Consumer Health Question Answering Using Off-the-Shelf Components

Alexander Pugachev[1(✉)], Ekaterina Artemova[2], Alexander Bondarenko[3], and Pavel Braslavski[1,4]

[1] HSE University, Moscow, Russia
apugachev@hse.ru
[2] Center for Information and Language Processing (CIS), MaiNLP Lab., LMU Munich, Munich, Germany
[3] Friedrich-Schiller-Universität Jena, Jena, Germany
[4] Ural Federal University, Yekaterinburg, Russia

Abstract. In this paper, we address the task of open-domain health question answering (QA). The quality of existing QA systems heavily depends on the annotated data that is often difficult to obtain, especially in the medical domain. To tackle this issue, we opt for PubMed and Wikipedia as trustworthy document collections to retrieve evidence. The questions and retrieved passages are passed to off-the-shelf question answering models, whose predictions are then aggregated into a final score. Thus, our proposed approach is highly data-efficient. Evaluation on 113 health-related yes/no question and answer pairs demonstrates good performance achieving AUC of 0.82.

Keywords: Health question answering · Medical information retrieval

1 Introduction

People actively seek answers to health-related questions online [13,15]. However, about half of top-ranked search engines' results may provide incorrect answers to such questions [6,30,31]. Consequently, there have been many research efforts to improve health-related search by ranking documents higher that contain relevant and correct information using, for example, a trustworthiness predictor or explicit expert relevance feedback [16,34]. Also, TREC Health Misinformation Track [9,10] addressed the task of ranking documents returned to health-related queries according to three dimensions: usefulness, credibility, and correctness. Submitted solutions utilized a wide range of IR and NLP methods such as: (1) the fusion of domain-specific representation models with neural quality estimators [27], (2) ensembles of BERT-based classifier built w.r.t. each target dimension [33], (3) continuous active learning to collect the datasets aimed at training T5-based classifier [1], and (4) axiomatic re-ranking [5], etc.

In this work, we take another perspective and move from the ranking task to open-domain question answering (OpenQA). Medical and health QA is an

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 571–579, 2023.
https://doi.org/10.1007/978-3-031-28238-6_48

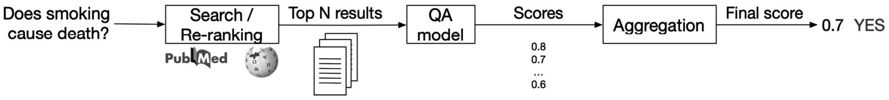

Fig. 1. Our proposed three-step open-domain question answering pipeline for health-related yes/no questions.

area of active research; a recent survey [19] provides a comprehensive overview of the field, including methods and datasets. For a historical perspective, we also refer the interested reader to a survey of pre-neural network methods in biomedical QA [3].

OpenQA aims to find an answer in a large document collection [26,35]. Traditionally, OpenQA pipeline has two components: (1) a *retriever* that returns relevant documents (or their parts – e.g. paragraphs) for the question and (2) a QA model (also referred to as *reader*) that infers the answer from the question-document pair obtained in the previous stage. In our study, we follow this architecture, but in contrast to medical QA systems consisting of dedicated components (see, for instance, [11]), we build our system from ready-to-use third-party blocks. At the retrieval stage, we do not assess information sources' credibility, but restrict the evidence search to PubMed and Wikipedia, both found to be reliable information sources in health and medical domains [21,23]. We use existing search APIs to retrieve documents, thus sparing indexing and ranker training. At the *reading* stage we use freely available QA models trained on existing data – either from the general or medical domain – thus making our approach very data-efficient.

The complete pipeline of our approach is presented in Fig. 1. Given a health-related question like "Does smoking cause death?", we first retrieve relevant documents from either PubMed or Wikipedia. Next, we use question answering models that output prediction probabilities of an answer for every pair of the question and retrieved document in the top-ranked results. In the final step, we aggregate the scores obtained for each question-document pair into a single final answer score.

To test our approach, we use a collection of 113 yes/no health questions like "Does celandine help with cancer?" with ground-truth expert answer annotations. Since, on the one hand, we address a binary prediction task, and on the other hand, we want our approach to inform the asker about to what degree the answer is conclusive, we use AUC as an evaluation measure. Our experiments show that using a mash-up of existing tools and solutions and sparing tailored training data achieves satisfactory results. The most effective combination of Google search over Wikipedia and RoBERTa model fine-tuned on general-domain BoolQ dataset achieves an AUC score of 0.82. Our proposed approach can serve as a strong baseline for health-related yes/no question answering. Our code and data are publicly available on GitHub.[1]

[1] https://github.com/apugachev/consumer-health-qa.

Table 1. The upper part of the table describes 113 test questions (keywords are in **bold**). The bottom part provides examples and statistics of BoolQ, PubMedQA, and BioASQ datasets the readers were trained on. Note that PubMedQA contains 55 questions with the answer *maybe*, while the rest of the data – only *yes/no* answers.

Source	#Questions (y/n)	Example
TREC [2,9]	84 (42/42)	Can **dupixent** treat **eczema**?
Yandex [6]	15 (7/8)	Does **celandine** help with **cancer**?
HBT [4]	14 (12/2)	Does **smoking** cause **death**?
BoolQ [8]	12,697 (7,907/4,790)	Is there a treatment for the bubonic plague?
PubMedQA [18]	500 (276/55/169)	Do mitochondria play a role in remodeling lace?
BioASQ [29]	742 (611/131)	Does metformin interfere thyroxine absorption?

2 Data

Document Collections. The idea of our approach is that we do not need to search for medical information in the wild but instead focus on trustworthy collections: PubMed and Wikipedia. PubMed[2] is a large collection of biomedical literature, comprised of 34 million items at the time of writing. Wikipedia[3] is a large online encyclopedia driven by massive community efforts. At the time of writing, English Wikipedia contains more than 6.5 million articles and is considered a valuable resource of health-related information [28]. An advantage of Wikipedia in the context of our study is that articles about diseases often contain a summary of related treatments, side effects, as well as misbeliefs.

Questions and Answers. The primary source of the test questions used for the evaluation of our proposed approach is the TREC 2019 Decision Track [2] and TREC 2021 Health Misinformation Track [9] data. All the questions have a similar structure: they ask whether a treatment/medicine is helpful for a disease/condition. The test suites contain a question, its corresponding keyword query, a narrative, an answer (*helpful/unhelpful*), and a link to a respective medical publication as evidence for the answer. In our experiments, we make use only of a query, a question, and an answer. To ensure a higher diversity of the test data, we added 15 questions from the *Yandex* log translated from Russian into English and provided with a grounded answer [6]. Finally, we added 14 questions from a study dealing with health beliefs in Twitter (hereafter *HBT*) [4]. These questions are generated from the verified statements from the paper and depart from the rest of the test questions following the *Does X help Y?* pattern and its variations. Since TREC 2019 data is the only one that contains questions with *inconclusive* labels, we removed such questions from the test set. In total, the final test set contains 113 questions, 94 of which are provided with PubMed document IDs as answer evidence (see details and examples in Table 1).

[2] https://pubmed.ncbi.nlm.nih.gov/.

[3] https://en.wikipedia.org/.

3 Approach

Evidence Search. We experiment with three different PubMed retrievers: (1) native PubMed search [24], (2) Google search over PubMed (implemented as a custom search engine restricted to the PubMed domain), and (3) Google's BioMed Explorer.[4] Native PubMed search [14] implements two-stage ranking: first, documents are retrieved based on BM25 scores and then re-ranked using a Lambda-MART-based model [7]. BioMed Explorer combines term- and BERT-based retrieval and is trained on a mix of human-annotated and automatically generated data from biomedical and general domains.[5] In each case, we perform separate searches for keyword and question query variants. In the case of PubMed keyword search, we create a conjunctive (AND) query and restrict the search to TITLE and ABSTRACT fields, thus aiming for high-precision results. In the rest of the configurations, we run a default search with the query as a string. Then, we fetch titles and abstracts of up to top 10 results for subsequent processing.

We search Wikipedia with keyword and question queries using (1) Wikipedia API[6] with default parameters and (2) Google custom search engine restricted to the English Wikipedia domain. We fetch up to 10 articles and split them into paragraphs (each paragraph is combined with the original Wikipedia page title), lemmatize, and rank based on query term occurrences. Top 10 ranked paragraphs are then passed to the readers.

Question Answering. We employ three third-party question answering models: a RoBERTa-large model fine-tuned on BoolQ[7] and two BioLinkBERT-large models – fine-tuned on PubMedQA and BioASQ.[8] RoBERTa-large [22] is a Transformer-based model with 355M parameters that follows BERT's [12] learning regime with some optimizations. BoolQ [8] is a QA dataset consisting of general-domain yes/no questions from Google search log, Wikipedia context paragraphs, and ground-truth answers. BoolQ is categorized into topics, the topic closest to the medical domain is "Nature/Science", which comprises about 20% of the dataset. Fine-tuned RoBERTa achieves an accuracy of 0.86 on the BoolQ test set, which is a good trade-off between the model's performance and size. LinkBERT [32] is also a BERT-like model with a document relation prediction as an auxiliary learning objective. BioLinkBERT-large with 340M parameters is pre-trained on PubMed corpus with citation links that demonstrated state-of-the-art on PubMedQA and BioASQ subsets of the BLURB benchmark for biomedical NLP [17] at the time of publication – 0.73 and 0.95 accuracy points, respectively. PubMedQA [18] is dataset with PubMed abstracts containing 1K expert-annotated yes/maybe/no questions along with a larger portion of unlabeled and automatically generated items. BioLinkBERT model that we use

[4] https://g.co/research/biomedexplorer/.

[5] https://ai.googleblog.com/2020/05/an-nlu-powered-tool-to-explore-covid-19.html.

[6] https://wikipedia.readthedocs.io/en/latest/code.html.

[7] https://huggingface.co/apugachev/roberta-large-boolq-finetuned.

[8] https://github.com/michiyasunaga/LinkBERT.

in our work utilizes only expert-annotated data for training. BioASQ[9] is a yearly challenge on biomedical question answering. At the time of writing the BioASQ training data comprises 4,719 questions of four types: factoid, yes/no, summary, and list questions. BioLinkBERT model employed in our study leverages only a subset of yes/no questions from the BioASQ 2019 edition [25].

The statistics of the data that was used for fine-tuning the readers are summarized in the bottom part of Table 1. The BoolQ dataset is significantly larger than the two medical QA datasets. Moreover, BoolQ questions from real users are "simpler", than the more specialized questions from PubMedQA and BioASQ. We pass questions and up to 10 retrieved PubMed abstracts or Wikipedia paragraphs to the readers that then return a continuous value from 0 to 1 (0 corresponds to a "no" answer and 1 – to "yes"). If no evidence was retrieved, we assign an inconclusive 0.5 score to the question answer.

Score Aggregation. Finally, we use three score aggregation methods: (1) the final score is derived solely from the top 1 evidence document, (2) plain average over the top 10 results (or less, if fewer documents are returned), and (3) weighted average (weights linearly decrease with the increased rank and sum up to one).

4 Results and Discussion

Table 2 reports the results of different configurations of our approach to answering health-related yes/no questions. Although we did not perform a thorough component-based evaluation, we can make some observations about the quality of components in our pipeline based on indirect indicators. For instance, Google retrieved the highest number of evidence documents from PubMed among top 10 results (see 'Hits' column in Table 2). However, we cannot unequivocally interpret these numbers — they can signal a higher search quality or also a search bias in the test collection: TREC annotators might have used Google or another major search engine to find evidence (most of our test questions come from the TREC tracks). We also applied the three readers to the PubMed abstracts available for 94 out of 113 questions (these abstracts come from the original data [2,6,9] and were manually selected by human annotators; one abstract per question). Readers fine-tuned on BoolQ, PubMedQA, and BioASQ achieved 0.88, 0.65, and 0.80 AUC points, respectively. These scores can be regarded as an upper limit estimate for these QA models applied to PubMed abstracts, i.e., the decline in the final scores can be attributed to retrievers' deficiencies. However, one should compare these values with caution, since evidence PubMed abstracts are available not for all questions, and the human bias in selecting these evidence documents may also play a role.

Using PubMed or Wikipedia only leads to a reduced recall and sometimes to no results at all (see '#0' column in Table 2). For example, three out of five search configurations in our experiments failed to find any documents for the

[9] http://www.bioasq.org/.

Table 2. AUC scores for different configurations. Hits: number of evidence PubMed documents in top 10 results; #0: number of queries with no evidence results. Final score aggregations variants: based on top 1 document, plain (avg), and weighted average (wavg) over top-ranked documents. Off-the-shelf readers: RoBERTa-large fine-tuned on BoolQ, BioLinkBERT models fine-tuned on PubMedQA and BioASQ. The overall best result is in **bold**, best results for each retriever are underlined.

	Retriever	Query	Hits	#0	RoBERTa-large (BoolQ)			BioLinkBERT (PubMedQA)			BioLinkBERT (BioASQ)		
					top 1	avg	wavg	top 1	avg	wavg	top 1	avg	wavg
PubMed	PubMed	keywords	10	31	0.65	0.66	0.66	0.61	0.62	0.61	0.56	0.58	0.58
		question	7	42	0.61	0.62	0.62	0.57	0.58	0.58	0.58	0.48	0.48
	Google	keywords	56	1	0.79	0.73	0.73	0.59	0.66	0.68	0.71	0.65	0.69
		question	39	2	0.71	0.76	0.77	0.64	0.64	0.67	0.65	0.57	0.62
	BioMed Explorer	keywords	41	0	0.72	0.75	0.74	0.63	0.64	0.66	0.58	0.74	0.71
		question	39	0	0.74	0.77	0.77	0.60	0.69	0.69	0.72	0.71	0.73
Wikipedia	Wikipedia	keywords	–	26	0.81	0.77	0.78	0.68	0.65	0.68	0.58	0.57	0.59
		question	–	56	0.65	0.68	0.68	0.57	0.56	0.58	0.48	0.48	0.47
	Google	keywords	–	19	0.80	0.79	**0.82**	0.63	0.72	0.68	0.55	0.57	0.57
		question	–	16	0.75	0.75	0.75	0.57	0.67	0.64	0.52	0.53	0.56

TREC 2021 question "Can I get rid of a pimple overnight by applying toothpaste?" and its corresponding keyword query "toothpaste pimple overnight", while Web results for these queries are abundant. Overall, native searchers of PubMed and Wikipedia suffer the most from the no returned results. This is due to restrictions imposed on a PubMed query search and its poor ability to handle question-like queries. Post-processing Wikipedia search results and re-ranking aiming for a higher precision also lead to a lower recall.

The evaluation results (see Table 2) show that RoBERTa fine-tuned on BoolQ significantly outperforms BioLinkBERT models in all configurations. We can conclude that the volume of data for fine-tuning the reader is more important than the in-domain pre-training of the language model. The best results are achieved on Wikipedia documents that often contain relevant information formulated in plain language. The impact of using up to top 10 retrieved results compared to just top 1 document is somewhat mixed: in some cases accounting for documents beyond top 1 improves results, but in other cases, the effect is the opposite. Overall, BioLinkBERT fine-tuned on PubMedQA outperforms its counterpart fine-tuned on BioASQ. General-domain Google search scores higher than other retrievers and Wikipedia is a more useful document collection for consumer health QA in our settings. Using keyword query searches often result in higher evaluation scores of our QA pipeline, although in the case of PubMed with BioMed Explorer (the latter is marketed as a question answering system) question queries outperform keyword variants in the majority of cases.

The highest evaluation results in our experiments are obtained using keyword queries (provided by human annotators), which can be seen as a limitation of our approach since we do not use automatic conversion of questions to queries. However, most natural language questions in the test collection can be transformed into keyword queries automatically by filtering out verbs, determiners, prepositions, and sometimes adverbs as the 'pimple–toothpaste' example suggests (see examples in Table 1).

5 Conclusion

Our solution exploits evidence search in PubMed and Wikipedia for open-domain health question answering. In our approach, we use different search tools to retrieve evidence documents and ready-to-use question answering models. Coupled with simple score aggregation heuristics, this combination delivers satisfactory results – best configuration achieves AUC of 0.82 on 113 test yes/no questions. The proposed approach does not use annotated data directly and does not require training on the target data or task. Thus, it can be considered a strong baseline. However, the main limitation of our work is a small test set, such that the evaluation results and the conclusions should be taken with a grain of salt.

There is ample room for future improvements within our proposed pipeline. In the future, we plan to elaborate on search results post-processing. In particular, we plan to investigate if evidence *sentences* in contrast to paragraphs can help to achieve better results. We also plan to explore if medical thesauri can help to increase recall of the Wikipedia search. Increasing the number and types of test questions, probably gleaning them from various existing question answering datasets, is another interesting avenue for future work.

Acknowledgments. A. Pugachev's research was supported in part through computational resources of HPC facilities at HSE University [20]. A. Bondarenko's work was supported by the Deutsche Forschungsgemeinschaft (DFG) in the project "ACQuA 2.0: Answering Comparative Questions with Arguments" (project 376430233) as part of the priority program "RATIO: Robust Argumentation Machines" (SPP 1999). P. Braslavski's work was supported in part by the Ministry of Science and Higher Education of the Russian Federation (project 075-02-2022-877).

References

1. Abualsaud, M., et al.: UWaterlooMDS at the TREC 2021 health misinformation track. In: Proceedings of the Thirtieth REtrieval Conference Proceedings (TREC 2021). National Institute of Standards and Technology (NIST), Special Publication (2021)
2. Abualsaud, M., Lioma, C., Maistro, M., Smucker, M.D., Zuccon, G.: Overview of the TREC 2019 decision track. In: Proceedings of the Twenty-Eigth Text REtrieval Conference (TREC 2019) (2019)

3. Athenikos, S.J., Han, H.: Biomedical question answering: a survey. Comput. Methods Program. Biomed. **99**(1), 1–24 (2010)
4. Bhattacharya, S., Tran, H., Srinivasan, P.: Discovering health beliefs in Twitter. In: Proceedings of the Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium. AAAI Technical Report, vol. FS-12-05. AAAI (2012)
5. Bondarenko, A., et al.: Webis at TREC 2021: deep learning, health misinformation, and podcasts tracks. In: The Thirtieth REtrieval Conference Proceedings (TREC 2021). National Institute of Standards and Technology (NIST), Special Publication (2021)
6. Bondarenko, A., Shirshakova, E., Driker, M., Hagen, M., Braslavski, P.: Misbeliefs and biases in health-related searches. In: Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM 2021), pp. 2894–2899. ACM (2021)
7. Burges, C.J.: From RankNet to LambdaRank to LambdaMART: an overview. Technical Report, Microsoft Research Technical Report MSR-TR-2010-82 (2010)
8. Clark, C., Lee, K., Chang, M.W., Kwiatkowski, T., Collins, M., Toutanova, K.: BoolQ: exploring the surprising difficulty of natural yes/no questions. 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), pp. 2924–2936 (2019)
9. Clarke, C.L.A., Maistro, M., Smucker, M.D.: Overview of the TREC 2021 health misinformation track. In: Proceedings of the Thirtieth Text REtrieval Conference, TREC 2021. NIST Special Publication, (NIST) (2021)
10. Clarke, C.L.A., Rizvi, S., Smucker, M.D., Maistro, M., Zuccon, G.: Overview of the TREC 2020 health misinformation track. In: Proceedings of the Twenty-Ninth Text REtrieval Conference, TREC 2020. NIST Special Publication, (NIST) (2020)
11. Demner-Fushman, D., Mrabet, Y., Ben Abacha, A.: Consumer health information and question answering: helping consumers find answers to their health-related information needs. J. Am. Med. Inf. Assoc. **27**(2), 194–201 (2019)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. 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), pp. 4171–4186. Association for Computational Linguistics (2019)
13. Finney Rutten, L.J., Blake, K.D., Greenberg-Worisek, A.J., Allen, S.V., Moser, R.P., Hesse, B.W.: Online health information seeking among US adults: measuring progress toward a healthy people 2020 objective. Public Health Rep. **134**(6), 617–625 (2019)
14. Fiorini, N., et al.: Best match: new relevance search for PubMed. PLoS Biol. **16**(8), e2005343 (2018)
15. Fox, S., Duggan, M.: Health online 2013. Health **2013**, 1–55 (2013)
16. Fröbe, M., Günther, S., Bondarenko, A., Huck, J., Hagen, M.: Using keyqueries to reduce misinformation in health-related search results. In: Proceedings of the 2nd Workshop Reducing Online Misinformation through Credible Information Retrieval 2022 co-located with The 44th European Conference on Information Retrieval ECIR 2022. CEUR Workshop Proceedings, vol. 3138, pp. 1–10. CEUR-WS.org (2022)
17. Gu, Y., et al.: Domain-specific language model pretraining for biomedical natural language processing. ACM Trans. Comput. Healthcare (HEALTH) **3**(1), 1–23 (2021)

18. Jin, Q., Dhingra, B., Liu, Z., Cohen, W., Lu, X.: PubMedQA: a dataset for biomedical research question answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 2567–2577 (2019)
19. Jin, Q., et al.: Biomedical question answering: a survey of approaches and challenges. ACM Comput. Surv. (CSUR) **55**(2), 1–36 (2022)
20. Kostenetskiy, P., Chulkevich, R., Kozyrev, V.: HPC resources of the higher school of economics. J. Phys. Conf. Ser. **1740**(1), 012050 (2021)
21. Laurent, M.R., Vickers, T.J.: Seeking health information online: does Wikipedia matter? J. Am. Med. Inf. Assoc. **16**(4), 471–479 (2009)
22. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
23. Morshed, T., Hayden, S.: Google versus PubMed: comparison of google and PubMed's Search tools for answering clinical questions in the emergency department. Ann. Emerg. Med. **75**(3), 408–415 (2020)
24. National Center for Biotechnology Information (US), Bethesda (MD): Entrez Programming Utilities Help (2010)
25. Nentidis, A., Bougiatiotis, K., Krithara, A., Paliouras, G.: Results of the seventh edition of the BioASQ challenge. in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pp. 553–568 (2020)
26. Prager, J.: Open-domain question answering. Found. Trends Inf. Retrieval **1**(2), 91–231 (2007)
27. Schlicht, I.B., de Paula, A.F.M., Rosso, P.: UPV at TREC health misinformation track 2021 ranking with SBERT and quality estimators. CoRR abs/2112.06080 (2021)
28. Smith, D.A.: Situating Wikipedia as a health information resource in various contexts: a scoping review. PloS One **15**(2), e0228786 (2020)
29. Tsatsaronis, G., et al.: An overview of the BioASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinf. **16**(1), 1–28 (2015)
30. White, R.: Beliefs and biases in web search. In: Jones, G.J.F., Sheridan, P., Kelly, D., de Rijke, M., Sakai, T. (eds.) Proceedings of the 36th International Conference on Research and Development in Information Retrieval (SIGIR 2013), pp. 3–12. ACM (2013)
31. White, R.W., Hassan, A.: Content bias in online health search. ACM Trans. Web (TWEB) **8**(4), 1–33 (2014)
32. Yasunaga, M., Leskovec, J., Liang, P.: LinkBERT: pretraining language models with document links. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 8003–8016 (2022)
33. Zhang, B., Naderi, N., Jaume-Santero, F., Teodoro, D.: DS4DH at TREC health misinformation 2021: multi-dimensional ranking models with transfer learning and rank fusion. arXiv preprint arXiv:2202.06771 (2022)
34. Zhang, D., Tahami, A.V., Abualsaud, M., Smucker, M.D.: Learning trustworthy web sources to derive correct answers and reduce health misinformation in search. In: Proceedings of the 45th International Conference on Research and Development in Information Retrieval (SIGIR 2022), pp. 2099–2104. ACM (2022)
35. Zhu, F., Lei, W., Wang, C., Zheng, J., Poria, S., Chua, T.S.: Retrieving and reading: a comprehensive survey on open-domain question answering. arXiv preprint arXiv:2101.00774 (2021)