal. 2022. Multi hash embeddings in spaCy.

Thomas Müller, Ryan Cotterell, Alexander Fraser, and Hinrich Schütze. 2015. Joint lemmatization and morphological tagging with lemming. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2268–2274, Lisbon, Portugal. Association for Computational Linguistics.

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.

A Appendix

A.1 Different sampling strategies on pretraining validation performance

We explored different sampling strategies and their effect on the pretraining validation loss curve as shown in Figure 3. We ran the pretraining pipeline for 20k steps (one-fifths of the final hyperparameter value) and measured the validation loss. The evaluation corpus was built from the validation set of the shared task, and we kept it the same throughout the experiment. We tested the following sampling strategies:

- **None:** we used the original dataset without any data sampling or augmentation.
- **Upsampling:** we upsampled each language to ensure that the number of their unique tokens is greater than or equal to the most dominant language.
- **Averaging:** we took the average number of unique tokens in the whole set and up-/downsampled each language based on this value.

Because any form of sampling resulted to unstable pretraining and higher validation loss, we decided to stick with the dataset's original data distribution. We highly recommend exploring alternative data mixes to ensure that all languages will be represented while keeping the training process stable.

A.2 Finetuning a model per language vs. monolithic system

We investigated if finetuning a model per language is more effective against a monolithic system, i.e., training on the full multilingual annotated corpora. Here, we combined the training corpora for all languages, then shuffled them before batching. The merged dataset has 194,281 documents for training and 26,954 documents for validation. This means that the downstream model sees a language mix per training epoch.

As shown in Figure 4, finetuning a model per language still yields the best results. One advantage of language-specific models is that we were able to set a different tokenizer per language—enabling us to get decent scores on LZH. Training the monolithic model is also sensitive to the training data distribution, as shown by the disparity in performance between majority languages (LATM, LAT) and minority ones (OHU, ORV, SAN). Due to these findings, we decided to train multiple models for our final system.

A.3 Alternative approach—multi-hash embeddings

We considered using multi-hash embeddings (Miranda et al., 2022) as an alternative approach. Instead of pretraining, these embeddings use orthographic features (e.g., prefix, suffix, norm, shape) to create a word vector table. This approach also applies the hashing trick, inspired by Bloom filters (Bloom, 1970), to decrease the vector table's memory footprint.

Figure 5 shows the results in comparison to our final system. It is notable that simple orthographic features are competitive with our transformer-based model. However, we chose to submit the transformer-based pipeline as our final system because it still outperforms the multi-hash embed method in the majority of our language-task pairs. We still recommend investigating this approach further because the hash-embed method has noticeable efficiency gains in terms of model size.

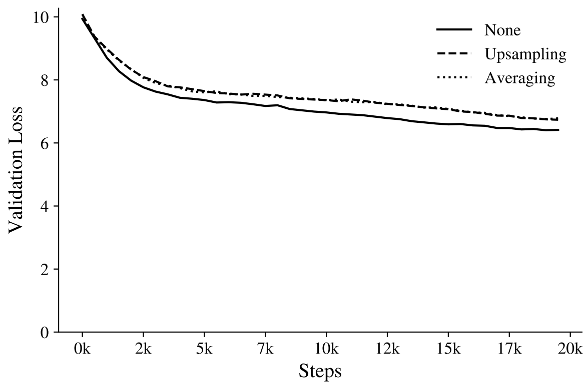

Figure 3: Validation loss curve for different sampling strategies in 20k steps.

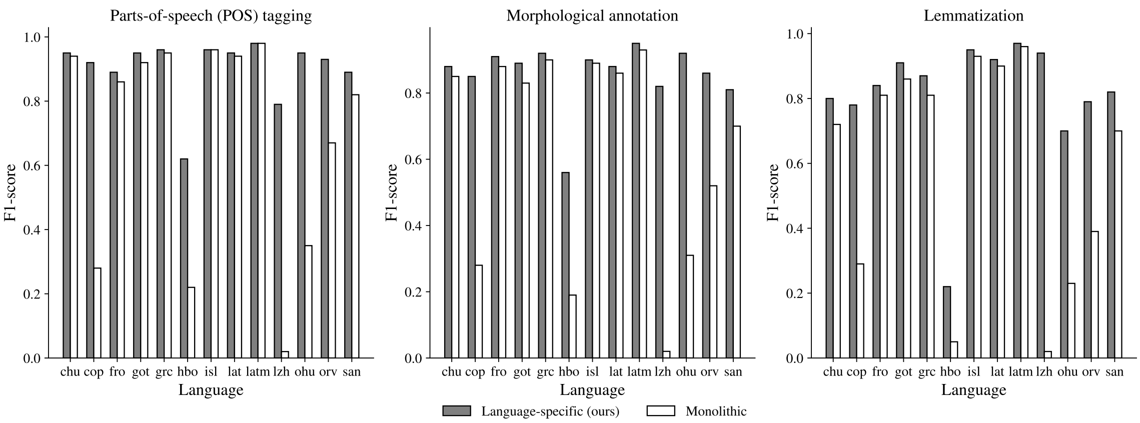

Figure 4: Comparison between training language-specific models versus a single monolithic model as evaluated on the validation set.

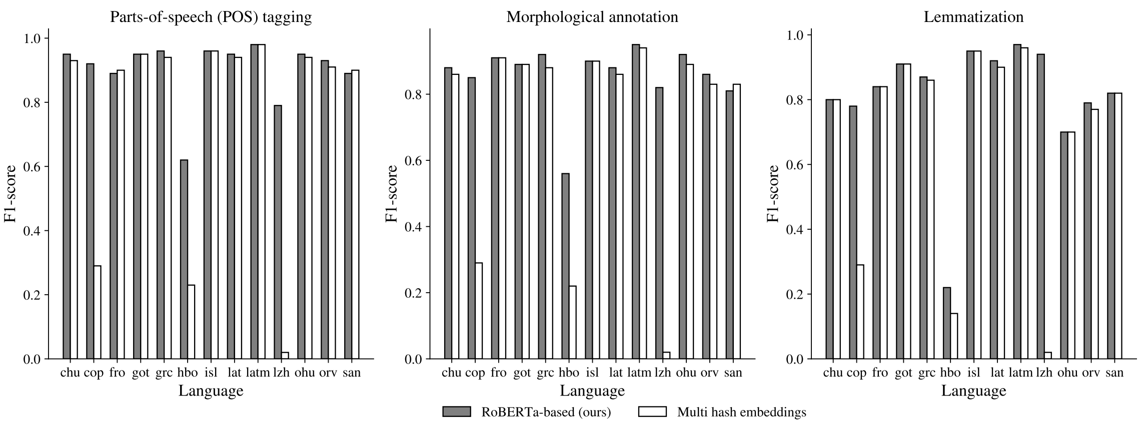

Figure 5: Comparison between a RoBERTa-based pretrained model and multi-hash embeddings (Miranda et al., 2022) as evaluated on the validation set.