Intent classification for the Polish language

Radosław Jurczak and **Marcin Malejky** and **Maria Szczerba**
Faculty of Mathematics, Informatics and Mechanics, University of Warsaw
{radoslaw.jurczak,m.malejky,m.szczerba7}@student.uw.edu.pl

Supervisor: **Jan Ludziejewski**

1 Introduction

Natural language understanding (NLU) is a branch of artificial intelligence that uses computer software to understand input in the form of sentences using text or speech. Its goal is to deal with machine reading comprehension. One of the earliest known attempts in computer-based natural language understanding took place in 1964, by Daniel Bobrow (Bobrow, 1964). Bobrow created a STUDENT program that solves algebra word problems.

Understanding the context and intent of the text is crucial for effective natural language understanding. Intent classification, a task in natural language processing, specifically focuses on this objective. By gaining a clear understanding of the intent, models can generate improved responses. This task is not as straightforward as it may seem but due to usefulness of intent classification in NLP there is high interest in developing better and better solutions.

NLU finds application in a wide range of tools and applications that involve the processing and interpretation of human language. It serves as a fundamental component in conversational AI systems such as voice recognition tools, chatbots or customer support chats. Additionally, NLU plays a vital role in language translation systems, where a deep understanding of the intended message is crucial. In each of these scenarios, the system's ability to comprehend and generate meaningful responses to user queries is of utmost importance.

In this paper, we are going to continue the work on intent classification for the Polish language using MASSIVE dataset (FitzGerald et al., 2022). MASSIVE dataset offers a challenge of modeling intents across 51 different languages. Each language consists of 19,521 datapoints spanning 18 domains with 60 intents to model. The authors of the testbed provided XLM and mT5 baselines that we are going to use as starting points for our experiments.

More specifically, we will train different intent classifier models on Polish section of the MASSIVE dataset, and then compare them (in terms of classification accuracy and F1 score) to each other, to the mT5 (Xue et al., 2021) and XLM (Conneau et al., 2020) models that were used as baselines in the MASSIVE paper, as well as to Bloom (authors), 2023), the largest open source multilingual model to date. We suggest that due to the benefits of language-specific pretraining, intent classifiers based on models designed specifically for the Polish language should perform better than classifiers that use generic multilingual models. Also, we expect the performance of the model trained on the Slavic language family to fall somewhere between the two, mostly because of linguistic similarity of the languages within a language family.

2 Related Work

There has been significant interest in using large language models for a range of NLU tasks, such as intent classification. Significant results were achieved, among others, using fine-tuned HuBERT (Wang et al., 2021) on SLURP dataset (Bastianelli et al., 2020) and by weak supervision approach to annotation of ORCAS dataset (Alexander et al., 2022).

Much of this works, however, focuses solely on the English language, while there is a growing need for solutions able to tackle multiple languages, not only outside of the Germanic language family, but also Indo-European family. This includes the increased focus on minority languages.

Recent advances in multilingual NLP resulted in multi-lingual models such as XLM-R (Conneau et al., 2020), mT5 (Xue et al., 2021), a multilingual variant of Text-to-Text Transfer Transformer (Raffel et al., 2019), M2M100 (Fan et al., 2020) or mBART (Liu et al., 2020). The paper introducing MASSIVE dataset provided evidence that these

models trained on massive multilingual datasets can exhibit relatively good performance across languages.

One of the latest papers working on the same problem is "Benchmarking Language-agnostic Intent Classification for Virtual Assistant Platforms" (Wang et al., 2022) from July 2022.

Recently, multiple neural language models for the Polish language have been presented: localized variants of the BERT architecture (Devlin et al., 2019) such as HerBERT (Mroczkowski et al., 2021), Polbert (Kłeczek, 2020) and TrelBERT (Szmyd et al., 2023), as well as the Polish RoBERTa and Longformer (Dadas et al., 2020). Moreover, a multilingual model designed specifically for the Slavic languages was proposed in (Arkhipov et al., 2019). These models show encouraging results on the comprehensive KLEJ benchmark for Polish NLP tasks (Rybak et al., 2020) and on BSNLP Multilingual NER dataset (Piskorski et al., 2017), respectively.

References

Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. Orcas-i: Queries annotated with intent using weak supervision.

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.

BigScience Workshop: (392 authors). 2023. Bloom: A 176b-parameter open-access multilingual language model.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package.

Daniel G. Bobrow. 1964. Natural language input for a computer problem solving system. Technical report, USA.

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.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Dariusz Kłeczek. 2020. Polbert: Attacking polish nlp tasks with transformers. In *Proceedings of the PolEval 2020 Workshop*. Institute of Computer Science, Polish Academy of Sciences.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.

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.

Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Josef Steinberger, and Roman Yangarber. 2017. The first cross-lingual challenge on recognition, normalization, and matching of named entities in Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 76–85, Valencia, Spain. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1191–1201, Online. Association for Computational Linguistics.

Wojciech Szmyd, Alicja Kotyla, Michał Zobniów, Piotr Falkiewicz, Jakub Bartczuk, and Artur Zygadło. 2023. TrelBERT: A pre-trained encoder for Polish Twitter. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 17–24, Dubrovnik, Croatia. Association for Computational Linguistics.

Gengyu Wang, Cheng Qian, Lin Pan, Haode Qi, Ladislav Kunc, and Saloni Potdar. 2022. Benchmarking language-agnostic intent classification for virtual assistant platforms. In *Proceedings of the Workshop on Multilingual Information Access (MIA)*, pages 69–76, Seattle, USA. Association for Computational Linguistics.

Yingzhi Wang, Abdelmoumene Boumadane, and Abdelwahab Heba. 2021. A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. 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. Association for Computational Linguistics.