Silver Retriever: Advancing Neural Passage Retrieval for Polish Question Answering

Piotr Rybak, Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences

ul. Jana Kazimierza 5, 01-248 Warsaw, Poland

`{firstname.lastname}@ipipan.waw.pl`

Abstract

Modern open-domain question answering systems often rely on accurate and efficient retrieval components to find passages containing the facts necessary to answer the question. Recently, neural retrievers have gained popularity over lexical alternatives due to their superior performance. However, most of the work concerns popular languages such as English or Chinese. For others, such as Polish, few models are available. In this work, we present Silver Retriever, a neural retriever for Polish trained on a diverse collection of manually or weakly labeled datasets. Silver Retriever achieves much better results than other Polish models and is competitive with larger multilingual models. Together with the model, we open-source five new passage retrieval datasets.

Keywords: Polish, Passage Retrieval, Question Answering, Semantic Search

1. Introduction

In modern open-domain question-answering systems, especially those based on Large Language Models (LLMs), it is crucial to effectively and accurately find relevant passages to correctly answer user questions and avoid hallucinations. Traditional lexical approaches such as BM25 (Robertson and Zaragoza, 2009) have long served as the basis for retrieval systems. However, these methods have limitations, such as being restricted to returning passages containing the exact query terms and ignoring word order.

Recently, the neural retrievers based on pre-trained language models and fine-tuned on large datasets of question-passage pairs, have significantly outperformed their lexical counterparts (Karpukhin et al., 2020). While progress has been focused on languages such as English and Chinese, the less popular languages such as Polish remain unexplored.

In response to this gap, we present the Silver Retriever,[1] a Polish neural retrieval model trained on a diverse collection of both manually and weakly labeled datasets. We evaluate the model on three retrieval tasks and demonstrate that the Silver Retriever shows significant performance improvements over existing Polish models, and competes favorably with larger multilingual models.

We also open-source five new passage retrieval datasets,[2] consisting of over 500,000 questions, which will enable further research and development in the field of Polish question answering.

2. Related Work

There are almost no models for Polish passage retrieval. The only available Polish neural retriever is HerBERT-QA (Rybak, 2023), published together with the MAUPQA dataset on which it was trained. Wojtasik et al. (2023) and Rybak et al. (2022) have trained neural retrievers on other Polish datasets, but haven't released them.

Dadas (2022) trained a family of sentence transformers (ST-DistilRoBERTa and ST-MPNet) on an automatically constructed dataset of paraphrase pairs from sentence-aligned bilingual text corpora. While these models were trained for sentence similarity, they can also be used for retrieval.

Apart from Polish resources, there are several multilingual models. LaBSE (Feng et al., 2022) is trained to produce similar representations for bilingual sentence pairs that are translations of each other. mContriever (Izacard et al., 2021) is a multilingual dense retriever trained with an unsupervised contrastive learning objective and fine-tuned on the MS MARCO dataset (Nguyen et al., 2016). The E5 model (Wang et al., 2022) is also trained with a contrastive objective, but on a weakly-labeled large-scale selection of text pairs and later fine-tuned on labeled datasets. SentenceTransformer (Reimers and Gurevych, 2019) is a library for training sentence transformers. While they focus mainly on English models, they also include several multilingual models, with MiniLM-L12-v2 being the most popular.[3]

---

[1]We publish the model at: `https://hf.co/ipipan/silver-retriever-base-v1`

[2]We publish the datasets at: `https://hf.co/datasets/ipipan/maupqa`

[3]`https://hf.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`

| Dataset | # Questions | | | # Passages | | |
	Positive	Negative	Total	Positive	Negative	Total
Existing datasets						
PolQA	4,591	4,961	5,000	27,131	34,904	62,035
MAUPQA	385,827	366,506	385,895	395,416	2,429,602	2,824,978
PoQuAD	46,161	51,967	56,588	46,187	299,865	346,052
PolevalPairs2021	1,347	1,913	1,977	2,088	17,608	19,690
New datasets						
1z10	18,593	21,353	22,835	22,014	139,471	160,850
GPT3.5-CC	29,591	29,484	29,591	29,720	251,959	281,679
GPT3.5-Wiki	29,674	24,437	29,674	29,748	115,564	145,312
MS MARCO	389,304	389,974	389,987	416,763	3,006,996	3,422,436
Multilingual-NLI	64,900	99,809	100,752	68,096	743,857	811,884
Total	969,988	823,490	1,022,299	1,037,163	7,039,826	8,074,916

Table 1: Basic statistics of the training datasets. *# Questions* refers to the number of unique questions with at least one positive/negative passage. *MAUPQA* is a collection of seven individual datasets, but we provide aggregate counts for brevity. *Total* represents the concatenation of all datasets.

3. Silver Retriever

3.1. Training Datasets

We use a collection of four existing and five newly created datasets for training. We summarize their basic statistics in Table 1 and give a short description of each of them below.

3.1.1. Existing datasets

PolQA (Rybak et al., 2022) is the first Polish dataset for open-domain question answering. It consists of 7,000 questions from the trivia domain and 87,525 manually matched Wikipedia passages. We use the training split of the dataset.

MAUPQA (Rybak, 2023) is a collection of seven datasets for Polish passage retrieval. Overall, it contains almost 400,000 question-passage pairs. The examples are weakly labeled, either automatically mined, generated, or machine-translated from English. The questions cover trivia, website FAQs, and various online texts.

PoQuAD (Tuora et al., 2023) is a Polish equivalent of the SQuAD (Rajpurkar et al., 2018). It consists of more than 70,000 question-passage pairs. Contrary to the PolQA dataset, the dataset was created by asking annotators to write a question about a given Wikipedia passage. We use the training split of the dataset.

PolEval 2021 Pairs During the 2021 PolEval Question Answering task (Ogrodniczuk and Przybyła, 2021), one of the task participants (Rybak, 2021) manually annotated over 10,000 question-passage pairs. We use it without any modification.

3.1.2. New datasets

1z10 *Jeden z dziesięciu* (Eng. 'One out of ten') is a popular Polish TV quiz show in which the contestants answer trivia questions about science, history, and the arts. We use the Whisper model (Radford et al., 2022) to transcribe 333 episodes of the show and the GPT-3.5[4] to extract question-answer pairs. To match questions with relevant passages, we first use SpaCy (Honnibal et al., 2020) to lemmatize questions and a corpus of Wikipedia passages distributed with the PolQA dataset. Then, the BM25 selects the top 100 passages which we re-rank using the mMiniLM-L6-v2 cross-encoder (Bonifacio et al., 2021).[5] Finally, we select the top 5 passages and verify them with GPT-3.5. In total, we get 22,835 questions matched with 22,014 relevant passages. We keep the irrelevant passages as negatives.

GPT-3.5-CC and GPT-3.5-Wiki Similar to the GenGPT3 dataset from MAUPQA, we prompt the generative language model to write a question about a given passage. However, we use GPT-3.5 instead of GPT-3 (Ouyang et al.). We sample passages either from Polish Wikipedia or from the Polish part of CCNet (Wenzek et al., 2020). We generate nearly 30,000 questions for each source.

Polish MS MARCO Similar to BEIR-PL (Wojtasik et al., 2023), we translate the training subset of the MS MARCO dataset (Nguyen et al., 2016) into Polish. We use the in-house Allegro[6] machine trans-

---

[4]https://platform.openai.com/docs/models/gpt-3-5
[5]https://hf.co/unicamp-dl/mMiniLM-L6-v2-mmarco-v2
[6]https://ml.allegro.tech

lation model. We only consider relevant question-passage pairs and ignore existing hard negatives. Instead, we mine for the hard negatives ourselves (see Section 3.2).

Multilingual NLI We convert the collection of machine-translated natural language inference (NLI) datasets (Laurer et al., 2022) into question-passage pairs. We convert the *premise* to a question by adding the prefix 'Czy' (Eng. 'Does') and use the *hypothesis* as a passage. Then, we consider *entailment* and *contradiction* labels as relevant (with answers 'Yes' and 'No' respectively) and *neutral* as negative.

3.2. Hard negatives

Hard negatives are negative (irrelevant) question-passage pairs that are difficult for the model to distinguish from the relevant pair, often due to the high lexical overlap between the question and the passage. Such pairs are crucial for training robust neural retrievers (Karpukhin et al., 2020).

We follow a similar strategy to Ren et al. (2021) for mining hard negatives. First, we lemmatize the questions and the corpus of passages and use the BM25 to select the top 10 candidate passages for each question. Then, we score them using the mMiniLM-L6-v2 cross-encoder and keep only the irrelevant passages.

We add hard negatives to all datasets except PolQA, which already contains hard negatives. As a source of passages, we use either the Wikipedia passages distributed with the PolQA dataset (for datasets that already use it) or the other passages of a given dataset (for all other datasets).

3.3. Denoising

To increase the quality of the training pairs, we apply several steps of filtering noisy pairs.[7] We remove questions/passages that are too short or too long, passages that are relevant to too many questions, and passages that are too similar to their questions. We also score pairs using the E5-Base bi-encoder and mT5-3B cross-encoder[8] to remove relevant pairs with scores lower than 10% and negative pairs with scores higher than 90%. Finally, we manually create a blacklist for both entire questions and individual words. In total, we discard 14% of the relevant question-passage pairs.

3.4. Training

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 HerBERT Base model (Mroczkowski et al.,

2021) for 15,000 steps, with batch size 1,024 and learning rate $2 \cdot 10^{-5}$. We leave the rest of the hyperparameters at their default values. We use all training datasets (see Section 3.1), but drop questions without relevant passages.

4. Evaluation

We evaluate the Silver Retriever on three passage retrieval datasets: PolQA, Allegro FAQ, and Legal Questions. The last two datasets were introduced in the 2022 PolEval Passage Retrieval task (Kobylinski et al., 2023). We compare the performance of our model with other state-of-the-art Polish and multilingual models.

For all tasks, we use 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).

4.1. PolQA

We use the test subset of the PolQA dataset, and unlike to the evaluation in the original work, we use all passages annotated as relevant (not just those found by human annotators). This gives a better estimate of the true performance but may overestimate the results for the BM25 baselines.[9] Similar, to the original evaluation, we ignore questions without relevant passages.

4.2. Allegro FAQ

Allegro FAQ[10] is a dataset of 900 frequently asked questions and 921 help articles about the large Polish e-commerce platform - Allegro.com. Each question-passage pair is manually checked and edited if necessary.

4.3. Legal Questions

Legal Questions[11] consists of 718 questions and 26,000 passages extracted from over 1,000 acts of law. We ignore the title of the passage because it degrades performance on all tested models.

4.4. Results

The BM25 retrievers, especially when using lemmas instead of word forms, achieve competitive performance on PolQA and outperform many neural models. However, the best model is either Silver

---

Model	PolQA		Allegro FAQ		Legal Questions		Average	
	Acc	NDCG	Acc	NDCG	Acc	NDCG	Acc	NDCG
Lexical models								
BM25	61.35	24.51	66.89	48.71	**96.38**	**82.21**	74.87	51.81
BM25 (lemma)	71.49	31.97	75.33	55.70	94.57	78.65	80.46	55.44
Multilingual models								
MiniLM-L12-v2	37.24	11.93	71.67	51.25	78.97	54.44	62.62	39.21
LaBSE	46.23	15.53	67.11	46.71	81.34	56.16	64.89	39.47
mContriever-Base	78.66	36.30	84.44	67.38	95.82	77.42	86.31	60.37
E5-Base	86.61	**46.08**	91.89	75.90	96.24	77.69	91.58	66.56
Polish models								
ST-DistilRoBERTa	48.43	16.73	84.89	64.39	88.02	63.76	73.78	48.29
ST-MPNet	56.80	21.55	86.00	65.44	87.19	62.99	76.66	49.99
HerBERT-QA	75.84	32.52	85.78	63.58	91.09	66.99	84.23	54.36
Silver Retriever (ours)	**87.24**	43.40	**94.56**	**79.66**	95.54	77.10	**92.45**	**66.72**

Table 2: Performance of passage retrieval for lexical and neural models. We use top 10 accuracy and NDCG@10 on the test splits of each dataset and bold the highest score for each metric.

Retriever (considering the accuracy of 87.24%) or E5-Base (NDCG of 46.08%).

For the Allegro FAQ, Silver Retriever scores much higher than E5-Base for both accuracy (94.56% vs 91.89%) and NDCG (79.66% vs 76.02%). The BM25 performs better than LaBSE, and MiniLM-L12-v2, but worse than the other models that achieve similar performance (around 85% accuracy and 65% NDCG).

For the Legal Questions, the BM25 achieves the best results (96.38% accuracy and 82.21% NDCG) and outperforms all neural models. This is not surprising since the passage length is often longer than 512 tokens and there is a high lexical overlap between questions and passages. Apart from the BM25 models, the best neural retriever is E5-Base, which is on par with BM25 in terms of accuracy (96.24% vs 96.38%) but achieves a much lower NDCG (77.79% vs 82.21%).

On average, the SilverRetriver performs best in both accuracy (92.45%) and NDCG (66.72%). The E5-Base model is a close second (91.63% and 66.63% respectively). The other models perform much worse, often worse than the BM25.

5. Ablation Study

We analyze the impact of training choices using only the PolQA validation set since two other datasets only contain a test set. Compared to the main training (see Section 3.3), we fine-tuned the HerBERT Base model for 5,000 steps, and a batch size of 256. The effect of adding hard negatives is positive but minimal (probably due to the short training). Denoising improves both accuracy and NDCG by 1.6 p.p. The effect of a larger batch size

(1024 vs. 256) is also positive for both metrics.

Variant	Acc@10	NDCG@10
base	79.09	35.51
+ hard negatives	79.20	36.64
+ denoising	80.89	38.28
+ batch size of 1024	**81.31**	**39.13**

Table 3: Impact of different design choices on the performance on the PolQA validation dataset. Subsequent models include all previous changes.

6. Conclusion

We present Silver Retriever, a neural passage retriever designed for Polish and trained on a variety of manually and weakly labeled datasets. The model's performance is evaluated on three retrieval tasks, on which it achieves strong results compared to several Polish and multilingual models, with the best scores on the Allegro FAQ and the best average score across all tasks.

In addition, the article contributes to the research community by open-sourcing five new passage retrieval datasets with a total of 500,000 questions.

In summary, the Silver Retriever and associated datasets represent a significant advance in the field of passage retrieval for Polish, and provide valuable resources for improving the accuracy and efficiency of open-domain question answering systems.

7. Acknowledgments

8. Bibliographical References

Luiz Henrique Bonifacio, Vitor Jeronymo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, , Roberto Lotufo, and Rodrigo Nogueira. 2021. mmarco: A multilingual version of ms marco passage ranking dataset.

Sławomir Dadas. 2022. Training effective neural sentence encoders from automatically mined paraphrases. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 371–378.

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.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python. 10.5281/zenodo.1212303.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

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.

Łukasz Kobylinski, Maciej Ogrodniczuk, Piotr Rybak, Piotr Przybyła, Piotr Pezik, Agnieszka Mikołajczyk, Wojciech Janowski, Michał Marcinczuk, and Aleksander Smywinski-Pohl. 2023. Poleval 2022/23 challenge tasks and results.

Moritz Laurer, Wouter van Atteveldt, Andreu Salleras Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *Preprint*. Publisher: Open Science Framework.

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.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset.

Maciej Ogrodniczuk and Łukasz Kobyliński, editors. 2021. *Proceedings of the PolEval 2021 Workshop*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Maciej Ogrodniczuk and Piotr Przybyła. 2021. PolEval 2021 Task 4: Question Answering Challenge. In (Ogrodniczuk and Kobyliński, 2021), pages 123–136.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the*

56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, QiaoQiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2021. RocketQAv2: A joint training method for dense passage retrieval and passage re-ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2825–2835, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stephen E. Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3:333–389.

Piotr Rybak. 2021. Retrieve and Refine System for Polish Question Answering. In (Ogrodniczuk and Kobyliński, 2021), pages 151–157.

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

Ryszard Tuora, Aleksandra Zwierzchowska, Natalia Zawadzka-Paluektau, Cezary Klamra, and Łukasz Kobyliński. 2023. Poquad - the polish question answering dataset - description and analysis. In *Proceedings of the 12th Knowledge Capture Conference 2023*, K-CAP '23, page 105–113, New York, NY, USA. Association for Computing Machinery.

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

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave.

2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Konrad Wojtasik, Vadim Shishkin, Kacper Wołowiec, Arkadiusz Janz, and Maciej Piasecki. 2023. Beir-pl: Zero shot information retrieval benchmark for the polish language.