Transfer Learning to the Rescue! Cross-Lingual Transfer for POS Tagging in an Endangered Language: Hamshetsnag

This study aims to develop a part-of-speech (POS) tagger for a low-resource and endangered language Hamshetsnag (Hms) spoken in Northeastern areas of Turkey, which is a dialect of Western Armenian. To accomplish this, we fine-tuned mBERT [1], a state-of-the-art multilingual encoder-only large language model, with cross-lingual joint and sequential transfer steps by using typologically similar languages. While supervised POS tagging is a solved problem for high-resource languages [2], it is indeed difficult for *truly* low-resource settings with accuracies below 50% [3]. To develop POS tagger for Hms, we use multilingual transfer learning methods [4] to transfer knowledge from typologically two similar languages Western Armenian (WArm) and Eastern (Standard) Armenian (Arm) that have more resources available. While some studies show that lexical overlap between languages can mitigate transfer in mBERT [5], others claim that little cross-sharing occurs [6].

Methodology. We created a Hms corpus of 907 sentences by scraping online dictionaries and collecting data from two native speaker consultants and grammar books, which we annotated the POS tags for each token (Figure 1). In addition, we formed 3 different hybrid datasets by using the WArm and Arm Universal Dependency (UD) treebank sentences [7], [8], with varying proportions of sentences (Table 1).[1] In the end, we had 6 different datasets to train the POS tagger (3 hybrid and 3 base).

Cross-Lingual Joint Transfer. We fine-tuned mBERT using Flair's [9] sequence tagger with our three hybrid datasets for 4 epochs on a T4 GPU with a learning rate of $10^{-}5$ and tested on Hms. We used cross-entropy as the loss function (equation 1), where $y_{i,c}$ is the cth element of the true label y_i and $p_{i,c}$ is the cth element of predicted probability distribution p_i. **Full Sequential Transfer.** We also fine-tuned mBERT with the entire UD dataset for each dialect (WArm and Arm) to adapt the model's parameters to the language family. After this step, we re-fine-tuned the model using our own annotated Hms sentences.

$$(1)\quad \text{CrossEntropyLoss}(y_i, p_i) = -\sum_{c=1}^{N} y_{i,c} \cdot \log(p_{i,c})$$

Results & Discussion. The models trained on higher-resource dialects received much higher POS tagging macro F1 scores than the default model trained on only Hms that received 27% accuracy (Table 2)[2]. When we used cross-lingual joint transfer with our hybrid datasets, we pushed Hms POS F1 to 70% when the proportion of WArm, which is typologically more similar to Hms, was greatest (Table 3). We also received similar results with sequential transfer when the model was fine-tuned on WArm following Arm, with a Hms POS F1 of 73%, a 46% increase over the default model (Table 4). Overall, these results demonstrate that it is possible to get acceptable sequence tagging scores for endangered languages like Hamshetsnag if typologically similar languages are utilized for multilingual transfer learning [10], [11].

[1] We transliterated WArm and Arm into Latin and normalized the scripts since Hms does not have a standardized script and we use Latin.

[2] We also trained Hidden Markov Models with the three baseline datasets and received very similar accuracies (48%, 85%, 88%, respectively).

Figure 1. Distribution of U-POS tags in the dataset.

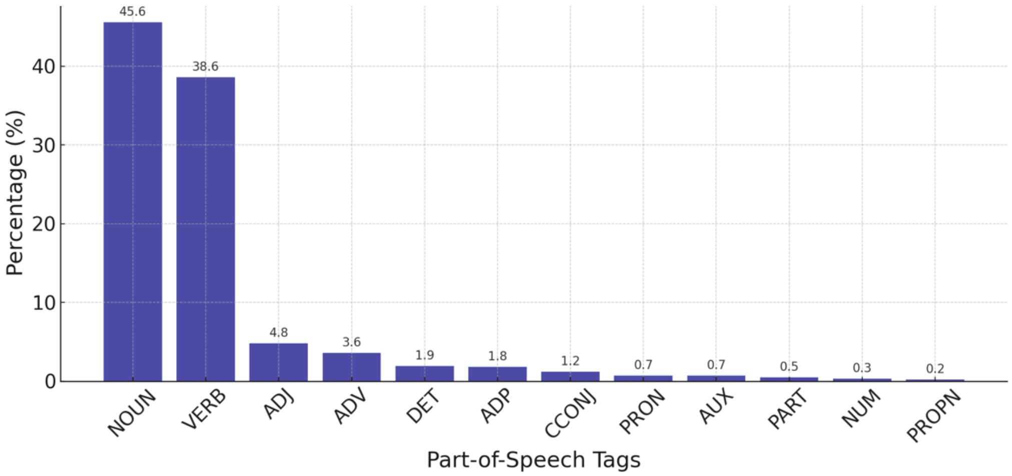

Table 1. Datasets used in this study.

Dataset	Train	Test	Dev	Tokens
WArm Treebank [8]	5282	680	694	121.583
Arm Treebank [7]	1974	277	249	52.220
Hms (our preliminary corpus)	727	90	90	2.205
Hybrid: Hms-WArm (15%-85%)	3285	330	330	44.503
Hybrid: Hms-WArm (50%-50%)	1394	140	140	11.421
Hybrid: Hms-Warm-Arm (30%-60%-10%)	1873	200	200	18.794

Table 2. Baseline model results.

Dataset	F1-Micro	F1-Macro
WArm Treebank[a]	0.95	0.87
Arm Treebank[b]	0.94	0.89
Hms	0.64	0.27

[a,b] We tested these models on WArm and Arm respectively to form a baseline.

Table 3. Cross-lingual joint/hybrid training results.

Dataset	F1-Micro	F1-Macro
Hms-WArm (15%-85%)	0.84	0.48
Hms-WArm (50%-50%)	0.82	0.51
Hms-WArm-Hms (30%-60%-10%)	**0.85**	**0.70**

Table 4. Sequential fine-tuning results.

Training Order	Micro-F1	Macro-F1
Arm-WArm-Hms	**0.81**	**0.73**
WArm-Arm-Hms	0.81	0.61

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." arXiv, May 24, 2019. Accessed: Mar. 14, 2024. [Online]. Available: http://arxiv.org/abs/1810.04805

[2] B. Bohnet, R. McDonald, G. Simoes, D. Andor, E. Pitler, and J. Maynez, "Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings." arXiv, May 21, 2018. Accessed: Mar. 15, 2024. [Online]. Available: http://arxiv.org/abs/1805.08237

[3] K. Kann, O. Lacroix, and A. Søgaard, "Weakly supervised pos taggers perform poorly on truly low-resource languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 8066–8073. Accessed: Mar. 15, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/6317

[4] J. Cho *et al.*, "Multilingual Sequence-to-Sequence Speech Recognition: Architecture, Transfer Learning, and Language Modeling," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, Dec. 2018, pp. 521–527. doi: 10.1109/SLT.2018.8639655.

[5] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is Multilingual BERT?" arXiv, Jun. 04, 2019. Accessed: Mar. 15, 2024. [Online]. Available: http://arxiv.org/abs/1906.01502

[6] J. Singh, B. McCann, R. Socher, and C. Xiong, "BERT is not an interlingua and the bias of tokenization," in *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, 2019, pp. 47–55. Accessed: Mar. 15, 2024. [Online]. Available: https://aclanthology.org/D19-6106/

[7] M. Yavrumyan, H. Khachatrian, A. Danielyan, and G. Arakelyan, "ArmTDP: Eastern Armenian treebank and dependency parser," in *XI International Conference on Armenian Linguistics, Abstracts. Yerevan*, 2017.

[8] M. M. Yavrumyan, "Universal dependencies for Armenian," in *International Conference on Digital Armenian, Inalco, Paris, October*, 2019, pp. 3–5.

[9] A. Akbik, T. Bergmann, D. Blythe, K. Rasul, S. Schweter, and R. Vollgraf, "FLAIR: An easy-to-use framework for state-of-the-art NLP," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 2019, pp. 54–59. Accessed: Mar. 15, 2024. [Online]. Available: https://aclanthology.org/N19-4010/

[10] Z. Wang, K. K, S. Mayhew, and D. Roth, "Extending Multilingual BERT to Low-Resource Languages." arXiv, Apr. 28, 2020. Accessed: Mar. 15, 2024. [Online]. Available: http://arxiv.org/abs/2004.13640

[11] D. Van Thin, H. Quoc Ngo, D. Ngoc Hao, and N. Luu-Thuy Nguyen, "Exploring zero-shot and joint training cross-lingual strategies for aspect-based sentiment analysis based on contextualized multilingual language models," *Journal of Information and Telecommunication*, vol. 7, no. 2, pp. 121–143, Apr. 2023, doi: 10.1080/24751839.2023.2173843.