Querying Across Genres for Medical Claims in News

Chaoyuan Zuo, Narayan Acharya, and **Ritwik Banerjee**
Computer Science, Stony Brook University
{chzuo, nacharya, rbanerjee}@cs.stonybrook.edu

Abstract

We present a query-based biomedical information retrieval task across two vastly different genres – newswire and research literature – where the goal is to find the research publication that supports the primary claim made in a health-related news article. For this task, we present a new dataset of 5,034 claims from news paired with research abstracts. Our approach consists of two steps: (i) selecting the most relevant candidates from a collection of 222k research abstracts, and (ii) re-ranking this list. We compare the classical IR approach using BM25 with more recent transformer-based models. Our results show that cross-genre medical IR is a viable task, but incorporating domain-specific knowledge is crucial.

1 Introduction

In recent years, the general population has increasingly sought out online sources for medical information (Fox, 2011; Fox and Duggan, 2013). Among the various types of sources, they mostly rely on online news articles, which often serve to disseminate medical findings from research studies (Medlock et al., 2015). It is, however, important to identify the source of a medical claim, especially during times of pervasive misinformation and during a pandemic, when people may not be able to visit a healthcare professional. When reporting a medical study, many news articles cite the original study either by embedding hyperlinks or explicitly showing a citation, thus providing the reader with critical markers of credibility (Fogg et al., 2009). Not all articles do this, however. Here, we present our work on finding scientific research publications that support the primary claims being made in a health-related news article. We design it as cross-genre query-based (or *ad hoc*) information retrieval (IR): given a medical claim made in a news article, retrieve the research publication supporting it.

(1a) *Tea drinkers live longer.*[†]

(1b) *Tea drinkers live longer, with the biggest boost linked to green variants.*[‡]

(2) *Tea consumption was associated with reduced risks of atherosclerotic cardiovascular disease and all-cause mortality, especially among habitual tea drinkers.*

[†] www.sciencedaily.com/releases/2020/01/200109105508.htm (accessed: May 31, 2020) cites the source and provides a hyperlink to it.

[‡] www.telegraph.co.uk/news/2020/01/09/tea-drinkers-live-year-half-longer (accessed: May 31, 2020) incomplete source information and no hyperlink to the original research.

Table 1: Cross-genre medical IR where the claims (1a and 1b) are presented in lay terms in the news and serve as queries. The support (2) is provided in a research publication, expressed in specialist language.

When scientific research makes its way out of conferences and journals into news meant for general consumption, the information is presented in a drastically different language. The general audience is often poorly equipped for specialist language comprehension, to the extent that changing domain-specific language to one meant for a general audience has been treated as a discipline by itself (Swales, 2000). So this change is necessary on one hand, but on the other hand, it also increases the difficulty of IR, especially so in token-based methods such as BM25 (Robertson et al., 2009).

In this work, we present a dataset (Sec. 2) of claims made in medical news articles, where each claim is associated with at least one peer-reviewed research publication supporting it. For each claim, we present an IR task in Sec. 3 – search for the corresponding publication from a large corpus of medical research literature. The task itself is divided into two stages: (i) retrieve a candidate list of 500 abstracts from a large corpus, and (ii) re-rank them to obtain the correct publication. After discussing our findings, we present an overview of related research in Sec. 4 before concluding.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1783–1789,
November 16–20, 2020.

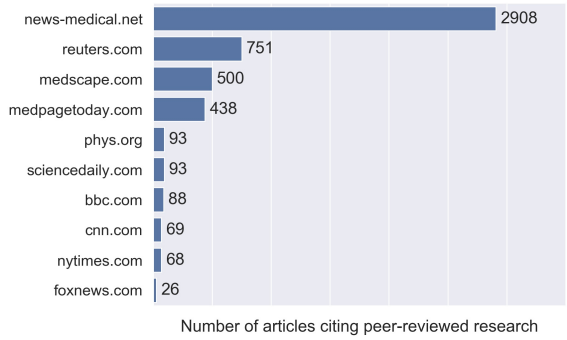

Figure 1: Distribution over sources of medical news articles that provide hyperlinks to cited research.

2 Dataset

Over a period of 18 months (Oct 2018 – March 2020), we collect 72,028 news articles from the RSS feeds of several medical news websites and also from the health category of popular general news websites. To ensure that only articles citing peer-reviewed scientific publications are retained, we check every document for hyperlinks to domains listed by Wikipedia as medical journals[1] and the list of top scientific publications on Alexa,[2] leaving 17,712 articles (24.6%) in our collection. Further, many articles were aggregations of disparate medical studies. We discard these using a combination of heuristics and manual verification, and retain only those articles that report on a single study or on a series of research studies that closely relate to each other. For articles retained after this step, the headline reflects the focal claim or finding of the cited research. This was observed by three independent readers who were given a random sample of 371 articles (7.4% of the dataset). All three agreed that for each one of these 371 articles, the headline did, indeed, present the main research finding. Since some articles cite using embedded hyperlinks, while others offer a reference section at the end of the article, we are able to collect the abstracts of the cited research.

Our final dataset[3] consists of tuples of the form $(h, \{a_i\})$, where h is the headline from a news article, and a_i are the abstracts of the research publications cited by that article. The publication titles are retained as well. There are 5,034 headlines and 4,566 abstracts (since some research publications are cited by multiple news articles). Fig. 1 shows the distribution of the news headlines over the top ten news domains in our collection.

Since not all research is open-access, we restrict ourselves to collecting the abstracts instead of the entire publication. We believe this does not prove to be a hindrance to the task, since it is reasonable to assume that the primary findings of a research study are mentioned in the abstract. We collect these abstracts through PubMed.[4] Further, to mimic the realistic scenario where a human reader or fact-checker needs to retrieve the correct publication (i.e., the research actually upholding the claim being made in a news article) from a vast collection, we also add 217,665 spurious abstracts from the biomedical research literature. We collect these abstracts from the non-commercial use open-access subset of PubMed Central,[5] to serve as the negative samples in our IR task.

3 Experiments

Our task is formulated in two stages, similar to other recent ad hoc IR (MacAvaney et al., 2019a; Yilmaz et al., 2019; Dai and Callan, 2019) – a token-based first step to obtain a candidate list, and then the final ranking by a transformer (Wolf et al., 2019). In spite of recent advances, the transformer-based models are large, and using them to compare each query with each document is computationally expensive even for a small corpus. Thus, the two-stage approach remains a prudent choice.

3.1 Candidate Selection

Given the size of the corpus of biomedical abstracts ($>$ 222k), our goal in this first stage is to reduce the search space for the final ranking task. For this, we consider the classical IR approach of token-based bag-of-words models (e.g., BM25) as well as embedding-based models that encode the claim (i.e., the news headline) and the research abstract in the same space. For the latter, we use the inner product of the embedded representations to measure the similarity between a headline and an abstract (Chang et al., 2020). Since most news articles cite only one research publication, and no article in our dataset cites more than three, precision is not an important measure for this task. Instead, we measure recall@k ($k = 1, 5, 20, 100, 500$). As argued in other recent two-stage approaches (Nie et al., 2019; Soleimani et al., 2020), a high recall is

[1] en.wikipedia.org/List_of_medical_journals
[2] www.alexa.com/topsites/category/Top/Science/Publications
[3] github.com/chzuo/emnlp2020-cross-genre-IR
[4] pubmed.ncbi.nlm.nih.gov
[5] www.ncbi.nlm.nih.gov/pmc

Model	R@1	R@5	R@20	R@100	R@500
Okapi BM25	0.295	0.428	0.538	0.653	0.761
BM25+	0.301	0.436	0.543	0.660	0.768
BM25+†	**0.376**	**0.530**	**0.630**	**0.738**	**0.830**
BERT	0.114	0.196	0.287	0.416	0.569
RoBERTa	0.105	0.191	0.289	0.421	0.576
BC-BERT	0.105	0.204	0.301	0.447	0.607
BC-BERT$^{\text{MED}}$	0.133	0.242	0.347	0.492	0.653
BC-BERT$^{\text{MED}}_{A}$	0.144	0.256	0.364	0.511	0.665
BC-BERT$^{\text{MED}}_{B}$	0.148	0.265	0.369	0.509	0.666

Table 2: **Candidate selection results.** The token-based model with preprocessing steps (†) achieves significantly better results compared to all other models. BC-BERT is the Bio+Clinical model, where $^{\text{MED}}$ denotes fine-tuning on the medical STS data, with A and B denoting the two modifications handling labeled abstracts: the entire abstract being encoded, and only the first and last three sentences being encoded.

crucial here, as the correct abstract will otherwise be left out from the final ranking.

As part of the token-based approaches, we use Okapi BM25 (Robertson et al., 2009) and a variant, BM25+ (Lv and Zhai, 2011a). We employ the Rank-BM25 tool,[6] based on Trotman et al. (2014). We evaluate these with and without preprocessing, where the preprocessing comprises converting the words into lowercase, removing function words, and stemming.[7] We also notice that several abbreviations are used in medical news that are not commonly found in the research literature (BP for "blood pressure", Tx for "treatment", etc.). If such an abbreviation appears more than twice in our dataset, we map it to its expansion, based on a dictionary of medical abbreviations.[8]

For the embedding-based approaches, we use two pre-trained models to encode the claim h and the abstracts a_i – BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019), from Sentence-BERT (Reimers and Gurevych, 2019). We obtain the ranked list of abstracts pertinent to the claim based on the inner products $\langle h, a_i \rangle$. The pre-trained models are fine-tuned on the Natural Language Inference (NLI) and the Semantic Textual Similarity (STS) benchmark datasets (Cer et al., 2017). Considering our dataset comprises medical news and biomedical literature while BERT and RoBERTa are trained on general texts, we also use the Bio+Clinical BERT (Alsentzer et al., 2019) model and tune it on the NLI and STS benchmark datasets. Additionally, we also tune the Bio+Clinical model on the medical STS dataset (Wang et al., 2018). It is worth noting that many medical research abstracts are further divided into labeled sections (e.g., 'Background', 'Results', 'Conclusion'). In our dataset, 36% of the abstracts featured such labels. We conduct three experiments where

(a) the whole abstract is encoded regardless of labeled sections being present (BC-BERT$^{\text{MED}}$),
(b) only the 'Background' and 'Conclusion' are encoded for abstracts with labeled sections (BC-BERT$^{\text{MED}}_{A}$), and
(c) identical to (b) when there are labeled sections, but only the first and last three sentences are encoded otherwise (BC-BERT$^{\text{MED}}_{B}$).

Table 2 shows that token-based models significantly outperform all embedding models in the candidate selection stage, with BM25+ achieving the best recall for all k when the preprocessing steps are included. Among the embeddings, fine-tuning on the medical STS data provides a significant improvement, which indicates the importance of domain-specific training. The BC-BERT$^{\text{MED}}_{B}$ experiment was conducted based on our observation that even in abstracts without labeled sections, the primary claims are seldom made in the middle region. The results appear to support this as well. Its improvement over the other variants of BC-BERT, however, is not significant.

3.2 Transformer-based Ranking

We keep $3,000$ headlines for training, $1,000$ for development, and $1,034$ for testing. We first use the best candidate selection model (BM25+†) to generate a list of 500 abstracts for each headline, and then concatenate a headline with an abstract. These concatenated strings serve as training data for our task. The ground-truth label is 1 for an input $h + a$ where a is, indeed, the abstract cited by the article with headline h. For other inputs, the label is 0. We use this labeled data to tune pre-trained transformer models. During prediction, we use the softmax probabilities of the classification scores to re-rank the abstracts for each headline, and calculate recall@k for $k = 1, 3, 5, 20$, as well as the mean reciprocal rank (MRR).

It is possible that the correct abstract was not retrieved during candidate selection. In that case, we add it back during training (but not testing). Since this data is highly imbalanced (roughly a $1:500$ ra-

[6] Rank-BM25: A two line search engine

[7] Lemmatization yields poorer results, omitted for brevity.

[8] abbreviations.yourdictionary.com/articles/medical-abbrev.html

Model	R@1	R@3	R@5	R@20	MRR
BM25+†	0.364	0.481	0.529	0.671	0.442
BERT$_{(20,50)}$	0.579	0.718	0.755	0.821	0.662
XLNet$_{(20,50)}$	0.543	0.697	0.735	0.804	0.628
DistilBERT$_{(20,50)}$	0.343	0.531	0.604	0.769	0.463
BC-BERT$_{(0,1)}$	0.311	0.527	0.601	0.775	0.447
BC-BERT$_{(4,10)}$	0.538	0.702	0.759	0.825	0.636
BC-BERT$_{(20,50)}$	**0.626**	**0.743**	**0.783**	**0.828**	**0.695**

Table 3: **Ranking results.** The Bio+Clinical model is denoted by BC-BERT. A model tuned on m positive (by augmentation) and n negative samples is shown by the subscript (m, n). The best performance is achieved by Bio+Clinical BERT with 1 epoch, batch size of 24 and maximum sequence length of 512 tokens.

tio for the classes labeled 1 and 0, respectively), we use natural language data augmentation (Ma, 2019) to oversample the positive class. These augmentations work by either inserting or substituting words that are highly likely based on distributional similarity. For training, we choose the augmentation parameters such that at most 10 but not exceeding 30% of the tokens in a sentence are augmented. We generate 4 augmented samples (2 insertions, 2 substitutions) and 20 augmented samples (10 insertions, 10 substitutions) when we use the top 10 and 50 negative samples, respectively, in the list of 500 abstracts for each headline.

As part of our experiments, we train different models – BERT, Bio+Clinical BERT, XLNet (Yang et al., 2019), and DistilBERT (Sanh et al., 2019) – with transformer. We train them on different versions of the datasets controlling for the number of negative samples per claim and the number of augmented positive samples. All models are trained for 1 and 2 epoch, batch size of 16 and 24, maximum sequence length of 256 and 512 tokens, and a learning rate of 5×10^{-5}. The final hyperparameters are manually chosen based on MRR achieved on the development set. All experiments are conducted on NVIDIA Tesla V110 GPUs.

3.3 Discussion

First, there is the existential question about candidate selection: why not simply train the final ranking algorithm with random negative samples instead of the token-based first step? With random negative sampling, we found that it was rather *obvious* for both human readers and learning algorithms that the negative samples did not support the claim, simply because random sampling often draws publications not related to the claim at all. This would defeat the objective of our work, which is to aid readers in attempting to fact-check a health-related claim based on the citation provided in a news article. It is unlikely that readers will compare a publication on a topic vastly different from the one being reported (*e.g.*, the news article makes a claim about COVID-19 while the research is about 'haemophilia'). Thus, even though random negative sampling is commonly used to train fact-checking systems (e.g., Hanselowski et al. (2018); Nie et al. (2019)), it is ill suited for the task presented here.

It is also worth pointing out that our evaluation relies on relevance labels obtained from citations from news articles. It is possible that some documents ranked higher are relevant and provide support to the medical claim, but were judged as irrelevant because they were not cited by the news article. Despite this, recall@k and MRR are meaningful. For instance, if the cited publication is ranked third, while two other relevant publications are ranked above it, recall@k will effectively find success at $k = 3$. With exhaustively verified non-relevance labels, this hypothetical scenario would yield $k = 1$. Obtaining these labels is a daunting task, however. Indeed, many IR benchmark datasets – *e.g.*, MS-MARCO (Nguyen et al., 2016) – do not provide strong non-relevance labels. In this general evaluation setup, the results may instead be viewed as a lower bound (*i.e.*, with exhaustive ground-truth labels of non-relevance, they are better, not worse).

BM25 is hard to beat as a baseline for candidate selection, but token-based methods err when the words in the news headline do not appear in the abstract, which is common when synonymous or similar meanings are expressed using different terms across two different genres. The best embedding-based model, BC-BERT$^{\text{MED}}_{B}$, was able to include 33% of the abstracts that BM25+† failed to retrieve in the top 500 candidates. This also indicates why contextual embeddings improve the ranking results (Table 3). [9] From candidate selection on the test set, the best recall@500 is 0.834, which serves as the upper bound for the ranking task. After training, the transformer-based models can nearly attain this bound for $k = 20$. This is true even for the general BERT embeddings tuned on just 20 positive and 50 negative samples. The Bio+Clinical variant outperforms the other models. The relative improvement over BERT, however, is not significant.

[9] Run times and results on the development set are provided in the Appendix.

Overall, our results show that these embeddings do not need much task-specific tuning on the final ranking. However, both token- and embedding-based approaches fail when the claim is fairly generic (e.g., "Research could help design better flu vaccines"), and these errors happen during candidate selection as well as the final ranking.

4 Related Work

Modern *ad hoc* IR systems are largely built upon bag-of-words representations, using term-weighting techniques like BM25 (Robertson et al., 2009) or its variants (Lv and Zhai, 2011a,b). Catena et al. (2019) used such a variation for query-based news retrieval, which focuses on specific regions in an article. They use the headlines as queries and formulate the task as retrieving the corresponding article. Such headline-content pairs from newswire have similarly been used in neural IR models as well (MacAvaney et al., 2019b).

Neural models have also recently been used in biomedical IR tasks, due to the availability of large datasets. Mohan et al. (2018) introduce a deep learning model to retrieve biomedical research literature. Further, deep neural architectures have been coupled with external knowledge bases (Zhao et al., 2019), where research documents are retrieved as part of a precision medicine task. In this body of work, the query is either an in-domain keyword, or structured information. As such, they cannot be readily used where the query may be expressed using complex linguistic structures found in the newswire. Example 1b in Table 1, for instance, stresses on a specific aspect of the claim using an adjectival clause as a modifier.

Given the success of BERT and its successors in natural language inference tasks, *ad hoc* IR systems have used them for claim verification (Hanselowski et al., 2018; Nie et al., 2019; Liu et al., 2019; Yang et al., 2019). Applications of such models to binary classification for query-based passage re-ranking suggest that contextual information can be valuable when re-ranking an initial list of possibly relevant documents retrieved by BM25 model (Nogueira and Cho, 2019). These approaches are not readily suitable for cross-genre IR, but they motivated some of our technical choices. For instance, our use of pointwise (instead of pairwise) loss was based on the discussion in Soleimani et al. (2020) regarding IR tasks with BERT-style models.

Fact-checking is a critical component in fighting misinformation, but medical misinformation is known to be nuanced. For example, instead of outright false claims, statements are known to undergo exaggeration. In this general context of thwarting medical misinformation, there is some notable work that, while being distinct from the IR task discussed here, complements our research. For instance, Sumner et al. (2014) studied the exaggeration of medical claims in the news vis-à-vis the original findings in research publications.

5 Conclusion

In contrast to recent research in ad hoc neural IR, which require large amounts of training data (Mitra and Craswell, 2018), we present a system that combines term-weighting techniques and neural models across two distinct linguistic genres. We also provide a novel dataset of medical newswire queries linked to research literature. Our results show that while neural models excel at re-ranking a small number of documents when pre-trained contextual embeddings are tuned on domain-specific data, classical token-based approaches remain difficult to beat in a cross-genre retrieval scenario when the search space is larger. Our data collection process also reveals that even in a domain as critically important as medical news, only a small fraction of news articles (24.6%) include a complete citation and a link to the original research. Thus, the presented task has utility in medical fact-checking, identifying health-related misinformation, and assessing some empirically verifiable aspects of health news reporting.

Acknowledgment

This work was supported in part by the Division of Social and Economic Sciences of the U.S. National Science Foundation (NSF) under the award SES-1834597.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proc. 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. ACL.

Matteo Catena, Ophir Frieder, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonellotto. 2019. Enhanced News Retrieval: Passages Lead the Way! In *Proc. 42nd International*

ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'19, pages 1269–1272, New York, NY, USA. ACM.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *SemEval-2017*, pages 1–14. ACL.

W C Chang, F Yu, Y W Chang, Y Yang, and S Kumar. 2020. Pre-training Tasks for Embedding-based Large-scale Retrieval. In *International Conference on Learning Representations*.

Zhuyun Dai and Jamie Callan. 2019. Deeper Text Understanding for IR with Contextual Neural Language Modeling. In *Proc. 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 985–988.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. ACL.

B. J Fogg, Gregory Cuellar, and David Danielson. 2009. Motivating, influencing, and persuading users: An introduction to captology. *The Human-Computer Interaction Handbook*, pages 109–122.

Susannah Fox. 2011. The Social Life of Health Information, 2011. Internet & American Life Project, Pew Research Center. Last accessed: May 31, 2020.

Susannah Fox and Maeve Duggan. 2013. Health Online 2013. Internet & Technology, Pew Research Center. Last accessed: May 31, 2020.

Andreas Hanselowski, Hao Zhang, Zile Li, Daniil Sorokin, Benjamin Schiller, Claudia Schulz, and Iryna Gurevych. 2018. UKP-athene: Multi-sentence textual entailment for claim verification. In *Proc. First Workshop on Fact Extraction and VERification (FEVER)*, pages 103–108, Brussels, Belgium. ACL.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692.

Yuanhua Lv and ChengXiang Zhai. 2011a. Lower-bounding term frequency normalization. In *Proc. 20th ACM International Conference on Information and Knowledge Management*, pages 7–16.

Yuanhua Lv and ChengXiang Zhai. 2011b. When documents are very long, BM25 fails! In *Proc. 34th international ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1103–1104.

Edward Ma. 2019. NLP Augmentation. https://github.com/makcedward/nlpaug. Last accessed: Oct 4, 2020.

S MacAvaney, A Yates, A Cohan, and N Goharian. 2019a. CEDR: Contextualized Embeddings for Document Ranking. In *Proc. 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1101–1104. ACM.

Sean MacAvaney, Andrew Yates, Kai Hui, and Ophir Frieder. 2019b. Content-Based Weak Supervision for Ad-Hoc Re-Ranking. In *Proc 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR'19, pages 993–996, New York, NY, USA. ACM.

S Medlock, S Eslami, M Askari, D L Arts, D Sent, S E de Rooij, and A Abu-Hanna. 2015. Health Information-Seeking Behavior of Seniors Who Use the Internet: A Survey. *J Med Internet Res*, 17(1):e10.

Bhaskar Mitra and Nick Craswell. 2018. An Introduction to Neural Information Retrieval. *Foundations and Trends in Information Retrieval*, 13(1):1–126.

Sunil Mohan, Nicolas Fiorini, Sun Kim, and Zhiyong Lu. 2018. A Fast Deep Learning Model for Textual Relevance in Biomedical Information Retrieval. In *Proc. 2018 World Wide Web Conference*, WWW '18, pages 77—-86, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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. *CoRR*, abs/1611.09268.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining Fact Extraction and Verification with Neural Semantic Matching Networks. In *Proc. AAAI Conference on Artificial Intelligence*, volume 33, pages 6859–6866.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. ACL.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

V Sanh, L Debut, J Chaumond, and T Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Amir Soleimani, Christof Monz, and Marcel Worring. 2020. BERT for Evidence Retrieval and Claim Verification. In *Advances in Information Retrieval*, pages 359–366, Cham. Springer International Publishing.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, Fred Boy, and Christopher D Chambers. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *BMJ*, 349.

John M Swales. 2000. Languages for Specific Purposes. *Annual Review of Applied Linguistics*, 20:59–76.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to BM25 and Language Models Examined. In *Proc. 2014 Australasian Document Computing Symposium*, pages 58–65.

Yanshan Wang, Naveed Afzal, Sunyang Fu, Liwei Wang, Feichen Shen, Majid Rastegar-Mojarad, and Hongfang Liu. 2018. MedSTS: a resource for clinical semantic textual similarity. *Language Resources and Evaluation*, pages 1–16.

T Wolf, L Debut, V Sanh, J Chaumond, C Delangue, A Moi, P Cistac, T Rault, R Louf, M Funtowicz, and J Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, abs/1910.03771.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, pages 5754–5764.

Z A Yilmaz, W Yang, H Zhang, and J Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proc. 2019 EMNLP-IJCNLP*, pages 3488–3494. ACL.

Sendong Zhao, Chang Su, Andrea Sboner, and Fei Wang. 2019. GRAPHENE: A Precise Biomedical Literature Retrieval Engine with Graph Augmented Deep Learning and External Knowledge Empowerment. In *Proc. 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 149–158, New York, NY, USA. ACM.

A Appendix

We present the ranking results on the development set here in Table 4, and the run-time of the experiments on the neural models in Table 5. In these tables, the notation is consistent with that used previously: BM25+† indicates that the text pre-processing steps described in Section 3 were included, BC-BERT denotes the Bio+Clinical model, and a model tuned on m positive (by augmentation) and n negative samples is indicated by the subscript (m, n).

	Development Set				
Model	**R@1**	**R@3**	**R@5**	**R@20**	**MRR**
BM25+†	0.370	0.472	0.523	0.633	0.444
BERT$_{(20,50)}$	0.582	0.724	0.762	0.829	0.665
XLNet$_{(20,50)}$	0.589	0.717	0.761	0.812	0.649
DistilBERT$_{(20,50)}$	0.357	0.548	0.625	0.784	0.480
BC-BERT$_{(0,1)}$	0.295	0.552	0.637	0.803	0.449
BC-BERT$_{(4,10)}$	0.564	0.716	0.762	0.830	0.654
BC-BERT$_{(20,50)}$	**0.649**	**0.756**	**0.786**	**0.833**	**0.713**

Table 4: The ranking results on the development set.

Model	**Train**	**Dev**	**Test**
BERT$_{(20,50)}$	2.15 hrs	2.23 hrs	2.57 hrs
XLNet$_{(20,50)}$	4.45 hrs	4.08 hrs	4.57 hrs
DistilBERT$_{(20,50)}$	1.24 hrs	1.55 hrs	1.58 hrs
BC-BERT$_{(0,1)}$	0.06 hrs	2.59 hrs	2.53 hrs
BC-BERT$_{(4,10)}$	0. 27 hrs	2.35 hrs	2.64 hrs
BC-BERT$_{(20,50)}$	2.35 hrs	2.56 hrs	2.59 hrs

Table 5: Run-time for the final ranking task on the training, test, and development sets.