R3 : Refined Retriever-Reader Pipeline for Multidoc2dial

Srijan Bansal* **Sumit Agarwal*** **Suraj Tripathi*** **Sireesh Gururaja***
Aditya Veerubhotla* **Ritam Dutt** **Teruko Mitamura** **Eric Nyberg**
{srijanb, sumita, surajt, sgururaj}@andrew.cmu.edu
{adityasv, rdutt, teruko}@andrew.cmu.edu, ehn@cs.cmu.edu
Language Technologies Institute, Carnegie Mellon University

Abstract

In this paper, we present our submission to the DialDoc shared task based on the MultiDoc2Dial dataset. MultiDoc2Dial is a conversational question answering dataset that grounds dialogues in multiple documents. The task involves grounding a user's query in a document followed by generating an appropriate response. We propose several improvements over the baseline's retriever-reader architecture to aid in modeling goal-oriented dialogues grounded in multiple documents. Our proposed approach employs sparse representations for passage retrieval, a passage re-ranker, the fusion-in-decoder architecture for generation, and a curriculum learning training paradigm. Our approach shows a 12 point improvement in BLEU score compared to the baseline RAG model.

1 Introduction

The task framework of document-grounded, conversational question answering unifies several closely related task frameworks, including open-domain question answering (QA), conversational QA, and knowledge-grounded generation. In open-domain question answering tasks, such as SQuAD (Rajpurkar et al., 2018), models are required to respond to a question with knowledge that may be located within a potentially large collection of documents. For conversational QA tasks like QuAC (Choi et al., 2018b), the queries posed to the model take the form of a dialogue, where previous dialogue turns contain necessary context to answer the current turn's question. Both of these task frameworks can be framed as either extractive QA or abstractive QA. Document-grounded conversational question answering tasks like CoQA (Reddy et al., 2019a) and Doc2Dial (Feng et al., 2020b) combine the above two frameworks. This setting requires models to understand user queries and their associated dialogue context, use them to find relevant grounding documents, and then generate coherent responses to user queries. This pipe-lined architecture forms the backbone of the baseline model, henceforth called retriever-reader.

In this paper, we present our approach to the MultiDoc2Dial (MDD) task (Feng et al., 2021), the successor to Doc2Dial, which complicates the Doc2Dial setting by constructing dialogues that are grounded in multiple documents. Each document is segmented into multiple passages, and thus document and passage are interchangeably used in this paper. As a result, models must retrieve the documents relevant to the current dialogue turn. These grounding documents could potentially be different from those grounded in previous dialogue turns.

We propose a model that improves over the baseline model by focusing on each component of its retriever-reader architecture. Firstly, we introduce sparse lexical representations in the retriever for matching, as outlined in Formal et al. (2021). Secondly, we rerank the retriever's results using techniques from Fajcik et al. (2021). Furthermore, we update the decoding process to incorporate the fusion-in-decoder (FiD) technique (Izacard and Grave, 2021). Finally, we use curriculum learning to train our models. We observe an improvement of 11.9 (BLEU) and 9.5 (F1) points on the validation; and an improvement of 9.5 (BLEU) and 10.3 (F1) points on test set in the MDD-SEEN setting compared to the baseline RAG model. We achieve 18.7 and 13.7 points improvement in the BLEU and F1 metric respectively on the MDD-UNSEEN test set compared to the baseline RAG model. Our submission (CMU-QA) stands 2^{nd} and 3^{rd} on the unseen and seen leaderboards [1] respectively.

*Equal contribution

[1] https://eval.ai/web/challenges/challenge-page/1437/leaderboard/3577

Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering, pages 148 - 154
May 26, 2022 ©2022 Association for Computational Linguistics

2 Related Works

The MultiDoc2Dial setting draws on related tasks like open-domain QA and conversational QA. Consequently, we investigate techniques that have shown success on those tasks. Conversational QA tasks, which typically assume that the grounding document is provided, use transformer-based architectures; the leading submissions to the QuAC and CoQA leaderboards use RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020), respectively. Retriever-reader architectures such as RAG (Lewis et al., 2020b) have become a popular choice for open-domain QA, increasingly using dense retrieval methods such as DPR (Karpukhin et al., 2020). We study works related to four areas for modeling improvement: retrieval, reranking, reader, and training.

Retriever : As the MultiDoc2Dial task is formulated in an open-domain setting, it requires the retrieval of relevant sources (passages) from a large pool of documents for generating the right output. Hence, we investigate the strides in information retrieval in recent years.

Recently, dense retrieval based approaches have shown competitive performance (Karpukhin et al., 2020; Xiong et al., 2020; Hofstätter et al., 2021) while also scaling to large corpora, like MS-MARCO dataset (Nguyen et al., 2016). They use a nearest neighbor index, such as FAISS (Johnson et al., 2019) to ensure scalability. Dense retrieval techniques aims to encode the query and passage into a shared semantic space where the relevance of a passage for a query can be computed by the inner product of their representations.

In contrast, sparse retrieval techniques perform exact token-level matching in the vocabulary space. There has been a growing interest in this field, with many advances achieving state-of-the-art results (Dai and Callan, 2020; Bai et al., 2020; Gao et al., 2021; Formal et al., 2021; MacAvaney et al., 2020). These models are advantageous due to their interpretable representations, efficient lookup, highly scalable inverted-list indexing, and excellent performance in exact term-based matching scenarios. Like dense retrieval based approaches, matches are computed via the dot product of the query and passage representations.

Reranking : While both dense and sparse retrieval methods have shown good progress, they must still embed the query and passage separately, because computing a match score between a query and every passage is computationally infeasible. As a compromise, re-ranking methods such as those in Fajcik et al. (2021) train a re-ranking module that can jointly embed the query and retrieved passages. Because the set of retrieved passages is significantly smaller than the whole corpus, re-ranking methods can model more complex relationships between the query and retrieved passages, and significantly boost retrieval performance.

Reader : Encoder-decoder based abstractive readers have been widely used in QA tasks. RAG (Lewis et al., 2020b) uses the BART-large model (Lewis et al., 2020a) which is pre-trained using a denoising objective and a variety of different noising functions. Moreover, RAG marginalizes output from each (query, passage) pair based on retrieval scores. It has obtained state-of-the-art results on a diverse set of generation tasks and outperforms comparably-sized T5 models.

Fusion in Decoder (FiD) (Izacard and Grave, 2021) performs well in extractive-based QA tasks like Natural Questions (Kwiatkowski et al., 2019). Unlike RAG model, the independent processing of the passages on the encoder side allows the FiD model to scale to a large number of passages, while the fusion in the decoder effectively combines evidence from multiple passages.

Training : Works such as Xu et al. (2020) have shown that fine-tuning a transformer model on examples, ordered on the basis of their difficulty, results in significant performance gains across different tasks. Kim et al. (2021) show more specifically that this type of curriculum design generalizes well to the document-grounded QA setting.

3 Task Description

MultiDoc2Dial is a conversational QA task that requires generating responses to user queries. In contrast to tasks like its predecessor, Doc2Dial (Feng et al., 2020a), and related tasks like QuAC (Choi et al., 2018a), ShARC (Saeidi et al.), and CoQA (Reddy et al., 2019b), which assume that the grounding document for the dialogue is given, MultiDoc2Dial constructs dialogues that are grounded in multiple documents. Each dialogue is constructed from a number of segments. Different segments are grounded in different documents; while all the dialogue turns within a segment are grounded in a single document. The dataset additionally marks the specific passage that is relevant to the current dialogue turn. However, the tran-

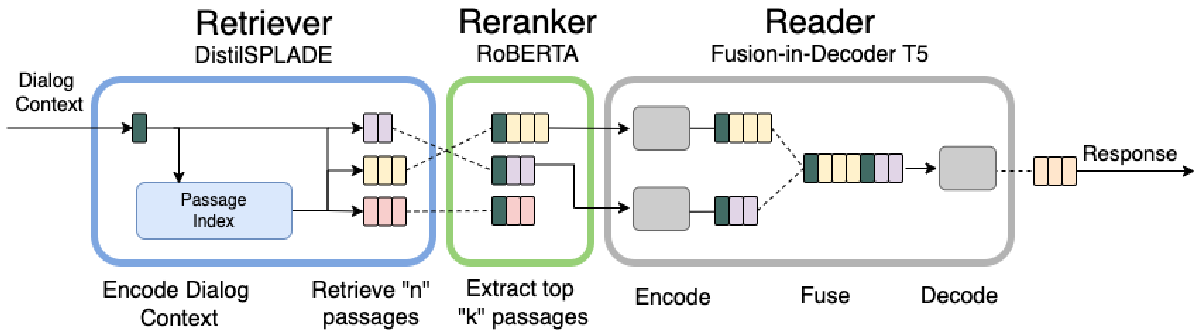

Figure 1: The proposed system architecture uses a bi-encoder (DistilSPLADE) retriever which fetches the top 100 relevant passages from the passage index, followed by a RoBERTA-based cross-encoder for reranking. The top 10 passages are passed to FiD with T5 to output the final response. This model is also used to perform curriculum learning as discussed in Section 4.

sitions between segments, which we refer to as topic shifts, are not marked. As a result, models are required not only to determine the grounding passages for each turn but also to determine which parts of the dialogue context continue to be relevant across topic shifts. The original dataset also presents several distinct domains of grounding documents from different public-facing websites that exhibit different writing styles. Each dialogue is grounded in documents drawn from the same domain.

The shared task defines two settings: one where all of the dialogues are grounded in documents from domains seen during training (MDD-SEEN), and another where the grounding domain is unseen (MDD-UNSEEN). Due to the open domain evidence retrieval and natural language response generation setting of the task, it lends itself well to a retriever-reader architecture. Broadly speaking, the MultiDoc2Dial task can be broken down into two distinct subtasks. Models must first retrieve the correct grounding passage from the provided corpora. They must then use the retrieved passages to generate a response to the user query in the most recent dialogue turn. While the MultiDoc2Dial paper defines both retrieval and a generation task, the shared task only evaluates reader outputs. Models are evaluated on the sum of different metrics: F1, BLEU (as implemented in Post 2018), METEOR (Banerjee and Lavie, 2005), and RougeL (Lin, 2004).

4 Methodology

For this task, we employ the standard retriever-reader architecture used in open-domain question answering. The model takes the user's current turn and dialogue context (previous turns) as the query. The query is then passed to the retriever which selects the top-n passages which are further passed to a reranker. The top-k (out of top-n) reranked passages are then fed to the reader along with the query to finally generate the agent's response.

In our experiments, we use DistilSPLADE (Formal et al., 2021) as our retriever, which augments the query and passages, subsequently projecting them to a sparse vector in the vocabulary space. Each coordinate in the projected vector represents the semantic importance of a term (also called "term impact" (Mallia et al., 2021)) for matching. The inputs are augmented by applying a sparsity-inducing activation function on the logits of a Masked Language Model such as BERT (Devlin et al., 2019), which selects the important words present in the passage and adds additional expansion to combat the vocabulary mismatch problem. The sparsity of the activation is complemented with the FLOPS regularizer (Paria et al., 2020) which minimizes the expected floating point operations required to perform matching. In addition to the training data provided in MS-MARCO (Campos et al., 2016), the model is trained using the pseudo-labels from a more expressive cross-encoder model, which improves the performance of the SPLADE model. This technique has shown state-of-the-art performance across several datasets and obtained the highest performance in our experiments.

The passages retrieved by the bi-encoder based retrieval are then passed through a RoBERTA (Liu et al., 2019) based cross-encoder. The RoBERTA model is trained to output a score that denotes the relevance of a passage to the given query. Due to the cross-attention between the query and the passage, the reranking proved to be effective by pulling up golden passages in the top-k documents that are passed on to the reader.

An abstractive reader is used to generate agent responses. We use a T5 based fusion in decoder (FiD) model which encodes all the top-k reranked passages one-by-one and concatenates them to form the input to the decoder. The decoder then learns to collect evidence from multiple passages to generate the response.

We also experiment with training our model using a curriculum learning approach originally proposed by Xu et al. (2020) and then implemented on Doc2Dial by Kim et al. (2021). To do so, we divide our training data randomly into 4 buckets, and train a teacher model on each bucket using FiD-T5. We then calculate each teacher model's performance (BLEU, RougeL and METEOR scores) on the other 3 buckets, which the teacher model has not seen during training. The training instances are then partitioned into "easy", "medium", and "hard" examples based on the scores chosen in Kim et al. (2021). We train in four phases, and each phase is trained until convergence. In the first phase, we train on a third of the easy examples; in the second, on a disjoint third of the easy examples, and a third of the medium examples; in the third phase, a disjoint third of all of the three partitions, and in the final phase, we train on the entire training set.

5 Experiments

Dataset : The MultiDoc2Dial dataset consists of 4796 dialogues, consisting 29,748 query turns and grounded in 4283 passages across 4 domains (Social Security Administration, Veteran Affairs, Student-Aid, and DMV). MDD-UNSEEN test corpus used in shared task is based on COVID domain.

5.1 Baseline

The proposed baseline for the MultiDoc2Dial shared task comprises a retrieval-augmented generator (RAG) model (Lewis et al., 2020b). The model uses a fine-tuned dense passage retrieval (DPR) model (Karpukhin et al., 2020) to find relevant passages and a pretrained sequence-to-sequence BART (Lewis et al., 2020a) to generate the response by marginalizing it according to document scores.

We use structure-based segmentation, with the original and reranking original scoring functions. We use DPR encoder finetuned on MultiDoc2Dial for retrieval, and a pretrained BART-large model.

5.2 Setup

Our experimental setup refines both the retriever and reader components of the existing architecture.

Retrieval We analyze the performance of different dense and sparse retrieval methods in a zero-shot setting on the MultiDoc2Dial dataset. For our dense retriever baselines, we conduct experiments with DPR, ANCE (Xiong et al., 2020) and TAS-B (Hofstätter et al., 2021). For sparse retrieval methods, we experiment with SPLADE-max and DistilSPLDAE (Formal et al., 2021). During training, we label the retrieved passages (excluding the golden passage) from BM25 as hard negatives. We also experiment with the finetuned DPR model to mine harder negatives.

Reranker Following (Fajcik et al., 2021), we select the top 100 passages from the DistilSPLADE retriever to be reranked using RoBERTA as a cross encoder. We use this reranking only during validation time. The top 10 reranked documents are passed to the reader.

Reader We experiment with both T5 and BART models as the reader. We use the T5 based reader model to circumvent the limited tokens used for BART along with the FiD model pretrained on natural questions [2]. We further experimented by placing the golden passage at the top-most position (Gold setting) in the retrieved passages before passing it to the reader during training. We also apply curriculum learning (CL) in the reader as per described in Section 4.

6 Results & Discussion

Table 1 shows our model's performance on the validation split. Applying DistilSPLADE as the retriever with FiD + T5 as the reader we saw a 10 point improvement in BLEU compared to the baseline. Reranking (RR) the retrieval outputs leads to further increase in the overall metrics. Additionally, curriculum learning (CL) boosts the model's performance. Setting M1 shows a BLEU score that is 1 point higher than the "DistillSplade + Fid + RR" model. We use the M1 setting for evaluation on the Test SEEN dataset. For the Gold setting, we saw a decrease in metrics for the RR and RR + CL settings.

6.1 Retrieval improvement

We present the results for different retriever configurations at Recall@10 and Recall@100 in Table

[2] https://github.com/facebookresearch/FiD

Model	Reader	EM	F1	BLEU	RougeL
Baseline	BART	3.6	33.8	19.2	31.4
DistilSPLADE + RAG	BART	4.8	38.5	23.7	36.2
DistilSPLADE + FiD	T5	5.1	42.3	29.7	40.2
DistilSPLADE + FiD + RR	T5	5.5	43.1	30.1	41.1
DistilSPLADE + FiD + RR + CL (**M1**)	T5	5.3	**43.3**	**31.1**	**41.4**
DistilSPLADE + FiD + Gold	T5	5.3	42.4	30.5	40.6
DistilSPLADE + FiD + Gold + RR	T5	5.5	42.5	30.4	40.7
DistilSPLADE + FiD + Gold + RR + CL (**M2**)	T5	**5.6**	43.0	30.5	41.0
M1 (on Shared Task MDD-SEEN test)	T5	-	46.2	31.8	44.2
M2 (on Shared Task MDD-UNSEEN test)	T5	-	33.0	25.0	32.0

Table 1: Model performance on the validation split for EM, F1, BLEU and RougeL. We see a consistent improvement across all metrics with DistilSPLADE as the retriever and FiD as the reader. Gold means the ground-truth passage was passed during training. Reranking (RR) and curriculum learning (CL) further boost performance on all metrics.

Model	R@10	R@100
DPR-PT	33.9	69.4
ANCE-PT	53.8	80.7
TAS-B-PT	53.9	85.0
SPLADE-max-PT	58.5	85.9
DistilSPLADE-PT	61.6	86.9
DPR-FT (Baseline)	73.2	92.8
SPLADE-max-FT	75.1	93.9
DistilSPLADE-FT	77.0	94.8
DistilSPLADE-FT+DPR-FT(Neg)	**78.6**	**94.9**
DistilSPLADE-FT+DPR-FT(Neg) + Reranker	**85.7**	**94.9**

Table 2: Performance of the retriever for different model configurations at Recall@10 and Recall@100. X-PT refers to the pretrained X model while X-FT implies that X was finetuned on MultiDoc2dial. DPR-FT was the retriever employed for the MultiDoc2Dial baseline.

2. It is evident that the pretrained sparse retrieval frameworks, Splade and DistilSPLADE, achieve better retrieval performance in comparison to the pretrained DPR model. This suggests that the exact matching over keywords and over the paraphrases generated for functional words achieves good retrieval performance. Unsurprisingly, the performance for all models improve significantly when they are fine-tuned on Multidoc2Dial dataset, with the sparse-retrievers still outperforming DPR. The performance shows a further boost when we use the fine-tuned DPR model to mine hard-negatives.

Reranking the validation passages increases the R@10 to 85% (Ref Table 2). This further leads to improvements in metrics in both the normal and the Gold setting.

6.2 Reader improvement

Our analysis, in Table 1, indicates that the FiD (T5-based) model outperforms the current BART-based baseline model on all the evaluation metrics. We observed an improvement of around 10 points in BLEU score in the FiD setting compared to the RAG model. FiD extracts relevant evidence from concatenated passages disregarding their retrieval scores, unlike RAG which uses them for marginalization. Reinforcing signals from the retriever for the reader component might be the cause of the dip in performance of RAG compared to FiD. We also observed that increasing the number of input tokens to the reader model helps capture dialogue and passage context relevant to the input query.

7 Conclusion

We introduced our submission (CMU-QA) for the Multidoc2Dial shared task. Our approach (R3) focuses on improving the overall retriever-reader pipeline using the sparse retriever DistilSPLADE and Fusion-in-decoder (FID) as the reader. We use a cross-attention based reranker to further boost recall scores. We refine the training process through curriculum learning to handle the diverse complexity of this dataset. For future work, we plan to improve results through better dialogue modelling and reducing noise or irrelevant information in the passages by taking top text spans. Further, we will aim to select the best of all candidate responses using a response re-ranker.

References

Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. Sparterm: Learning term-based sparse representation for fast text retrieval. *ArXiv*, abs/2010.00768.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. Ms marco: A human generated machine reading comprehension dataset. *ArXiv*, abs/1611.09268.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018a. Quac: Question answering in context. In *EMNLP*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018b. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Zhuyun Dai and Jamie Callan. 2020. *Context-Aware Document Term Weighting for Ad-Hoc Search*, page 1897–1907. Association for Computing Machinery, New York, NY, USA.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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.

Martin Fajcik, Martin Docekal, Karel Ondrej, and Pavel Smrz. 2021. R2-d2: A modular baseline for open-domain question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 854–870.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020a. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Sankalp Patel, Sachindra Joshi, and Luis A Lastras. 2020b. doc2dial: A goal-oriented document-grounded dialogue dataset. *arXiv preprint arXiv:2011.06623*.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. Coil: Revisit exact lexical match in information retrieval with contextualized inverted list. In *NAACL*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *EACL 2021-16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 874–880. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

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.

Boeun Kim, Dohaeng Lee, Sihyung Kim, Yejin Lee, Jin-Xia Huang, Oh-Woog Kwon, and Harksoo Kim. 2021. Document-grounded goal-oriented dialogue systems on pre-trained language model with diverse input representation. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 98–102, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020b. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Chin-Yew Lin. 2004. Looking for a few good metrics: Rouge and its evaluation.

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 preprint arXiv:1907.11692*.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Expansion via prediction of importance with contextualization. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pages 1573–1576.

Antonio Mallia, O. Khattab, Nicola Tonellotto, and Torsten Suel. 2021. Learning passage impacts for inverted indexes. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

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.

Biswajit Paria, Chih-Kuan Yeh, Ian En-Hsu Yen, Ning Xu, Pradeep Ravikumar, and Barnabás Póczos. 2020. Minimizing flops to learn efficient sparse representations. *CoRR*, abs/2004.05665.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019a. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019b. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *CoRR*, abs/2007.00808.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. 2020. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104.