Combining Automated and Manual Data for Effective Downstream Fine-Tuning of Transformers for Low-Resource Language Applications

Ulyana Isaeva[1], Danil Astafurov[1,2], Nikita Martynov[1,3],

[1]Sber, [2]ITMO University, [3]New Economic School,

Correspondence: ulyana.isaeva20@gmail.com

Abstract

This paper addresses the constraints of downstream applications of pre-trained language models (PLMs) for low-resource languages. These constraints are pre-train data deficiency preventing a low-resource language from being well represented in a PLM and inaccessibility of high-quality task-specific data annotation that limits task learning. We propose to use automatically labeled texts combined with manually annotated data in a two-stage task fine-tuning approach. The experiments revealed that utilizing such methodology combined with vocabulary adaptation may compensate for the absence of a targeted PLM or the deficiency of manually annotated data. The methodology is validated on the morphological tagging task for the Udmurt language. We publish our best model that achieved 93.25% token accuracy on HuggingFace Hub[1] along with the training code[2].

1 Introduction

The evolution of transformer-based pre-trained language models (PLMs) has enabled leveraging them as a basis to fine-tune for numerous downstream tasks, including morphological analysis (Baxi and Bhatt, 2024). The pipeline is complicated for low-resource languages (LRLs), which are rarely included in the pre-training data of PLMs, primarily due to the scarcity of data available (Imani-Googhari et al., 2023). When tackling a downstream task for a LRL without a PLM, one approach to address the data deficiency is to scale up the volume of high-quality task-specific data annotation.

Annotated texts in LRLs are contributed mainly by field linguists, who indicate the primary demand for such tools. However, these specialists do not necessarily own the technical skills required to utilize state-of-the-art deep learning-based approaches. Thus, rule-based algorithms have become the typical approach to developing morphological tools for LRLs. They face limitations for languages with morphological form ambiguity, where they predict multiple morphological descriptions for a single word. Proper disambiguation requires costly manual annotation by rare specialists. Given these constraints, ambiguous annotation is more accessible and scalable than manually disambiguated labels.

Addressing these considerations, we propose a two-stage fine-tuning methodology using automatically ambiguously annotated data combined with manually labeled data to achieve optimal performance in the morphological analysis task. Our experiments focus on the Udmurt language, which, while not entirely low-resource in terms of available data, was not included in the pre-training data of open-source multilingual PLMs until recently. The resulting morphological tagging tool performance is comparable to that of an alternative approach on the basis of a massively multilingual PLM. Thus, the proposed method is supposed to compensate for the absence of a PLM for a LRL. We also show that this approach can reach baseline performance with up to 3 times less manually annotated data.

To put our findings into practice, we open-source a morphological analyzer for Udmurt with an accuracy of 93.25% on all test tokens and 85.7% on tokens with ambiguous labels. Of all our experiments, the maximum performance was achieved using a recently introduced Glot500-m model (Imani-Googhari et al., 2023), which, among other 500+ languages, was pre-trained on texts in Udmurt.

[1]https://huggingface.co/ulyanaisaeva/bert-morph-tagger-udmurt

[2]https://github.com/ulyanaisaeva/bert-morph-tagger-udmurt

Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025), pages 86–90

2 Methodology

We model the morphological analysis task as a token classification problem, where each label is a concatenation of a part-of-speech (POS) tag and morphological features of the word.

The architecture consists of a transformer encoder and a dense projection layer, predicting label probabilities for input words. It outputs a tensor of shape $L \times K$ where L is sequence length (i.e., the number of words) and K is the number of unique labels. If a word is tokenized into multiple subtokens, we assign the label to the first one and mask out all the subsequent word subtokens during loss calculation.

Applying transformer-based pre-trained encoder models to downstream classification tasks has proven effective in numerous studies. For LRLs that commonly lack a specialized PLM, the PLMs of first choice are multilingual ones, like mBERT, which inherits the original BERT architecture (Devlin et al., 2019) and has been pre-trained on the top 100 languages with the largest Wikipedias. Ács et al. (2021) investigates the transferability of BERT-like models to unseen languages (i.e., languages the model has not been pre-trained on) via fine-tuning on limited training data. The authors observe that high-resource monolingual models, though effective in their specific language, show worse cross-language transferability than multilingual models in token classification tasks such as POS tagging and named entity recognition. Importantly, Ács et al. (2021) showed that monolingual models for genetically unrelated languages can transfer more efficiently than multilingual ones in cases where the languages share the same script, e.g., ruBERT for Russian performed better than multilingual BERT applied to Uralic languages with Cyrillic script (Erzya, Moksha, Komi Permyak).

2.1 Tokenizer adaptation

The observations related to script similarity are attributed to the impact of tokenization on model performance. The more a tokenizer is relevant to a given language, the less a word is split into pieces during tokenization. Since multilingual models' tokenizers are trained on languages with various scripts, their vocabularies tend to contain shorter subwords and thus have higher fertility, defined as the average number of word pieces per word.

Presumably, for token classification tasks like morphological tagging or named entity recognition, a more targeted tokenizer (i.e., with lower fertility) would be more optimal. This suggestion is tested by Wang et al. (2020), showing that adapting a model's tokenizer to an unseen language improves downstream zero-shot performance in NER tasks in that language. The methodology implies adding 30K new targeted items to the vocabulary while randomly initializing the corresponding model's embedding weights.

In this study, we utilize tokenizer vocabulary adaptation (VA) to improve morphological tagging accuracy. As an adaptation technique, we leverage the Vocabulary Initialization with Partial Inheritance approach (Samenko et al., 2021). It aims at preserving the model's knowledge from the pre-training stage instead of learning all embedding weights from scratch. Original model embedding weights are inherited for tokens in the new vocabulary, which are also found in the initial one; the other weights are randomly initialized.

To find the optimal vocabulary size, we fitted several WordPiece (following mBERT) tokenizers on ***Train-AML*** with sizes ranging from 1K to 128K (log step with base 2) and measured fertility on the ***Valid-AML***. At the size of 32K, the fertility plateaus around 1.18, and so does the ratio of tokens not split into subwords (85.93%); this vocabulary size is selected for future experiments.

2.2 Combining automated and manual annotation

Morphological form ambiguity (homonymy) is a phenomenon where the same word form may be attributed with different morphological description depending on the context, e.g., English 'records' is a plural noun in 'This song sets *records* for popularity' and a 3rd person singular present tense verb in 'He *records* and plays ten instruments'. The disambiguation of such labels requires word context understanding. Yet, classifying a token accurately only to a group of ambiguous labels is a task achievable even by simple context-unaware algorithms. In the example above, it would mean reducing the space of possible labels to 'NOUN,pl', 'VERB,3sg,prs' without selecting the single correct label. The two-step fine-tuning approach proposed in this work leverages this mechanism to improve morphological tagging accuracy, including for words with ambiguity.

The first step is to pre-fine-tune (PFT) the classifier using ambiguously annotated data (e.g.,

with a context-unaware analyzer) with multiple pseudo-correct labels for words with morphological homonymy. This stage's learning objective is to narrow the set of most probably predicted labels to a group of labels that correspond to ambiguous word forms. We hypothesize that such pre-fine-tuning would provide the model with an initial intuition about the homonymous nature of morphological labels.

Seemingly, this PFT could be modeled as a multi-label classification problem. In fact, by the nature of the task, only one of a word's homonymous forms is actually correct. This is why we model this pre-training stage as single-label multi-class classification with a softmax for class probabilities, though it requires additional changes to how we treat multiple pseudo-correct labels during loss calculation.

The basic loss function for multiclass classification is cross-entropy, defined as a sum of negative predicted log probabilities of positive labels.

$$CE = \sum_{\{i|K_i \in correct\}} -\log \hat{p}_i$$

The minimum of this loss function is achieved when these probabilities are equal and sum into 1, while the others are all equal to 0. Given the nature of morphological homonymy, it is suboptimal to teach the classifier to equalize probabilities in the set of pseudo-correct labels with only one being actually true. Taking this into account, we propose to calculate the PFT loss function as a negative logarithm of the sum of predicted probabilities for positive classes.

$$MLCE = -\log \sum_{\{i|K_i \in correct\}} \hat{p}_i$$

This function would still penalize models for predicting high probabilities for wrong labels, and vice versa, yet remain indifferent to how the probabilities for pseudo-correct labels are mutually distributed.

The second training step is task fine-tuning (FT), which requires reliable manually disambiguated annotation to finally learn to precisely select from a homonymic group of tags. The model is still offered to choose from the full set of all possible labels, yet it is supposed to rely on positive bias to ambiguous labels acquired during the PFT step. Similarly to the PFT, the FT is done using the softmax activation function at the last projection layer, which outputs label probabilities that sum into 1.

2.3 Data

The proposed approach is relevant in the case of presence of 2 types of task-specific data:

- AML: automatically labeled texts where a word may contain more than one label in case further label selection is constrained by morphological ambiguity and requires context analysis or manual disambiguation.
- MDL: manually labeled (or disambiguated after automatic annotation) texts with a single label per word.

See Appendix A for the data origin details.

2.4 Metrics

To evaluate the proposed morphological tagging pipelines, we use the following metrics:

- *Token accuracy* (TAcc) is the ratio of tokens with correctly predicted tags.
- *Token accuracy (homonymous)* (TAccH) is the same metric, but calculated only on tokens with morphological form ambiguity.

3 Experiments

Model selection. We focus on the morphological analysis for the Udmurt language. Until recently, and by the time this research was planned, there had not been a multilingual model pre-trained in the Udmurt language until Glot500-m (ImaniGooghari et al., 2023) was published. Since the absence of a targeted PLM is still the case for numerous LRLs, we chose the multilingual BERT (mBERT[3]) as the baseline model. Referring to previous findings on the transferability of monolingual models sharing the same script as the target language, we also experimented with the BERT model for the Russian language (ruBERT[4], (Zmitrovich et al., 2024)), since Udmurt uses Cyrillic script too.

To keep up with the updates in the area of multilingual models, we provide a comparison with Glot500-m[5] which is pre-trained in 500+ languages, including Udmurt.

Experimental setup. The three above-mentioned models are tested in 4 main setups:

1. FT: only fine-tuning on ***Train-MDL***
2. VA+FT: vocabulary adaptation on ***Train-MDL*** and FT

[3] https://huggingface.co/google-bert/bert-base-multilingual-cased

[4] https://huggingface.co/ai-forever/ruBert-large

[5] https://huggingface.co/cis-lmu/glot500-base

3. PFT+FT: pre-fine-tuning on ***Train-AML*** and FT
4. VA-PFT-FT: VA and PFT and FT

The Udmurt language, which is our focus in this work, is not an extremely low-resource language since there are available text corpora and NLP tools. Emulating the MDL-data scarcity setup common for low-resource languages, we fine-tune the best-performing setup for ruBERT and mBERT on a reduced subset from ***Train-MDL*** (100, 200, 500, 1000, 5000 sentences of the original 10000 sentences).

4 Results & Discussion

The results of the described experiments are provided in Table 1. Every row section compares baseline (i.e., with fine-tuning only) performance to that of models with vocabulary adaptation and/or pre-training on ambiguously annotated data.

Model	TAcc	TAccH
mBERT-FT (Devlin et al., 2019)	86.28	77.04
VA-FT	87.55	77.49
PFT-FT	87.02	78.66
VA-PFT-FT	91.38	81.54
ruBERT-FT (Zmitrovich et al., 2024)	86.35	77.02
VA-FT	87.89	77.65
PFT-FT	87.32	77.87
VA-PFT-FT	91.24	81.00
Glot500-FT (ImaniGooghari et al., 2023)	92.44	85.34
VA-FT	85.63	85.63
PFT-FT	**93.25**	**85.70**
VA-PFT-FT	91.17	81.52

Table 1: Models' performance on ***Test-MDL***. See subsection 2.4 for the evaluation details.

The baseline models achieved 86.3 and 86.4 token accuracy with mBERT and ruBERT, respectively. Applying the VA procedure before the FT brings an improvement of 1.3 and 1.5 pp while PFT on the model with the original tokenizer before the FT increases the performance at 0.7 and 1.0 pp, respectively, for mBERT and ruBERT. However, the improvement brought by the cumulative usage of both VA and PFT over the FT-only baseline performance is approximately 5 pp for both backbone models. Thus, these two procedures appear far more effective when applied jointly rather than separately.

Glot500-m baseline showed the best baseline performance across our experiments and was further improved when pre-fine-tuned on ambiguous annotated data. Yet adapting the vocabulary of Glot500-m both with and without PFT decreased the overall performance.

The pipelines with VA and PFT based on mBERT and ruBERT perform worse yet comparably to FT-only Glot500-m baseline. This is important evidence suggesting that the utilization of the proposed two-stage training pipeline may be seen as an effective compensatory approach in cases when there is no available model pre-trained on the target LRL.

To address the cases of extremely LRLs where manual annotation is scarse, we trained the baseline and the enhanced pipelines on reduced train data subsets, the results are provided in Figure 1.

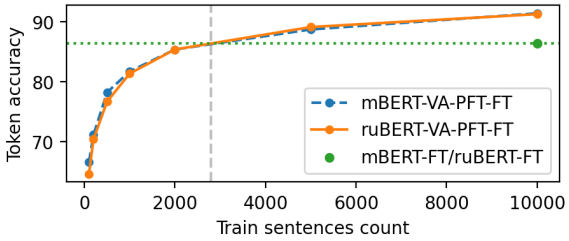

Figure 1: Model performance (token accuracy) on ***Test-MDL*** with reduced train data.

It can be observed that the proposed pipeline with VA and PFT may compensate for up to 3x less manually annotated data, i.e., utilizing the VA-PFT-FT pipeline with a 3 times reduced manually labeled train data can achieve performance comparable to that of the FT-only baseline on full-volume train data.

Despite the previous findings, the results of our experiments do not provide any evidence to choose ruBERT over mBERT since they share similar scores across all setups.

5 Conclusion

In this work, we present a two-stage fine-tuning procedure that leverages both automatically and manually annotated task-specific train data. The proposed approach combined with vocabulary adaptation increased morphological tagging accuracy by 5 pp in our experiments with the Udmurt language. We show that this improvement may compensate for train data deficiency and the absence of a specialized PLM, which are two major stumbling blocks in low-resource classification problems. As a practical outcome of the study, we open-source the best-performing morphological tagging model based on Glot500-m. We also publish the training code to facilitate the application of the methodology to other LRLs.

6 Limitations

While this study provides insights into choosing the backbone model and fine-tuning procedure for morphological analysis for low-resource languages, there are several limitations that should be considered when interpreting the results.

First, this methodology has so far been validated on only one language. We encourage future research on its applicability to different low-resource setups.

Second, in our experiments, ***AML*** and ***x-MDL*** datasets shared the same annotation scheme. Presumably, this will often be the case in the setups when the manual annotation is done over the automatic pre-labeling. Yet our experiments do not provide evidence to the contrary cases of mismatching annotation schemes.

7 Acknowledgments

The authors would like to thank Alexey Sorokin for sharing the initial idea that inspired this research.

References

Timofey Arkhangelskiy. 2019. Corpora of social media in minority Uralic languages. In *Proceedings of the Fifth International Workshop on Computational Linguistics for Uralic Languages*, pages 125–140, Tartu, Estonia. Association for Computational Linguistics.

Jatayu Baxi and Brijesh Bhatt. 2024. Recent advancements in computational morphology : A comprehensive survey. *arXiv preprint*. ArXiv:2406.05424 [cs].

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.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Yu. V. Normanskaja, O. D. Borisenko, I. B. Beloborodov, and A. I. Avetisyan. 2022. The Software System LingvoDoc and the Possibilities It Offers for Documentation and Analysis of Ob-Ugric Languages. *Doklady Mathematics*, 105(3):187–206.

Igor Samenko, Alexey Tikhonov, Borislav Kozlovskii, and Ivan P. Yamshchikov. 2021. Fine-Tuning Transformers: Vocabulary Transfer. *arXiv:2112.14569 [cs]*. ArXiv: 2112.14569.

Zihan Wang, Karthikeyan K, Stephen Mayhew, and Dan Roth. 2020. Extending Multilingual BERT to Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2649–2656, Online. Association for Computational Linguistics.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 507–524, Torino, Italia. ELRA and ICCL.

Judit Ács, Dániel Lévai, and András Kornai. 2021. Evaluating Transferability of BERT Models on Uralic Languages. *arXiv preprint*. ArXiv:2109.06327 [cs].

A Data origin

As a source of ambiguously annotated data, we utilize Tsakorpus (Arkhangelskiy, 2019) of standard written literary Udmurt language. This corpus is not public but is provided by the maintainer for research purposes. We annotate the texts using an open-source rule-based morphological analyzer[6] which does not conduct contextual disambiguation, i.e., it outputs all possible labels for words.

Filtering out the sentences with at least one word without a morphological label resulted in a dataset of approximately 558K words (64K sentences). Further in this paper, we refer to the corpus as the ***Train-AML***.

Manually annotated data was derived from LingvoDoc, a system for collaborative language documentation (Normanskaja et al., 2022), the data volume is 100K words (12K sentences). This data was processed with the same analyzer, and as a result, every word was attributed with both automatic labels (without disambiguation) and a manual one (which is always one of the ambiguous labels). We randomly partition this dataset into ***Train-MDL***, ***Valid-MDL*** and ***Test-MDL*** splits in a ratio 80-10-10, with the corresponding volumes of approx. 10K, 1.2K, and 1.2K sentences, respectively.

[6] https://github.com/timarkh/uniparser-grammar-udm/