Multi-Task Retrieval-Augmented Text Generation with Relevance Sampling

Sebastian Hofstätter [1] **Jiecao Chen** [2] **Karthik Raman** [2] **Hamed Zamani** [3]

Abstract

This paper studies multi-task training of retrieval-augmented generation models for knowledge-intensive tasks. We propose to clean the training set by utilizing a distinct property of knowledge-intensive generation: The connection of query-answer pairs to items in the knowledge base. We filter training examples via a threshold of confidence on the relevance labels, whether a pair is answerable by the knowledge base or not. We train a single Fusion-in-Decoder (FiD) generator on seven combined tasks of the KILT benchmark. The experimental results suggest that our simple yet effective approach substantially improves competitive baselines on two strongly imbalanced tasks; and shows either smaller improvements or no significant regression on the remaining tasks. Furthermore, we demonstrate our multi-task training with relevance label sampling scales well with increased model capacity and achieves state-of-the-art results in five out of seven KILT tasks.

1. Introduction

Retrieval augmented generation models are trained as a unit consisting of retrieval and generation modules (Lewis et al., 2020). The knowledge base accessed by the retriever module offers many benefits for practical use, such as maintainability through updates and domain adaptations. On the other hand, this setup brings additional complexity to the text generation tasks, as we now administer connections of a query-answer pair to relevant items in the knowledge base, for a more holistic view including retrieval performance (Zamani et al., 2022). The coverage sparsity of relevance judgements of large collections and the resulting reliability issues are well studied, yet still a timely problem in the retrieval community (Zobel, 1998; Voorhees, 2001; Craswell et al., 2021; Hofstätter et al., 2021). This challenge is exacerbated when tasks are retroactively expanded (Kwiatkowski et al., 2019), re-purposed (Bajaj et al., 2016) or adapt the collection (Petroni et al., 2021).

We propose a simple yet effective approach for training retrieval-augmented models for knowledge-intensive tasks with noisy labels. We use a confidence score for query-answer pairs and items in the knowledge base. This confidence can be sourced from manually annotated, heuristic, or model generated aspects. We filter training examples via a threshold of confidence on the relevance labels, whether a pair is answerable by the knowledge base or not. With this we aim to reduce noise in the training process, and produce better results with fewer training examples.

To study our training approach, we use a fixed T5-based dense retrieval module (Ni et al., 2021) and train a Fusion-in-Decoder (FiD) generator (Izacard & Grave, 2020) on multiple tasks of the KILT benchmark (Petroni et al., 2021). KILT aggregates and heuristically maps many different English Wikipedia-based generation tasks to a single Wikipedia snapshot, which introduces considerable noise in the label quality, due to the time-shifted nature of the task creations.

We apply our confidence threshold on relevance label filtering to remove a training example if no knowledge item could be identified as sufficiently relevant from the existing labels. Because of the time shifted knowledge base, if an answer is not available in the new passage text anymore, we have a lower confidence, that the query can be answered at all given the new passages. After this step, we apply downsampling on imbalanced tasks for a balanced multi-task training on all the seven tasks of KILT that have passage mappings; spanning open domain QA, fact verification, slot filling, and dialogue categories: HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), T-REx (Elsahar et al., 2018), Zero Shot RE (zsRE) (Levy et al., 2017)), FEVER (Thorne et al., 2018), and Wizard of Wikipedia (WoW) (Dinan et al., 2018).

Furthermore, we demonstrate the robustness of our sampling strategy by creating an alternative to the prevalent original, aggregation method from Wikipedia paragraphs to retrievable units. Finally, we study the impact of our training method on increased capacities of the generator backbone.

[1]TU Wien, Austria (work conducted during an internship at Google) [2]Google, USA [3]University of Massachusetts Amherst, USA. Correspondence to: Sebastian Hofstätter <s.hofstaetter@tuwien.ac.at>.

ICML workshop on Knowledge Retrieval and Language Models, Baltimore, Maryland, USA, 2022.

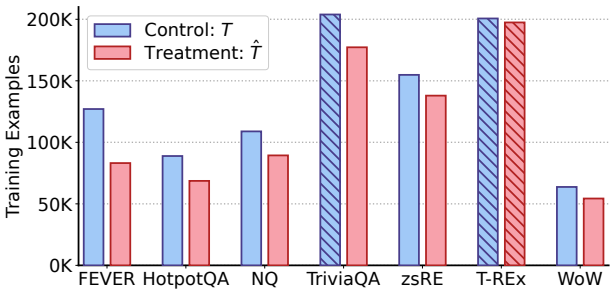

Figure 1. Training examples per task and sampling method. *Hatched bars indicate downsampling with potentially more training data available.*

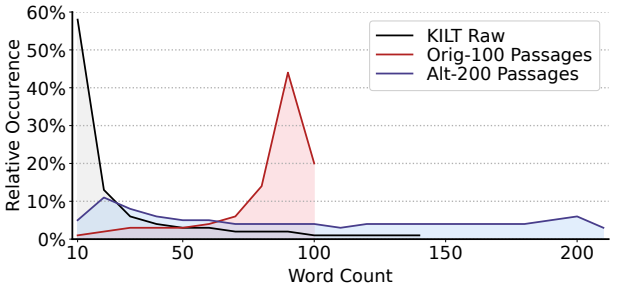

Figure 2. Statistics of the passage lengths of the raw KILT texts, its original chunking (Orig-100) and our alternative approach (Alt-200). *The word counts are binned to 10 words.*

We find that our training strategy significantly improves the effectiveness on the two strongly imbalanced datasets: TriviaQA (+ 12.7 EM) and T-REx (+4.9 Accuracy). This leads to a new state-of-the-art in TriviaQA, and is competitive in T-REx compared to more specialized models. It also statistically significant improves two out of the remaining five tasks, albeit at a smaller rate. When scaling up the FiD backbone with our multi-task sampling technique from T5-Base to T5-Large and T5-XL, we observe expected quality gains across all our evaluated tasks and outperform the state-of-the-art on a total of five out of seven KILT tasks on the official leaderboard.

2. Relevance-Based Confidence Sampling

The main goal in retrieval-augmented generation is to generate an answer string a given a query q; with a secondary goal of identifying a set of relevant passages P from a collection C, which are the source of the answer. In a dataset, the relevant passage set $P_t^{(q,a)}$ using a threshold t, is:

$$P_t^{(q,a)} = \{p \mid \Phi(p, q, a) > t, \forall\, p \in C\} \tag{1}$$

where Φ is a mapping function between a passage, query and answer triple, returning a confidence value, whether this passage is relevant or not. Only if the confidence is higher than our set threshold t, do we include the passage in the set. From the view of a dataset creator, it is usually unfeasible to conduct annotations for all possible pairs, therefore, those pairings without annotations return a null confidence for relatedness, even if it might be related.

As dataset creation is a very costly operation, many works adapt and evolve existing datasets. When knowledge intensive datasets are evolved, the query-answer pair may stay the same while the confidence values of the passage connections change. Therefore, we hypothesise it is beneficial not to include all possible training examples, rather only take into account query-answer pairs, where a higher confidence threshold on the relevance label is set, to reduce noise. A low or no confidence value might indicate a low quality query answer pair.

Starting from the training set T, which includes all possible pairs (q, a), we define a filtered version $\hat{T}_t$ as follows:

$$\hat{T}_t = \{(q, a) \mid \exists\, p \in P_t^{(q,a)}, \forall\, (q, a) \in T\} \tag{2}$$

where we need to define a threshold t as our sampling boundary. The boundary needs to be adapted to the properties of the applied task.

In this work we apply our sampling approach on KILT which conducted a heuristic mapping process of passages for given query-answer pairs. We implement the confidence mapping Φ as the BLEU score in their mapping and set the threshold t to be > 0, filtering all pairs, where no overlap in the previously annotated document was found. Our sampling is not limited to KILT and could be extended to other resources with a similar setup or by mining weakly-supervised relevance signals, as proposed by Asai et al. (2021), and filtering for example based on the confidence of the labelling model.

3. Experiment Design

KILT multi-task training. We train a single generator model on multiple tasks of the KILT benchmark, most of which already provide training sets of similar magnitude (50 to 150 thousand), except for: TriviaQA (1.8 million) and T-REx (12.5 million), accounting for 96% of training examples. FiD training requires considerable hardware resources, therefore we decided to downsample the oversized datasets, rather than upsample the others.[1]

Figure 1 shows the number of query-answer pairs available per sampling method: our control T and treatment $\hat{T}$. We apply our filtering before downsampling oversized tasks to balance our multi-task training set. For our unified training we downsample oversized tasks to 200K examples, then combine and shuffle all tasks. By applying the relevance-label filter method $\hat{T}$ we reduce our training set size by 25% (or 140K fewer examples) to 808K examples compared to our control.

[1] A T5-Base training run until convergence requires roughly one TPU month of compute with our downsample strategies.

Table 1. Comparing sampling strategies and model capacity scaling for multi-task training on the KILT dev set. Highest result in bold. Our results are averaged over the last 10 checkpoints with a 95% confidence interval shown in gray.

	Model	LM	Open Domain QA *EM* NQ	HotpotQA	TriviaQA	Fact *Accuracy* FEVER	Slot Filling *Accuracy* T-REx	zsRE	Dialog *F1* WOW
Related Methods									
1	RAG (Petroni et al., 2021)	BART-L	44.4	27.0	71.3	86.3	59.2	44.7	13.1
2	DPR+FiD (Piktus et al., 2021)	T5-Base	55.0	38.0	71.4	90.9	80.9	72.4	16.1
3	KGI (Glass et al., 2021)	BART-L	–	–	–	–	84.0	71.3	–
4	Re2G (Anonymous, 2022)	BART-L	46.7	–	74.0	91.1	**86.6**	–	19.4
Ours (DPR-100 passages)									
5	GTR + FiD with control T	T5-Base	54.1 ±.3	31.1 ±.2	65.0 ±.6	89.8 ±.2	78.0 ±.5	70.7 ±.4	19.8 ±.2
6	GTR + FiD with treatment $\hat{T}$	T5-Base	54.4 ±.3	31.0 ±.2	78.1 ±.2	89.6 ±.4	82.9 ±.1	71.6 ±.3	19.5 ±.2
Ours (Alt-200 passages)									
7	GTR + FiD with control T	T5-Base	55.1 ±.3	31.6 ±.2	65.7 ±.6	89.8 ±.4	77.6 ±.3	70.2 ±.2	20.1 ±.2
8		T5-Base	56.0 ±.3	31.8 ±.2	78.4 ±.2	89.6 ±.4	82.5 ±.1	71.4 ±.3	19.9 ±.2
9	GTR + FiD with treatment $\hat{T}$	T5-Large	60.7 ±.5	35.4 ±.2	81.8 ±.2	92.1 ±.2	82.9 ±.1	72.9 ±.4	19.9 ±.2
10		T5-XL	**62.9** ±.4	**39.0** ±.2	**84.3** ±.2	**92.8** ±.4	84.1 ±.2	**75.2** ±.3	**21.0** ±.3

Alternative retrievable units. The initial KILT release offers a fine-granular view on the Wikipedia collection. The raw paragraph collection contains 111.4 million items, with a strong concentration of very short sequences (< 20 words), as shown in Figure 2. This is not a practical number of passages to index with dense retrieval methods (as the memory requirement is determined by the number of passages). A standard aggregation approach, proposed by Karpukhin et al. (2020) for DPR is to aggregate the raw data to passages of up to 100 words. It adds as many words to a new passage until 100 words or a changed document are reached, which breaks most paragraphs at the boundaries in half. Therefore, we propose an alternative aggregation strategy to relax the strict length requirement and favoring not to break up paragraphs. We aggregate whole raw-paragraphs until they reach 200 words, or start a new passage if they do not fit. This change results in a very different length distribution, as shown in Figure 2. It results in only 27.7 million passages, compared to 35.7 million of the original chunking. In both cases the title of the page is added to all passages.

Implementation. All our experiments are based on the T5X framework (Roberts et al., 2022). We use a fixed GTR-Base dense retrieval model (Ni et al., 2021), which is pre-trained on the MSMARCO passage retrieval task (Bajaj et al., 2016) and has been shown to generalize well on the BEIR benchmark (Thakur et al., 2021). We train an FiD model (Izacard & Grave, 2020) using T5 v1.1 as language model backbone (Raffel et al., 2020) on TPUs. We attach task specific markers to the input for the multi-task training. We cap the input at 384 tokens (combined query and passage) and a maximum of 64 output tokens. For training we use a batch size of 128 with 50 retrieved passages, and a learning rate of 10^{-3} with the Adafactor optimizer (Shazeer & Stern, 2018). We do not tune our models to a specific checkpoint, rather train them all for 50K steps. The only special case is T5-XL, which uses a learning rate of $5 * 10^{-4}$ and is trained for 30K steps. We use beam search with a beam size of 4 for the decoding.

Evaluation. To reduce the noise in our results, we present the mean and a 95% confidence interval measured with a t-statistic of the last 10 checkpoints (every thousand steps from 40K to 50K training steps).

4. Results

In this section we present and discuss our experimental results. An important note with every use of the KILT benchmark is that the numbers presented here are only comparable to other works also based on the KILT benchmark and not the original versions of the individual tasks. This is due to the changed collection as well as changed query sets, as described by Petroni et al. (2021).

The results are shown in Table 1. In the first section (lines 1-4) we show related works, which also report KILT-based scores: RAG (Lewis et al., 2020), as evaluated by Petroni et al. (2021); DPR + FiD (Piktus et al., 2021); KGI (Glass et al., 2021); and Re2G (Anonymous, 2022). We present our results using the original passage units in the second (lines 5 & 6) and our alternative retrieval units in the third section (lines 7 & 8). In both sections we compare the random downsampling and our proposed relevance-label guided sampling strategy.

Sampling strategies. First, we focus on the two strongly imbalanced tasks (TriviaQA and T-REx), which had their

Table 2. Comparing our models with related work on the KILT test set via the leaderboard (as of July 3rd 2022). Highest result in bold.

	Model	Generator	Open Domain QA *EM* NQ	HotpotQA	TriviaQA	Fact *Acc.* FEVER	Slot Filling *Accuracy* T-REx	zsRE	Dialog *F1* WOW
Top Leaderboard Entries									
1	RAG (Petroni et al., 2021)	BART-Large	44.4	27.0	71.3	86.3	59.2	44.7	13.1
2	DPR + FiD (Piktus et al., 2021)	T5-Base	51.6	38.3	72.7	89.0	82.2	74.0	15.7
3	KGI (Glass et al., 2021)	BART-Large	45.2	–	61.0	85.6	84.4	72.6	18.6
4	Re2G (Anonymous, 2022)	BART-Large	51.7	–	76.3	89.6	**87.7**	–	18.9
5	Hindsight (Paranjape et al., 2021)	BART-Large	–	–	–	–	–	–	19.2
6	SEAL+FiD (Bevilacqua et al., 2022)	T5-?	53.7	**40.5**	70.9	89.5	83.7	74.7	18.3
Ours (Alt-200 passages)									
7	GTR + FiD with treatment $\hat{T}$	T5-Base	52.4	30.1	78.9	87.1	83.4	81.5	18.4
8		T5-XL	**61.2**	39.1	**84.6**	**92.3**	85.2	**83.7**	**20.6**

training examples change the most under our relevance-label sampling strategy: We see that for both passage variants, both tasks improve considerably with the proposed sampling. Comparing lines 7 & 8 we observe a gain for TriviaQA of 12.7 EM and 4.9 Accuracy for T-REx. For the other tasks on our alternative passage-units, we observe small, but significant gains on NQ, and zsRE. The other tasks FEVER, WOW, and HotpotQA only result in non-significant changes inside the 95% confidence interval.

Retrievable units. To observe the impact of our alternative passage aggregation strategy we need to compare the pairs of lines 5 & 7 as well as lines 6 & 8. Even though the properties of the two passage collections are very different, the results of the retrieval augmented generation are very similar. Our alternative approach is slightly better on NQ, TriviaQA, and WOW; a virtual draw on ZsRE, T-REx, and HotpotQA, but worse on FEVER. Overall, we also notice, that our passage sampling strategy works slightly better on the alternative passages, resulting in the best overall results of our ablation. Therefore, we select this combination (line 8) for the following experiments.

Scaling the generator capacity. In most NLP settings, increasing the capacity of a pre-trained model leads to effectiveness gains, at the cost of efficiency. Given how the related methods use varying generator capacities (such as BART-Large (Lewis et al., 2019)), we want to understand and measure the implications of scaling up the generator for our alternative passage aggregation and relevance-label sampling strategy $\hat{T}$. We show these results in Table 1 for T5-Base (line 8), T5-Large (line 9), and T5-XL (line 10).

We find that scaling the model size and compute resource consistently improve results over all tasks. This is an expected result. Nevertheless, we wanted to confirm that our sampling improvement is not just beneficial in a smaller setting.

Leaderboard comparison. We submitted a T5-Base and T5-XL version of the FiD model with our relevance sampling to the official KILT leaderboard[2] for a blind evaluation and present the results in Table 2. Compared to related methods our FiD model with T5-Base is already state of the art on two tasks (TriviaQA and zsRE). Our T5-XL version sets a new state of the art ceiling on a total of five KILT tasks. We outperform the previous best methods by: NQ +7.4 EM, TriviaQA +6.4 EM, FEVER +2.7 Accuracy, ZS-RE +9 Accuracy, and WoW +0.05 F1. We only come in second place on HotpotQA (-1.4 EM) and T-REx (-2.5 Accuracy). This might be attributable to our handicapped zero shot retriever, as HotpotQA is challenging for retrieval models; and downsampling of the T-REx training data, as the related methods are trained exclusively on the single task, without the need for training data adjustments.

Overall, these results are a strong indicator for the viability and usefulness of our relevance-label sampling strategy considering that it has access to 140K fewer training examples than the baseline. We want to emphasize that when we compare our already competitive results to related work our approach is handicapped in a few key areas: **1)** we are not training the retriever (which is out of scope, but orthogonal to our work and should lead to further improvements); **2)** we are training a single model, which gives us less chance to overfit on a single task; **3)** we do not employ multiple training loops, index updates, or knowledge distillation. Therefore, we conclude that multi-task training is a viable option for the community to build upon going forward.

Are we just gaming the benchmark? A valid concern we need to raise is whether we are really improving the quality of the model, or simply moving the training set construction closer to the way the tests sets have been con-

[2]The leaderboard is available at:
https://eval.ai/web/challenges/challenge-page/689

structed by Petroni et al. (2021). The KILT test sets filter an average 18% of queries compared to their original task versions. Petroni et al. (2021) removed a query if not at least one of the answers could be mapped to a passage at least once. Crucially, if one of the answers is partially mappable, all the other answers for this query were also kept as valid. Our analysis shows that, while the average ratio of mapped answers increases compared to the raw training data, especially exact mapped answers still only account for 10% to 67% of available answers. Therefore, we argue that we are not gaming the benchmark, as we exclusively select mapped query-answer pairs for our training, which differs from the test set construction. For a conclusive answer to this question future work should evaluate our trained models on other, independently created, evaluation tasks. A setup which is increasingly common in the neural retrieval community (Ni et al., 2021; Hofstätter et al., 2022).

5. Related Work

Multi-task training. To the best of our knowledge, the multi-task focus of the KILT community so far has been on the retriever module and not the answer generator. The foundational retrieval augmented architectures FiD (Izacard & Grave, 2020), RAG (Lewis et al., 2020), and REALM (Guu et al., 2020) are trained on individual tasks. In their initial baseline setup Petroni et al. (2021) already studied the impact of multi-task retrieval training; Maillard et al. (2021) continued to study various configurations for KILT multi-task single-model retrieval. Lewis et al. (2021) trained the RePAQ-retriever system on multiple tasks, but for their FiD defer mechanism used task-specific FiD checkpoints. For context-given question answering Khashabi et al. (2020) trained UnifiedQA on multiple QA tasks.

Improved RAG training. Many of the recent papers improving RAG-style models optimized end-to-end processes (f.e. EMDR2 (Singh et al., 2021)), ensembling multiple modules (f.e. R2-D2 (Fajcik et al., 2021)), or creating multiple training loops to update the indexed documents multiple times (f.e. Hindsight (Paranjape et al., 2021)). Our approach differs, as we focus on the selection of the available training data in a multi-task setting. For more information on retrieval-enhanced machine learning models, we refer the reader to Zamani et al. (2022).

6. Conclusion

We proposed a simple yet effective approach for multi-task training of the FiD retrieval-augmented generation model on the KILT benchmark. We cleaned (and downsampled were necessary) the training set by removing query-answer pairs with low relevance confidence. We demonstrated that this approach substantially improves two imbalanced tasks, and has a smaller benefit on two of the remaining five tasks. By scaling the model capacity we achieve state-of-the-art results on five KILT tasks evaluated by the leaderboard.

Acknowledgements

This research was supported in part by the Google Visiting Scholar program and in part by the Center for Intelligent Information Retrieval. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsors. We would like to thank Jianmo Ni for helping us setting up T5X retrieval.

References

Anonymous. Re2g: Retrieve, rerank, generate. *ACL Submission (OpenReview)*, 2022.

Asai, A., Gardner, M., and Hajishirzi, H. Evidentiality-guided generation for knowledge-intensive nlp tasks. *arXiv preprint arXiv:2112.08688*, 2021.

Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

Bevilacqua, M., Ottaviano, G., Lewis, P., Yih, W.-t., Riedel, S., and Petroni, F. Autoregressive search engines: Generating substrings as document identifiers. *arXiv preprint arXiv:2204.10628*, 2022.

Craswell, N., Mitra, B., Yilmaz, E., Campos, D., Voorhees, E. M., and Soboroff, I. Trec deep learning track: Reusable test collections in the large data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2369–2375, 2021.

Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*, 2018.

Elsahar, H., Vougiouklis, P., Remaci, A., Gravier, C., Hare, J., Laforest, F., and Simperl, E. T-rex: A large scale alignment of natural language with knowledge base triples. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

Fajcik, M., Docekal, M., Ondrej, K., and Smrz, P. R2-d2: A modular baseline for open-domain question answering. *arXiv preprint arXiv:2109.03502*, 2021.

Glass, M., Rossiello, G., Chowdhury, M. F. M., and Gliozzo, A. Robust retrieval augmented generation for zero-shot slot filling. *arXiv preprint arXiv:2108.13934*, 2021.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*, 2020.

Hofstätter, S., Sertkan, M., and Hanbury, A. Tu wien at trec dl and podcast 2021: Simple compression for dense retrieval. 2021.

Hofstätter, S., Khattab, O., Althammer, S., Sertkan, M., and Hanbury, A. Introducing neural bag of whole-words with colberter: Contextualized late interactions using enhanced reduction. 2022. doi: 10.48550/ARXIV.2203.13088.

Izacard, G. and Grave, E. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

Joshi, M., Choi, E., Weld, D. S., and Zettlemoyer, L. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*, 2020.

Khashabi, D., Min, S., Khot, T., Sabharwal, A., Tafjord, O., Clark, P., and Hajishirzi, H. Unifiedqa: Crossing format boundaries with a single qa system. *arXiv preprint arXiv:2005.00700*, 2020.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Levy, O., Seo, M., Choi, E., and Zettlemoyer, L. Zero-shot relation extraction via reading comprehension. *arXiv preprint arXiv:1706.04115*, 2017.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020.

Lewis, P., Wu, Y., Liu, L., Minervini, P., Küttler, H., Piktus, A., Stenetorp, P., and Riedel, S. Paq: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, 9:1098–1115, 2021.

Maillard, J., Karpukhin, V., Petroni, F., Yih, W.-t., Oğuz, B., Stoyanov, V., and Ghosh, G. Multi-task retrieval for knowledge-intensive tasks. *arXiv preprint arXiv:2101.00117*, 2021.

Ni, J., Qu, C., Lu, J., Dai, Z., Ábrego, G. H., Ma, J., Zhao, V. Y., Luan, Y., Hall, K. B., Chang, M.-W., et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021.

Paranjape, A., Khattab, O., Potts, C., Zaharia, M., and Manning, C. D. Hindsight: Posterior-guided training of retrievers for improved open-ended generation. *arXiv preprint arXiv:2110.07752*, 2021.

Petroni, F., Piktus, A., Fan, A., Lewis, P. S. H., Yazdani, M., Cao, N. D., Thorne, J., Jernite, Y., Karpukhin, V., Maillard, J., Plachouras, V., Rocktäschel, T., and Riedel, S. 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, NAACL-HLT 2021, Online, June 6-11, 2021*, 2021.

Piktus, A., Petroni, F., Karpukhin, V., Okhonko, D., Broscheit, S., Izacard, G., Lewis, P., Oğuz, B., Grave, E., Yih, W.-t., et al. The web is your oyster–knowledge-intensive nlp against a very large web corpus. *arXiv preprint arXiv:2112.09924*, 2021.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21: 1–67, 2020.

Roberts, A., Chung, H. W., Levskaya, A., Mishra, G., Bradbury, J., Andor, D., Narang, S., Lester, B., Gaffney, C., Mohiuddin, A., Hawthorne, C., Lewkowycz, A., Salcianu, A., van Zee, M., Austin, J., Goodman, S., Soares, L. B., Hu, H., Tsvyashchenko, S., Chowdhery, A., Bastings, J., Bulian, J., Garcia, X., Ni, J., Chen, A., Kenealy, K., Clark, J. H., Lee, S., Garrette, D., Lee-Thorp, J., Raffel, C., Shazeer, N., Ritter, M., Bosma, M., Passos, A., Maitin-Shepard, J., Fiedel, N., Omernick, M., Saeta, B., Sepassi, R., Spiridonov, A., Newlan, J., and Gesmundo, A. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*, 2022.

Shazeer, N. and Stern, M. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Con-*

ference on Machine Learning, pp. 4596–4604. PMLR, 2018.

Singh, D., Reddy, S., Hamilton, W., Dyer, C., and Yogatama, D. End-to-end training of multi-document reader and retriever for open-domain question answering. *Advances in Neural Information Processing Systems*, 34, 2021.

Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., and Gurevych, I. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*, 2021.

Thorne, J., Vlachos, A., Christodoulopoulos, C., and Mittal, A. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.

Voorhees, E. M. The philosophy of information retrieval evaluation. In *Workshop of the cross-language evaluation forum for european languages*, pp. 355–370. Springer, 2001.

Yang, Z., Qi, P., Zhang, S., Bengio, Y., Cohen, W. W., Salakhutdinov, R., and Manning, C. D. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*, 2018.

Zamani, H., Diaz, F., Dehghani, M., Metzler, D., and Bendersky, M. Retrieval-enhanced machine learning. *arXiv preprint arXiv:2205.01230*, 2022.

Zobel, J. How reliable are the results of large-scale information retrieval experiments? In *Proc. of SIGIR*, 1998.