A Simple Approach to Jointly Rank Passages and Select Relevant Sentences in the OBQA Context

Anonymous ACL submission

Abstract

In the open book question answering (OBQA) task, selecting the relevant passages and sentences from distracting information is crucial to reason the answer to a question. HotpotQA dataset is designed to teach and evaluate systems to do both passage ranking and sentence selection. Many existing frameworks use separate models to select relevant passages and sentences respectively. Such systems not only have high complexity in terms of the parameters of models but also fail to take the advantage of training these two tasks together since one task can be beneficial for the other one. In this work, we present a simple yet effective framework to address these limitations by jointly ranking passages and selecting sentences. Furthermore, we propose consistency and similarity constraints to promote the correlation and interaction between passage ranking and sentence selection.The experiments demonstrate that our framework can achieve competitive results with previous systems and outperform the baseline by 28% in terms of exact matching of relevant sentences on the HotpotQA dataset.

Question: The football manager who recruited David Beckham managed Manchester United during what timeframe?

Passage1, 1995–96 Manchester United F.C. season: The1995-96 season was Manchester United's fourth season in the Premier League, and their 21st consecutive season in the top division of English football. United finished the season by becoming the first English team to win the Double (league title and FA Cup) twice. *Their triumph was made all the more remarkable by the fact that Alex Ferguson had sold experienced players Paul Ince, Mark Hughes and Andrei Kanchelskis before the start of the season, and not made any major signings.Instead, he had drafted in young players like Nicky Butt, David Beckham, Paul Scholes and the Neville brothers, Gary and Phil.*

passage2, Alex Ferguson: *Sir Alexander Chapman Ferguson,CBE (born 31 December 1941) is a Scottish former football manager and player who managed Manchester United from1986 to 2013.* He is regarded by many players, managers and analysts to be one of the greatest and most successful managers of all time.

Answer: from 1986 to 2013

Supporting facts: [["1995-96 Manchester United F.C.season",2],["1995-96 Manchester United F.C. season",3],["AlexFerguson",0]]

Figure 1: An example from the HotpotQA dataset, where the question should be answered by combining supporting facts(SP) from two passages. In the SP, the first string refers to the title of passage, and the second integer means the index of the sentence.

1 Introduction

Open book question answering (OBQA) requires a system to find the relevant documents to reason the answer to a question. It has wide and practical Natural Language Processing (NLP) applications such as search engines (Kwiatkowski et al., 2019) and dialogue systems (Reddy et al., 2019; Choi et al., 2018). Among several OBQA datasets (Dhingra et al., 2017; Mihaylov et al., 2018; Khot et al., 2020), HotpotQA (Yang et al., 2018) is more challenging because it requires a system not only to find the relevant passages from large corpus but also find the relevant sentences in the passage which eventually reach to the answer. Such a task also increases the interpretability of the systems.

To address this challenge, most of the previous work (Nie et al., 2019; Fang et al., 2020; Tu et al., 2019; Groeneveld et al., 2020) use two-step pipeline: identify the most relevant passage by one model and then match each question with a single sentence in the corresponding passage by another model. Such systems are heavy in terms of the size of the models which requires long training and inference time. Green AI has recently been advocated to against the trend of building large models which are both environmentally unfriendly and expensive, raising barriers to participation in NLP research (Schwartz et al., 2020). Apparently, systems using multiple models to solve HotpotQA task do not belong to the family of Green AI. Furthermore, the benefits of learning from passage ranking

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 181 - 187
July 10-15, 2022 ©2022 Association for Computational Linguistics

and selecting relevant sentences are not well utilized by these systems. Intuitively, if a passage is ranked high, then some sentences in the passage should be selected as relevant. On the other hand, if a passage is ranked low, then all sentences in the passage should be classified as irrelevant.

To build a Green AI system and take advantage of multi-task learning, we introduce a Two-in-One model, a simple model trained on passage ranking and sentence selection jointly. More specifically, our model generates passage representations and sentence representations simultaneously, which are then fed to a passage ranker and sentence classifier respectively. Then we promote the interaction between passage ranking and sentence classification using consistency and similarity constraints. The consistency constraint is to enforce that the relevant passage includes relevant sentences, while the similarity constraint ensures the model to generate the representation of relevant passages more closer to the representations for relevant sentences than irrelevant ones. The experiments conducted on the HotpotQA datasets demonstrate that our simple model achieves competitive results with previous systems and outperforms the baselines by 28%.

2 Related Work

HotpotQA Systems A straightforward way to solve the HotpotQA challenge is to build a hierarchical system (Nie et al., 2019), meaning a system first ranks relevant passages and then identifies relevant sentences from the selected passages. Such a hierarchical system involves multiple models thus requires long inference time. More importantly, such a system only leverages the impact of passage ranking on sentence selection but ignores the influence of the sentence selection on the passage ranking. Our framework achieves these two tasks by one model and facilitates the interaction by two constraints. Groeneveld et al. (2020) proposes a pipeline based on three BERT models (Devlin et al., 2019) to solve the HotpotQA challenge. The system first selects relevant sentences and then detects the answer span, finally, identifies the relevant sentences according to the answer span. Though the pipeline is strong, the way it solves the problem is opposite to human beings. We, humans, identify the relevant sentences, and then give the answer span. Many existing works demonstrate the effectiveness of graph neural networks(GNN) on HotpotQA challenge (Fang et al., 2020; Tu et al., 2019). Since GNN is out of the scope of this work, we do not compare it with these frameworks.

Joint Model for QA Joint learning has been studied in Question Answering Tasks. Deng et al. (2020) proposes a joint model to tackle community question answering such that the model can simultaneously select the set of correct answers from candidates and generate an abstractive summary for each selected answer. Sun et al. (2019) proposes a generative collaborative network to answer questions and generate questions. The main difference between our work and previous ones are in two sense (1) our proposed model uses the shared encoder to tackle two classification tasks (2) besides the loss function to optimize individual tasks, we also propose two constraints that utilize the relation between these two tasks.

3 HotpotQA Dataset

HotpotQA dataset (Yang et al., 2018) is designed for multi-hop reasoning question answering tasks, i.e. to reason over multiple documents and answer questions (see Figure 1). Particularly, HotpotQA challenge requires reasoning over two passages. Furthermore, to guide the system to perform meaningful and explainable reasoning, the dataset also provides supporting facts (SP) that reach the answer to the question. HotpotQA provide two challenging settings: in **Fullwiki setting**, a system needs to rank passage from the entire wiki corpus; in **Distractor setting**, 10 distracting passages (including relevant ones) are given for each question. In this work, we mainly focus on the latter setting. From the training set, we find that 70.4% questions have exactly two supporting facts (SP), and 60.0% of SP are the first sentence of passages.

4 Method

We aim to jointly conduct two tasks, passage ranking and supporting facts selection for HotpotQA. Given a question Q, the goal is to simultaneously rank the set of candidates A = $\{a_1, ..., a_i\}$ and identify the supporting facts for the TopK[1] passages.

4.1 Model: Two-in-One Framework

We introduce the proposed joint model for passage ranking and support fact selection, Two-in-One, which uses state-of-the-art transformer-based

[1] The value of K depends on the task, and for HotpotQA, K is 2.

model (Vaswani et al., 2017) to encode questions and contexts. In this work, we use RoBERTa (Liu et al., 2019), however, any other variants like ELECTRA (Clark et al., 2020) can be applied in this framework. The model architecture is given in Figure 2. On top of the encoder, there are two MLP layers to score passages and sentences respectively. In details, given a question and a passage, we firstly create an input to feed through RoBERTa (Liu et al., 2019) by concatenating the question and the passage as follows, $\langle s \rangle Q \langle /s \rangle S_1 \langle /s \rangle S_2 ... \langle /s \rangle S_k \langle /s \rangle$ where $\langle s \rangle$ and $\langle /s \rangle$ are special tokens in RoBERTa, S_i is the i^{th} sentence from a passage. We take $\langle s \rangle$ as the contextual representation for passage ranking and the $\langle /s \rangle$ in front of each sentence for sentence selection. The passage ranker and the sentence classifier have identical structure (two-layer Multiple-Layer Perceptron(MLP)) but different weights.

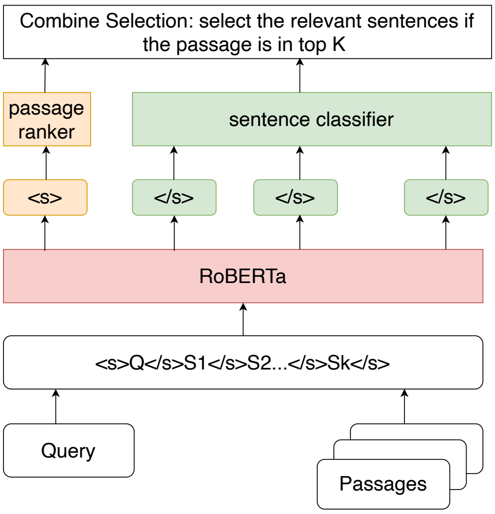
Combine Selection: select the relevant sentences if the passage is in top K
passage ranker
sentence classifier
<s>
</s>
</s>
</s>
RoBERTa
<s>Q</s>S1</s>S2...</s>Sk</s>
Query
Passages

Figure 2: The architecture of Two-in-One model for passage ranking and relevant sentence selection. For HotpotQA dataset, K is two.

The model is jointly trained by passage loss and sentence loss. In detail, during the training time, we assign the relevant passages and sentences with ground truth score 1 while irrelevant passages and sentences with ground truth score -1. Then, Mean Square Error(MSE) loss is applied to calculate the passage and sentence loss as follows,

$$\begin{aligned} \mathcal{L}^{pass} &= (\hat{y} - y)^2, \\ \mathcal{L}^{sent} &= \sum_{i=1}^{K} (\hat{x}_i - x_i)^2, \\ \mathcal{L}^{joint} &= \mathcal{L}^{pass} + \mathcal{L}^{sent}, \end{aligned} \tag{1}$$

where $\hat{y}$ is the predicted passage score, y is the ground truth score of the passage, $\hat{x}_i$ and x_i are the predicted sentence score and ground truth score of S_i, respectively, and K is the total number of sentences in the passage. We simply sum up the passage loss and sentence loss to jointly update model parameters.

During the inference time, passages are ranked based on the logits given by the passage ranker. For the sentence classification, we take 0[2] as the threshold to classify the relevance of each sentence: if the score given by the sentence classifier is larger than 0, then it is relevant; otherwise, irrelevant.

Next, we introduce two constraints to facilitate the interaction between these two tasks.

4.2 Consistency Constraint

Intuitively, if a passage is relevant to the question, then there are some sentences from the passages that are relevant; on the other hand, if a passage is not relevant to the answer, then there should not be relevant sentences inside the passage. Thus, we propose a consistency constraint over the passage ranker and sentence classifier to minimize the gap between the passage score and the maximum sentence score. The loss function is as follows:

$$\mathcal{L}^{con} = (\hat{y} - max(\mathbf{x}))^2, \tag{2}$$

where $\mathbf{x} = [\hat{x}_1 \dots \hat{x}_n]$ denotes a stack of predicted sentence scores.

4.3 Similarity Constraint

As we have shown at the beginning of this section, token $\langle s \rangle$ is used to get the passage score, and each token $\langle /s \rangle$ is used to get the sentence score. Intuitively, the similarity between token $\langle s \rangle$ of a relevant passage is more close to token $\langle /s \rangle$ of a relevant sentence than to $\langle /s \rangle$ of any irrelevant sentence. To enforce this constraint, we use triplet as follows:

$$\begin{aligned} \mathcal{L}^{sim} = \frac{1}{N \cdot M} \sum_{i=1}^{N} \sum_{j=1}^{M} (max\{d(v^p, v_i^r) \\ - d(v^p, v_j^n) + m, 0\}), \end{aligned} \tag{3}$$

where $d(\cdot, \cdot)$ is the Euclidian similarity, N is the number of relevant sentences, M is the number of irrelevant sentences, v^p, v^r, v^n is the vector representation of the relevant passage, relevant sentence,

[2] The reason for threshold "0" is that it is the middle value of 1 and -1, which are labels for relevant and irrelevant sentences in the training time.

and irrelevant sentence respectively. Equation 3 enforces that all the relevant sentences should have higher similarity with the passage than all the irrelevant sentences by a margin m; otherwise, the model would be penalized. In practice, we set the margin m at 1 and find optimum results. We train our model in an end-to-end fashion by combining $\mathcal{L}^{joint}$, $\mathcal{L}^{con}$ and $\mathcal{L}^{dis}$.

5 Experiment

In this section, we first describe the training setup, and then introduce two baselines. We evaluate the two baselines and our proposed joint model on the HotpotQA dataset. Yang et al. (2018) provides two metrics for supporting facts evaluation, exact matching (EM) and F1 score. We also present the precision and recall of SP, and the exact matching of passages for detailed comparison. Meanwhile, we compare our model with the QUARK system (Groeneveld et al., 2020). Lastly, we conduct an ablation study to show the effectiveness of the proposed similarity loss and consistent loss.

5.1 Experiment Setup

We use Huggingface (Wolf et al., 2020) and Pytorch (Paszke et al., 2019) libraries to implement each model. We use 4 TX1080 and V100 NVIDIA to train models in 5 epochs with a learning rate of 1e-5, batch size of 32. We set the maximum input length in training to be 512.

5.2 Baseline

To have comparable size of the model, two baselines have similar structure as our Two-in-One model. Our model has two classification heads, whereas each of the baselines has one classification head. One baseline is to select relevant sentences, and the other one is to rank passages.

Sentence Selection Baseline The first baseline is to select relevant sentence, and particularly, we use a RoBERTa-large with an additional MLP trained on question and a single sentence: $\langle s\rangle Q\langle /s\rangle S\langle /s\rangle$, where Q is a question and S is a sentence. Although this model can not predict the relevant passage directly, based on the assumption that relevant passages include relevant sentences, we pick up two relevant passages based on the top2 sentence scores. When the top1 and the top2 sentences are from the same passage, we continue searching based on the ranking sentence scores until the second document comes up. Then the supporting facts are those sentences from the relevant documents with a score larger than 0.

Passage Selection Baseline In the second baseline, again, we use RoBERTa-large but with the goal of passage selection. The input to the model is a question and a passage: $\langle s\rangle Q\langle /s\rangle P\langle /s\rangle$. Since such a model can not predict sentence relevancy score, based on the statistic of HotpotQA that majority of training set has two supporting facts and the most of them are the first sentences in a paragraph (see Section 3), we select supporting facts by the first sentence of the top1 and top2 passages.

5.3 Result

As we see from Table 1, Two-in-One framework outperforms two baselines with large-margin improvement in all metrics, especially we see a significant improvement on the EM of SP. Our framework outperforms the Sentence Selection Baseline by 20% and 4.5% improvement on the precision and recall of SP, respectively, which demonstrates that jointly learning is beneficial for sentence classification. Also, jointly learning benefits for the passage ranking by comparing Two-in-One with Passage Selection Baseline on the EM of passage. Besides, we also compare Two-in-One with QUARK (Groeneveld et al., 2020), a framework involving three BERT models, (roughly three times larger than ours). Two-in-One achieves comparable results in terms of F1 and EM of SP regardless of much less parameters in our system. Notice that we do not have the other three values because they are not presented in their original paper.

5.4 Ablation

To evaluate the impacts of the consistency constraint and the similarity constraint, we conduct experiments with and without constraints. From Table 2, we see that both consistency constraint and similarity constraint improve F1 and EM of SP and the similarity constraint also improves the EM of passages. We found that without any constraint, though the model can rank the passages well, it suffers from distinguishing between close sentences. The similarity constraint addresses this issue in some sense by maximizing the distance between relevant and irrelevant sentences.

To better understand the impact of consistency constraint, we analyze the consistency between the passage score and the sentence score. The predic-

Model	# Parameters	SP Precision	SP Recall	SP F1	SP EM	Passage EM
Sentence Selection Baseline	~330M	67.96	81.05	72.02	28.12	69.70
Passage Selection Baseline	~330M	66.43	56.55	60.20	27.30	90.44
Two-in-One + sim (Ours)	~330M	**88.06**	**85.68**	**85.82**	**59.17**	**91.11**
QUARK	~1020M*	N/A	N/A	86.97	60.72	N/A
SAE(RoBERTa)	~660M+*	N/A	N/A	87.38	**63.30**	N/A
HGN(RoBERTa)	~330M+*	N/A	N/A	**87.93**	N/A	N/A

Table 1: The Results for two baselines and Two-in-One model with similarity constraint on dev set of HotpotQA distracting dataset. SP stands for supporting facts and EM for Exact Match. * refers to estimation. The bottom systems have much larger model size than our method, where QUARK (Groeneveld et al., 2020), is the result of a framework with 3 BERT models, SAE (Tu et al., 2019) uses two large language models and an GNN model, and HGN (Fang et al., 2020) uses a large language model, a GNN model and other reasoning layers.

Model	SP F1	SP EM	Passage EM
Two-in-One	85.52	58.67	90.93
Two-in-One + con	85.55	58.98	90.29
Two-in-One + sim	**85.82**	**59.17**	**91.11**
Two-in-One + con + sim	85.63	58.74	90.78

Table 2: The results for Two-in-One model with or without consistency and similarity constraints.

tion of a model is consistent if the passage score agrees with the sentence scores and the agreement can be measured by the gap between the passage score and the maximum sentence score among all sentences in that passage. We observe that by adding the consistency constraint, the gap between the passage score and the sentence score is much smaller than without the consistency constraint, i.e. 0.03 v.s. 0.11. It demonstrates that the constraint is beneficial for consistent prediction.

6 Future Work

While in this work, we show the initial and promising results of the Two-in-One model on one single dataset, there are a couple of directions we can explore in the future such as those discussed below.

Model Architecture It is easy to extend the Two-in-One model to Three-in-One model such that besides the passage ranking and sentence selection modules, a third module can predict the answer span. Like the simple extractive QA model based on RoBERTa, where a linear layer or an MLP can predict the start and end position of the answer span. A restricted inference procedure can be enforced that the answer span should be predicted from the selected sentence given by the previous model. One benefit is to reduce the difficulty for the answer selection model since less sentences will be seen by the model and the second benefit is to increase the interpretability of the model. On the other hand, if the sentence selection model makes mistakes, then such errors will carry to the answer span model which yields the wrong answer eventually.

Passage and Sentence Representation We use the contextual vector of a special token in front of each sentence to represent the sentence; we can also try to use the average pooling of every token in the sentence to get the representation of a sentence. Similar for the passage representation.

Evaluate on More Dataset To show that the generalization of the proposed model, it can also evaluate on more datasets, such as NaturalQuestion (NQ) dataset (Kwiatkowski et al., 2019). Although the NQ dataset does not have annotated support sentences, the sentence which contains the answer can be taken as the support sentence to train the sentence selection model. It is worth mentioning that in the HotpotQA dataset, there are multiple support sentences while the NQ only has one, thus, if the Two-in-One model is trained on a single dataset, then one model might not generalize well to other dataset. A simple solution might be to train the Two-in-One model on multi-datasets.

Zero-shot Testing It is also interesting to see if Two-in-One model can generalize better to unseen domains than simple baselines without any fine-tuning. To verify this, we can compare the Two-in-One model and baselines models trained on the HotpotQA dataset to other datasets.

7 Conclusion

In this work, we present a simple model, Two-in-One, to rank passage and classify sentence together. By jointly training with passage ranking and sentence selection, the model is capable of capturing the correlation between passages and sentences. We show the effectiveness of our proposed framework by evaluating the model performance on the HotpotQA datasets, concluding that jointly modeling passage ranking and sentence selection is beneficial for the task of OBQA. Compared to the existing QA systems, our model, with fewer parameters and more green than previous models, can achieve competitive results. We also propose multiple future directions to improve our model such as exploring the relationship among passages, supporting sentences, and answers in modeling and generalizing our method on more datasets.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pretraining text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yang Deng, Wai Lam, Yuexiang Xie, Daoyuan Chen, Yaliang Li, Min Yang, and Ying Shen. 2020. Joint learning of answer selection and answer summary generation in community question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7651–7658. AAAI Press.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Kathryn Mazaitis, and William W. Cohen. 2017. Quasar: Datasets for question answering by search and reading. *ArXiv preprint*, abs/1707.03904.

Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical graph network for multi-hop question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8823–8838, Online. Association for Computational Linguistics.

Dirk Groeneveld, Tushar Khot, Mausam, and Ashish Sabharwal. 2020. A simple yet strong pipeline for HotpotQA. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8839–8845, Online. Association for Computational Linguistics.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2020. QASC: A dataset for question answering via sentence composition. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8082–8090. AAAI Press.

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.

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*, abs/1907.11692.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566, Hong Kong, China. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

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

Roy Schwartz, Jesse Dodge, Noah A Smith, and Oren Etzioni. 2020. Green ai. *Communications of the ACM*, 63(12):54–63.

Yibo Sun, Duyu Tang, Nan Duan, Tao Qin, Shujie Liu, Zhao Yan, Ming Zhou, Yuanhua Lv, Wenpeng Yin, Xiaocheng Feng, et al. 2019. Joint learning of question answering and question generation. *IEEE Transactions on Knowledge and Data Engineering*, 32(5):971–982.

Ming Tu, Kevin Huang, Guangtao Wang, Jing Huang, X. He, and Bowen Zhou. 2019. Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents. *ArXiv preprint*, abs/1911.00484.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

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.