BERT-based models are currently used for solving nearly all Natural Language Processing (NLP) tasks and most often achieve state-of-the-art results. One of them is HerBERT, the state-of-the-art language model for Polish, which outperforms other Polish language models on KLEJ Benchmark (Rybak et al., 2020) and POS tagging (Wróblewska, 2020). The HerBERT model and the experiments were presented in Mroczkowski et al., (2021):

HerBERT: Efficiently Pretrained Transformer-based Language Model for Polish

Robert Mroczkowski[1] **Piotr Rybak**[1] **Alina Wróblewska**[2] **Ireneusz Gawlik**[1,3]

[1]ML Research at Allegro.pl
[2]Institute of Computer Science, Polish Academy of Sciences
[3]Department of Computer Science, AGH University of Science and Technology
{firstname.lastname}@allegro.pl, alina@ipipan.waw.pl

Abstract

BERT-based models are currently used for solving nearly all Natural Language Processing (NLP) tasks and most often achieve state-of-the-art results. Therefore, the NLP community conducts extensive research on understanding these models, but above all on designing effective and efficient training procedures. Several ablation studies investigating how to train BERT-like models have been carried out, but the vast majority of them concerned only the English language. A training procedure designed for English does not have to be universal and applicable to other especially typologically different languages. Therefore, this paper presents the first ablation study focused on Polish, which, unlike the isolating English language, is a fusional language. We design and thoroughly evaluate a pretraining procedure of transferring knowledge from multilingual to monolingual BERT-based models. In addition to multilingual model initialization, other factors that possibly influence pretraining are also explored, i.e. training objective, corpus size, BPE-Dropout, and pretraining length. Based on the proposed procedure, a Polish BERT-based language model – HerBERT – is trained. This model achieves state-of-the-art results on multiple downstream tasks.

1 Introduction

Recent advancements in self-supervised pretraining techniques drastically changed the way we design Natural Language Processing (NLP) systems. Even though, pretraining has been present in NLP for many years (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017), only recently we observed a shift from task-specific to general-purpose models. In particular, the BERT model (Devlin et al., 2019) proved to be a dominant architecture and obtained state-of-the-art results for a variety of NLP tasks.

While most of the research related to analyzing and improving BERT-based models focuses on English, there is an increasing body of work aimed at training and evaluation of models for other languages, including Polish. Thus far, a handful of models specific for Polish has been released, e.g. Polbert[1], first version of HerBERT (Rybak et al., 2020), and Polish RoBERTa (Dadas et al., 2020).

Aforementioned works lack ablation studies, making it difficult to attribute hyperparameters choices to models performance. In this work, we fill this gap by conducting an extensive set of experiments and developing an efficient BERT training procedure. As a result, we were able to train and release a new BERT-based model for Polish language understanding. Our model establishes a new state-of-the-art on the variety of downstream tasks including semantic relatedness, question answering, sentiment analysis and part-of-speech tagging.

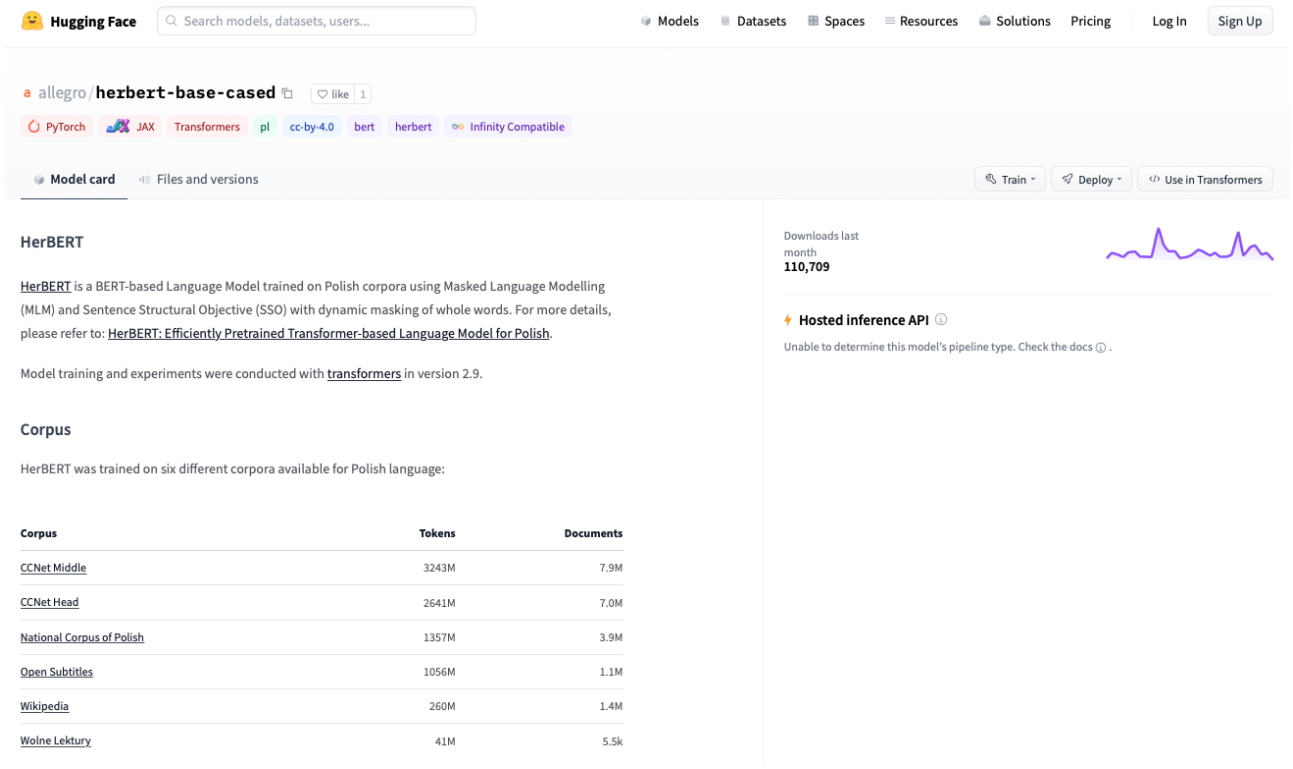
To summarize, our contributions are:

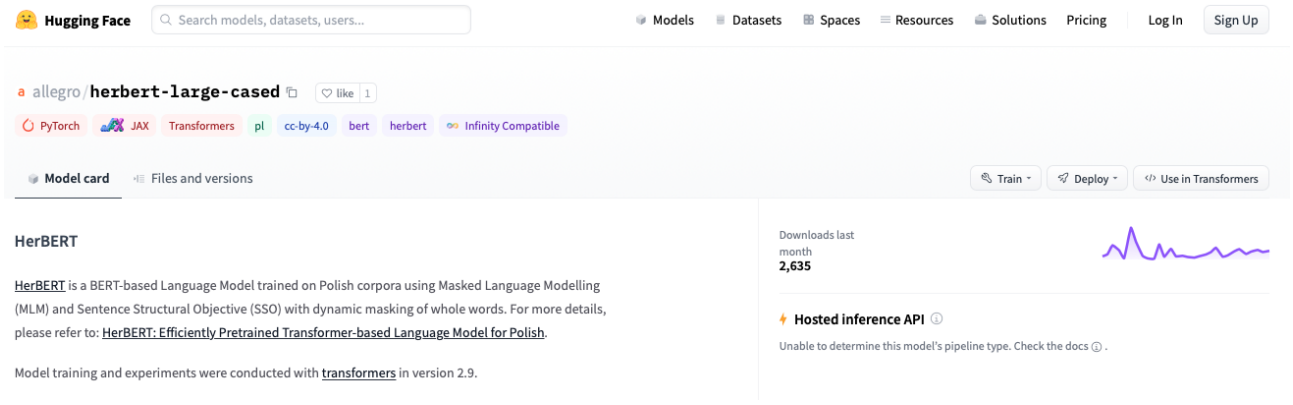
1. development and evaluation of an efficient pretraining procedure for transferring knowledge from multilingual to monolingual language models based on work by Arkhipov et al. (2019),
2. detailed analysis and an ablation study challenging the effectiveness of Sentence Structural Objective (SSO, Wang et al., 2020), and Byte Pair Encoding Dropout (BPE-Dropout, Provilkov et al., 2020),
3. release of HerBERT[2] – a BERT-based model for Polish language understanding, which achieves state-of-the-art results on KLEJ Benchmark (Rybak et al., 2020) and POS tagging task (Wróblewska, 2020).

[1]https://github.com/kldarek/polbert
[2]https://huggingface.co/allegro/herbert-large-cased

Proceedings of the 8th BSNLP Workshop on Balto-Slavic Natural Language Processing, pages 1–10
April 20, 2021.

HerBERT has two model versions: HerBERT$_{\text{SMALL}}$ and HerBERT$_{\text{LARGE}}$. HerBERT$_{\text{SMALL}}$ is trained on high quality corpora, i.a. NKJP, a well balanced collection of Polish texts. NKJP accounts for 74% of this training data set. HerBERT$_{\text{LARGE}}$ is trained on a five times larger corpus including primarily texts of a lower quality. HerBERT$_{\text{LARGE}}$ slightly outperforms HerBERT$_{\text{SMALL}}$ on various NLP tasks, as HerBERT$_{\text{SMALL}}$ has much less parameters. Due to the smaller number of parameters, HerBERT$_{\text{SMALL}}$ is much smaller in size and as a more practical model, it is used more often, as evidenced by the number of model downloads from the HuggingFace repository: 110.7K vs. 2.6K:





The HerBERT model gained international recognition, especially due to cooperation between academia (ICS PAS) and industry (Allegro.pl), as indicated in the article about HerBERT published on European Language Resource Coordination web page:

The state-of-the-art NLU model for Polish: successful cooperation between academia and industry

Too often, research is hindered by a lack of cooperation between academia and industry. Different goals, priorities, the way of working, and last but not least financial conditions make it difficult to reconcile these two worlds and work together on a single project. However, when they overcome the obstacles synergy appears and great results are achieved.

This kind of collaboration recently happened in the Polish NLP community. The ML Research team at Allegro.pl (a popular e-commerce marketplace and the third largest company on the Warsaw Stock Exchange) has started work on developing a BERT-based model for Polish language understanding (NLU) as a part of their NLP infrastructure. The main issue that arose was the lack of a large, diverse, and high-quality corpus that could be used to train the model. Such criteria are met by the National Corpus of Polish (NKJP), which consists of texts from many different sources, such as classic literature, books, newspapers, journals, transcripts of conversations, and texts crawled from the Internet.

The R&D NKJP project was a joint initiative of four scientific institutions: Institute of Computer Science at the Polish Academy of Sciences (ICS PAS, coordinator), Institute of Polish Language at the Polish Academy of Sciences, Polish Scientific Publishers PWN, and the Department of Computational and Corpus Linguistics at the University of Łódź, and was financed by the Ministry of Science and Higher Education.

NKJP can be explored in a dedicated search engine. However, the collection of source texts is not publicly available due to copyright reasons and may only be used by these four members of the consortium. Thanks to the joint work of Allegro and ICS PAS legal teams, as well as obtaining consent from PWN, the owner of a large part of the texts, all formal obstacles in using the corpus were overcome.

The cooperation resulted in training and open-sourcing HerBERT, a BERT-based model for Polish language understanding. The conducted experiments confirmed its high performance on a set of eleven diverse linguistic tasks, as HerBERT turned out to be the best on eight of them. In particular, it is the best model for Polish NLU model according to the KLEJ Benchmark. The model and its empirical evaluation are presented in the article by Mroczkowski et al. (2021, to appear at BSNLP).

Both HerBERT Base and HerBERT Large are released under CC BY-SA 4.0 licence as a part of the transformers library. Since its appearance in the HuggingFace repository, the model has been very popular. HerBERT Base has been downloaded over 13,500 times in the last month.

2021-03-17