Generating Multiple-Choice Questions for Medical QA with Distractors and Cue-Masking

Damien Sileo[1], Kanimozhi Uma[2], Marie-Francine Moens[2]

[1]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[2]Department of Computer Science, KU Leuven, Belgium
[1]`damien.sileo@inria.fr`

Abstract

Medical multiple-choice question answering (MCQA) is a challenging evaluation for medical natural language processing and a helpful task in itself. Medical questions may describe patient symptoms and ask for the correct diagnosis, which requires domain knowledge and complex reasoning. Standard language modeling pretraining alone is not sufficient to achieve the best results with BERT-base size encoders. Jin et al. (2020) showed that focusing masked language modeling on disease name prediction when using medical encyclopedic paragraphs as input leads to considerable MCQA accuracy improvement. In this work, we show that (1) fine-tuning on generated MCQA dataset outperforms the masked language modeling based objective and (2) correctly masking the cues to the answers is critical for good performance. We release new pretraining datasets and achieve state-of-the-art results on 4 MCQA datasets, notably +5.7% with base-size model on MedQA-USMLE.

1. Introduction

The multiple-choice question answering (Rogers et al., 2021) task can be formulated with $\{Q, \{O_1...O_N\}\}$ examples where Q represents a question and O the candidate options. The goal is to select the correct answer from the options. Medical multiple-choice question answering (MCQA) has valuable applications for patient or physician assistance but requires very high accuracy to be useful in practice. State of the art on MedQA-USMLE was around 40% (Jin et al., 2020) before the advent of large language models (LLMs) (Kung et al., 2023), but this task remains challenging for text encoders. Medical knowledge is key to this task which can ask questions about patient diagnosis or the appropriate treatment. Medical knowledge graphs, such as UMLS (Schuyler et al., 1993) and SnomedCT (Donnelly et al., 2006) mainly encode terminological knowledge (Schulz and Hahn, 2001). They are sparse when it comes to the practical medical knowledge which is instead available as text in encyclopedias. Various training methodologies allow text encoders to absorb external knowledge. Text encoders acquire some factual knowledge via masked language modeling (MLM) (Petroni et al., 2019), but Jin et al. (2020) showed that MLM objective solely focused on disease names significantly enhances downstream task accuracy.

We compare targeted MLM with auxiliary pretraining on a generated MCQA dataset constructed with medical concepts as answers, associated paragraphs as questions Q, and generated distractors as other options. In particular, we show that we can leverage *differential diagnoses* to obtain distractors. Strictly speaking, differential diagnosis is the process of differentiating several conditions by examining the associated clinical features with additional tests. The term differential diagnosis is also used to denote commonly associated conditions that often need to be distinguished – *bronchitis* is a differential diagnosis of *common cold*. We assemble a dataset of differential diagnoses and show that they provide helpful distractors. We also show that we can find differential diagnoses with a model trained to retrieve incorrect options based on the correct option from an existing MCQA dataset.

We then analyze the importance of properly masking the cues[1] to the correct option O*, i.e., parts of the answer that are present in Q. The DiseaseBERT pretraining masks all the tokens from the disease name to incentivize the model to look at the symptoms. We show that token-level masking is sub-optimal, as the masking of some tokens can also give away the answer. We propose a new masking scheme tailored to MCQA pretraining, called probability-matching cue masking, to prevent both present and masked token from giving away the answer. We also collect new sources of encyclopedic text for medical pretraining. Our contributions are: (i) We compare MLM, targeted MLM and auxiliary fine-tuning on generated MCQA data; (ii) We identify issues with previous cue-masking techniques and propose a new masking strategy; (iii) We propose and distribute[2] new pretraining datasets; and (iv) We perform controlled comparison experiments for our contributions and achieve

---

[1]A cue is the presence of a set of tokens that can help the prediction of the correct answer.

[2]`hf.co/sileod/datasets/wikimedqa`, also in tasksource (Sileo, 2023)

state-of-the-art on 4 datasets for 120M parameters models. Our datasets can also provide new instruction-tuning data for LLMs.

2. Related work

Medical text encoders pretraining: Numerous models (Lewis et al., 2020; Lee et al., 2020; Michalopoulos et al., 2021; Lewis et al., 2020; Kanakarajan et al., 2021; Gu et al., 2021; Yasunaga et al., 2022) adapt BERT pretraining to the biomedical domain to derive domain-specific text encoders. Our work is close to DiseaseBERT (He et al., 2020) which builds upon these encoders as an additional pretraining stage (Phang et al., 2018) to improve their knowledge. Other work focus on external knowledge extraction (Chen et al., 2019; Xia et al., 2021) and integration (Guu et al., 2020; Wang et al., 2021), but knowledge augmented models still rely on pretrained text encoders.

Medical question answering: Multiple datasets were proposed for medical MCQA for English (Jin et al., 2020; Pal et al., 2022; Hendrycks et al., 2021), Spanish (with English translations) (Vilares and Gómez-Rodríguez, 2019) and Chinese (Jin et al., 2020; Li et al., 2021). PubMedQA (Jin et al., 2019) and emrQA (Pampari et al., 2018) are other large-scale biomedical QA datasets, but they address extractive question answering, i.e., cases where the answer to a question is explicitly in the text. Our work is the first to generate MCQA data for medical domain pretraining.

Distractor prediction: Our work is related to the problem of generating distractors for multiple-choice questions. These models use the existing answers to derive other answers that are plausible yet wrong. We distinguish two strands of approaches. Retrieval-based models use the correct answer as a query to retrieve related yet wrong alternatives among the answers to other questions (Ha and Yaneva, 2018). Generation-based models (Chung et al., 2020) learn to generate distractors with language models and focus on the diversity and adequacy of the generated distractors. Here, we tailor distractor prediction to medical MCQA and also draw a new parallel between distractor generation and differential diagnosis.

3. Improving knowledge infusion

DiseaseBERT infuses knowledge by predicting title tokens based on paragraphs where title tokens are masked. We replace that stage with a fine-tuning stage with synthetic MCQA data and provide new analyses on token masking. Figure 1 illustrates the MCQA generation process and the problem of extraneous masking.
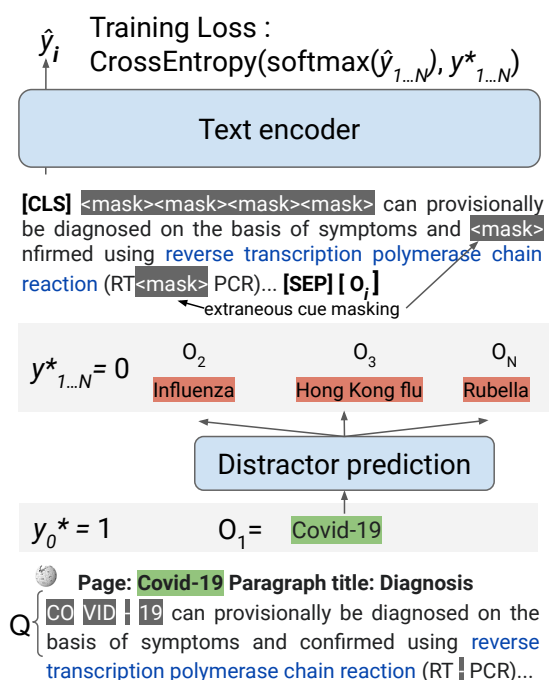

Figure 1: Overview of our data generation technique. Here, we illustrate the naive cue masking scheme used by (Jin et al., 2020), where tokens from the correct answers are masked. The second maskings of *co* and - can be cues to the answer.

3.1. MCQA data generation with differential diagnoses and distractor retrieval

We generate data by using paragraphs content as questions Q, the title as a correct option O^*, and to generate interesting data, we look for related but distinct options, and we mask the direct cues to the correct options in Q. We noticed that some Wikipedia pages were associated with differential diagnoses cross-linked on DBPedia. We collect associations between pages and differential diagnoses and use them as negative examples. In section 5, we will show that they constitute high-quality negative examples. Since these differential diagnoses are not available for all pages, - especially pages that are not related to diseases, but procedures -, we derive negative examples by using a retrieval model. We will train a retrieval model on previous, smaller MCQA datasets, and evaluate the retriever's ability to find differential diagnoses. We then use the retrieval models to find the most related titles on the same encyclopedia.

We also experimented with distractor generation by using generative models, but obtained unconvincing results. One advantage of using retrieval, is that because of editorial choices, each page covers different information. This prevents retrieved negative examples from being too similar to the

correct answer, and this feature is not exploited by distractor LM-based generations models.

3.2. Cue masking

Naive masking: (Jin et al., 2020) masks tokens that are in the answer. However, masking some particular tokens is not necessarily concealing the information about the targeted disease. Some specific tokens are masked everywhere (e.g., a dash -). If these tokens can be predicted based on surrounding terms, the correct disease can be guessed without actually using useful medical knowledge. We call the masking that helps easy guesses *extraneous masking*, an example of which is illustrated in Figure 1, where the masked dash can give away the answer if the model has learned that a dash is plausible between *RT* and *PCR*. To address this problem, we will evaluate word-level masking, which necessarily leads to less extraneous masking.

Probability-matching masking (ours): Another problem with naive masking is that if tokens from the correct answer are necessarily masked, a model can detect an incorrect option when it contains a word w that is in the question as w would be masked if the option was correct. We propose a new strategy that takes advantage of the negative examples to alleviate this phenomenon. Instead of always masking a word when it is in the correct answer, we mask a word w with the following probability:

$$p_w = \frac{1}{|\{O_i, w \in O_i, i \in 1..N\}|} \quad (1)$$

where O denotes the other options. This masking scheme ensures that no cue-based classifier can predict the correct answer based on neither presence nor absence of specific tokens. It also prevents common tokens from being unnecessarily masked.

4. Datasets

4.1. Pretraining data

We collect new pretraining data from three open-source websites:

Wikipedia Medicine Portal: We crawl pages from the Wikipedia Medicine projects which indexes medical pages.[3] We remove pages that match persons or organizations according to WikiData, and pages referring to years. We obtain a total of $75k$ paragraphs.

Wikidoc: We crawl overview pages from the Wiki-Doc specialized encyclopedia[4] which leads to $28k$ paragraphs.

Loss/Masking strategy	MedQA-USMLE Accuracy
BioLinkBERT-base (Yasunaga et al., 2022)	40.0
+MLM/Token (Jin et al., 2020)	40.1
+MLM/Word	41.2
+Discriminative/Word	42.5
+Discriminative/Probability-Matching	**43.6**
- Differential diagnoses	42.6

Table 1: MEDQA-USMLE validation-set accuracy percentage of BioLinkBERT after knowledge infusion, with varying losses and masking strategies.

WikEM: We crawl content pages from the the WikEM encyclopedia[5] which is an open source medical encyclopedia targeted for emergency medicine, and we obtain $15k$ paragraphs.

4.2. Downstream tasks

We use 4 medical MCQA datasets to perform evaluation. These datasets contain a question and four options, one of them being correct.

MedQA-USMLE: (Jin et al., 2020) gathers 10k/1.2k/1.2k train/validation/test medical MCQA examples collected from training questions for medical entrance exams found on the Web.

MedMCQA: (Pal et al., 2022) contains 182k/6.2k/4.2k train/validation/test medical entrance exam training questions.

HEAD-QA: (Vilares and Gómez-Rodríguez, 2019) we focus on the medical questions translated to English with 0.2k 0.4k validation/test examples, and use the MedMCQA train set as our train set.

MMLU: (Hendrycks et al., 2021) is the professional medicine subset of the MMLU language understanding benchmark, which contains 272 test examples. Following (Yasunaga et al., 2022), we used MedQA-USMLE as a training set.

5. Experiments

We generate MCQA examples with the aggregated paragraphs of the pretraining data from Section 4.1, with the distractor generation of section and masking strategies of section 3. We first compare cue masking schemes and pretraining objectives, with ablations on the MedQA-USMLE dataset, then show overall results with the other datasets. We fine-tune BioLinkBERT[6] on WikiMedQA then evaluate BioLinkBERT+WikiMedQA on each downstream task with fine-tuning. We always use standard hyperparameters (5 epochs, sequence length

---

[3] en.wikipedia.org/...medicine_articles
[4] www.wikidoc.org/...:AllPages

[5] wikem.org/...?title=Special:AllPages
[6] WikiMedQA fine-tuning also improves the accuracy of PubMedBERT (Gu et al., 2021) and BioElectra (Kanakarajan et al., 2021) but both still underperform BioLinkBERT on downstream tasks

	HEAD-QA+K	MedMCQA	MedMCQA+K	MedQA	MMLU
Previous sota under 340M parameters	42.4	40.0	43.0	44.6	50.7
BioLinkBERT-base (Yasunaga et al., 2022)	40.0	43.4	49.1	40.6	44.5
BioLinkBERT-large (Yasunaga et al., 2022)	44.1	48.3	52.8	44.6	50.6
BioLinkBERT-base+DiseaseBERT	37.8	47.8	44.5	41.2	50.5
BioLinkBERT-base+WikiMedQA	41.0	46.9	49.0	45.7	49.6
BioLinkBERT-large+WikiMedQA	**44.5**	**50.8**	**53.9**	**47.2**	**51.1**

Table 2: MCQA test accuracy of previous state-of-the-art models and ours. DiseaseBERT (He et al., 2020) uses word-level masking. $\mathcal{D}+K$ refers to the dataset $\mathcal{D}$ with retrieved external knowledge concatenated to the question (see section 5.1). HEAD-QA previous sota is from (Liu et al., 2020). MedMCQA sota are from (Pal et al., 2022), and MedQA and MMLU sota are from (Yasunaga et al., 2022); these sota (state of the art) results use large-sized models).

of 256, learning rate of 2.10^{-5} batch size of 16). We use a multiple-choice-question answering setup (we predict logit scores for each option by concatenating the question and the option, then use a softmax and optimize the likelihood of the correct option).

5.1. Knowledge augmentation

We also evaluate the pretrained models in a setting where retrieved external knowledge is concatenated to the question. We index previously mentioned Wikipedia medical articles with a BM25[7] search engine, using ElasticSearch 8.0 default hyperparameters (Robertson et al., 2009), and we concatenate the 10 most relevant passages[8].

5.2. Distractor prediction

To perform distractor prediction, we optimize the `MultipleNegatives` ranking loss (Henderson et al., 2017) on the SentenceBERT framework (Reimers and Gurevych, 2019) using a BioLinkBERT-base (Yasunaga et al., 2022) text encoder and default parameters. We train the ranking model on the MedMCQA training examples, using the correct answer as a query, the associated noncorrect options as relevant and answers to other questions as irrelevant distractors. We evaluate distractor prediction on 3446 differential diagnoses collected on DBPedia, and found out that using a disease as a query returns a correct differential diagnosis with precision@3/recall@3 of $11\%/15\%$[9].

We generate 7 incorrect options for each title associated with the paragraphs of texts from the section 4.1, and we use probability-matching cue masking to build the WikiMedQA dataset. We use

differential diagnoses and retrieved distractors as additional options.

5.3. Cue masking and distractors retrieval

Table 1 compares fine-tuning on the WikiMedQA Wikipedia part to the DiseaseBERT infusion, and shows the impact of masking strategies. Word-level masking outperforms token-level masking, which shows that less masking leads to less extraneous masking and better knowledge infusion. Probability-matching masking also outperforms naive masking at the word level which further validates the importance of addressing extraneous masking. Finally, removing the differential diagnoses from the options substantially decreases accuracy, which showcases the value of differential diagnoses as natural distractors. From now on, we refer to the generated data with differential diagnoses and probability-matching masking as WikiMedQA.

5.4. Overall results

Table 2 shows the test accuracy of BioLinkBERT fine-tuned on WikiMedQA then on various datasets, compared to the task-specific state-of-the-art. Fine-tuning on WikiMedQA leads to considerable accuracy improvements on all tasks, whether external knowledge is available or not, which shows that this pretraining leads to generalizable text representations for medical question answering.

6. Conclusion

We proposed a new dataset for Medical MCQA in English pretraining by leveraging distractor retrieval and cue masking. We identified the problem of extraneous masking, proposed probability-matching masking, demonstrated its advantage, and showed that differential diagnoses were helpful distractors. Fine-tuning on WikiMedQA leads to considerable

---

[7]We also experimented with Dense Passage Retrieval (Karpukhin et al., 2020) but obtained inferior results.

[8]The concatenated knowledge is truncated if it leads to overflow of text encoder input maximum sequence length.

[9]Random chance scores less than $0.2\%/0.2\%$. BM25 scores $0.6\%/0.7\%$

improvement on several datasets, and this method can be ported to other languages.

7. Acknowledgements

8. Ethical Considerations

Data Privacy: Our data sources include Wikipedia, WikiDoc, and WikiEm, all of which are public and contain little to no private patient data. Although this minimizes privacy concerns, caution is warranted if the methodology is adapted to more sensitive data sources.

Information Accuracy: While Wikipedia, WikiDoc, and WikiEm are peer-reviewed, they are not necessarily reviewed by medical experts. This presents a risk of inaccurate information, which could be harmful in a medical setting. The models good performance on MCQA benchmarks should not eclipse the need for cross-referencing with authoritative medical sources.

Utility Versus Risk: The models strong performance in academic settings does not immediately translate to clinical safety or efficacy. Further validation is essential before considering real-world healthcare applications.

Responsibility: The models are intended to supplement, rather than replace, human medical expertise. They are not an alternative to professional medical advice.

9. Bibliographical References

Jun Chen, Jingbo Zhou, Zhenhui Shi, Bin Fan, and Chengliang Luo. 2019. Knowledge abstraction matching for medical question answering. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 342–347.

Ho-Lam Chung, Ying-Hong Chan, and Yao-Chung Fan. 2020. A BERT-based distractor generation scheme with multi-tasking and negative answer training strategies. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4390–4400, Online. Association for Computational Linguistics.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pretraining. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Le An Ha and Victoria Yaneva. 2018. Automatic distractor suggestion for multiple-choice tests using concept embeddings and information retrieval. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 389–398, New Orleans, Louisiana. Association for Computational Linguistics.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *ArXiv*, abs/1705.00652.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2020. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *arXiv preprint arXiv:2009.13081*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. 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)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELEC-TRA:pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154, Online. Association for Computational Linguistics.

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.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020. 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*, pages 146–157, Online. Association for Computational Linguistics.

Jing Li, Shangping Zhong, and Kaizhi Chen. 2021. MLEC-QA: A Chinese Multi-Choice Biomedical Question Answering Dataset. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8862–8874, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ye Liu, Shaika Chowdhury, Chenwei Zhang, Cornelia Caragea, and Philip S Yu. 2020. Interpretable multi-step reasoning with knowledge extraction on complex healthcare question answering. *arXiv preprint arXiv:2008.02434*.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

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.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2021. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *arXiv preprint arXiv:2107.12708*.

Stefan Schulz and Udo Hahn. 2001. Medical knowledge reengineering—converting major portions of the umls into a terminological knowledge base. *International Journal of Medical Informatics*, 64(2):207–221.

P Schuyler, W Hole, Mark Tuttle, and David Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81:217–22.

Damien Sileo. 2023. tasksource: A dataset harmonization framework for streamlined nlp multi-task learning and evaluation. *arXiv preprint arXiv:2301.05948*.

David Vilares and Carlos Gómez-Rodríguez. 2019. HEAD-QA: A healthcare dataset for complex reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 960–966, Florence, Italy. Association for Computational Linguistics.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Yuan Xia, Chunyu Wang, Zhenhui Shi, Jingbo Zhou, Chao Lu, Haifeng Huang, and Hui Xiong. 2021. Medical entity relation verification with large-scale machine reading comprehension. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD)*, pages 3765–3774.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. Linkbert: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*.