Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4569–4576
Marseille, 20-25 June 2022

Evaluating Retrieval for Multi-domain Scientific Publications

Nancy Ide*, Keith Suderman, Jingxuan Tu*, Marc Verhagen*, Shanan E. Peters‡, Ian Ross‡, John Lawson*, Andrew Borg*, James Pustejovsky***
*Brandeis University, **Johns Hopkins University, ‡University of Wisconsin-Madison
*Waltham, Massachusetts, USA, **Baltimore Maryland, USA, ‡Madison, Wisconsin, USA
ide@cs.vassar.edu, suderman@jhu.edu, {jxtu, marc}@cs.brandeis.edu, peters@geology.wisc.edu,
iross@cs.wisc.edu, lawsonj@gmail.com, borgchampa@gmail.com, jamesp@cs.brandeis.edu

Abstract

This paper provides an overview of the xDD/LAPPS Grid framework and provides results of evaluating the AskMe retrieval engine using datasets included in the BEIR benchmark. Our primary goal is to determine a solid baseline of performance to guide further development of our retrieval capabilities. Beyond this, we aim to dig deeper to determine when and why certain approaches perform well (or badly) on both in-domain and out-of-domain data, an issue that has to date received relatively little attention.

Keywords: document retrieval, evaluation, retrieval benchmarks

1. Introduction

Recently, retrieval and mining of scientific publications has emerged as an area of growing interest, in part due to the need to rapidly and effectively exploit the increasingly large body of COVID-19 literature for information relevant to the current pandemic. Prior to the current crisis, scientists had already sought means to navigate the rapidly growing body of literature in domains beyond bioscience, which due to its exponentially increasing size has become unmanageable without automated assistance. The fundamental need is for efficient and accurate retrieval engines to locate articles relevant to a given topic, or which contain mentions of entities of interest to the researcher. Beyond this, scientists want to be able to extract specific data from massive publication databases that cover a wide range of areas in a way that is accurate, repeatable, and leads to the discovery of new and related information.

We are developing an end-to-end framework to ultimately accomplish this goal that builds on and integrates two major projects: (1) xDD[1] (formerly known as GeoDeepDive) (Peters et al., 2017), one of the world's largest and fastest-growing collections of scientific publications; and (2) the Language Applications (LAPPS) Grid[2] (Ide et al., 2014), which provides an extensive suite of fully interoperable natural language processing (NLP) modules. A critical component of this framework is the AskMe retrieval engine (Suderman et al., 2020)[3], which serves as a bridge between the xDD publication database and the LAPPS Grid suite of NLP tools.

AskMe's fundamental function is to rapidly search xDD's massive publication database in response to a user's query, and subsequently pass all or some of the results directly into the LAPPS Grid for processing by NLP tools and predefined workflows in a one-stop, seamless operation. To that end, AskMe must be suitably robust to handle massive datasets and, especially, flexible enough to provide good results over documents from a wide range of scientific domains. Automated means to evaluate AskMe's performance in the xDD environment is therefore a crucial necessity for its effective development. However, retrieval benchmarks using heterogeneous data are virtually nonexistent; existing benchmarks tend to focus on a particular task or domain (e.g., MultiReQA (Guo et al., 2020), KILT (Petroni et al., 2021)), TREC-COVID (Roberts et al., 2020)), and/or rely on relatively small corpora.

The recently developed BEIR retrieval benchmark (Thakur et al., 2021) seeks to address this gap by providing several retrieval datasets representing diverse retrieval tasks and domains, in order to evaluate the ability of a given approach to generalize across these variables. Although the focus is currently on evaluation of trained neural models, BEIR is currently the best existing benchmark dataset for evaluation of AskMe's performance.

This paper provides an overview of the xDD/LAPPS Grid framework and provides preliminary results of evaluating the AskMe retrieval engine using BEIR benchmark datasets. Our primary goal is to determine a solid baseline of performance to guide further development of our retrieval capabilities. Beyond this, we aim to ultimately dig deeper to determine when and why certain approaches perform well (or badly) on both in-domain and out-of-domain data, an issue that has to date received relatively little attention.

2. Background

xDD/COSMOS. xDD (eXtract Dark Data) is one of the world's largest single repositories of scientific publications that spans all domains of knowledge, automatically and continuously incorporates new documents,

[1] https://geodeepdive.org
[2] https://www.lappsgrid.org
[3] http://askme-ng-lapps.duckdns.org:8080/ask

and updates API endpoints every hour. xDD has accumulated millions of documents from multiple commercial and open-access publishers (at this writing, over 15M publications), and negotiations with additional commercial and open access publishers are ongoing. Users do not have direct access to original documents, but they do have full access to document metadata and several different levels of access to a variety of data products and services that incorporate full-text content, including snippets of text surrounding user-specified search strings and more restricted programmatic access to the output generated by various software tools, such as NLP. API access is provided by an ElasticSearch[4] index that is created over the full text of all documents; updates to the index occur hourly as new documents are acquired and enter the system (several thousand per day). UW-Madison's Center for High Throughput Computing (CHTC)[5] supplies the computational power for processing documents, which also allows for deploying new software tools quickly against all existing documents or subsets of documents defined by any number of different criteria, ranging from full-text content to reference metadata.

The xDD infrastructure is an integral part of the developing UW-COSMOS pipeline, which uses AI-methods to visually segment PDFs to locate and extract figures, tables, and equations, as well as body text elements, headers, and footers. The COSMOS system creates searchable Anserini[6] and ElasticSearch indexes over table, figure and equation objects and their associated text (i.e., captions and full-text content) to enable rich, context-based retrieval (e.g., the retrieval of figures that are pertinent to a given user-generated search string). COSMOS is currently deployed over multiple different subsets of documents in xDD, known as "sets." Sets are first-class data objects in xDD, meaning that any user query can be optionally restricted to a given subset of the 14.5M documents in xDD that are identified as potentially relevant to a given research problem. In addition to locating and extracting indexed visual objects from set documents, COSMOS generates multiple different embedding models using Word2Vec. COSMOS output and embedding models are updated episodically as set definitions are revised and as new documents are acquired and automatically identified by xDD as potentially relevant to a given set.

LAPPS Grid/Galaxy. The LAPPS Grid is a platform enabling complex NLP analyses while hiding the complexities associated with the underlying infrastructure. It provides seamless access to a wide range of NLP tools and resources and provides for their use interoperably, thereby eliminating the effort required to convert input and output formats to use a set of tools and/or resources together. In addition to a wide range of basic NLP tools, the Grid includes several machine learning tools such as deep learning (neural) modules from the Weka machine learning platform (Hall et al., 2009) and spaCy (Honnibal and Montani, 2017). The Grid is flexible and extensible, as tools and datasources are routinely added to the LAPPS Grid as required by various researchers and projects.

The LAPPS Grid uses the Galaxy framework[7] (Goecks et al., 2010) as its workflow and data management system. LAPPS Grid services are also directly accessible via API; several affiliated platforms provide access to LAPPS NLP tools via this channel. The LAPPS Grid also provides secure access to licensed data and software (e.g., resources stored at the Linguistics Data Consortium) as well as authentication via SAML2 where needed, which was developed in a collaboration with major EU CLARIN projects[8] (Hinrichs et al., 2018).

AskMe. The AskMe application is an open-source, ElasticSearch-based microservice architecture providing access to several publicly available scientific publication repositories. It is tightly coupled with the LAPPS Grid and, specifically, the Grid's Galaxy workflow engine, to enable processing of retrieval results using the Grid's suite of NLP tools. The service runs in a Docker Swarm on Jetstream (Stewart et al., 2015; Towns et al., 2014), an NSF-funded compute cluster. In response to a natural language query, AskMe applies several simple re-ranking algorithms to hone the results of ElasticSearch's BM25 similarity processing, as follows:

1. The total number of search terms that appear, including duplicates–e.g., if the search terms are *X Y* and the section contains *...X...Y...X...* (where periods represent tokens that are not search terms), the score is 0.2 (3/15).
2. The percentage of the search terms that appear. In the above example the score would be 1.0 since all search terms appear in the section.
3. The order of terms in the search query, e.g., if the search terms are *X Y* then *...X...Y...* will score higher than *...Y...X...*.
4. The total number of sentences that contain one or more search terms.
5. The percentage of search terms that appear in the first sentence.
6. Sequences of consecutive of search terms, e.g., if the search terms are *X Y* and the section contains *...X.YYX..XY* the score would be 0.4167 (5/12) as there are two sequences of consecutive terms (*YYX* and *XY*, with an aggregate length of 5) in the 12 tokens.

A section's score is the sum of the component algorithm scores; the overall document score is computed

[4]https://www.elastic.co

[5]http://chtc.cs.wisc.edu/

[6]https://github.com/castorini/anserini

[7]https://galaxyproject.org; funded by NSF and other national funding sources.

[8]Funded by the Andrew K. Mellon Foundation.

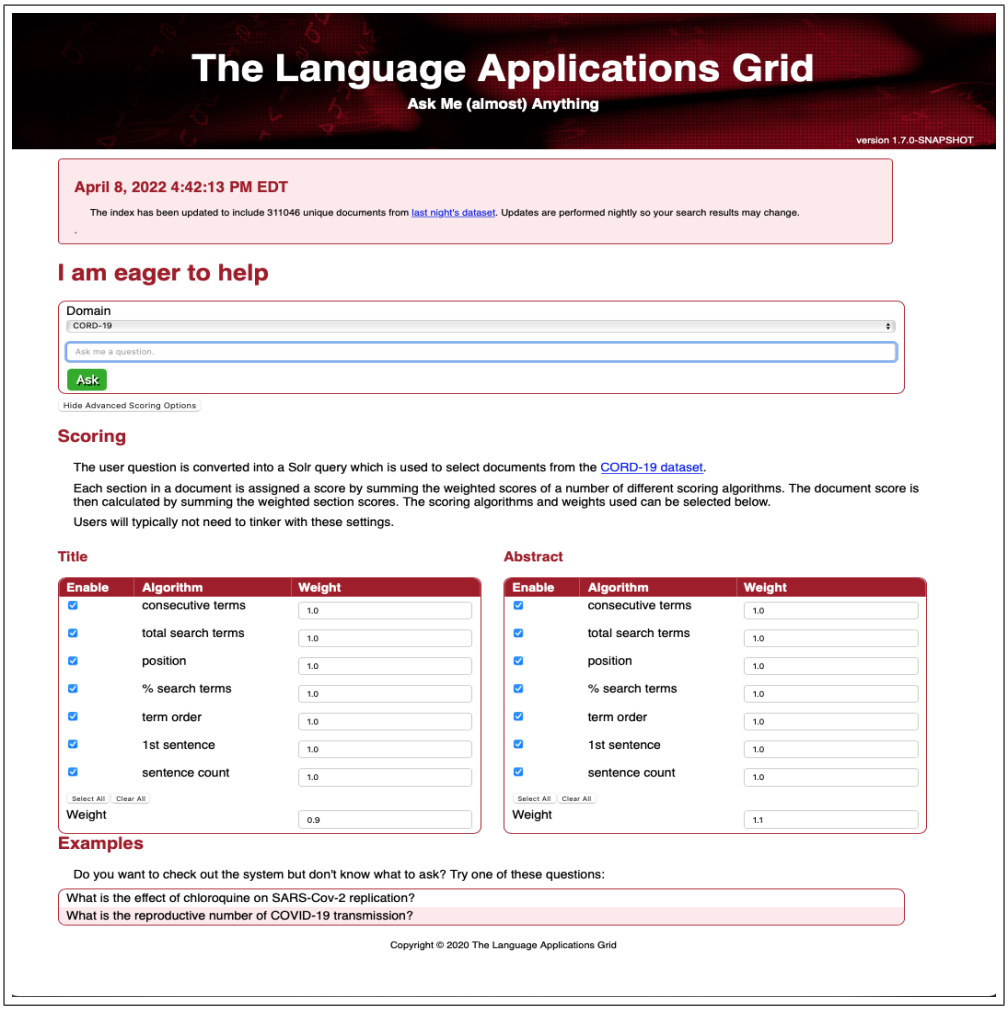

Figure 1: AskMe Query Interface

by adding the scores for the individual sections. The main AskMe page (Figure 1) includes options to adjust the weights for individual scoring algorithms over the title and abstract of a given document, to suit specific document types or interests or simply fine-tune results. Once generated, the user can directly import results into LAPPS/Galaxy for processing with a user-defined or pre-defined NLP workflow, for example to identify entities, relations, dependencies, etc. for inclusion in a knowledge base which can in turn be used to augment subsequent searches in xDD. A preliminary study compared AskMe's performance to a few commonly used, lexical-based Covid literature search engines (Suderman et al., 2020) indicated that AskMe produced results on a par with LitCovid and several other Covid publication search engines.[9]

xDD/LAPPS collaboration. With funding from the US National Science Foundation, xDD and LAPPS Grid developers have joined forces to develop and deploy an end-to-end supporting sophisticated search and retrieval from the xDD holdings, use and augmentation of facilities for advanced and well-established NLP and machine learning tools, and extracting and aggregating data from scientific publications. The infrastructure will incorporate state-of-the-art language models created from domain-specific documents in order to automate knowledge base construction and question answering with little supervision. The goal is to do this in bulk and for a daily stream of documents that are processed and read as they are acquired. The infrastructure will also include metadata repositories/indexes for all publications, including not only bibliographic information but also information such as the entities contained in each document, etc.

Figure 2 provides an example of flow through the xDD/LAPPS Grid framework. Documents returned to AskMe for a given user query are transduced for ingestion into LAPPS/Galaxy to extract information, which could then be added to a knowledge base and reused to enhance domain-specific retrieval, generate training data for machine learning, or feed into additional applications such as analytic tools, etc. The process is iterative, and the workflow can be run incrementally as

[9]Due to a lack of benchmark retrieval datasets for Covid publications at the time, the evaluation was performed manually.

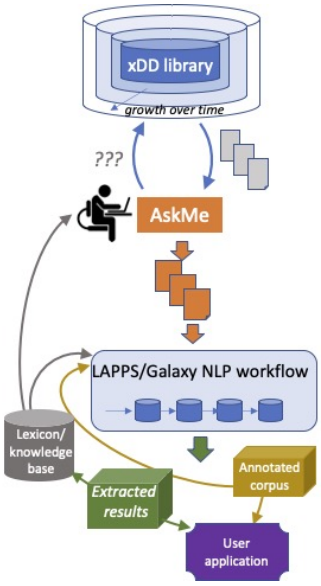

Figure 2: Schematic of a simple flow of processes in the xDD/LAPPS framework.

xDD grows and new relevant documents are ingested.

3. BEIR Benchmark

The BEIR benchmark includes eighteen retrieval datasets designed to evaluate the performance of trained neural models on out-of-domain texts and different retrieval tasks in a zero-shot setting. Pre-trained transformer models such as BERT (Devlin et al., 2019) and similar neural systems (Yates et al., 2021) have been shown to out-perform traditional traditional lexical approaches (e.g., TD-IDF and BM25 (Robertson and Zaragoza, 2009b)) for retrieval tasks such as question-answering (Karpukhin et al., 2020a) and passage retrieval (Nogueira and Cho, 2019a; Iyer et al., 2021). However, despite often being referred to as "open domain QA", prior evaluations typically address a given task or a limited domain and/or are performed over relatively small datasets. Also, there has been little attention paid to performance in *zero-shot* situations where training data is unavailable. BEIR addresses many of these issues by providing a diverse evaluation benchmark, including texts from a broad range of domains and created using a variety of annotation strategies, as well as involving different tasks with varying degrees of difficulty.

The BEIR datasets were used to evaluate ten diverse retrieval methods from five broad architectures: lexical, sparse, dense, late interaction, and re-ranking. Overall, the study found that no single approach consistently out-performs other approaches on all datasets. They noted a trade-off between performance and computational cost, finding that computationally expensive models such as re-ranking and late interaction returned the most accurate results. More efficient embedding-based approaches not only performed more poorly, but in many instances under-performed lexical models like BM25. For our purposes, the BEIR benchmark study provides evidence that traditional lexical retrieval approaches, and BM25 in particular, remain a strong baseline for zero-shot retrieval in settings that involve diverse domains and tasks.

AskMe currently relies on ElasticSearch's BM25 similarity metric augmented with simple heuristics to refine results. UW-Madison's facilities for high performance computing provide means to rapidly generate language models and periodically generating new models as xDD's collection grows, which can ultimately enhance AskMe's performance. However, it is not at all clear to what extent the addition of more computationally expensive approaches will improve performance, or, more importantly, which approaches would best serve the needs of xDD users. Therefore, to establish a solid baseline and starting point for further development, we need to evaluate AskMe's current performance over data that is sufficiently diverse and task-agnostic. BEIR is at present the best benchmark with which to perform such an evaluation.

4. Experiment

As an initial experiment, we evaluated AskMe's performance on the three biomedical datasets used in the BEIR evaluation: (1) TREC-COVID (Voorhees et al., 2021), containing CORD-19 scientific articles related to the COVID-19 pandemic; (2) NFCorpus (Boteva et al., 2016), which includes queries harvested from `NutritionFacts.org` (NF) that are linked to research articles on PubMed; and (3) BioASQ (Tsatsaronis et al., 2015), a ~15M PubMed document database used in a biomedical semantic questioning challenge. The remaining BEIR datasets are geared toward tasks such as tweet retrieval, argument retrieval, and duplicate question retrieval, and are therefore unsuitable for our purposes.

Following the BEIR study, we use Normalized Cumulative Discount Gain (nDCG@k, where k indicates the number of top hits that are included) as an evaluation metric (see (Wang et al., 2013)) and utilize the Python interface of the official TREC evaluation tool (Van Gysel and de Rijke, 2018) for our computations. Figure 4 shows comparative results for AskMe against other systems in the BEIR study. The results show that AskMe's lexical plus rule-based re-ranking approach out-performs all models apart from BM25+CE (Wang et al., 2020) and docT5query (Nogueira et al., 2019) on the TREC-COVID data, and out-performs all but BM25+CE on the NFCorpus data. For the BioASQ dataset, AskeMe out-performs all but COlBERT and BM25+CE. These results are notable, especially considering the greater resource demands of models such as BM25+CE and ColBERT.

In the full BEIR study involving 18 datasets, BM25+CE out-performed all others on all but two of the additional 16 datasets, but BM25 generally out-performed most of the neural approaches, thus under-

Dataset	#Queries	#Documents	Avg. D/Q	Avg. Word Length	
				Query	Document
TREC-COVID	50	171,332	493.5	10.60	160.77
BioASQ	500	14,914,602	4.7	8.05	202.61
NFCorpus	323	3,633	38.2	3.30	232.26

Figure 3: Basic statistics for the datasets used in this study, based on BEIR test results (see (Thakur et al., 2021), Table 1, for details for all BEIR datasets). The TREC-COVID and NFCorpus datasets contain 3-level relevancy judgements, and BioASQ uses binary (relevant, not relevant) judgements. Avg. D/Q indicates average relevant documents per query.

Dataset	AskMe	BM25	DCT	SP	dT5q	DPR	ANCE	TAS	GenQ	CB	BM+CE
TREC-COVID	**0.685**	0.656	0.406	0.538	**0.713**	0.332	0.654	0.481	0.619	0.677	**0.757**
BioASQ	**0.469**	0.465	0.407	0.351	0.431	0.127	0.306	0.383	0.398	**0.474**	**0.523**
NFCorpus	**0.339**	0.325	0.283	0.301	0.328	0.189	0.237	0.319	0.319	0.305	**0.350**

Figure 4: Performance on three BEIR benchmark datasets for AskMe and the other models used in the BEIR study. Askme's scores in red; scores in blue out-perform Askme. Scores are computed with Normalized Cumulative Discount Gain applied to the top 10 hits. Name expansions and citations for each model are shown in Figure 5.

lining the inability of more complex approaches to adequately generalize to out-of-domain data.[10]

5. Discussion and Future Work

The most interesting result of our preliminary study is that AskMe, which augments BM25 with simple, manually-produced re-ranking heuristics, seemingly performs at a level comparable to many far more complex approaches. Neural re-ranking approaches, such as those relying on the cross-attentional mechanisms in BERT (Nogueira and Cho, 2019b) are the best performers to date, but these approaches have notoriously high computational overhead (see, e.g., (Reimers and Gurevych, 2019)) and are thus impractical for use in many scenarios. For example, in the BEIR study, ColBERT required about 900GB to store the 15M document BioASQ index, which is comparable in size to the xDD data, whereas AskMe requires around 20GB. While further investigation is required, our experiment suggests that reasonable results can be produced with low-expense approaches, which is an important finding for systems that operate in real-time, possibly low-resource situations.

The xDD system covers a broad range of domains and its uses run across the gamut of potential tasks and expected results. Also, xDD's already enormous body of texts is constantly growing. Given this relatively unique set of circumstances, it is necessary to carefully explore our options for efficient and effective document retrieval. In particular, we need to explore retrieval performance in a wide range of scenarios that xDD users could require, which may involve retrieval across very broad domain data or data from a highly specific domain; involve different tasks (e.g., question answering, passage retrieval, but also retrieval of potential data for text mining); and/or demand different degrees of recall vs. precision. Therefore, in addition to comparing AskMe's performance over benchmark data to other approaches, we want to dig deeper into when and why certain approaches perform well (or badly), and how we can use that information to advantage. Despite the flurry of contemporary research efforts focused on retrieval, very little attention has been given to exploring the strengths and weaknesses of different retrieval approaches in a diverse setting.

Our initial study verifies the BEIR results, which show that relatively straightforward lexical approaches perform as well in general as more complex neural models when applied to out-of-domain data. This observation suggests several avenues for continued study and experimentation, and raises the question of whether any approach can perform consistently well over diverse data, or if it is necessary to rely on domain-specific models to achieve highly accurate results.

BM25+CE's performance compared to that of other neural approaches suggests a path forward for retrieval from multi-domain data. Answer re-ranking has been shown to address the problem of high recall vs. low accuracy among the top results of many QA models (Wang et al., 2018a; Wang et al., 2018b) and remains a promising avenue for enhancing the precision of retrieval results. Another interesting avenue for further study is to evaluate BM25+CE against more complex approaches that utilize re-ranking, e.g., recent work such as (Iyer et al., 2021), to assess the necessity of building sophisticated neural models. At the same time, we need to determine when and where the advantages of re-ranking outweigh its high computational costs, and where simple lexical approaches may suffice.

Another avenue we are exploring is the application of semantic visualization tools and environments over the AskMe search results. For example, (Tu et al., 2021)

[10] See (Thakur et al., 2021) for results of the full BEIR study.

Model Type	Abbr	Name	Reference
Lexical	BM25	BM25	(Robertson and Zaragoza, 2009a)
Sparse	DCT	DeepCT	(Dai and Callan, 2020)
	SP	SPARTA	(Zhao et al., 2021)
	dT5q	docT5query	(Nogueira et al., 2019)
Dense	DPR	DPR	(Karpukhin et al., 2020b)
	ANCE	ANCE	(Xiong et al., 2020)
	TAS	TAS-B	(Hofstätter et al., 2021)
	GenQ	GenQ	Model trained on synthetically generated data
Late Interaction	CB	ColBERT	(Khattab and Zaharia, 2020)
Re-ranking	BM+CE	BM25+CE	(Wang et al., 2020)

Figure 5: Models used in the BEIR study, by type. The *Abbr* column provides the short name used in Figure 4 for each model.

has been successfully deployed over the COVID-19 Literature (Wang et al., 2021). Semantic visualization techniques, a set of text processing and visualization techniques and tools for enhanced knowledge exploration, can enhance scientific discovery over complex corpora, by transforming large datasets of complex networks into rich semantic-aware text data; processes text data in a hierarchical manner; and provides visualizations for the indexed data.

6. Conclusion

We describe a framework under development by teams at University of Wisconsin-Madison and Brandeis University to provide means to explore and exploit a huge repository of scientific publications using state-of-the-art NLP technology. A fundamental capability within this framework is efficient and effective retrieval and, potentially, the ability to adapt to specific retrieval scenarios. Our initial experiment shows that AskMe in its current instantiation performs competitively with other approaches.

We continue our work on evaluating AskMe; our work so far has revealed that, for evaluation of retrieval in a real-world, multi-domain setting, existing benchmarks are largely inadequate. The BEIR benchmark is, to our knowledge, the only effort so far that begins to address a need for more comprehensive evaluation of current methods, and we hope to see this and similar efforts expanded in the future. More importantly, our need for robust and flexible retrieval has opened up a wide spectrum of questions concerning the scalability, adaptability, and in particular the relative strengths and weaknesses of various approaches, both lexical and neural. We feel that this is an area ripe for further investigation that can guide our enhancement of retrieval from xDD.

7. Acknowledgements

This work is supported by US National Science Foundation grant 2104025.

8. Bibliographical References

Boteva, V., Ghalandari, D. G., Sokolov, A., and Riezler, S. (2016). A full-text learning to rank dataset for medical information retrieval. In Nicola Ferro, et al., editors, *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy, March 20-23, 2016. Proceedings*, volume 9626 of *Lecture Notes in Computer Science*, pages 716–722. Springer.

Dai, Z. and Callan, J., (2020). *Context-Aware Term Weighting For First Stage Passage Retrieval*, page 1533–1536. Association for Computing Machinery, New York, NY, USA.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). 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)*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Goecks, J., Nekrutenko, A., and Taylor, J. (2010). Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11:R86.

Guo, M., Yang, Y., Cer, D., Shen, Q., and Constant, N. (2020). Multireqa: A cross-domain evaluation for retrieval question answering models.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Hinrichs, E., Ide, N., Pustejovsky, J., Hajič, J., Hinrichs, M., Elahi, M. F., Suderman, K., Verhagen, M., Rim, K., Straňák, P., and Mišutka, J. (2018). Bridging the LAPPS Grid and CLARIN. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Hofstätter, S., Lin, S., Yang, J., Lin, J., and Hanbury, A. (2021). Efficiently teaching an effective dense retriever with balanced topic aware sampling. *CoRR*,

abs/2104.06967.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Ide, N., Pustejovsky, J., Cieri, C., Nyberg, E., Wang, D., Suderman, K., Verhagen, M., and Wright, J. (2014). The Language Applications Grid. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Iyer, S., Min, S., Mehdad, Y., and Yih, W. (2021). RECONSIDER: improved re-ranking using span-focused cross-attention for open domain question answering. In *NAACL-HLT*, pages 1280–1287. Association for Computational Linguistics.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020a). 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, November. Association for Computational Linguistics.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. (2020b). 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, November. Association for Computational Linguistics.

Khattab, O. and Zaharia, M., (2020). *ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT*, page 39–48. Association for Computing Machinery, New York, NY, USA.

Nogueira, R. and Cho, K. (2019a). Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Nogueira, R. and Cho, K. (2019b). Passage re-ranking with BERT. *CoRR*, abs/1901.04085.

Nogueira, R., Lin, J., , and Epistemic, A. (2019). From doc2query to docTTTTTquery. Online preprint.

Peters, S., Ross, I., Czaplewski, J., Glassel, A., Husson, J., Syverson, V., Zaffos, A., and Livny, M. (2017). A new tool for deep-down data mining. *EOS*, 98.

Petroni, F., Piktus, A., Fan, A., Lewis, P., Yazdani, M., De Cao, N., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. (2021). KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online, June. Association for Computational Linguistics.

Reimers, N. and Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Roberts, K., Alam, T., Bedrick, S., Demner-Fushman, D., Lo, K., Soboroff, I., Voorhees, E., Wang, L. L., and Hersh, W. R. (2020). TREC-COVID: rationale and structure of an information retrieval shared task for COVID-19. *Journal of the American Medical Informatics Association*, 27(9):1431–1436, 07.

Robertson, S. and Zaragoza, H. (2009a).

Robertson, S. and Zaragoza, H. (2009b). The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Stewart, C. A., Cockerill, T., Foster, I., Merchant, N. C., Skidmore, E., Taylor, J., Tuecke, S., Turner, G., Vaughn, M., , and Gaffney, N. (2015). Jetstream–a self-provisioned, scalable science and engineering cloud environment-nsf acceptance report. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*.

Suderman, K., Ide, N., Verhagen, M., Cochran, B., and Pustejovsky, J. (2020). AskMe: A LAPPS Grid-based NLP Query and Retrieval System for Covid-19 Literature. In *NLP COVID-19 Workshop 2020*. Association for Computational Linguistics.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. (2021). BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Towns, J., Cockerill, T., Dahan, M., Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V., Lathrop, S., Lifka, D., Peterson, G. D., Roskies, R., Scott, J. R., and Wilkins-Diehr, N. (2014). Xsede: Accelerating scientific discovery. *Computing in Science Engineering*, 16(5):62–74.

Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M. R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga, A., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., and Paliouras, G. (2015). An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138.

Tu, J., Verhagen, M., Cochran, B., and Pustejovsky, J. (2021). Exploration and discovery of the covid-19 literature through semantic visualization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 76–87.

Van Gysel, C. and de Rijke, M. (2018). Pytrec-eval: An extremely fast python interface to trec-eval. *The 41st International ACM SIGIR Conference on Research and Development in Information Retrieval*, June.

Voorhees, E., Alam, T., Bedrick, S., Demner-Fushman, D., Hersh, W. R., Lo, K., Roberts, K., Soboroff, I., and Wang, L. L. (2021). Trec-covid: Constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1), feb.

Wang, Y., Wang, L., Li, Y., He, D., and Liu, T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. In Shai Shalev-Shwartz et al., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 25–54, Princeton, NJ, USA, June. PMLR.

Wang, S., Yu, M., Jiang, J., Zhang, W., Guo, X., Chang, S., Wang, Z., Klinger, T., Tesauro, G., and Campbell, M. (2018a). Evidence aggregation for answer re-ranking in open-domain question answering. In *ICLR (Poster)*. OpenReview.net.

Wang, Y., Liu, K., Liu, J., He, W., Lyu, Y., Wu, H., Li, S., and Wang, H. (2018b). Multi-passage machine reading comprehension with cross-passage answer verification. In *ACL (1)*, pages 1918–1927. Association for Computational Linguistics.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *CoRR*, abs/2002.10957.

Wang, Q., Li, M., Wang, X., Parulian, N., Han, G., Ma, J., Tu, J., Lin, Y., Zhang, R. H., Liu, W., and Aabhas Chauhan, Yingjun Guan, B. L. R. L. X. S. Y. F. H. J. J. H. S.-F. C. J. P. J. R. D. L. A. E. M. P. C. V. C. S. B. O. (2021). Covid-19 literature knowledge graph construction and drug repurposing report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 66–77.

Xiong, L., Xiong, C., Li, Y., Tang, K., Liu, J., Bennett, P. N., Ahmed, J., and Overwijk, A. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval. *CoRR*, abs/2007.00808.

Yates, A., Nogueira, R., and Lin, J., (2021). *Pretrained Transformers for Text Ranking: BERT and Beyond*, page 1154–1156. Association for Computing Machinery, New York, NY, USA.

Zhao, T., Lu, X., and Lee, K. (2021). SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online, June. Association for Computational Linguistics.