Data-Efficient Autoregressive Document Retrieval for Fact Verification

James Thorne
KAIST AI
thorne@kaist.ac.kr

Abstract

Document retrieval is a core component of many knowledge-intensive natural language processing task formulations such as fact verification and question answering. Sources of textual knowledge, such as Wikipedia articles, condition the generation of answers from the models. Recent advances in retrieval use sequence-to-sequence models to incrementally predict the title of the appropriate Wikipedia page given a query. However, this method requires supervision in the form of human annotation to label which Wikipedia pages contain appropriate context. This paper introduces a distant-supervision method that does not require any annotation to train autoregressive retrievers that attain competitive R-Precision and Recall in a zero-shot setting. Furthermore we show that with task-specific supervised fine-tuning, autoregressive retrieval performance for two Wikipedia-based fact verification tasks can approach or even exceed full supervision using less than $1/4$ of the annotated data indicating possible directions for data-efficient autoregressive retrieval.

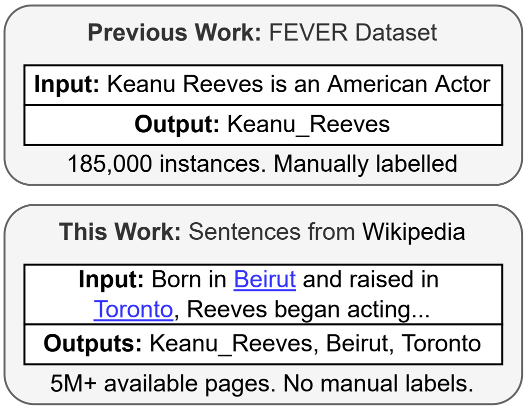

Figure 1: We present a distantly supervised pre-training objective for autoregressive information retrieval. Only using sampled sentences from Wikipedia, without labels, competitive scores can be attained for entity retrieval.

1 Introduction

Conditioning answer generation on knowledge from textual sources is a common component of many well-studied natural language processing tasks. For example, in the SQuAD (Rajpurkar et al., 2016) question answering task, a passage of text is used as a source of information to generate this answer. To enable machine-reading at scale, recent studies combine retrieval with reasoning (Chen et al., 2017; Roller et al., 2021) mandating that systems select appropriate passages from a corpus, such as Wikipedia, to condition answer generation. Furthermore, tasks such as fact verification (Thorne et al., 2018; Wadden et al., 2020; Diggelmann et al., 2020) use evidence retrieved from a corpus and consider both the label and the retrieved passages for evaluation.

Recent advances have been made in neural retrieval models, exploiting the structure of these tasks. De Cao et al. (2020) model retrieval as entity grounding (Bunescu and Paşca, 2006; Le and Titov, 2018): the retriever is trained to predict the title of the Wikipedia document for a given input and is built on a seq2seq architecture. Even though GENRE (De Cao et al., 2020) yields improvements for many retrieval-oriented NLP tasks in the KILT benchmark (Petroni et al., 2021), the model requires supervision during training with labeled data that contains the document titles for a given input.

In this paper, we present a method for training a high-precision and high-recall retrieval system on Wikipedia data in a self-supervised manner. Our system can be trained in less than 6 hours on a single GPU without human-annotated data. The recall far exceeds conventional retrieval methods such as TF-IDF and BM25. Compared to GENRE (De Cao et al., 2020), which is trained with 11 annotated datasets for thousands of GPU-hours, our self-supervised approach for **D**ata **E**fficient

https://github.com/j6mes/sustainlp2022-deardr

Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP), pages 44 - 51
December 7, 2022 ©2022 Association for Computational Linguistics

Auto**R**egressive **D**ocument **R**etrieval, DEARDR, performs within an R-Precision that is 6.36% lower and a Recall@10 that is 0.91% lower for the FEVER Shared Task. We additionally show that with task-specific fine-tuning, DEARDR can attain precision and recall exceeding a fully supervised baseline for FEVER (and on-par with GENRE), with just 16K annotated instances rather than 109K in the full dataset. Similar findings are observed for the HoVer fact verification task (Jiang et al., 2020).

2 Background

Conventional information retrieval methods, such as TF-IDF and BM25, have been applied many knowledge-intensive NLP tasks (Chen et al., 2017; Thorne et al., 2018), with reasonable success. These methods do not require supervision: instead, query-document similarity is estimated based on token-level frequency information from observations on a fixed corpus. At test time, sparse discrete vector encodings of documents and the query are compared to return documents with the highest similarity to the query. While this aids application to new tasks and settings, recall can be low, especially when there are variations in phrasing due to the sparse encoding of which tokens (or variants considering n-grams or subtokens) are present. Neural retrieval (Hanselowski et al., 2018; Karpukhin et al., 2020) in contrast, uses neural networks to generate dense encodings of the query and passages. These models are trained with supervision: the training data contains lists of appropriate passages for a given query, but typically does not contain negative instances. How negative instances are sampled in training influences the suitability of the retrieved documents for the downstream task (Cohen et al., 2019; Karpukhin et al., 2020). Variants of training regimes for dense retrieval also use Cloze-task (Lee et al., 2019), Wikipedia revision information (Chang and Kao, 2012), contrastive (Izacard et al., 2021) learning, or multi-task learning (Maillard et al., 2021) to improve performance.

2.1 Autoregressive Document Retrieval

In contrast to the previous approaches, where the content of a passage is scored for a query using it's content, autoregressive document retrieval (De Cao et al., 2020) uses a seq2seq model that is trained to predict a relevant document title, such as a Wikipedia page. Tokens are decoded incrementally left to right where the and scored with $p(\mathbf{y}|\mathbf{x}) = \prod_{i=1} p(y_i|y_{<i}, \mathbf{x})$ where the decoded document title $\hat{\mathbf{y}} \in \mathcal{E}$ exists in a corpus $\mathcal{E}$. To ensure this constraint is satisfied, constrained decoding sets $p(y_i|y_{<i,x}, \mathbf{x}) = 0$ for token sequences $(y_1, \dots y_i)$ that do not occur in the index. In practice this works well in the Wikipedia domain where document titles are simple canonical descriptors of an entity or concept. An extension, mGENRE (Cao et al., 2021), has been trained for multi-lingual entity linking using Wikipedia hyperlinks and internationalized versions of the pages from the Wikidata graph as supervision targets in other languages. This has not been applied to an entity linking task, but not evaluated for document retrieval.

Similarly, GENRE did use hyperlink-based information by incorporating data from BLINK during training. However, its contribution to system performance appears low (De Cao et al., 2020, Table 8), warranting further investigation. While the use of pre-training with hyperlink information in retrieval has shown promise (Ma et al., 2021) in other formulations, the use of distant-supervision in autoregressive retrieval, using the article titles and hyperlinks in training is emerging and has been studied in contemporaneous work (Chen et al., 2022). Lee et al. (2022) train autoregressive models for multi-hop retrieval, with a data augmentation strategy. Alternative autoregressive retrieval formulations are designed to predict document sub-strings (Bevilacqua et al., 2022): this obviates the need to have unique document identifiers.

3 Data Efficient Document Retrieval

The primary objective of this paper is to reduce the dependency on supervised instances and exploit *distant supervision* to train an autoregressive document retrieval system. Distant supervision for DEARDR exploits the structural aspects of the Wikipedia corpus: specifically, the page titles (denoted PT) and hyperlinks (denoted HL) from sentences to other pages. Sentences from Wikipedia documents are sampled from the corpus as input and DEARDR is trained to decode the page title or hyperlinks, or both (denoted PTHL).

Even though the DEARDR has only been pre-trained with distant supervision without exposure to annotated training data for a knowledge intensive NLP task, we hypothesize that training to predict a Wikipedia page title or hyperlinks acts a a reasonable analog that simulates a common component of many retrieval oriented tasks. This should be sufficient to allow zero-shot application to retrieve relevant documents without the need for human annotations enabling application of knowledge-intensive NLP tasks to new domains or languages. In contrast, the GENRE model (De Cao et al., 2020) is trained with data from eleven Wikipedia-based NLP tasks with millions of annotated instances.

Constrained decoding At test time, a result set is decoded by aggregating the results from a beam search (Sutskever et al., 2014) with constrained decoding (De Cao et al., 2020). However, in contrast to GENRE, which predicts a single entity per beam, DEARDR is trained to predict a sequence of all the hyperlinked page names (illustrated in Figure 1).

Self-supervised vs task-specific retriever Once the DEARDR retriever has been pre-trained with self-supervision on Wikipedia data, it can be applied in a *zero-shot* setting to the knowledge-intensive task of fact verification. Some aspects of the test-task formulation, may have patterns that differ to what DEARDR is exposed to during the pre-training. Using small numbers of task-specific training data, the pre-trained DEARDR will be fine-tuned evaluated on downstream tasks. We hypothesize that the pre-training regimen for DEARDR will reduce the number of instances needed to train the system and attain similar performance to a system with full supervision.

4 Experimental Setup

Three different pre-training regimens for DEARDR, based on page title (PT), sentence hyperlinks (HL) and a combination of both (PTHL), are performed using the snapshot of Wikipedia from June 2017. This was the snapshot used for the FEVER shared task. Document-level retrieval for two fact verification tasks will be evaluated: FEVER (Thorne et al., 2018) and HoVER (Jiang et al., 2020).

Zero-Shot Document Retrieval: Without exposure to the underlying test task, DEARDR will be pre-trained using unlabeled instances from English Wikipedia articles (pre-trained with PT, HL or PTHL), and then applied to instances from these retrieval-based NLP tasks. From Wikipedia, we generate 16.8M distant-supervision instances.

Data-Efficient Document Retrieval: The DEARDR model will be fine-tuned using a low number of labeled instances from the target task. During training, we sample instances uniformly at random. We optimize training for *Recall* with early stopping. This occurred after 12,500 steps (100K instances in total).

Supervised Baseline: For a controlled baseline system that DEARDR can be compared against, we train a document retriever for the target task using all available data in a fully supervised setting.

Previous Work: We compare to GENRE (De Cao et al., 2020) which was trained with data from *eleven* retrieval tasks. We also compare to sparse-vector retrieval methods such as BM25 and TF-IDF. Finally, for dense-vector retrieval, we compare against DPR (Karpukhin et al., 2020). Because contrastive retrieval (Izacard et al., 2021) does not offer significant advantages over BM25 for FEVER, we do not evaluate against it.

4.1 Evaluation

Document retrieval is evaluated using two performance evaluation metrics: R-Precision and Recall@k. R-Precision is the precision of retrieved documents@R where R is the number of expected elements labeled for the instance. If the test set only specifies 1 valid document, this is equivalent to Precision@1. However, datasets sometimes require a multi-hop combination of pages for inference, requiring multiple documents to be considered for evaluation. Recall@k is the proportion of the gold documents present in the first k elements predicted by the model. This metric is a useful indicator of potential upper-bound system performance for some tasks, such as FEVER, which considers up to 5 retrieval results for scoring claim veracity. Where multiple answer sets are present for instances, we consider each answer set independently and return the max score over all the sets to allow comparison to the KILT methodology (Petroni et al., 2021).

4.2 Implementation

We use the HuggingFace (Wolf et al., 2020) implementation of T5-base (Raffel et al., 2020). This is fine-tuned using data as outlined in the previous section. We optimize hyper-parameters by sweeping the learning rate and scheduler (documented in Appendix A.2) and maximizing R-Precision on the dev split. The index for constraining decoding is constructed from all subtokens generated by the T5 Tokenizer for article titles from the Wikipedia version for the test task.

Trainer	**R-Precision (%)**		**Recall@10 (%)**	
	Page	**Link**	**Page**	**Link**
GENRE	*26.51*	*38.45*	*36.28*	*55.31*
PT	33.12	21.10	40.04	29.95
HL	2.65	**71.91**	17.49	84.35
PTHL	**33.17**	38.91	**37.70**	**84.89**

Table 1: R-Precision and Recall of page titles (page), and hyperlink destinations (link) of sentences sampled from Wikipedia using our training approaches (PT, HL, PTHL) compared to a contemporary supervised approach which only underwent task-specific traiing and did not undergo the pre-training.

Approach	**FEVER Retrieval (%)**	
	R-Prec	**Recall@10**
DearDr (PT) ZS	77.66	91.95
DearDr (HL) ZS	56.55	89.94
DearDr (PTHL) ZS	75.89	88.18
DearDr (PT) 16K	82.49	**94.85**
Supervised	81.36	94.28
GENRE (11 tasks)	**84.02**	92.86
TF-IDF	29.89	68.57
BM25	40.42	70.58
DPR	55.98	77.53

Table 2: Without exposure to training instances from the FEVER task, DearDr attains high recall and R-Precision (R-Prec) for document retrieval for FEVER in a zero-shot (ZS) setting and can be further improved with fine-tuning on 16K (16,000) data.

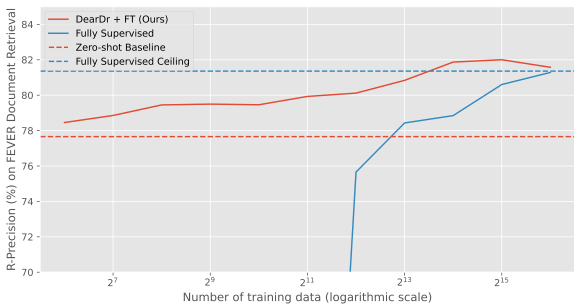

Figure 2: Learning curve showing greater R-Precision when training with fewer instances using DearDr pre-training compared to conventional supervised training

5 Results and Discussion

5.1 Pre-training Intrinsic Evaluation

DearDr was optimized by selecting the model with the highest R-precision on the FEVER shared task. For the R-Precision on the PT and HL components of the pre-training task, we provide the following intrinsic evaluation listed in Table 1 to evaluate the pre-training objectives. Without pre-training on hyperlinks, recall is low indicating that hyperlink pre-training may be beneficial to multi-entity retrieval needed for some FEVER instances.

5.2 Downstream Extrinsic Evaluation

FEVER: For the FEVER shared task, we trained DearDr with instances sampled from the Wikipedia snapshot for the task without using any human-annotated data. Table 2 highlights the retriever's R-Precision and Recall@10 in comparison to a fully supervised system showing that in the zero shot setting (without exposure to any labeled data) document retrieval scores are adequate and far exceed retrieval from token-based similarity methods such as TF-IDF and BM25. Because FEVER is a claim verification task, the claims are similar in nature to sentences sampled from Wikipedia pages. The similarity between the claims and the Wikipedia sentences the zero-shot system was exposed to during training mean that this system is able to apply well to this task.

While GENRE was trained on 11 tasks with over 100K fact verification instances and over 500K question answering instances, the R-Precision of our zero-shot system is only 6.36% lower with Recall@10 less than one percent lower. Given that most modern fact verification approaches perform supervised re-ranking of sentences from these documents, we do not foresee this lower precision having such a large impact on the final task score.

With full supervision from the FEVER task, our control model for comparison attains an R-Precision of 81.36%. However, with small numbers of instances for fine-tuning on FEVER, higher recalls can be attained. With 16,384 instances (less than 1/8 of the dataset), an R-Precision and Recall@10 exceeding this supervised baseline can be achieved. Furthermore, with only 2048 instances (2% of the dataset), an R-Precision of at least 80% is attained. In contrast, without pre-training R-Precision for both of these models with the same number of data is less than 30%. Learning curves are plotted in Figure 2.

HOVER: The multi-hop nature of HoVer presents more complex reasoning challenges than

Approach	HOVER Retrieval (%)	
	R-Prec	Recall@10
GENRE (Transfer)	43.28	49.41
DEARDR (PT) ZS	32.83	36.43
DEARDR (HL) ZS	42.22	43.78
DEARDR (PTHL) ZS	38.94	47.44
DEARDR(PTHL) 4K	45.62	49.14
DEARDR(PTHL) All	**46.23**	50.33
Supervised 4K	29.24	35.22
Supervised	46.22	**50.38**

Table 3: The multi-hop aspect of HoVer presents new challenges. Despite this, DearDr attains higher R-Prec with fewer training data than supervised baselines

FEVER and is reported in Table 3. Our zero-shot model has R-Precision that is less than 1% of GENRE. While GENRE wasn't trained on HoVer, it was trained on HotpotQA (Yang et al., 2018) which the HoVer dataset is derived from.

The benefit of pre-training with hyperlinks becomes apparent for multi-hop challenges as R-Precision for HL and PTHL exceed PT. With limited fine-tuning, using $1/4$ of the dataset, R-Precision with DEARDR is less than 1% away from a fully supervised model, despite using fewer data. Without pre-training, R-Precision is unsatisfactory. With all data, DEARDR performs as good as a model without pre-training. While DEARDR is beneficial for this task with fewer data, there are clearly more complex challenges with multi-hop reasoning that require further data augmentation, such as (Lee et al., 2022), to be solved solved by autoregressive methods for retrieval.

Question Answering: The similarity between fact verification and DEARDR pre-training is similar, aiding retrieval. However, application to question answering (TriviaQA (Joshi et al., 2017, TQA), HotpotQA (Yang et al., 2018, HPQA) and NaturalQuestions (Kwiatkowski et al., 2019, NQ)) requires further study. Table 4 shows pre-training does offer a benefit, but with fewer data for fine-tuning, similar gains in retrieval cannot be attained.

6 Conclusions and Future Work

We show that distant supervision and pre-training enables high precision autoregressive document retrieval with fewer annotated training data. While previous work has studied the utility of pre-training

Approach	R-Precision (%)		
	TQA	HPQA	NQ
GENRE*	69.2	51.3	60.3
DEARDR (PTHL) ZS	44.24	42.44	18.05
DEARDR (PTHL) 32K	53.98	43.54	36.94

Table 4: Application to question answering highlights further challenges (* reported by De Cao et al. (2020)).

for dense-retrieval, this work aids understanding of sparse autoregressive retrieval. In application to fact verification, fewer labeled training data were required. However, when we applied this method to question answering, satisfactory results were not obtained due to the domain shift between the two tasks. Better understanding this limitation would be required to adapt DEARDR pre-training to a wider range of tasks and multi-hop reasoning.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)).

References

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Wen-tau Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *arXiv preprint arXiv:2204.10628*.

Razvan Bunescu and Marius Paşca. 2006. Using encyclopedic knowledge for named entity disambiguation. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 9–16, Trento, Italy. Association for Computational Linguistics.

Nicola De Cao, Ledell Wu, Kashyap Popat, Mikel Artetxe, Naman Goyal, Mikhail Plekhanov, Luke Zettlemoyer, Nicola Cancedda, Sebastian Riedel, and Fabio Petroni. 2021. Multilingual autoregressive entity linking. In *arXiv pre-print 2103.12528*.

Yang-Jui Chang and Hung-Yu Kao. 2012. Link prediction in a bipartite network using wikipedia revision information. In *2012 Conference on Technologies and Applications of Artificial Intelligence*, pages 50–55. IEEE.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual*

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Yiqun Liu, Yixing Fan, and Xueqi Cheng. 2022. Corpusbrain: Pre-train a generative retrieval model for knowledge-intensive language tasks. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 191–200.

Daniel Cohen, Scott M. Jordan, and W. Bruce Croft. 2019. Learning a better negative sampling policy with deep neural networks for search. In *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '19, page 19–26, New York, NY, USA. Association for Computing Machinery.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. pages 1–16.

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

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Towards unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Phong Le and Ivan Titov. 2018. Improving entity linking by modeling latent relations between mentions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1595–1604, Melbourne, Australia. Association for Computational Linguistics.

Hyunji Lee, Sohee Yang, Hanseok Oh, and Minjoon Seo. 2022. Generative multi-hop retrieval. *arXiv preprint arXiv:2204.13596*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, Florence, Italy. Association for Computational Linguistics.

Zhengyi Ma, Zhicheng Dou, Wei Xu, Xinyu Zhang, Hao Jiang, Zhao Cao, and Ji-Rong Wen. 2021. Pre-training for ad-hoc retrieval: hyperlink is also you need. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 1212–1221.

Jean Maillard, Vladimir Karpukhin, Fabio Petroni, Wen-tau Yih, Barlas Oguz, Veselin Stoyanov, and Gargi Ghosh. 2021. Multi-task retrieval for knowledge-intensive tasks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1098–1111, Online. Association for Computational Linguistics.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 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. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

A Appendix

A.1 Hardware Requirements

Experiments were performed on a single workstation with a single NVIDIA GTX 1080 Ti GPU.

A.2 Implementation

A.2.1 Pre-training

The following parameters were adjusted as part of the hyper-parameter optimization with the best parameters for all experiments indicated in bold. For HL, using learning rate of 5e-6 was more beneficial.

- Learning rate: 1e-4, **5e-5**, 1e-5, *5e-6*, 1e-6.
- Scheduler: **constant with warmup**, constant, linear.

A.2.2 Fine-tuning + Supervised

The following parameters were adjusted as part of the hyper-parameter optimization with the best parameters for all experiments indicated in bold. For fine-tuning, using a dropout of 0.2 was more beneficial.

- Learning rate: 1e-4, 5e-5, **1e-5,** 5e-6, 1e-6.
- Scheduler: **constant with warmup**, constant, linear.
- Dropout: **0.1**, 0.2, 0.3

B Licenses

The dataset released with this paper makes use of data from Wikipedia which is licensed under creative commons CC-BY-SA 4.0 license.