Boosting the quality of morphological segmentation using transfer learning: a case study on Czech and Slovak

1 Abstract

Automatic morphological segmentation, which is the decomposition of words into morphemes, the smallest linguistic units carrying meaning, has not yet been fully resolved. While linguistically precise morphological dictionaries exist, they are available for only a limited number of languages and contain a restricted vocabulary for each (Slavíčková, 1975; Ološtiak et al., 2015). For these languages, Various algorithms have been developed that can process all available words computationally (Batsuren et al., 2022; Garipov et al., 2023), but their results are not always linguistically accurate. Additionally, there are languages for which no annotated data exists, or if it does, it is available only in very limited quantities. In such cases, transfer learning appears to be a suitable method for improving the accuracy of morphological segmentation (Liu et al., 2021).

Our research focuses on applying transfer learning, a machine learning approach where knowledge from one task is utilized to solve a similar task. We applied this method to the automatic morphological segmentation of words in Czech and Slovak, which are closely related languages. As the core algorithm for morphological segmentation, we used a convolutional neural network (CNN), specifically a ResNet architecture (He et al., 2016) with 15 layers. This architecture achieved the best results compared to other methods. The input for the neural network consisted of words converted into one-hot encoded sequences of characters without diacritics, with an additional binary feature indicating the presence of diacritics. The labels were binary sequences corresponding to the word length, where "0" indicated the absence of a segmentation boundary between adjacent characters, and "1" indicated the presence of a boundary. Linguistically annotated datasets are available, for both languages, allowing for the simulation of different scenarios regarding data availability in the target languages.

Several experiments were conducted. First, we examined how the size of the training data influenced the success rate of monolingual models for both Czech and Slovak. This established baseline values for comparison with transfer learning. As expected, the model's accuracy increased with the size of the training dataset (see Table 1). In all experiments, the performance was evaluated using morph F1 score and word accuracy.

Next, we investigated the impact of transfer learning on morphological segmentation, both from Czech to Slovak and vice versa. In each case, the model was pre-trained on all available data from the "parent" language and then fine-tuned on the dataset of the "child" language. As in the monolingual experiment, we compared how different sizes of the "child" dataset influenced model performance.

As expected, the accuracy of the cross-lingual model increased with the size of the "child" dataset. More interestingly, the impact of transfer learning was in most cases greater than merely doubling the dataset size for a monolingual model (see Figure 1 and 2). Another notable result came from the "zero-shot" experiment, where the model was trained only on the source language and applied to the target language without fine-tuning. Finally, we tested a bilingual model trained on all available Czech and Slovak data. This bilingual model performed slightly better than monolingual models trained on the full dataset for each language. These results are summarized in Table 1. All results were evaluated on test sets of size 4000 words for each language.

2 Tables and Graphs

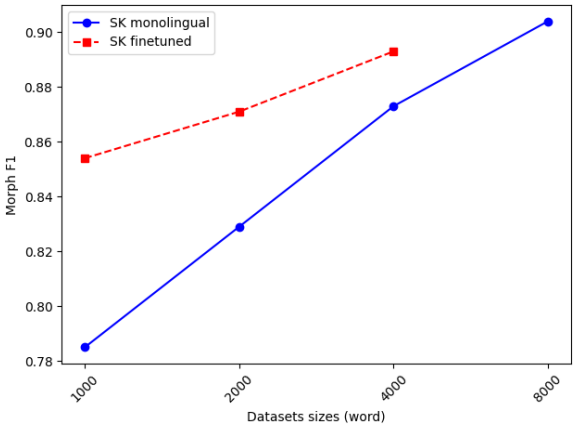

Figure 1: Different training dataset sizes for monolingual Slovak model versus fine-tuned model from the "parent" model trained on the full Czech dataset = 52459 words

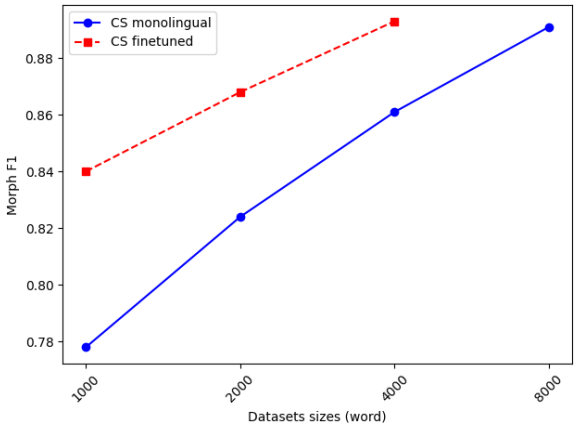

Figure 2: Different training dataset sizes for monolingual Czech model versus fine-tuned model from the "parent" model trained on the full Slovak dataset = 65847 words

	Czech		**Slovak**	
Train size	**Morpheme F1**	**Word accuracy**	**Morpheme F1**	**Word accuracy**
1000	0.778	0.573	0.785	0.538
2000	0.824	0.645	0.829	0.615
4000	0.861	0.708	0.873	0.704
8000	0.891	0.766	0.904	0.773
16000	0.920	0.827	0.933	0.842
32000	0.935	0.855	0.965	0.916
full train size	0.944	0.878	0.982	0.957
joint model	0.949	0.887	0.983	0.959
zero-shot	0.475	0.220	0.502	0.222

Table 1: Word accuracy and morpheme F1 score for increasing train sizes of monolingual models up to the full size of training data. The full data size for Czech is 52,459 words and it is 65,847 words for Slovak.

References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinović, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, et al. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. *arXiv preprint arXiv:2206.07615* .

Timur Garipov, Dmitry Morozov, and Anna Glazkova. 2023. Generalization ability of cnn-based morpheme segmentation. In *2023 Ivannikov Ispras Open Conference (ISPRAS)*. IEEE, pages 58–62.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 770–778.

Zoey Liu, Robert Jimerson, and Emily Prud'Hommeaux. 2021. Morphological segmentation for Seneca. In *First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Martin Ološtiak, Ján Genči, and Soňa Rešovská. 2015. *Retrográdny morfematický slovník slovenčiny*. Filozofická fakulta Prešovskej univerzity v Prešove.

Eleonora Slavíčková. 1975. *Retrográdní morfematický slovník češtiny*. Academia, Prague.