Proceedings of the 18th Conference on Computer Science and Intelligence Systems pp. 297–302
DOI: 10.15439/2023F1627
ISSN 2300-5963 ACSIS, Vol. 35

Reranking for a Polish Medical Search Engine

Jakub Pokrywka, Krzysztof Jassem, Piotr Wierzchoń
Adam Mickiewicz University
Faculty of Mathematics and Computer Science,
Email: {firstname.lastname}@amu.edu.pl

Piotr Badylak, Grzegorz Kurzyp
WN PWN,
Email: {firstname.lastname}@pwn.pl

***Abstract*—Healthcare professionals are often overworked, which may impair their efficacy. Text search engines may facilitate their work. However, before making health decisions, it is important for a medical professional to consult verified sources rather than unknown web pages. In this work, we present our approach for creating a text search engine based on verified resources in the Polish language, dedicated to medical workers. This consists of collecting and comprehensively analyzing texts annotated by medical professionals and evaluating various neural reranking models. During the annotation process, we differentiate between an abstract information need and a search query. Our study shows that even within a group of trained medical specialists there is extensive disagreement on the relevance of a document to the information need. We prove that available multilingual rerankers trained in the zero-shot setup are effective for the Polish language in searches initiated by both natural language expressions and keyword search queries.**

I. Introduction

WHEN seeking content in a domain-specific text, a medical professional is faced with the dilemma of whether to consult a work published by a verified source or to query the Internet. Often, verified documents are published only in print, and so browsing them is time-consuming. On the other hand, a lot of Internet content is created by non-professionals and is not error-free, thus finding accurate data is difficult. This is especially true in the case of non-English online resources. However, querying Google or Wikipedia is tempting when one has to act under time constraints, for example, during a medical appointment. Considering the workload of healthcare workers [1], this statement holds even more significance. To address this issue, medical publishers attempt to provide online access to their domain-specific resources.

This paper describes the results of a project aimed at creating an intuitive search engine encompassing 852 books on medicine published in the Polish language. The tool is designed to find a book passage (usually a paragraph) relevant to a question posed in natural language and to present it to the user.

We present our novel approach to data annotation, which distinguishes between information needs and term queries. Our annotation data analysis shows extensive disagreement between trained specialists regarding document relevance. We evaluate several rerankers for the domain-specific task in the Polish language, for which currently only zero-shot rerankers are available. Our experiments prove the superiority of such reranking models to a strong BM25-based baseline. It is found that rerankers trained on vast multilingual data in a zero-shot setup perform better than a language-specific model fine-tuned to minor domain reranking data.

The rest of the paper is organized as follows. Section II concerns related work in biomedical and medical natural language processing, especially information retrieval. In Section III, we explain our task as a reranking problem, differentiating it from a full retrieval setup, and provide an overview of our search engine configuration. In Section IV we describe a typical use case for our system, which determines our annotation process presented in Section V. In Section VI we report statistics and the conclusion of the collected dataset and present our dataset preparation steps. Then, in Section VII the reranking task setup is described, which leads to Section VIII, where reranking models are presented, and Section IX, where their results are reported. In sections X and XI the possible future work and conclusions of this paper are presented.

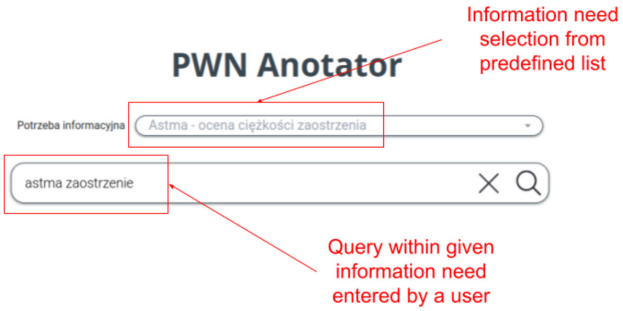

Fig. 1. Information Need selection for annotators and sample query.

II. Related work

Recent findings in Natural Language Processing, particularly Large Language Models (LLMs), have significantly increased the level of language understanding not only in general-oriented tasks, but also in biomedical and medical tasks [2], [3], [4], [5], [6]. In [7] the authors present a benchmark for a question-answering task in the medical domain, and show that the answers provided by LLMs are in agreement with expert knowledge in 93% of cases. Another medical benchmark is introduced in [8], where the authors conclude that pretraining language models from scratch results in gains over fine-tuned general-domain language models. However, a comprehensive survey on biomedical question-answering [9] shows the immaturity of such systems

in a real-life scenario. All the above-mentioned models and benchmarks presented are in English, and no such corpora and models exist for the Polish language, which differentiates this work from the others.

To make a binding decision on a medical case, a human expert (for example, a doctor) prefers to rely on verified medical knowledge sources, rather than one (even precise) answer generated by a language model. This is mainly due to the phenomenon of artificial hallucination [10], [11], [12]. Relevant information may be found in a digital resource by a ranking function (e.g. BM25), optionally modified by a reranker. One existing benchmark for the reranking task [13] is available for the English language. [14] reports on the machine translation of the MS MARCO dataset [15] into multiple languages. The authors claim that their reranking models perform well even in non-English languages when fine-tuned in a zero-shot manner. Healthcare decision-making based on a search engine are examined in [16], [17], and some medical search models and datasets are proposed in [18], [19], [20].

III. Search engine setup

According to [13], models based on reranking are superior to full retrieval models. Moreover, it is easier to perform automatic evaluation on reranking models than full retrieval models, because such evaluation avoids cases when the model retrieves a document unseen by any human annotator. For these reasons, we decided to formulate our task in a reranking setup.

In order to meet commercial expectations, we needed to craft as strong baseline as possible. We started with the SOLR engine, equipped with the Polish Morfologik [21] lemmatizer. We handcrafted the scoring function, awarding full n-gram matches higher scores than word matches. Moreover, we used carefully adjusted weights to ensure case sensitivity, as this is crucial for the recognition of medical abbreviations (AED, DIC, etc.).

IV. Medical search case scenario

To mirror user needs we fabricated case scenarios, namely real-world situations that may cause an Information Need (IN) on the part of the system user. A case scenario consists of an event description, initial conditions, and the Information Need, represented in two forms: a natural language expression and a term query. We define an IN as an abstract term: the knowledge that a user wants to acquire from the system.

An example scenario is shown here:

- Event description: *A 30-year-old female patient presents to a PCP (Primary Care Provider) in a small town. She has severe sore throat and a high temperature.*
- Initial conditions:
  - The doctor measured the patient's temperature (38.5 °C).
  - The doctor confirmed characteristic symptoms of tonsillitis: distended and reddened mucous membrane of the tonsils and palate.
  - The patient reported that she is breastfeeding.
- Natural language description of the IN: *I want to learn how to treat tonsillitis in a breastfeeding woman.*
- Term query: *tonsillitis in a breastfeeding woman - treatment*

TABLE I
Statistics for annotators

	mean	stdev
Information Needs annotated	39	43
total queries used	264	230
total passages annotated	12,068	10,662
time for an Information Need	67 min	55 min
time for a query	6 min	2 min
time for an annotation	8 sec	3 sec
relevant/all annotations ratio	0.29	0.17

V. Annotation process

We hired 21 medical workers (doctors, paramedics, and medical students) for consultation on system requirements and for the annotation process. Initially, they were asked to propose some INs that they may encounter in their work. Additionally, to collect other potential INs, we used the website https://konsylium24.pl/, which is a Polish web forum for medical staff. The website verifies whether users are listed in Polish doctors' registers.

Once the set of INs had been established, we started the annotation process. After logging in, an annotator chooses an IN which he/she feels familiar with. The selection window is presented in Figure 1.

The annotator inputs a number of **queries** for each IN to a SOLR-based search engine, so that for one IN there are always multiple queries. The user may input any words that may help them find a relevant document (synonyms, hyperonyms, etc.).

Example queries for the above-mentioned IN may be: *how to treat tonsillitis in a breastfeeding woman?; tonsillitis breastfeeding treatment; breastfeeding medicines tonsilits; breastfeed woman amigdalitis;* etc.

The annotators were advised not to exceed 20 queries for an IN, and to stop when further enquiry was unlikely to return new relevant passages. The annotation platform returned a maximum of 5 pages, with 10 passages per page, for a query, as in Figure 2. The annotators were asked to read all returned passages and to tag them as relevant/irrelevant to the IN only, regardless of the input search query. If the same passage was returned again within an IN in response to a different query, the annotator would tag it once more. In total, the annotators spent over 478 hours actively tagging the passages. Statistics on their work are given in Table I.

The aim of the procedure was to acquire a more accurate dataset for training and evaluation than a simple query–passage relevancy dataset, which would be limited by the top documents returned by SOLR for one query. The dataset should help the reranker learn semantic structures such as synonyms and hyperonyms.

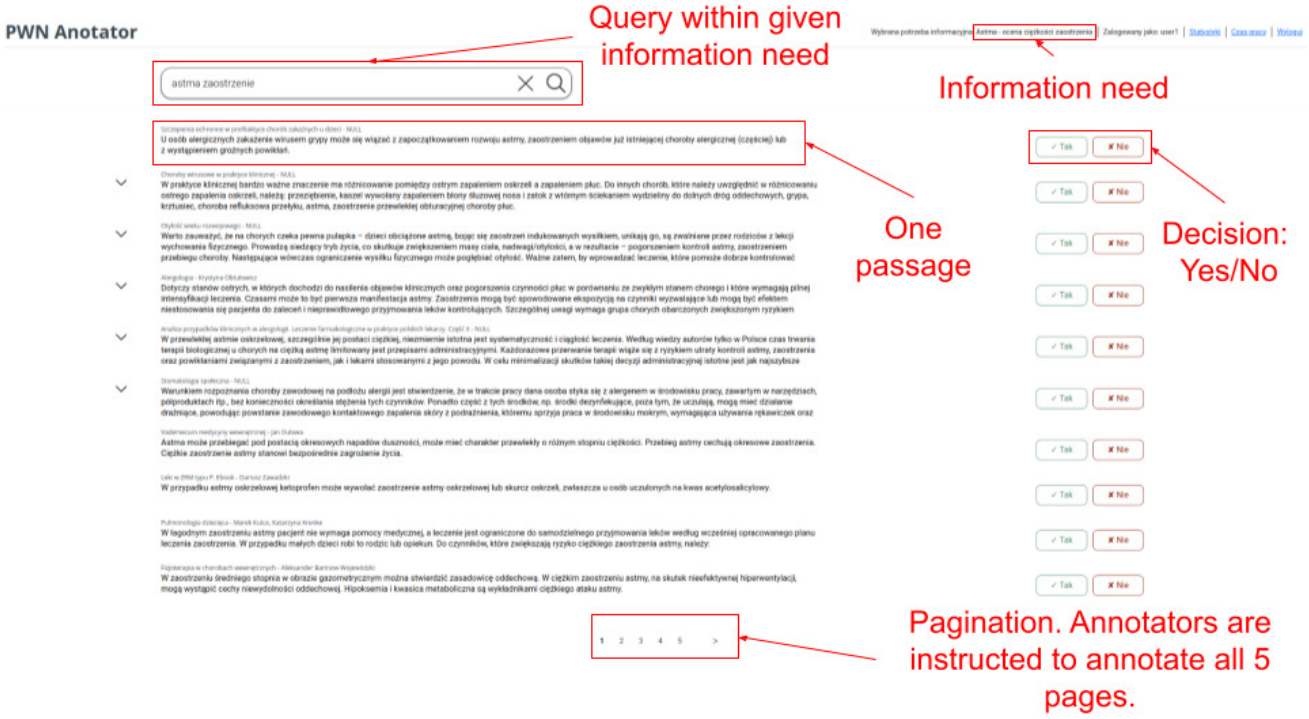

Fig. 2. Passages annotation view for a given Information Need.

VI. Dataset preparation

We rejected INs with fewer than 100 annotations. The dataset consists of 231 INs, each of them represented by a natural language description and one best-fitted term query. In total, we obtained 230,808 triplets of the form (IN, passage, relevancy annotation). Of these, 50,608 decisions were positive and 180,200 were negative, which means that about 22% of annotations were marked as relevant. A passage could be annotated more than once with different tags within the same IN, for example, if two or more annotators worked with the same IN. An analysis of mismatched annotations is presented in Table III. The results show weak agreement on IN–passage relevance between annotators. The mixed opinion percentages range from 15.6% (for two annotations) to above 50% (seven annotations and more). Carelessness on the part of the annotators is probably not the main reason for the disagreement, as we monitored their activity in the annotation platform.

VII. Task setup

We carried out experiments with Information Needs being represented firstly by a term query and then by a natural language expression.

For each IN, we queried the backbone SOLR-based search engine. The returned documents (no more than 500 for each IN) formed an input sample for the reranking model. The proposed model was expected to return the same set of documents sorted in order of decreasing relevance.

TABLE II
Information Need statistics. Annotations concern relevance for (IN, passage) pairs. There may be multiple annotations for one pair.

items per IN	minimum	maximum	mean	median
queries	1	109	21	15
returned passages	50	2279	439	325
annotations	100	4491	999	714
decision: relevant	1	1868	219	119
decision: irrelevant	23	4396	780	562

The golden truth relevance of a document is binary. The document is regarded as relevant to an IN if it was tagged as positive by at least 50% of annotators. NDCG@10 and NDCG@50 are used as evaluation metrics. NDCG metric is well defined in [22].

VIII. Models

For a baseline, we used the SOLR-based model described in section III.

HerBERT [23] (huggingface: allegro/herbert-base-cased) is a Polish-language model which achieves good results in Polish language understanding tasks [24]. We fine-tuned it on our training dataset. Unfortunately, there are no mass Polish corpora for reranking to use along with our data.

There exist some multilingual neural cross-encoder rerankers running on Polish texts. They use an mT5 [25] or mMiniLM [14] backbone, also trained on Polish texts. These models are further fine-tuned for document reranking on multilingual MS MARCO datasets using one or more

TABLE III
STATISTICS ON ANNOTATIONS FOR (INFORMATION NEED, PASSAGE) PAIRS. ONE PAIR MAY BE ANNOTATED BY MULTIPLE ANNOTATORS, **K** REPRESENTS HOW MANY ANNOTATORS ANNOTATED GIVEN (INFORMATION NEED, PASSAGE) PAIR. COLUMN **ANNOTIATIONS** EQUALS TO **K*PAIRS WITH K ANNOTATIONS**. COLUMN **ALL ANNOTATIONS RELEVANT** IS SIMPLY A NUMBER OF ANNOTATIONS IN WHICH ALL THE ANNOTATORS AGREE THAT A GIVEN PAIR IS RELEVANT. COLUMN $\geq$ **0.5 ANNOTATIONS RELEVANT** STANDS FOR A NUMBER OF PAIRS IN WHICH AT LEAST HALF OF THE ANNOTATORS AGREE THAT A PAIR IS RELEVANT.

k	pairs with k annotations	annotations	annotations %	all annotations relevant	all annotations irrelevant	annotations mixed	all annotations relevant %	all annotations irrelevant %	annotations mixed %	$\geq$ 0.5 annotations relevant	$\geq$0.5 annotations relevant %
1	70301	70301	30.5 %	10572	59729	0	15.0 %	85.0 %	0.0 %	10572	15.0%
2	15189	30378	13.2 %	1850	10970	2369	12.2 %	72.2 %	15.6 %	4219	27.8%
3	5452	16356	7.1 %	540	3478	1434	9.9 %	63.8 %	26.3 %	1112	20.4%
4	3607	14428	6.3 %	275	2055	1277	7.6 %	57.0 %	35.4 %	1009	28.0%
5	2134	10670	4.6 %	136	1069	929	6.4 %	50.1 %	43.5 %	438	20.5%
6	1510	9060	3.9 %	86	752	672	5.7 %	49.8 %	44.5 %	330	21.9%
7	1024	7168	3.1 %	43	458	523	4.2 %	44.7 %	51.1 %	222	21.7%
8	863	6904	3.0 %	36	377	450	4.2 %	43.7 %	52.1 %	197	22.8%
9	583	5247	2.3 %	30	255	298	5.1 %	43.7 %	51.1 %	141	24.2%
10	589	5890	2.6 %	34	224	331	5.8 %	38.0 %	56.2 %	151	25.6%
>10	3057	54406	23.6 %	117	910	2030	3.8 %	29.8 %	66.4 %	818	26.8%
total	104309	230808	100%	13719	80277	10313	13.2 %	77.0 %	9.9 %	19209	18.4%

languages other than English (but not including Polish). The authors of [14] proved that these models learn to rerank documents in this zero-shot setup. We used mT5-based rerankers (huggingface: unicamp-dl/mt5-base-mmarco-v2, unicamp-dl/mt5-3B-mmarco-en-pt, unicamp-dl/mt5-13b-mmarco-100k) and mMiniLM-based rerankers (huggingface: cross-encoder/mmarco-mMiniLMv2-L12-H384-v1), which vary in terms of number of parameters and inference time. We used these rerankers in two setups: without fine-tuning (no-ft) and with additional fine-tuning to our training data (ft). We did not fine-tune the mT5-based rerankers because of the long inference time, which meant that they would not be useful as production models. We also tested several multilingual bi-encoders, among which mpnet (huggingface: paraphrase-multilingual-net-base-v2) [26] performed best. All fine-tuned models were trained separately on term queries and natural language queries.

IX. RESULTS

The results are given in Table IV. Almost all cross-encoder rerankers achieve better results than the SOLR baseline. Only HerBERT performs worse, probably due to its being trained with only 100 samples of INs, in contrast to other transformer models that were trained previously on the multilingual MS MARCO dataset containing millions of samples. Cross-encoder rerankers based on mMiniLM are the fastest as regards inference time and achieve results that are much better than the baselines, but not as good as those of the larger models, especially the reranker based on mT5 13B. Further fine-tuning of the reranker based on mMiniLM on our 100 samples dataset improves its quality on natural language queries, but not on term queries. All of the cross-encoder models produce better results when trained on term queries than when trained on natural language queries. This also holds for HerBERT, although that model did not see multilingual MS MARCO or another reranking dataset with short queries. For the bi-encoder mpnet the opposite is true, possibly because of the similarity of the natural language sentences in the corpus on which it was trained.

In our opinion, the ft mMARCO MiniLM appears to be the best model for production applications. Its inference time is satisfactory and increases the NDCG@10 from 34.30 to 43.76 in term queries, and from 28.18 to 40.52 in natural language queries. The NDCG@50 gains are not less resounding-respectively from 32.72 to 34.80 and from 25.16 to 32.53. However, in terms of business terms, we value NDCG@10 over NDCG@50, since we expect a user to be more likely to browse only the top ten search results than 50.

X. FUTURE WORK

The next step is to perform an automatic translation of the MS MARCO dataset into the Polish language and to fine-tune a Polish or multilingual model. It would be beneficial to test models that have also been pre-trained on Polish medical text corpora. Another suggestion is to replace the SOLR search system with a fast bi-encoder network or late interaction transformer [27] in order to enrich the reranker input with passages using synonyms in the medical domain, which are difficult to create manually. After releasing the product for commercial use, we will collect real users' logs for the model training dataset, and run A/B tests.

TABLE IV
MODELS' RESULTS ON THE TEST DATASET. FINE-TUNED MODELS ARE TRAINED SEPARATELY ON TERM QUERIES AND NATURAL LANGUAGE QUERIES. THE INFERENCE TIME IS AVERAGED FOR ONE NATURAL LANGUAGE IN QUERY WITH UP TO 500 DOCUMENTS FOR BATCH SIZE 30 AND THE NVIDIA A100 80GB MODEL CARD. FOR BI-ENCODERS, DOCUMENT ENCODING IS NOT INCLUDED IN THE INFERENCE TIME, AS IT MAY BE DONE OFFLINE. THE ABBREVIATION FT INDICATES FINE-TUNING ON OUR TRAINING DATASET, AND NO-FT INDICATES NO FINE-TUNING.

	term query		natural language query			
method	NDCG@10	NDCG50	NDCG@10	NDCG50	inference time [s]	params
random baseline	6.19	7.12	4.58	5.99	-	-
SOLR	34.40	32.72	28.18	25.16	-	-
no-ft mmpnet bi-encoder	26.02	23.46	29.30	24.38	0.05	278M
ft HerBERT base	30.86	28.44	14.36	13.83	2.08	124M
no-ft mMARCO MiniLM	43.41	34.69	35.64	28.55	0.96	118M
ft mMARCO MiniLM	43.76	34.80	40.52	32.53	0.96	118M
no-ft mT5 base	41.17	33.81	36.94	29.70	3.90	582M
no-ft mT5 3B	44.78	38.02	43.13	34.11	27.60	3742M
no-ft mT5 13B	45.45	39.97	44.87	36.25	93.47	12921M

XI. CONCLUSIONS

In this paper, we have described the process of collecting datasets for a search engine for healthcare professionals. We placed emphasis on cooperation with specialized end users. We built models for queries formulated either in natural language or by means of keywords. We distinguished between an information need and a query that serves to satisfy such a need. We fine-tuned and evaluated several rerankers, which turned out to perform better than the baselines. In our experiments, searching with term queries yielded slightly better results than the use of natural language queries. Moreover, we observed a considerable lack of consent in annotations between qualified medical workers.

Our work is based on Polish medical texts, for which no mass reranker corpora or reranker models are available, except for those fine-tuned in a zero-shot manner. We have shown that the described setup is sufficient for creating a production-ready reranker for Polish medical texts and that zero-shot trained multilingual reranker models perform better than rerankers trained on a language-specific model fine-tuned on only a small number of INs.

REFERENCES

[1] I. Portoghese, M. Galletta, R. C. Coppola, G. Finco, and M. Campagna, "Burnout and workload among health care workers: the moderating role of job control," *Safety and Health at Work*, vol. 5, no. 3, pp. 152–157, 2014.

[2] P. Lewis, M. Ott, J. Du, and V. Stoyanov, "Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art," in *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, (Online), pp. 146–157, Association for Computational Linguistics, Nov. 2020.

[3] H.-C. Shin, Y. Zhang, E. Bakhturina, R. Puri, M. Patwary, M. Shoeybi, and R. Mani, "BioMegatron: Larger biomedical domain language model," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, (Online), pp. 4700–4706, Association for Computational Linguistics, Nov. 2020.

[4] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, pp. 1234–1240, 09 2019.

[5] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," 2020.

[6] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "Biogpt: generative pre-trained transformer for biomedical text generation and mining," *Briefings in Bioinformatics*, vol. 23, no. 6, 2022.

[7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, P. Payne, M. Seneviratne, P. Gamble, C. Kelly, N. Scharli, A. Chowdhery, P. Mansfield, B. A. y. Arcas, D. Webster, G. S. Corrado, Y. Matias, K. Chou, J. Gottweis, N. Tomasev, Y. Liu, A. Rajkomar, J. Barral, C. Semturs, A. Karthikesalingam, and V. Natarajan, "Large language models encode clinical knowledge," 2022.

[8] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, oct 2021.

[9] Q. Jin, Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu, "Biomedical question answering: A survey of approaches and challenges," *ACM Comput. Surv.*, vol. 55, jan 2022.

[10] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, nov 2022. Just Accepted.

[11] Y. Xiao and W. Y. Wang, "On hallucination and predictive uncertainty in conditional language generation," *arXiv preprint arXiv:2103.15025*, 2021.

[12] N. Dziri, S. Milton, M. Yu, O. Zaiane, and S. Reddy, "On the origin of hallucinations in conversational models: Is it the datasets or the models?," *arXiv preprint arXiv:2204.07931*, 2022.

[13] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models," 2021.

[14] L. Bonifacio, V. Jeronymo, H. Q. Abonizio, I. Campiotti, M. Fadaee, R. Lotufo, and R. Nogueira, "mmarco: A multilingual version of the ms marco passage ranking dataset," 2021.

[15] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, *et al.*, "Ms marco: A human generated machine reading comprehension dataset," *arXiv preprint arXiv:1611.09268*, 2016.

[16] A. Bondarenko, E. Shirshakova, M. Driker, M. Hagen, and P. Braslavski, "Misbeliefs and biases in health-related searches," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, (New York, NY, USA), p. 2894–2899, Association for Computing Machinery, 2021.

[17] D. Cohen, K. Du, B. Mitra, L. Mercurio, N. Rekabsaz, and C. Eickhoff, "Inconsistent ranking assumptions in medical search and their downstream consequences," in *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, (New York, NY, USA), p. 2572–2577, Association for Computing Machinery, 2022.

[18] N. Rekabsaz, O. Lesota, M. Schedl, J. Brassey, and C. Eickhoff, "Tripclick: The log files of a large health web search engine," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, (New York, NY, USA), p. 2507–2513, Association for Computing Machinery, 2021.

[19] J. Jimmy, G. Zuccon, J. Palotti, L. Goeuriot, and L. Kelly, "Overview of the clef 2018 consumer health search task," *International Conference of the Cross-Language Evaluation Forum for European Languages*, vol. 2125, 2018.

[20] K. Roberts, D. Demner-Fushman, E. M. Voorhees, W. R. Hersh, S. Bedrick, A. J. Lazar, S. Pant, and F. Meric-Bernstam, "Overview of the trec 2019 precision medicine track," in *Proceedings of the Text Retrieval Conference (TREC)*, vol. 1250, NIH Public Access, 2019.

[21] M. Miłkowski and P. IFiS, "Morfologik," *Web document: http://morfologik. blogspot. com*, 2007.

[22] Y. Wang, L. Wang, Y. Li, D. He, and T.-Y. Liu, "A theoretical analysis of ndcg type ranking measures," in *Conference on learning theory*, pp. 25–54, PMLR, 2013.

[23] 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*, (Kiyv, Ukraine), pp. 1–10, Association for Computational Linguistics, Apr. 2021.

[24] P. Rybak, R. Mroczkowski, J. Tracz, and I. Gawlik, "Klej: Comprehensive benchmark for polish language understanding," *arXiv preprint arXiv:2005.00630*, 2020.

[25] 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*, (Online), pp. 483–498, Association for Computational Linguistics, June 2021.

[26] N. Reimers and I. Gurevych, "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, 11 2019.

[27] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 39–48, 2020.