MAUPQA: Massive Automatically-created Polish Question Answering Dataset

Piotr Rybak

Institute of Computer Science,
Polish Academy of Sciences
piotr.cezary.rybak@gmail.com

Abstract

Recently, open-domain question answering systems have begun to rely heavily on annotated datasets to train neural passage retrievers. However, manually annotating such datasets is both difficult and time-consuming, which limits their availability for less popular languages. In this work, we experiment with several methods for automatically collecting weakly labeled datasets and show how they affect the performance of the neural passage retrieval models. As a result of our work, we publish the MAUPQA dataset, consisting of nearly 400,000 question-passage pairs for Polish, as well as the HerBERT-QA neural retriever.

1 Introduction

Open-domain question answering (OpenQA) systems aim to provide answers to questions from a variety of topics, using a large collection of passages as a knowledge base. Recently, the development of such systems has been accelerated by the release of several large-scale question-passage datasets, such as MS MARCO (Nguyen et al., 2016), TriviaQA (Joshi et al., 2017), and Natural Questions (Kwiatkowski et al., 2019, NQ). These datasets enabled the training of neural passage retrieval models (e.g. Dense Passage Retrieval, Karpukhin et al., 2020), which can select passages from a knowledge base that are the most likely to contain the answer to the question.

However, the annotation of such datasets is a time-consuming and expensive process, which limits their availability for less popular languages (Rogers et al., 2022). Another limiting factor is the availability of real questions. Datasets like MS MARCO or Natural Question consist of real questions asked by search engine users. For less popular languages (like Polish), such a source of questions is not available. This leads to two alternatives: either to train a system on a small dataset (which might not be sufficient for the model to reach its full potential) or to create a dataset automatically. The first approach was recently described by Rybak et al. (2022) who published the PolQA dataset which consists of 7,000 trivia questions and 87,525 manually annotated passages.

In this work, we experiment with the latter approach and show how different methods for automatic data collection can impact the performance of the neural passage retrieval models. Our contributions can be summarized as follows:

1. We experiment with several methods for automatically collecting weakly-labeled question-passage pairs, and show their impact on the performance of the retrieval models.
2. We publish the MAUPQA dataset consisting of almost 400,000 question-passage pairs for Polish.[1]
3. We release the HerBERT-QA neural retriever, which achieves the best results on the PolQA dataset.[2]

2 Related Work

Weakly-labeled datasets Over the years, many techniques were developed for the automatic creation of weakly-labeled datasets. One general idea is to use a weak model to automatically label the unlabeled dataset (Lee, 2013). In the case of OpenQA, either simple lexical models like BM-25 (Robertson and Zaragoza, 2009) or more powerful neural models are used to retrieve relevant passages for given questions. To further improve the accuracy of retrieved examples the passages can be filtered out using cross-encoders (Ren et al., 2021) or answers (if available, Karpukhin et al., 2020).

However, the above method can only be used if the source of questions is available. If that is

[1] https://hf.co/datasets/ipipan/maupqa
[2] https://hf.co/ipipan/herbert-base-qa-v1

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 11–16

Dataset	Questions	Passages	Answers	Correct	Unambiguous	Relevant	Overall
PolQA	4,591	57,921	5,634	99%	99%	92%	90%
CzyWiesz-v2	29,078	29,078	-	100%	92%	73%	70%
GenGPT3	10,146	10,177	10,146	92%	44%	89%	33%
MKQA	4,036	4,036	4,036	73%	73%	21%	15%
MTNQ	135,781	142,008	-	60%	78%	80%	41%
MFAQ	172,768	178,937	-	81%	84%	55%	43%
Templates	15,993	15,993	14,520	88%	100%	89%	78%
WikiDef	18,093	18,093	18,093	95%	77%	88%	65%
All	385,895	398,322	46,795	76%	82%	69%	46%

Table 1: Basic statistics for all used datasets. *All* represents the concatenation of all MAUPQA datasets (i.e. without PolQA). *PolQA* refers to the training part of the PolQA dataset. *PolQA* dataset has more answers than questions since it might contain multiple answer variants for a single question (e.g. *7* and *seven*). Some datasets don't have any answers due to the way they were created.

not the case, then questions can be automatically created. Either using templates (Fabbri et al., 2020) or trained models (Lewis et al., 2021).

Another line of work takes advantage of existing datasets and translates them automatically to other languages (Lewis et al., 2020). The quality of the machine translation model directly impacts the quality of the created dataset (Bonifacio et al., 2021).

Polish OpenQA datasets Few datasets exist for Polish OpenQA. The first published dataset for passage retrieval was the *Czy wiesz?* dataset (Marcińczuk et al., 2013). It is a collection of 4,721 questions from the *Did you know?* section on Polish Wikipedia out of which only 250 questions were manually labeled with a relevant passage. Rybak et al. (2020) later annotated an additional 1,070 questions with relevant passages.

The PolQA dataset (Rybak et al., 2022) is a recently introduced dataset for Polish OpenQA. It consists of 7,000 trivia questions and 87,525 manually annotated passages (both positive and hard-negative). Even though the number of question-passage pairs is impressive for a less popular language like Polish, the number of unique questions is still rather limited.

3 MAUPQA Dataset

The MAUPQA dataset consists of seven parts. Four of them are created from scratch (Czy wiesz?, GenGPT3, Templates, WikiDef), and the other three are based on existing resources (MKQA, MTNQ, MFAQ).

3.1 Quality Assessment

To assess the quality of MAUPQA datasets, we sample and manually annotate 100 question-passage pairs for each dataset. Our manual verification consists of three aspects:

Correct We check if the question is a valid, grammatically correct question written in Polish.

Unambiguous We define that the question is ambiguous if it cannot be answered without providing additional information. For example, the question "Where is the headquarter of the company?" is ambiguous because it doesn't specify the name of the company and thus makes it impossible to answer the question.

Relevant The final aspect is the relevance of the passage to the question, i.e. whether the passage contains the answer to the question.

We also calculate the **overall** correctness of the example as the proportion of examples that satisfy all three of the above aspects. We show the results of the quality evaluation in the Table 1 together with the sizes of each dataset.

3.2 Datasets

Below, we describe each of the seven MAUPQA datasets:

CzyWiesz-v2 Similarly to the original *Czywiesz?* dataset, we first gather all questions from the *Did you know?* section on Polish Wikipedia together with a link to the relevant Wikipedia article. To select the relevant passage, we score all passages within this article using a multilingual cross-encoder (Bonifacio et al., 2021)[3] and choose the one with the highest score. We use a few simple heuristics to filter out questions regarding images (e.g. "Who is the famous general *in the photo*?"). Additionally, we remove questions from the KLEJ benchmark test set (Rybak et al., 2020).

The final dataset consists of 29,078 questions. They are grammatically correct, mostly unambiguous, and have a high rate of relevant passages (73%, see Table 1). Manual inspection shows that irrelevant passages are the result of the cross-encoder errors. In most cases, the relevant passage exists in the matching article but it was not selected.

GenGPT3 In the GenGPT3 dataset, we explore the application of the *text-davinci-003* model (Ouyang et al.) for generating question-answer pairs based on a given passage. To obtain passages, we use the Polish subset of CCNet (Wenzek et al., 2020). These passages turned out to be very diverse, covering domains such as news, legal, technical, etc. To guide the model in generating relevant questions, we use the prompt: *Napisz pytanie i odpowiedź do poniższego paragrafu. Pytanie musi mieć przynajmniej pięć słów. Odpowiedź może mieć najwyżej pięć słów* (Write a question and answer for the following passage. The question must be at least five words. The answer can be up to five words). In addition, we provide two examples within the prompt to help the model learn to generate appropriate question-answer pairs.

Through our experiments, we observe that the generated questions are grammatically correct in 92% of the cases and highly relevant (89% of the cases). However, we also find that the questions are often ambiguous, with 56% of them requiring a contextual understanding of the passage to answer.

MKQA The MKQA (Longpre et al., 2021) dataset consists of 10,000 questions sampled from the NQ dataset and manually translated into 25 languages (including Polish). We clean MKQA dataset by removing questions without answers, requiring long answers (*Why?* and *How?* questions), and ambiguous ones ("Who is the *current* president?"). We end up with 4,036 questions.

Since the original dataset doesn't include matching passages, we use the BM-25 algorithm (Robertson and Zaragoza, 2009) to select the top 100 candidate passages which we re-rank using a multilingual cross-encoder. In either case, we append the answer to the query to increase the performance of the passage retrieval. However, it still proved to be difficult to retrieve relevant passages and only 21% of them are correct.

MTNQ To create the machine-translated NQ dataset (MTNQ) we select all questions with relevant passages from the NQ dataset and split those passages into sentences. Then, we translate both questions and sentences into Polish using Allegro[4] machine translation model.

Even though the translation model is high quality (similar to Google Translate), the translations still contain many errors. Two main reasons are incorrectly translating named entities (e.g. movie titles) and very noisy input (NQ questions are Google search phrases). It is worth noting that MKQA, which is a manually translated subset of NQ, also has a high ratio of ungrammatical questions.

MFAQ The MFAQ dataset (De Bruyn et al., 2021) contains 234 million multilingual (4 million Polish) questions scraped from FAQ websites. However, many of them are artificially created, e.g. "What is the best hotel in *city*?" for hundreds of different *cities*. To clean the data, we cluster lexically similar questions and passages and remove clusters with over 5 questions. Additionally, some of the questions are not in Polish. We filter them using the fasttext language-id model (Joulin et al., 2017, 2016).

After filtering, the dataset contains 178,937 passages, i.e. less than 5% of the original dataset. This shows the risk of using questions extracted directly from crawled websites. The cleaned dataset has rather high quality, in terms of grammatical correctness, unambiguity, and relevance of passages. The MFAQ is much more diverse than other datasets (except for *GenGPT3*) and contains questions from a wide range of domains (customer support, lifestyle, technical, etc.).

Templates We take advantage of the Wikipedia structure to generate questions using predefined

[3] https://hf.co/unicamp-dl/mMiniLM-L6-v2-mmarco-v2

[4] https://ml.allegro.tech/

templates. For example, list pages group together similar entities (e.g. "Writers born in Poland") which allows generating questions like "Where was Zbigniew Herbert born?". We also use tables (e.g. "What is the capital of Poland?") and chronologies (e.g. "In which year World War 2 started?"). In total, we use 33 templates to generate questions. Since each question has a link to the relevant Wikipedia article, we use the same method as in the *CzyWiesz-v2* dataset to select the most relevant passage from the relevant article.

Overall, we created 15,993 questions from templates. They are high quality but the process of creating templates was surprisingly time-consuming and took a few hours per template.

WikiDef We use Wiktionary[5] to generate questions based on word definitions. Some definitions have links to Wikipedia articles which we use to create the question-passage pairs. For example, the definition of "Monday" is "the first day of the week". Based on it, we generate the question "What is the name of *the first day of the week*?". Then, we select the first passage from the linked Wikipedia article as the relevant passage. We remove short definitions (less than 5 words) containing names of 23 predefined "categories" (e.g. city) to avoid ambiguous questions (e.g. "What is the name of *a city in Poland*?").

We end up with 18,093 questions asking for word definitions. This is the least diverse dataset of all as all questions follow the same template. Even though we tried to filter unambiguous questions there are still 23% of them in the final dataset.

4 Evaluation

We use the Tevatron library (Gao et al., 2022) to train the neural retriever. For each dataset, we fine-tune the HerBERT Base model (Mroczkowski et al., 2021) for 2,000 steps, with batch size 128 and learning rate 10^{-5}. Otherwise, we use default parameters. We experimented with training models for 5,000 steps but it didn't increase the performance. We use a single hard-negative per question when training on PolQA dataset. For other datasets, we only use in-batch random negatives as they don't contain hard-negatives.

For evaluation, we use Accuracy@10 (i.e. is there at least one relevant passage within the top 10 retrieved passages) and NDCG@10 (i.e. score of each relevant passage within the top 10 retrieved passages depends descending on its position, Järvelin and Kekäläinen (2002)). Each model is evaluated on the PolQA development dataset. We use provided Polish Wikipedia dump as a knowledge base.

[5] https://www.wiktionary.org/

5 Results

The baseline retriever trained using manually annotated PolQA dataset achieves 60.8% accuracy@10 (see Table 2). Individually, none of the automatically created datasets has a comparable score.

As expected, the best model is *MTNQ* with an accuracy of 58.5%. It is the second largest dataset, similarly to PolQA it contains mostly trivia questions, and is based on manually labeled question-passage pairs. Comparably large *MFAQ* dataset obtains much lower performance (38.7%), probably due to domain mismatch as otherwise, its quality is higher than *MTNQ*.

The *MKQA*, which is a manually translated subset of *NQ* dataset achieves surprisingly good results (51.5%). It is unexpected considering that only 21% of its passages are actually relevant.

The second best result (54.2%) is achieved by the *GenGPT3* dataset. Despite the diverse nature of the questions from different domains, and the relatively modest size of the dataset, it exhibits a remarkable level of quality that allows it to serve as a reliable source for training a passage retriever.

The third best result (54.1%) is scored by *CzyWiesz-v2* dataset. The other two datasets created based on Wikipedia perform much worse, *Templates* obtains accuracy of 45.9% and *WikiDef* only 19.9%. It is also the lowest result of all datasets, probably due to its low diversity.

None of the datasets is perfect and each of them has its own disadvantages. However, the retriever trained on all of them results in better performance than the manually annotated dataset (61.2% vs 60.8%). If we further fine-tune the retriever pretrained on MAUPQA, we obtain the state-of-the-art result for Polish passage retrieval of 62.7%. We name this retriever HerBERT-QA and release it alongside the created datasets.

6 Conclusion

In this work, we present MAUPQA, the largest Polish QA dataset with almost 400k question-passage pairs. Even though the dataset is created automatically it achieves competitive results on the Polish

Dataset	Acc@10	NDCG@10
PolQA	60.8%	26.9%
CzyWiesz-v2	54.1%	22.0%
GenGPT3	54.2%	22.1%
MKQA	51.5%	21.6%
MTNQ	58.5%	24.1%
MFAQ	38.7%	14.0%
Templates	45.9%	16.9%
WikiDef	19.9%	7.7%
All	61.2%	25.2%
All → PolQA	**62.7%**	**27.4%**

Table 2: Passage retriever performance trained on different datasets. We use top-10 accuracy and NDCG@10 on the PolQA development set. *All* represents the concatenation of all MAUPQA datasets (i.e. without PolQA). *All → PolQA* is a model first trained on the MAUPQA dataset and then fine-tuned on the PolQA dataset.

passage retrieval task and after fine-tuning on the PolQA dataset sets a new state-of-the-art performance.

Each of the seven datasets which make up MAUPQA has different properties and results in the vastly different performance of passage retrievers. Thanks to recent advancements of machine translation models, we recommend translating existing English datasets as the best way to cheaply obtain competitive QA datasets. Otherwise, generating questions using GPT-3 model proves to work well and can be applied to multiple different domains (for which there might not be an English dataset). If a set of questions already exists for a given language, then using pseudo-labeling also results in a surprisingly good dataset. However, to get the best performance, it is useful to combine multiple different datasets.

We believe our work will benefit the Polish NLP community, both by publishing a MAUPQA dataset, as well as the state-of-the-art passage retrieval model. Our study also lays a path for other languages on how to construct similar datasets.

Limitations

The MAUPQA dataset focuses only on the Polish language and the drawn conclusions might not hold for other languages. For example, the format of sentences in the *Did you know?* section of Polish Wikipedia makes it very easy to transform them into questions. This is not the case for other languages. Some of them don't even have the *Did you know?* section.

Except for choosing the number of training steps (2,000 or 5,000), we didn't perform any additional hyper-parameter search and used the default Tevatron values. We also tested only one encoder architecture (HerBERT Base). The results for other setups might be different.

Except for GenGPT3 and MFAQ, all datasets (including the evaluation dataset) use Wikipedia as a knowledge base. This might negatively impact the perceived performance of the retrievers trained on GenGPT3 and MFAQ. We suspect that those retrievers might generalize better to other domains but there are no Polish QA datasets on which we could have tested it.

7 Acknowledgments

We thank the Allegro.com Machine Learning Research team for giving us access to their machine translating model.

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.

Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2021. MFAQ: a multilingual FAQ dataset. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 1–13, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alexander Fabbri, Patrick Ng, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2020. Template-based question generation from retrieved sentences for improved unsupervised question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4508–4513, Online. Association for Computational Linguistics.

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

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. 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.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Dong-Hyun Lee. 2013. Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. 2021. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2021. MKQA: A linguistically diverse benchmark for multilingual open domain question answering. *Transactions of the Association for Computational Linguistics*, 9:1389–1406.

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.

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.

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*.

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.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2022. QA dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*

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.

Piotr Rybak, Piotr Przybyła, and Maciej Ogrodniczuk. 2022. Improving question answering performance through manual annotation: Costs, benefits and strategies.

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 12th Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.