Neural Retrieval for Question Answering with Cross-Attention Supervised Data Augmentation

Yinfei Yang[a], **Ning Jin**[b], **Kuo Lin**[b], **Mandy Guo**[a], **Daniel Cer**[a]

[a]Google Research
Mountain View, CA, USA

[b]Google Cloud AI
Sunnyvale, CA, USA

Abstract

Early fusion models with cross-attention have shown better-than-human performance on some question answer benchmarks, while it is a poor fit for retrieval since it prevents pre-computation of the answer representations. We present a supervised data mining method using an accurate early fusion model to improve the training of an efficient late fusion retrieval model. We first train an accurate classification model with cross-attention between questions and answers. The cross-attention model is then used to annotate additional passages in order to generate weighted training examples for a neural retrieval model. The resulting retrieval model with additional data significantly outperforms retrieval models directly trained with gold annotations on Precision at N (P@N) and Mean Reciprocal Rank (MRR).

1 Introduction

Open domain question answering (QA) involves finding answers to questions from an open corpus (Surdeanu et al., 2008; Yang et al., 2015; Chen et al., 2017; Ahmad et al., 2019). The task has led to a growing interest in scalable end-to-end retrieval systems for question answering.

When QA is formulated as a reading comprehension task, cross-attention models like BERT (Devlin et al., 2019) have achieved *better-than-human* performance on benchmarks such as the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016). Cross-attention models are especially well suited for problems involving comparisons between paired textual inputs, as they provide *early* fusion of fine-grained information within the pair. This encourages careful comparison and integration of details across and within the two texts.

However, early fusion across questions and answers is a poor fit for retrieval, since it prevents pre-computation of the answer representations. Rather, neural retrieval models independently compute embeddings for questions and answers, typically using dual encoders for fast scalable search (Henderson et al., 2017; Gillick et al., 2018; Yang et al., 2019b; Karpukhin et al., 2020). Using dual encoders results in *late* fusion within a shared embedding space.

For machine reading, early fusion using cross-attention introduces an inductive bias to compare fine grained text spans within questions and answers. This inductive bias is missing from the single dot-product scoring operation of dual encoder retrieval models. Thus, late fusion is expected to require more training data to learn the necessary representations for fine grained comparisons.

To support learning improved representations for retrieval, we explore a supervised data augmentation approach leveraging a complex classification model with cross-attention between question-answer pairs. Given gold question passage pairs, we first train a cross-attention classification model as the supervisor. Then any collection of questions can be used to mine potential question passage pairs under the supervision of the cross-attention model. The retrieval model training benefits from additional training pairs annotated with the graded predictions from the cross-attention model augmenting the existing gold data. Experiments on MultiReQA-SQuAD and MultiReQA-NQ establish significant improvements on Precision at N (P@N) and Mean Reciprocal Rank (MRR).

The supervised mining approach is closely connected to the recently studied hard negative mining for neural retrieval models (Xiong et al., 2020; Lu et al., 2020). The key differences is that the proposed approach finds the positive training examples, while the negative mining approaches find the negative examples for training. The two approaches are complementary and can be combined.

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Short Papers), pages 263–268
August 1–6, 2021.

2 Neural Passage Retrieval for Open Domain Question Answering

Open domain question answering systems usually follow a two-step approach: first retrieve question relevant passages, and then scan the returned text to identify the answer span using a reading comprehension model (Jurafsky and Martin, 2018; Kratzwald and Feuerriegel, 2018; Yang et al., 2019a). Prior work has focused on the answer span annotation task and has even achieved super human performance on some datasets. However, the evaluations implicitly assume the trivial availability of passages for each question that are likely to contain the correct answer. While the retrieval task can be approached using traditional keyword based retrieval methods such as BM25, there is a growing interest in developing more sophisticated neural retrieval methods (Lee et al., 2019; Guu et al., 2020; Karpukhin et al., 2020).

3 Retrieval Question-Answering (ReQA)

Ahmad et al. (2019) introduced Retrieval Question-Answering (ReQA), a task that has been rapidly adopted by the community (Guo et al., 2020; Chang et al., 2020; Ma et al., 2020; Zhao and Lee, 2020; Roy et al., 2020). Given a question, the task is to retrieve the answer sentence from a corpus of candidates. ReQA provides direct evaluation of retrieval, independent of span annotation. Compare to Open Domain QA, ReQA focuses on evaluating the retrieval component and, by construction, avoids the need for span annotation.

We explore the proposed approach on MultiReQA-NQ and MultiReQA-SQuAD (Guo et al., 2020).[1] MultiReQA (Guo et al., 2020) established standardized training / dev / test splits. Statistics for each tasks are listed in Table 1.

Dataset	Training Pairs	Test	
		Questions	Candidates
NQ	106,521	4,131	22,118
SQuAD	87,133	10,485	10,642

Table 1: Statistics of MutiReQA NQ and SQuAD tasks: # of training pairs, # of questions, # of candidates.

4 Methodology

In this section we describe the proposed approach using a neural retrieval model augmented with supervised data mining. Figure 2 illustrates our approach using a cross-attention classifier to supervise the data augmentation process for training a retrieval model. After training the cross-attention model, we retrieve additional potential answers to questions using an off-the-shelf retrieval system[2]. The predicted scores from our classifier with cross-attention are then used to weight and filter the retrieved candidates with positive examples serving as additional training data for the dual encoder based retrieval model.

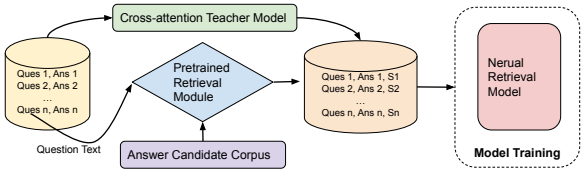

Figure 1: Use of a cross-attention model for the supervised mining of additional QA pairs. Our accurate cross-attention model supervises the mining process by identifying new previously unannotated positive pairs. Mined QA pairs augment the original training data for the dual encoder based neural passage retrieval model.

4.1 BERT Classification Model

Cross-attention models like BERT are often used for re-ranking after retrieval and can significantly improve performance as they allow for fine-grained interactions between paired inputs (Nogueira et al., 2019; Han et al., 2020). Here we formalize a binary classification task for predicting question answer relatedness. We use the question-answer pairs from the training set as our positive examples. Negatives are sampled for each question using the following strategies with a 1:1:1 ratio: *(1) A sentence from the top 10 nearest neighbors returned by a term based BM25 (Robertson and Zaragoza, 2009) over a sentence pool containing all supporting documents in a corpus. (2) A sentence from the top 10 nearest neighbors using the Universal Sentence Encoder - QA (USE-QA) (Yang et al., 2019b). (3) A sentence randomly sampled from its supporting documents, excluding the question's gold answer.* The sampled non-answer sentences are paired with their questions as negative examples. A BERT model is fine-tuned following the default setup from the Devlin et al. (2019).

[1] https://github.com/google-research-datasets/MultiReQA

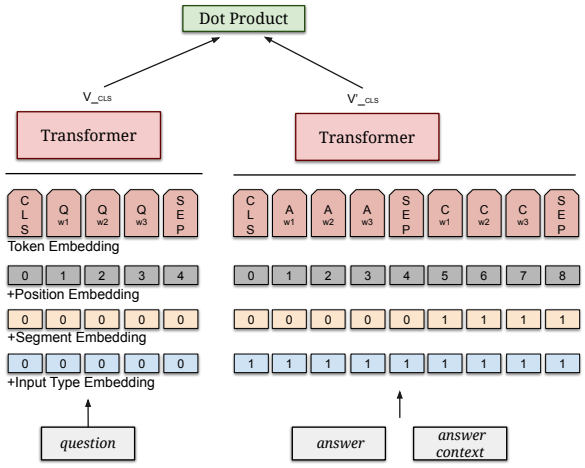

Figure 2: The BERT dual encoder architecture. The answer and context are concatenated and fed into the answer encoder. Figure from (Guo et al., 2020).

4.2 Dual-Encoder Retrieval Model

We follow Guo et al. (2020) and employ a BERT based dual-encoder model for retrieval. The model architecture is illustrated in figure **??**. The dual-encoder model critically differs from the cross-attention model in that there is no early interactions (cross-attention) between the question and answer. The resulting independent encodings are only combined in the final dot-product scoring a pair. The same BERT encoder is used for questions and answers with the output of the CLS token taken as the output encoding. For answers, the answer and context are concatenated and segmented using the segment IDs from the original BERT model. A learned *input type embedding* is added to each input token representation to distinguish questions and answers within the encoding model.

The BERT dual-encoder model can be fine-tuned using the in batch sampled softmax loss (Gillick et al., 2018):

$$\mathcal{J} = \sum_{(x,y)\in Batch} \frac{e^{\phi(x,y)}}{\sum_{\bar{y}\in\mathcal{Y}} e^{\phi(x,\bar{y})}} \quad (1)$$

Where x is the question, y is the correct answer, $\mathcal{Y}$ is all answers in the same batch that are used as sampled negatives, and $\phi(x, y)$ is the dot product of question and answer representations. Note that the dot product is scaled by X100 during training, which is a critical component when applying l_2 normalization to the embeddings.

[2]Note the approach can also be applied to any collection of questions, even for those without ground truth answers.

4.3 Mining Augmented Training Pairs

We create an augmented training set for the retrieval model using our cross-attention based QA model. For each question in the training set, we employ USE-QA to mine the top 10 nearest neighbors from the entire training set, and then remove those retrieved pairs which are true positives. Next the cross-attention based QA model is used to score the retrieved pairs. The dual-encoder based neural retrieval model is then trained on the combination the additional scored positive pairs and the original QA pairs from the training set. The original pairs are assigned a score 1.

4.4 Weighted In-batch Softmax for Dual-Encoder Retrieval Model

The neural retrieval model is trained using the batch negative sampling loss (Gillick et al., 2018) in equation 2. We modify the standard formulation to include a weight, $w(x, y)$, for each pair.

$$\mathcal{J}' = \sum_{(x,y)\in Batch} w(x,y) \frac{e^{\phi(x,y)}}{\sum_{\bar{y}\in\mathcal{Y}} e^{\phi(x,\bar{y})}} \quad (2)$$

We set $w(x, y)$ to 1 if (x, y) is a ground truth positive pair and $p(x, y)^2$, otherwise, whereby $p(x, y)$ is the probability from the cross-attention model.

5 Evaluation

In this section we evaluate the proposed approach using the MultiReQA evaluation splits for NQ and SQuAD. Models are assessed using Precision at N (P@N) and Mean Reciprocal Rank (MRR). Following the ReQA setup (Ahmad et al., 2019), we report P@N for N=$[1, 5, 10]$. P@N evaluates whether the true answer sentence appears in the top-N ranked candidates. MRR is calculated as MRR $= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank_i}$, where N is the total number of questions, and $rank_i$ is the rank of the first correct answer for the ith question.

5.1 Configurations

Our cross-attention QA models are fine-tuned from the public English BERT for 10 epochs, using a batch size of 256 and a weighted Adam optimizer with learning rate 3e-5. We experiment with both $\text{BERT}_{\text{Base}}$ and $\text{BERT}_{\text{Large}}$. All hyper-parameters are set using a dev set split out from the training data (10%). When mining for silver data, we only keep candidate pairs with positive cross-attention QA model scores (≥ 0.5).

Models	NQ		SQuAD	
	ACC	AUC-PR	ACC	AUC-PR
Majority	73.7	–	74.8	–
$\text{BERT}_{\text{dual\_encoder}}$	75.8	49.3	80.3	62.0
(x-attn) $\text{BERT}_{\text{Base}}$	84.3	92.8	92.6	96.5
(x-attn) $\text{BERT}_{\text{Large}}$	84.9	93.5	93.6	97.1

Table 2: Accuracy (ACC) and area under the precision-recall curve (AUC-PR) for the classification task. **Majority** is a simple baseline that always predicts false. **(x-attn)** indicates cross-attention QA models.

The $\text{BERT}_{\text{Base}}$ model is used to initialize the dual encoder retrieval model. During training we use a batch size of 64, and a weighted Adam optimizer with learning rate 1e-4. The maximum input length is set to 96 for questions and 384 for answers. Models are trained for 200 epochs. The embeddings are l_2 normalized. Hyper-parameters are manually tuned on a held out development set.

5.2 Performance for the Classification Task

The classification data created using the method from section 4.1 contains a total of 531k and 469k training examples for NQ and SQuAD, respectively. Test sets extracted from the SQuAD and NQ test splits contain 15k and 41k examples.[3]

Table 2 provides the performance of the cross-attention models, compared to a majority baseline which always predict false and a $\text{BERT}_{\text{dual\_encoder}}$ retrieval model without any mined examples that uses cosine similarity for prediction. Cross-attention based models outperform the baselines by a wide margin,[4] with $\text{BERT}_{\text{Large}}$ achieving the highest performance on all metrics. This is consistent with our hypothesis that early fusion models outperform late fusion based retrieval models. Both models achieve better performance on SQuAD than NQ. The SQuAD task has higher token overlap, as described in section 3, making the task somewhat easier. We use the $\textbf{BERT}_{\textbf{Large}}$ model to supervise the data augmentation in the next section.

5.3 Mined Examples

We mined the SQuAD and NQ training data to construct additional QA pairs. After collecting and scoring addition pairs using the method described in section 4.3, we obtained 53% (56,148) and 12% (10,198) more examples for NQ and SQuAD, respectively. Table 4 illustrated the examples retrieved by USE-QA and predicted as positive examples by our cross-attention QA classification model. Both examples are clear positive QA pairs.

Much less data is mined for SQuAD then NQ. We believe it is because of the way SQuAD was created, whereby workers write the questions based on the content of a particular article. The resulting questions are much more specific and biased toward a particular question types, e.g. *what* questions Ahmad et al. (2019). Additionally, the candidate pool for SQuAD is only half that of NQ, resulting in questions having fewer opportunities to be matched to good additional answers.

5.4 Results on the Retrieval QA

Table 3 gives P@N and MRR@100 for retrieval models on MultiReQA-SQuAD and MultiReQA-NQ. The first two rows show the result from two simple baselines: BM25 (Robertson and Zaragoza, 2009), USE-QA, and $\text{USE-QA}_{\text{finetune}}$ reported by Guo et al. (2020). BM25 remains a strong baseline, especially with 62.8% P@1 and 70.5% MRR for SQuAD. BM25's performance on NQ is much lower, as there is much less token overlap between NQ questions and answers. USE-QA matches the performance of BM25 on NQ but performs worse on SQuAD.[5] $\text{BERT}_{\text{dual\_encoder}}$ performs well compared to other baselines, especially on NQ with a +6.6 point improvement compared to the $\text{USE-QA}_{\text{finetune}}$ model.[6] Its P@1 on SQuAD performs better than USE-QA and BM25, but -3.1 points MRR worse than $\text{USQ-QA}_{\text{finetune}}$. On average, $\text{BERT}_{\text{dual\_encoder}}$ is the best among those baselines.

Performance improves by a large margin using augmented training data from our cross-attention QA model, obtaining a +8.6 and +7.0 improvement on NQ P@1 and MRR. Compare to NQ, the improvement on SQuAD is rather marginal. The augmented $\text{BERT}_{\text{dual\_encoder}}$ retrieval model only achieves slightly improved performance on SQuAD, with +1 points for both P@1 and MRR. As discussed in section 5.3, we mine much less data from SQuAD compare to NQ, with only 10% more data than the original training set. As demonstrated by the strong BM25 performance and shown in (Guo et al., 2020), the SQuAD QA pairs have high token overlap between question and answers,

[3] The positive / negative ratio is roughly 1:3.

[4] The poor performance of $\text{BERT}_{\text{dual\_encoder}}$ is also aligned with the hypothesis that cosine similarity score is not a globally consistent measurement of how good a pair (Guo et al., 2018).

[5] USE-QA can be fine-tuned, which usually significantly outperforms the default USE-QA model (Guo et al., 2020).

[6] Our $\text{Bert}_{\text{dual\_encoder}}$ performs better than the one reported in Guo et al. (2020), likely due to additional training epochs.

Models	NQ				SQuAD			
	P@1	P@5	P@10	MRR	P@1	P@5	P@10	MRR
BM25	24.7	–	–	36.6	62.8	–	–	70.5
USE-QA	24.7	–	–	34.7	51.0	–	–	62.1
USE-QA$_{\text{finetune}}$	38.0	–	–	52.3	**66.8**	–	–	**75.9**
BERT$_{\text{dual_encoder}}$	44.7	77.1	85.1	58.9	62.8	85.4	91.0	72.8
BERT$_{\text{dual_encoder}}$ Augmented	**53.3**	**82.3**	**88.5**	**65.9**	63.8	86.1	91.6	73.7

Table 3: Precision at N(P@N) (%) N=$[1, 5, 10]$ and Mean Reciprocal Rank (MRR) (%) on the MultiReQA tasks.

Score	Silver QA Pair
0.92	**Q:** what are the names of the two old muppets in the balcony that heckle everyone ? **A:** Statler and Waldorf are a pair of Muppet characters known for their cantankerous opinions and shared penchant for heckling.
0.90	**Q:** where the phrase dressed to the nines come from **A:** It appears in book six of Jean - Jacques Rousseau 's Confessions , his autobiography ...

Table 4: Mined positive examples identified using our cross-attention QA classification model.

minimizing the advantage of the neural methods in capturing more complex semantic relationships.

Effectiveness of Weighted Softmax. We further experimented the Retrieval QA tasks using the model with the non-modified softmax using the augmented data. All other configurations are keep the same. The MRR of the model using non-modified softmax is 60.1 on MultiReQA-NQ and 71.9 on MultiReQA-SQuAD, which are much worse than the model using weighted softmax. This result indicates the weighted softmax is important for the proposed approach.

6 Conclusion

In this paper, we propose a novel approach for making use of an early fusion classification model to improve late fusion retrieval models. The early fusion model is used for data mining to augment the training set for the late fusion model. The proposed approach mines 53% (56,148) and 12% (10,198) more examples for MultiRQA-NQ and MultiRQA-SQuAD, respectively. Compared to the models directly trained with gold annotations, the resulting retrieval models improve +8.6% and +1.0% P@1 on NQ and SQuAD respectively. The current pipeline assumes there exists annotated in-domain question answer pairs to train the cross-attention model. With a strong general purpose cross-attention model, our method could be modified to train in-domain retrieval models without gold data. We leave this to the future work.

References

Amin Ahmad, Noah Constant, Yinfei Yang, and Daniel Cer. 2019. ReQA: An evaluation for end-to-end answer retrieval models. In *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pages 137–146, Hong Kong, China. Association for Computational Linguistics.

Wei-Cheng Chang, Felix X. Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

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.

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.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 165–176, Brussels, Belgium. Association for Computational Linguistics.

Mandy Guo, Yinfei Yang, Daniel Cer, Qinlan Shen, and Noah Constant. 2020. MultiReQA: A cross-domain evaluation for retrieval question answering models. *arXiv preprint arXiv:2005.02507*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*.

Matthew Henderson, Rami Al-Rfou, Brian Strope, Yun-Hsuan Sung, László Lukács, Ruiqi Guo, Sanjiv Kumar, Balint Miklos, and Ray Kurzweil. 2017. Efficient natural language response suggestion for smart reply. *arXiv preprint arXiv:1705.00652*.

Daniel Jurafsky and James H. Martin. 2018. *Speech and Language Processing (3rd Edition, in draft)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

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.

Bernhard Kratzwald and Stefan Feuerriegel. 2018. Adaptive document retrieval for deep question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 576–581, Brussels, Belgium. Association for Computational Linguistics.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *ArXiv*, abs/1906.00300.

Jing Lu, Gustavo Hernandez Abrego, Ji Ma, Jianmo Ni, and Yinfei Yang. 2020. Neural passage retrieval with improved negative contrast.

Ji Ma, Ivan Korotkov, Yinfei Yang, Keith Hall, and Ryan McDonald. 2020. Zero-shot neural retrieval via domain-targeted synthetic query generation.

Rodrigo Nogueira, Wei Yang, Kyunghyun Cho, and Jimmy Lin. 2019. Multi-stage document ranking with bert. *arXiv preprint arXiv:1910.14424*.

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 Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Uma Roy, Noah Constant, Rami Al-Rfou, Aditya Barua, Aaron Phillips, and Yinfei Yang. 2020. Lareqa: Language-agnostic answer retrieval from a multilingual pool.

Mihai Surdeanu, Massimiliano Ciaramita, and Hugo Zaragoza. 2008. Learning to rank answers on large online QA collections. In *Proceedings of ACL-08: HLT*, pages 719–727, Columbus, Ohio. Association for Computational Linguistics.

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

Wei Yang, Rui Qiao, Haocheng Qin, Amy Sun, Luchen Tan, Kun Xiong, and Ming Li. 2019a. End-to-end neural context reconstruction in Chinese dialogue. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 68–76, Florence, Italy. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019b. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Tianchang Zhao and Kyusong Lee. 2020. Talk to papers: Bringing neural question answering to academic search. *arXiv preprint arXiv:2004.02002*.