Intent classification for Polish language

Grzegorz Przybylski
grzegorz.przybylski@student.uj.edu.pl

Paweł Fornalik
pawel.fornalik@student.uj.edu.pl

Szymon Dziedzic
szymon.stanslaw.dziedzic@student.uj.edu.pl

Supervisor: Andrii Krutsylo
andrii.krutsylo@ipipan.waw.pl

Abstract

In this paper, we check, how different approaches and architectures for the task of intent classification in the Polish language yield different results. We compare the performance of multilingual models (XLM-R and mT5) with Polish-oriented models (Polbert and HerBERT) and a translation approach using M2M100 as a translator and RoBERTa as a classifier. We conclude that multilingual models perform better in the intent classification task in the Polish language when compared to Polish-specific models and the translation approach, even though the accuracy difference between them isn't significant.

1 Introduction

Natural Language Processing (NLP) has witnessed significant advancements in recent years, enabling machines to understand and process human language more effectively. One crucial NLP task is intent classification, which involves determining the underlying intent or purpose behind a given user query or statement. Intent classification has numerous practical applications, such as chatbots, virtual assistants, customer support systems, and information retrieval systems. While significant progress has been made in intent classification for English, there is a notable scarcity of resources and research in intent classification for the Polish language. This project aims to bridge this gap by developing an intent classification model specifically tailored for the Polish language.

The ability to accurately classify user intents in the Polish language is essential for building efficient and user-friendly NLP systems that can cater to Polish-speaking users. By understanding user intent, systems can provide relevant and appropriate responses, improving user satisfaction and overall system performance. Moreover, with the increasing demand for multilingual NLP applications, it is crucial to expand the scope of research beyond English and include languages such as Polish to ensure inclusivity and accessibility for a broader user base.

For this project, we will leverage the existing research and methodologies in intent classification and transfer learning while adapting them to the unique characteristics and linguistic nuances of the Polish language.

In particular, we will use well-researched models trained in the Polish language, like HerBERT [12], Polbert [9], and a classifier trained for this task. The third approach we will test is a combination of M2M100 [5] and RoBERTa [11] models with classifier head. M2M100 will translate Polish queries to English and RoBERTa, which is trained solely on English corpora, will process these translated sentences. For training purposes, we will use the Polish section of the MASSIVE [6] dataset published in 2022, which is a multilingual dataset for training intent classification models. The performance results achieved by our models expressed as intent accuracy will be compared against a common baseline. As our baseline, we will use mT5 [19] and XLM-R [2] multilingual models. Our experiments aim to compare the results obtained from fine-tuned models with the performance reported in the MASSIVE paper [6].

The main research questions we would like to answer in our paper are:

1. Can we train the classifier based on the Polish-specific BERT model with a classification head, so it performs better than known multilingual intent classification models on the task intent classification in the Polish language?

2. How does the translation approach (M2M100

and RoBERTa) perform in comparison with the models mentioned above?

3. How does the performance of the Polish intent classification model vary with different model architectures or training strategies?

2 Related work

In recent years several models were developed on the transformer architecture [17], achieving notable performance in several NLP tasks, for example, BERT [3] trained on slot filling or GPT [14].

While intent classification for English has been extensively studied, the research on intent classification for the Polish language is limited. However, some studies have explored related tasks in Polish NLP, such as sentiment analysis, named entity recognition, and part-of-speech tagging.

Regarding intent classification, a few notable works have focused on multilingual intent classification, which can be adapted to Polish. For example, [4] proposed Multilingual BERT (M-BERT), a transformer-based model pre-trained on multilingual text, which achieved state-of-the-art performance on various NLP tasks, including intent classification.

Another relevant work [8] introduced a Polish corpus called PolEmo 2.0, which includes sentiment labels for Polish sentences. Although their work is primarily centered around sentiment analysis, the dataset can be leveraged for intent classification, as understanding sentiment is often intertwined with determining the underlying intent.

However, for intent classification in languages other than English, there was a deficit of large multilingual datasets. Of those that were, notable examples are SLURP [1], NLU Evaluation Data [18], Airline Travel Information System (ATIS) [7], Multilingual Task-Oriented Semantic Parsing (MTOP) [10] or Cross-lingual Multilingual Task-Oriented Dialog [16].

In 2022 Amazon presented the MASSIVE dataset [6] — Multilingual Amazon SLURP (spoken language understanding resource package) for Slot-filling, Intent classification, and Virtual assistant Evaluation. MASSIVE contains 1M realistic, parallel, labeled virtual assistant utterances spanning 51 languages. They also presented modeling results on XLM-R [2] and mT5 [19], including exact match accuracy, intent classification accuracy, and slot-filling F1 score. Our Polish-specific models will be compared with these results. Amazon has released this dataset, modeling code, and models publicly.

3 Models

For our experiments, we chose to utilize 3 different architectures; HerBERT [12] with classification head, Polbert [9] with classification head and M2M100 with RoBERTa and classification head.

3.1 HerBERT with classification head

HerBERT is a language model developed by Allegro [12] for their newly developed KLEJ benchmark [15]. As such we chose it in hopes for effective extractions of important features useful in intent classification. Its outputs consist of a 768-dimensional pooled output vector. Additionally, we used its intermediate states aggregated with the average function. The concatenation of these two makes an input for the classification head, consisting of three linear hidden layers, with 1536 (2*768), 768, and 384 neurons and the output softmax layer of 60 neurons (same as the number of intent classes in MASSIVE [6]) in the version which performed the best in our tests. We experimented with both fine-tuning and freezing the weights of HerBERT, but in the best version, we perform only transfer learning and freeze the HerBERT model weights leaving only the classification head for training.

3.2 Polbert with classification head

Polbert [9] is a language model based on BERT, created by Dariusz Kłeczek for the tasks presented in the PolEval 2020 [13]. As this required extraction of key information from the text we chose to adopt Polbert for use in intent classification with the expectation of good results. To do this we used the same classifier architecture as for HerBERT [12], i.e. taking its 768-wide pooling output and mean of hidden states as input, three linear layers with output size of 60 (same as intent classes in MASSIVE [6]) which we wrap in softmax. Similarly, as in HerBERT, in the best version, we freeze the weights of the transformer and train only the classification head.

3.3 M2M100 and Roberta with classification head

As an additional model, we chose a different paradigm, than with the HerBERT [12] and Polbert [9] with classification head. RoBERTa [11] is a model based on BERT, but with better-tuned hyperparameters. Because RoBERTa is trained solely on

English sources, we used the M2M100 [5] model to translate Polish prompts to English and input them into Roberta. Then we used the same type of classifier head architecture (3 hidden layers and 60-neuron softmax output layer) as was mentioned in the HerBERT and Polbert sections.

4 Experiments

We performed multiple experiments to find which classification head architecture and hyperparameters lead to the best results achieved by the considered models. The dataset we used to test our models was a Polish subset of the MASSIVE [6] dataset. The task on which our models were tested (as a benchmark, so we can compare their scores to our baselines) was single-label Polish language intent classification. The loss function we used for this problem was cross-entropy. The parameters under our investigation involved variations in the learning rate, batch size, number of hidden layers, neurons in each layer, and optimization algorithms, amongst others. Data from each trial was meticulously logged, tracking key performance metrics using a monitoring system of our choice (Neptune).

After running multiple tests we came to the conclusion that:

- batch size = 32
- dropout = 0.2
- learning rate = 0.0003
- AdamW optimizer
- 3 linear layers in classification head

yields the best results for all our tested architectures. This is the set of parameters that were used to achieve the best results presented in the next section. We also considered both the transfer-learning and fine-tuning approaches to this problem, but after some testing, we noticed that the version in which we freeze base model weights and train only the classifier performs better.

To save time and resources during experiments with our third approach, we first translated the Polish subset of MASSIVE using M2M100 [5] to English. Then we used these translations as input for separate runs with RoBERTa [11] and classification head.

The code which we used to run all of these experiments and implement our models is publicly available `https://github.com/szdziedzic/intent-classification-for-polish-language`.

5 Results

On its best run using the parameters specified above HerBERT [12] model with classification head was able to achieve the maximum validation set accuracy of 0.8435. Shortly after the training started around the 100th epoch both accuracy and loss stabilized and have not significantly improved ever since, even though we ran 1000 epochs total.

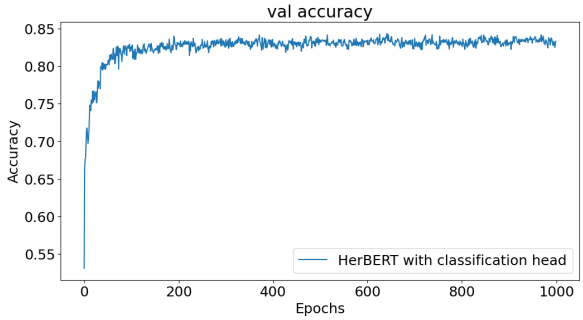

Figure 1: HerBERT with classification head validation set accuracy

When it comes to architecture utilizing the Polbert [9] model it was able to achieve a maximum of 0.8539 accuracy on the validation set. The loss function and accuracy stabilized around the 100th epoch as well in this case. We can observe that this architecture was learning slightly faster than the one utilizing HerBERT [12] and its accuracy curve is steeper.

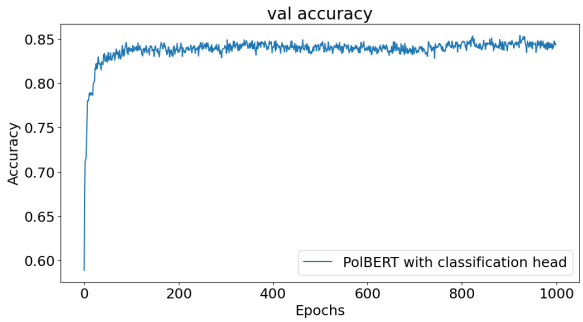

Figure 2: PolBERT with classification head validation set accuracy

M2M100 [5] and RoBERTa [11] with classification head performed the worst out of these three tested architectures hitting a maximum of around 0.8081 validation accuracy during the 788th epoch.

The comparison between validation accuracy curves visually indicates that the Polbert [9] with classification head performed the best of the three considered models.

In the table below we can see the results of all

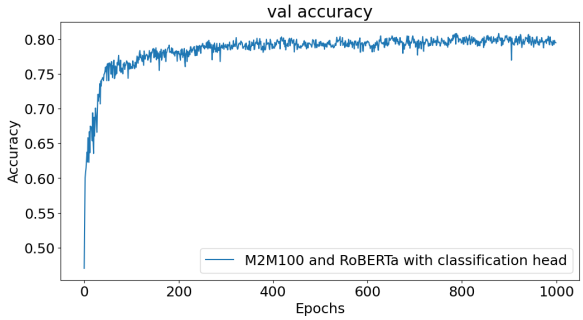

Figure 3: M2M100 and RoBERTa with classification head validation set accuracy

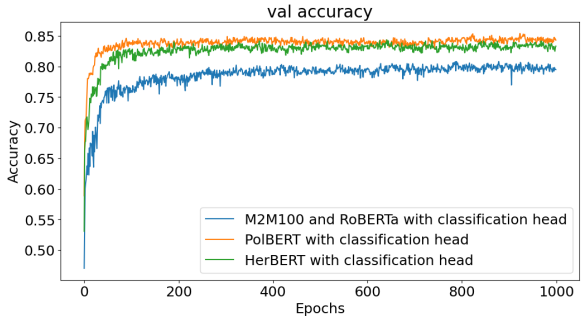

Figure 4: Validation set accuracy comparison between the models

of our considered models in comparison to our baselines (MT5 [19], XLM-R [2]).

Model	Accuracy
MT5	0.871
XLM-R	0.858
PolBERT with classification head	0.854
HerBERT with classification head	0.844
M2M100 and RoBERTa with classification head	0.808

6 Conclusions

In this paper we trained and evaluated three models based on transformer architecture for intent classification in Polish, using the MASSIVE [6] dataset: HerBERT [12] with classification head, Polbert [9] with classification head, and the translation approach using M2M100 [5] together with RoBERTa [11] with classification head.

In the experiments we ran, we have clearly shown the ability of these models to convey intent from natural language prompts in Polish, notably our best performing model; Polbert with classification head achieved over 85% accuracy on this task. We expected this transferability, as transformers have been shown before to contain general information about the text, useful in a variety of tasks.

However, our results were generally worse than those achieved by comparable multilingual models assessed in the MASSIVE paper: MT5 [19] and XLM-R [2]. That points to the high viability of these multilingual models in the intent classification task and is supported by the theory that deep language patterns generalize and reinforce across different training languages, improving overall performance.

References

[1] E. Bastianelli, A. Vanzo, P. Swietojanski, and V. Rieser. SLURP: A Spoken Language Understanding Resource Package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[2] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Multilingual bert. https://github.com/google-research/bert/blob/master/multilingual.md, 2018.

[5] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin. Beyond english-centric multilingual machine translation, 2020.

[6] J. FitzGerald, C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, and P. Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages, 2022.

[7] C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. The atis spoken language systems pilot corpus. In *Human Language Technology - The Baltic Perspectiv*, 1990.

[8] J. Kocoń, P. Miłkowski, and M. Zaśko-Zielińska. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[9] D. Kłeczek. Polbert: Attacking polish nlp tasks with transformers. In M. Ogrodniczuk and Łukasz Kobyliński, editors, *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences, 2020.

[10] H. Li, A. Arora, S. Chen, A. Gupta, S. Gupta, and Y. Mehdad. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online, Apr. 2021. Association for Computational Linguistics.

[11] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[12] R. Mroczkowski, P. Rybak, A. Wróblewska, and I. Gawlik. 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, Apr. 2021. Association for Computational Linguistics.

[13] M. Ogrodniczuk and Łukasz Kobyliński, editors. *Proceedings of the PolEval 2020 Workshop*, Warsaw, Poland, 2020. Institute of Computer Science, Polish Academy of Sciences.

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.

[15] P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online, July 2020. Association for Computational Linguistics.

[16] S. Schuster, S. Gupta, R. Shah, and M. Lewis. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. 2017.

[18] P. S. Xingkun Liu, Arash Eshghi and V. Rieser. Benchmarking natural language understanding services for building conversational agents. In *Proceedings of the Tenth International Workshop on Spoken Dialogue Systems Technology (IWSDS)*, pages xxx–xxx, Ortigia, Siracusa (SR), Italy, April 2019. Springer.

[19] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.