Transferring BERT Capabilities from High-Resource to Low-Resource Languages Using Vocabulary Matching

Piotr Rybak
Institute of Computer Science, Polish Academy of Sciences
ul. Jana Kazimierza 5, 01-248 Warsaw, Poland
piotr.rybak@ipipan.waw.pl

Abstract

Pre-trained language models have revolutionized the natural language understanding landscape, most notably BERT (Bidirectional Encoder Representations from Transformers). However, a significant challenge remains for low-resource languages, where limited data hinders the effective training of such models. This work presents a novel approach to bridge this gap by transferring BERT capabilities from high-resource to low-resource languages using vocabulary matching. We conduct experiments on the Silesian and Kashubian languages and demonstrate the effectiveness of our approach to improve the performance of BERT models even when the target language has minimal training data. Our results highlight the potential of the proposed technique to effectively train BERT models for low-resource languages, thus democratizing access to advanced language understanding models.

Keywords: bert, silesian, kashubian

1. Introduction

The field of natural language processing (NLP) has made remarkable progress in recent years, largely due to the rise of pre-trained language models. These models, most notably BERT (Bidirectional Encoder Representations from Transformers, Devlin et al., 2019), leverage unsupervised pre-training on massive text corpora to learn meaningful contextual representations of words. As a result of pre-training, such models have significantly reduced the data and computational requirements for further fine-tuning to a specific downstream task and have demonstrated exceptional capabilities in a wide range of NLP tasks, from text classification, named entity recognition, question answering to information retrieval.

While pre-trained language models have ushered in a new era of NLP, a significant challenge remains in extending their benefits to low-resource languages. The limited availability of text data and linguistic resources makes it difficult or impossible to pre-train models such as BERT using traditional methods. Moreover, these languages are often overlooked by the NLP community, leaving their speakers at a disadvantage in the era of AI-driven technologies.

In response to this challenge, our work proposes a simple approach of transferring BERT capabilities from high-resource and low-resource languages. The core idea is to use an external dictionary to adapt the BERT tokenizer of the high-resource model to properly initialize the BERT model for the low-resource language. This process makes further fine-tuning much more efficient and allows the BERT model to be trained with minimal training data.

We demonstrate the effectiveness of our approach by conducting experiments on two low-resource languages: Silesian and Kashubian. These languages, spoken by relatively small communities, face significant data scarcity and have been underrepresented in the NLP research community. Silesian is a West Slavic ethnolect spoken primarily in the region of Upper Silesia in Poland and is considered either one of the four major dialects of Polish or a separate regional language distinct from Polish. Kashubian is also a West Slavic language but is spoken primarily in the Pomeranian Voivodeship of Poland. Similar to Silesian, Kashubian is recognized either as a Polish dialect or as a separate language.

To summarize, our contributions are:

1. Release of the first BERT models for Silesian and Kashubian,[1]
2. A simple yet effective method to transfer BERT capabilities from high-resource to low-resource languages.

2. Related Work

2.1. Training BERT models for low-resource languages

The development of multilingual models, including BERT, has been an important step in addressing language diversity. However, training multilingual models does not necessarily ensure good performance for low-resource languages (Wu and Dredze, 2020; Lauscher et al., 2020) or robust alignment of word embeddings across languages (Cao et al., 2020).

[1] The models are available at: https://hf.co/ipipan/kashubian-herbert-base and https://hf.co/ipipan/silesian-herbert-base

LREC-COLING 2024, pages 16745–16750

A significant line of research has focused on training cross-lingual models, using both supervised and unsupervised methods. Supervised approaches use parallel data to train models, either by simply pre-training on parallel data (Lample and Conneau, 2019) or by aligning the output representation of the parallel sentences (Feng et al., 2022). Unsupervised approaches often mine word or sentence pairs to be used later in the training phase (Schuster et al., 2019; Hangya et al., 2022).

Techniques to improve performance specifically for low-resource languages include increasing the BERT vocabulary (Wang et al., 2020), reducing the number of parameters, and pre-training with multiple targets (Gessler and Zeldes, 2022). Other approaches focus on adapting the BERT model embeddings either using parallel data (Tran, 2020) or simply by matching tokens between tokenizers' vocabularies (Arkhipov et al., 2019). Others use dictionaries to weakly translate datasets used for pre-training (Wang et al., 2022b).

2.2. Silesian and Kashubian resources

Silesian and Kashubian, our focus languages, have very limited linguistic resources and training data. The only publicly available corpus is Wikipedia.

WikiANN (Rahimi et al., 2019), a multilingual named entity recognition dataset automatically generated from Wikipedia articles, contains 300 examples for both Silesian and Kashubian. The manual inspection shows its low quality, as most of the examples are just article titles.

Tatoeba[2] is a crowd-sourced collection of user-provided translations in a large number of languages, including 61 sentences for Silesian and 962 sentences for Kashubian. They are useful for evaluating machine translation models, but not language models.

3. Method

3.1. Data collection and cleaning

We use the `20230901` snapshot of Silesian and Kashubian Wikipedia as a corpus for pre-training the BERT model. To improve the data quality, we perform several steps to clean the raw text. First, we parse each article using WikiExtractor (Attardi, 2015), but contrary to the default settings, we keep lists as valid text. We remove short and repetitive articles about villages, cities, provinces, etc., as well as empty articles about a particular day or year. For Silesian, we also remove over 40,000 automatically generated articles about plants and convert articles written in *Steuer* script to *Ślabikorz* script using the Silling converter.[3] The final corpus consists of 485,736 words for Silesian and 300,497 words for Kashubian (see Table 1).

Language	# Docs	# Words	# Dict
Silesian	5,068	485,736	23,934
Kashubian	3,061	300,497	23,762

Table 1: Number of documents and words in the training corpus, and the size of the bilingual dictionary used for vocabulary matching.

3.2. Vocabulary matching

Pre-trained BERT models serve as a perfect initialization for training models for new languages, especially when the languages are closely related. While the weights for the self-attention layers can be copied directly from the source to the target model, the challenge arises for the embedding layer and the final masked language modeling (MLM) classification head, as both are tightly coupled with the tokenizer.

To overcome the difference in the tokenizers' vocabularies, we used a method similar to (Arkhipov et al., 2019), but extended it to the cross-lingual scenario. If a token from the target model vocabulary is present in the bilingual dictionary and its translation is present in the source model vocabulary, we directly copy its weights. For example, the token `tósz` (dog in Kashubian) is present in the Kashubian-Polish dictionary and its translation is `pies`. Since the token `pies` exists in the vocabulary of the source model, we copy the token embedding directly from the source model to the target model.

If the target token is not present in the bilingual dictionary, but is present in the source vocabulary, we also copy its weights directly.

If the target token is not present in either the bilingual dictionary or the source vocabulary, we split the token from the target vocabulary using the source tokenizer. As a result we get the sequence of tokens from the source vocabulary. We obtain the embedding by averaging the token embeddings from the source model.

We use existing bilingual dictionaries for vocabulary matching between the target language and Polish, which we use as the source language. For Silesian, we use the Silling dictionary[4] and for Kashubian, we use the Sloworz dictionary.[5] Both dictionaries contain about 23 thousand pairs (see Table 1).

3.3. Pre-training

We train the BERT base model using an MLM objective implemented in the HuggingFace library

[2] https://tatoeba.org/

[3] https://silling.org/konwerter-ze-steuera-na-slabikorz/

[4] https://silling.org/slownik/

[5] https://sloworz.org/

(Wolf et al., 2020). We either train the model from scratch or fine-tune the HerBERT base model (Mroczkowski et al., 2021) for 150 epochs, with a batch size of 720 examples[6] and a learning rate of $5 \cdot 10^{-4}$. We use a vocabulary of 8,000 tokens for both languages. We leave the rest of the hyperparameters at their default values.

4. Evaluation

4.1. Masked words prediction

We evaluate the different variants of pre-training BERT models on the task of predicting masked words. To obtain comparable results between models with different tokenizer vocabulary sizes, we mask and predict the whole word rather than individual tokens. When a word consists of multiple tokens, we predict them causally token by token. We report the accuracy for all words within the validation set of 100 randomly selected Wikipedia articles.

4.1.1. Results

We use the multilingual models mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020) as baselines. XLM-R performs the worst, with 16.15% for Silesian and 14.95% for Kashubian. Multilingual BERT performs much better with 32.58% and 31.10% respectively.

Since both Silesian and Kashubian share many similarities with Polish, we use HerBERT, a BERT model trained for the Polish language, as another baseline. Without any fine-tuning on the Silesian/Kashubian corpus, it achieves a low accuracy of 33.03% for Silesian and 34.69% for Kashubian (see Table 2). After fine-tuning, the accuracy increases to 59.27% and 55.88% respectively.

Next, we train the randomly initialized BERT model with a Silesian or Kashubian tokenizer (and not Polish like in HerBERT). The accuracies of 44.83% for Silesian and 40.23% for Kashubian are better than the zero-shot HerBERT baseline, but much lower than the fine-tuned HerBERT.

When the model is initialized with the HerBERT weights, the accuracy increases from 44.83% to 48.71% for Silesian and from 40.23% to 46.08% for Kashubian. It is still lower than the fine-tuned HerBERT, which illustrates the importance of matching the tokenizer vocabulary between the source and target models and allowing to leverage also the embedding layer of the source pre-trained model.

After matching the vocabularies with bilingual dictionaries (see section 3.2), the performance of the models increases and surpasses the fine-tuned HerBERT. For Silesian, the accuracy is 1 p.p. higher than with fine-tuning HerBERT (60.27% vs 59.27%), and the increase for Kashubian is more than 3 p.p. (59.13% vs 55.88%).

[6]Except for training from scratch, when we use 240 examples. Otherwise the training degenerates.

#	Tokenizer	Model	Silesian	Kashubian
		Zero-shot		
1	mBERT	mBERT	32.58	31.10
2	XLM-R	XLM-R	16.15	14.95
3	HerBERT	HerBERT	33.03	34.69
4	Matched	HerBERT	25.39	28.41
		Fine-tuned		
5	HerBERT	HerBERT	59.27	55.88
6	Not matched	Random	44.83	40.23
7	Not matched	HerBERT	48.71	46.08
8	Matched	HerBERT	**60.27**	**59.13**

Table 2: The accuracy of predicting the masked words for the validation set. Models can use either the *HerBERT* tokenizer or the tokenizer trained on the analyzed language. The latter can be either *matched* using a bilingual dictionary and thus able to use the embedding layer of the HerBERT model, or *not matched* where the embedding layer is randomly initialized. The weights of the BERT model itself can be initialized either *randomly* or using the *HerBERT* model.

4.2. Passage retrieval

Additionally, we evaluate the Silesian model on the passage retrieval task. We create a small test set of 50 question-passage pairs from the *Did you know?* section of the Silesian Wikipedia (similar to Marcińczuk et al., 2013). We use the whole Silesian Wikipedia as a passage corpus (10,605 passages) and evaluate the models using both Accuracy@10 (i.e. there is at least one relevant passage within the top 10 retrieved passages) and NDCG@10 (i.e. the score of each relevant passage within the top 10 retrieved passages depends descendingly on its position, Järvelin and Kekäläinen, 2002).

4.2.1. Evaluated models

First, we evaluate two existing state-of-the-art models in a zero-shot setup, i.e. without any additional fine-tuning on Silesian data. The Multilingual E5 Base model is a powerful neural retriever trained on both weakly labeled and supervised datasets for over 100 languages (but not Silesian) (Wang et al., 2022a). Silver Retriever is the best Polish model for passage retrieval (Rybak and Ogrodniczuk, 2023). Next, we compare two of our models fine-tuned on the Silesian corpus. The standard HerBERT model and the model initialized with HerBERT weights using vocabulary matching (models in rows 5 and

8 of Table 2). Due to the lack of a Silesian training set for this task, we fine-tune the models on the two Polish datasets, PolQA (Rybak et al., 2022) and MAUPQA (Rybak, 2023; Rybak and Ogrodniczuk, 2023). We use a standard Dense Passage Retriever (DPR, Karpukhin et al., 2020) architecture implemented in the Tevatron library (Gao et al., 2022). We fine-tune the models for 1,250 steps, with a batch size of 1,024 and a learning rate of $2 \cdot 10^{-5}$. We leave the rest of the hyperparameters at their default values.

#	Model	Acc@10	NDCG@10
	Zero-shot		
1	Multilingual E5 Base	80.00	74.79
2	Silver Retriever	84.00	73.59
	Fine-tuned		
3	HerBERT	92.00	79.72
4	Matched HerBERT	**96.00**	**81.79**

Table 3: The performance of the passage retrieval task for Silesian. *HerBERT* refers to the HerBERT model fine-tuned to the Silesian corpus and *Matched HerBERT* refers to the model that uses vocabulary matching (see Section 3.2).

4.2.2. Results

The E5 models score 80% for accuracy and 74.79% for NDCG. Silver Retriever achieves comparable results, it scores higher for accuracy (84%), but lower for NDCG (73.59%).

Both fine-tuned models outperform the zero-shot retrievers, but the HerBERT model with matched vocabulary achieves much higher results, 96% vs 92% for accuracy and 81.79% vs 79.72% for NDCG. It proves the effectiveness of the vocabulary-matching technique.

5. Conclusion

In this paper, we propose a simple method to extend the capabilities of pre-trained language models, in particular BERT, to low-resource languages through vocabulary matching. Experiments conducted on Silesian and Kashubian languages demonstrate the effectiveness of the proposed method on masked word prediction and passage retrieval tasks.

The contributions of this work include the publication of the first BERT models for the Silesian and Kashubian languages, marking a significant advance in linguistic resources for these languages.

6. Acknowledgments

This work was supported by the European Regional Development Fund as a part of 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19. The work was financed as part of the investment: "CLARIN ERIC – European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024-2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

7. Bibliographical References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Giusepppe Attardi. 2015. Wikiextractor. `https://github.com/attardi/wikiextractor`.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations. In *International Conference on Learning Representations*.

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.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume*

1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. 2022. Tevatron: An efficient and flexible toolkit for dense retrieval.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the The 2nd Workshop on Multi-lingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Viktor Hangya, Hossain Shaikh Saadi, and Alexander Fraser. 2022. Improving low-resource languages in pre-trained multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11993–12006, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20:422–446.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Michał Marcińczuk, Adam Radziszewski, Maciej Piasecki, Dominik Piasecki, and Marcin Ptak. 2013. Evaluation of baseline information retrieval for Polish open-domain question answering system. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 428–435, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. Association for Computational Linguistics.

Piotr Rybak. 2023. MAUPQA: Massive automatically-created Polish question answering dataset. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 11–16, Dubrovnik, Croatia. Association for Computational Linguistics.

Piotr Rybak and Maciej Ogrodniczuk. 2023. Silver retriever: Advancing neural passage retrieval for polish question answering.

Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk. 2022. Polqa: Polish question answering dataset.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. 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 1599–1613, Minneapolis, Minnesota. Association for Computational Linguistics.

Ke Tran. 2020. From english to foreign languages: Transferring pre-trained language models.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022a. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022b. Expanding pretrained models to thousands more languages via lexicon-based adaptation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.