Complementary Evidence Identification in Open-Domain Question Answering

Xiangyang Mou
Rensselaer Polytechnic Institute
moux4@rpi.edu

Mo Yu
IBM Research
yum@us.ibm.com

Shiyu Chang
MIT-IBM Watson AI Lab
shiyu.chang@ibm.com

Yufei Feng
Queen's University
feng.yufei@queensu.ca

Li Zhang
Amazon Web Services
lzhangza@amazon.com

Hui Su
Fidelity
Hui.Su@fmr.com

Abstract

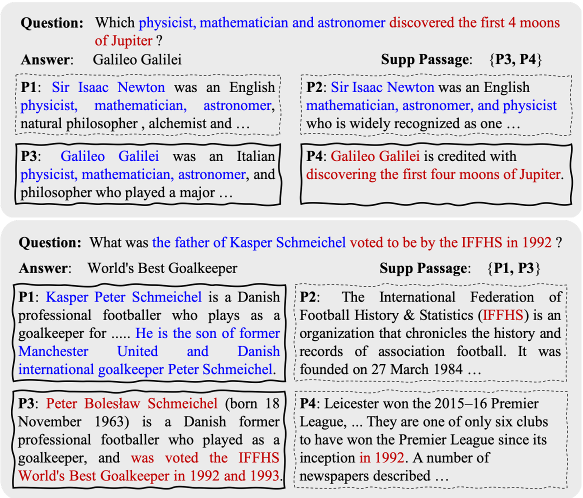
This paper proposes a new problem of complementary evidence identification for open-domain question answering (QA). The problem aims to efficiently find a small set of passages that covers full evidence from multiple aspects as to answer a complex question. To this end, we proposes a method that learns vector representations of passages and models the sufficiency and diversity within the selected set, in addition to the relevance between the question and passages. Our experiments demonstrate that our method considers the dependence within the supporting evidence and significantly improves the accuracy of complementary evidence selection in QA domain.

1 Introduction

In recent years, significant progress has been made in the field of open-domain question answering (Chen et al., 2017; Wang et al., 2017, 2018; Clark and Gardner, 2018; Min et al., 2018; Asai et al., 2019). Very recently, some works turn to deal with a more challenging task of asking complex questions (Welbl et al., 2018; Clark et al., 2018; Yang et al., 2018) from the open-domain text corpus. In the open-domain scenario, one critical challenge raised by complex questions is that each question may require multiple pieces of evidence to get the right answer, while the evidence usually scatters in different passages. Examples in Figure 1 shows two types of questions that require evidence from multiple passages.

To deal with the challenging multi-evidence questions, an open-domain QA system should be able to (1) efficiently retrieve a small number of passages that cover the full evidence; and (2) accurately extract the answer by jointly considering the candidate evidence passages. While there have been several prior works in the latter direction (Wang et al., 2017; Clark and Gardner, 2018; Lin et al., 2018), the solutions to the first problem still rely on traditional or neural information retrieval (IR) approaches, which solely measure the relevance between the question and each individual paragraph, and will highly possibly put the wrong evidence to the top.[1] For example in Figure 1 (top), **P1** and **P2** are two candidate evidence passages that are closely related to the question but only cover the same unilateral fact required by the question, therefore leading us to the wrong answer *Newton*.

This paper formulates a new problem of **complementary evidence identification** for answering complex questions. The key idea is to consider the problem as measuring the properties of the selected passages, more than the individual relevance. Specifically, we hope the selected passages can serve as a set of spanning bases that supports the



Question: Which physicist, mathematician and astronomer discovered the first 4 moons of Jupiter ?
Answer: Galileo Galilei **Supp Passage:** {P3, P4}

P1: Sir Isaac Newton was an English physicist, mathematician, astronomer, natural philosopher , alchemist and …

P2: Sir Isaac Newton was an English mathematician, astronomer, and physicist who is widely recognized as one …

P3: Galileo Galilei was an Italian physicist, mathematician, astronomer, and philosopher who played a major …

P4: Galileo Galilei is credited with discovering the first four moons of Jupiter.

Question: What was the father of Kasper Schmeichel voted to be by the IFFHS in 1992 ?
Answer: World's Best Goalkeeper **Supp Passage:** {P1, P3}

P1: Kasper Peter Schmeichel is a Danish professional footballer who plays as a goalkeeper for ..... He is the son of former Manchester United and Danish international goalkeeper Peter Schmeichel.

P2: The International Federation of Football History & Statistics (IFFHS) is an organization that chronicles the history and records of association football. It was founded on 27 March 1984 …

P3: Peter Bolesław Schmeichel (born 18 November 1963) is a Danish former professional footballer who played as a goalkeeper, and was voted the IFFHS World's Best Goalkeeper in 1992 and 1993.

P4: Leicester won the 2015–16 Premier League, ... They are one of only six clubs to have won the Premier League since its inception in 1992. A number of newspapers described …

Figure 1: Examples of complex questions involving two facts of a person. Different facts are color-coded. **P#** are all relevant passages, while only the ones with solid-line boxes are the true supporting passages.

[1] (Min et al., 2019) pointed out the shortcut problem in multi-hop QA. However, as some works (Wang et al., 2019) show that even a better designed multi-hop model can still benefit from full evidence in such situation.

question. The selected passage set thus should satisfy the properties of (1)*relevancy*, i.e., they should be closely related to the question; (2) *diversity*, i.e., they should cover diverse information given the coverage property is satisfied; (3) *compactness*, i.e., the number of passages to satisfy the above properties should be minimal. With these three defined properties, we hope to both improve the selective accuracy and encourage the interpretability of the evidence identification. Note that complementary evidence identification in QA is different from Search Result Diversification (SRD) in IR on their requirement of compactness. The size of the selected set is constrained in QA tasks by the capability of downstream reasoning models and practically needs to be a small value, whereas it is not the case in SRD.

To achieve the above goals, a straightforward approach is to train a model that evaluates each subset of the candidate passages, e.g., by concatenating passages in any subsets. However, this approach is highly inefficient since it requires to encode $O(K^L)$ passage subsets, where K is the total number of candidates and L is the maximum size of subsets. Thus, a practical complementary evidence identification method needs to be computationally efficient. This is especially critical when we use heavy models like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), where passage encoding is time and memory consuming.

To this end, we propose an efficient method to select a set of spanning passages that is sufficient and diverse. The core idea is to represent questions and passages in a vector space and define the measures of our criterion in the vector space. For example, in the vector space, sufficiency can be defined as a similarity between the question vector and the sum of selected passage vectors, measured by a cosine function with a higher score indicating a closer similarity; and diversity can be defined as ℓ_1 distance between each pair of passages. By properly training the passage encoder with a loss function derived by the above terms, we expect the resulted vector space satisfies the property that the complementary evidence passages lead to large scores. In addition, our method only encodes each passage in the candidate set once, which is more efficient than the naive solution mentioned above. To evaluate the proposed method, we use the multi-hop QA dataset HotpotQA (the full wiki setting) since the ground-truth of evidence passages are provided. Experiments show that our method significantly improves the accuracy of complementary evidence selection.

2 Proposed Method

2.1 Task Definition

Given a question q and a mixture set of paragraphs $\mathcal{P} = \mathcal{P}^+ \cup \mathcal{P}^-$ with some paragraphs $p \in \mathcal{P}^+$ relevant to q and some $p \in \mathcal{P}^-$ irrelevant. Our goal is to select a small subset of paragraphs $\mathcal{P}_{sel} \subset \mathcal{P}$, such that every $p \in \mathcal{P}_{sel}$ satisfies $p \in \mathcal{P}^+$ (relevancy), and all $p \in \mathcal{P}_{sel}$ can jointly cover all the information asked by q (complementary). The off-the-shelf models select relevant paragraphs independently, thus usually cannot deal with the complementary property. The inner dependency among the selected $\mathcal{P}_{sel}$ needs to be considered, which will be modeled in the remaining of the section.

2.2 Model and Training

Vector Space Modeling We apply BERT model to estimate the likelihood of a paragraph p being the supporting evidence to the question q, denoted as $P(p|q)$. Let q and p_i denote the input texts of a question and a passage. We feed q and the concatenation of q and p_i into the BERT model, and use the hidden states of the last layer to represent q and p_i in vector space, denoted as $\boldsymbol{q}$ and $\boldsymbol{p_i}$ respectively. A fully connected layer $f(\cdot)$ followed by sigmoid activation is added to the end of the BERT model, and outputs a scalar $P(p_i|q)$ to estimate how relevant the paragraph p_i is to the question. Note that in our implementation $\boldsymbol{p}_i$ is based on both q and p_i, but we omit the condition on q for simplicity.

Complementary Conditions Previous works extract evidence paragraphs according to $P(p|q)$, which is estimated on each passage separately without considering the dependency among selected paragraphs. To extract complementary evidence, we propose that the selected passages $\mathcal{P}_{sel}$ should satisfy the following conditions that intuitively encourage each selected passage to be a basis to support the question:

• **Relevancy:** $\mathcal{P}_{sel}$ should have a high probability of $\sum_{p_i \in \mathcal{P}_{sel}} P(p_i|q)$;

• **Diversity:** $\mathcal{P}_{sel}$ should cover passages as diverse as possible, which can be measured by the average distance between any pairs in $\mathcal{P}_{sel}$, e.g., maximizing $\sum_{i,j \in \{i,j|p_i,p_j \in \mathcal{P}_{sel}, i \neq j\}} \ell_1(\boldsymbol{p}_i, \boldsymbol{p}_j)$. Here

$\ell_1(\cdot,\cdot)$ denotes L_1 distance;

• **Compactness:** $\mathcal{P}_{sel}$ should optimize the aforementioned conditions while the size being minimal. In this work we constrain the compactness by fixing $|\mathcal{P}_{sel}|$ and meanwhile maximizing $cos(\sum_{i\in\{i|p_i\in\mathcal{P}_{sel}\}}\boldsymbol{p}_i,\boldsymbol{q})$. We use $\cos(\cdot,\cdot)$ to encourage the collection of evidence covers what needed by the question.

Training with Complementary Regularization We propose a new supervised training objective to learn the BERT encoder for QA that optimizes the previous conditions. Note that in this work we assume a set of labeled training examples are available, i.e., the ground truth annotations contain complementary supporting paragraphs. Recently there was a growing in such datasets (Yang et al., 2018; Yao et al., 2019), due to the increasing interest in model explainability. Also, such supervision signals can also be obtained with distant supervision.

For each training instance $(q,\mathcal{P})$, we define

$$\{\boldsymbol{p}_i\}^+ = \{\boldsymbol{p}_i\}, \quad \forall i \in \{i|p_i \in \mathcal{P}^+\} \tag{1}$$

$$\{\boldsymbol{p}_i\}^- = \{\boldsymbol{p}_i\}, \quad \forall i \in \{i|p_i \in \mathcal{P}^-\} \tag{2}$$

$$\{\boldsymbol{p}_i\} = \{\boldsymbol{p}_i\}^+ \cup \{\boldsymbol{p}_i\}^- \tag{3}$$

Denoting $y_{p_i} = 1$ if $p_i \in \mathcal{P}^+$ and $y_{p_i} = 0$ if $p_i \in \mathcal{P}^-$, we have the following training objective function:

$$\begin{aligned}\mathcal{L}(\{\boldsymbol{p}_i\};\boldsymbol{q};y) &= \mathcal{L}_{sup}(\{\boldsymbol{p}_i\};\boldsymbol{q};y)\\ &+ \alpha\mathcal{L}_d(\{\boldsymbol{p}_i\}^+) + \beta\mathcal{L}_c(\{\boldsymbol{p}_i\};\boldsymbol{q};y)\end{aligned} \tag{4}$$

where

$$\mathcal{L}_{sup}(\{\boldsymbol{p}_i\};\boldsymbol{q};y) = -\sum_i y_{p_i}\log(f(\boldsymbol{p}_i)), \tag{5}$$

$$\mathcal{L}_d(\{\boldsymbol{p}_i\}^+) = \sum_{\boldsymbol{p_i},\boldsymbol{p_j},i\neq j}(1-\ell_1(\boldsymbol{p_i},\boldsymbol{p_j})). \tag{6}$$

$$\mathcal{L}_c(\{\boldsymbol{p}_i\};\boldsymbol{q};y) = \begin{cases} 1-\cos(\boldsymbol{q},\sum_i \boldsymbol{p}_i), \\ \quad \text{if} \quad \Pi_{p_i}y_{p_i} = 1 \\ \max(0,\cos(\boldsymbol{q},\sum_i \boldsymbol{p}_i)-\gamma), \\ \quad \text{if} \quad \Pi_{p_i}y_{p_i} = 0 \end{cases} \tag{7}$$

where α and β are the hyperparameter weights and $\ell_1(\cdot,\cdot)$ denotes L1 loss between two input vectors. Eq 5 is the cross-entropy loss corresponding to relevance condition; Eq 6 regularizes the diversity condition; Eq 7 is the cosine-embedding loss[2] for the compactness condition and $\gamma > 0$ is the margin to encourage data samples with better question coverage.

2.3 Inference via Beam Search

Score Function During inference, we use the following score function to find the best paragraph combination:

$$\begin{aligned}g(\mathcal{P}_{sel};q;\{\boldsymbol{p}_i\}) &= \sum_{p_i}P(p_i|q) + \alpha\cos(\sum_{\boldsymbol{p_i}}\boldsymbol{p}_i,\boldsymbol{q})\\ &+\beta\sum_{\boldsymbol{p_i},\boldsymbol{p_j},i\neq j}\ell_1(\boldsymbol{p_i},\boldsymbol{p_j})\end{aligned} \tag{8}$$

where α and β are hyperparameters similar to Eq 4. Note that our approach requires to encode each passage in $\mathcal{P}$ only once for each question, resulting in an $O(K)$ time complexity of encoding ($K = |\mathcal{P}|$); and the subset selection is performed in the vector space, which is much more efficient than selecting subsets before encoding.

Beam Search In a real-world application, there is usually a large candidate set of $\mathcal{P}$, e.g., retrieved passages for q via a traditional IR system. Our algorithm requires $O(K)$ time encoding, and $O(K^L)$ time scoring in vector space when ranking all the combinations in L candidates. Thus when K becomes large, it is still inefficient even when $L = 2$. We resort to beam search to deal with scenarios with large Ks. The details can be found in Appendix A.

3 Experiments

3.1 Settings

Datasets Considering the prerequisite of sentence-level evidence annotations, we evaluate our approach on two datasets, a synthetic dataset **MNLI-12** and a real application **HotpotQA-50**. Data sampling is detailed in Appendix B.

• **MNLI-12** is constructed based on the textual entailment dataset MNLI (Williams et al., 2018), in order to verify the ability of our method in finding complementary evidence. In original MNLI, each premise sentence corresponds to three hypotheses sentences: entailment, neutral and contradiction. To generate complementary pairs for each premise sentence, we split each hypothesis sentence into two segments. The goal is to find the segment combination that entails premise sentence, and our dataset, by definition, ensures that only the combination of two segments from the entailment hypothesis can entail the premise, not any of its subset or other combinations. The original train/dev/test splits from MNLI are used.

[2] Refer to CosineEmbeddingLoss in PyTorch.

• **HotpotQA-50** is based on the open-domain setting of the multi-hop QA benchmark HotpotQA (Yang et al., 2018). The original task requires to find evidence passages from abstract paragraphs of all Wikipedia pages to support a multi-hop question. For each q, we collect 50 relevant passages based on bigram BM25 (Godbole et al., 2019). Two positive evidence passages to each question are provided by human annotators as the ground truth. Note that there is no guarantee that $\mathcal{P}_{50}$ covers both evidence passages here. We use the original development set from HotpotQA as our test set and randomly split a subset from the original training set as our development set.

3.2 Settings

Baseline We compare with the BERT passage ranker (Nie et al., 2019) that is commonly used on open-domain QA including HotpotQA. The baseline uses the same BERT architecture as our approach described in Section 2.2, but is trained with only the relevancy loss (Eq 5) and therefore only consider the relevancy when selecting evidence.

We also compare the DRN model from (Harel et al., 2019) which is designed for the SRD task. Their ensemble system first finds the most relevant evidence to the given question, and then select the second diverse evidence using their score function. The major differences from our method are that (1) they train two separate models for evidence selection; (2) they do not consider the compactness among the evidences. It is worth mentioning that we replace their LSTM encoder with BERT encoder for fair comparison.

Metric During the evaluation we make each method output its top 2 ranked results[3] (i.e. the top 1 ranked pair from our method) as the prediction. The final performance is evaluated by exact match (EM), i.e., whether both true evidence passages are covered, and the F1 score on the test sets.

3.3 Results

In the experiments, we have $M = 3$, $N = 4$ for MNLI-12 and $M = 4$, $N = 5$ for HotpotQA-50 with our method. The values are selected according to development performance. We follow the settings and hyperparameters used in (Harel et al., 2019) for the DRN model. Table 1 shows the performance. The upper-bound measures how many pieces of true evidences enclosed by the complete set of candidate passages where our proposed ranker selects from. For HotpotQA dataset, we use a bi-gram BM25 ranker to collect top 50 relevant passages and build the basis for the experiments[4], which inevitably leads some of the true evidences to be filtered out and makes its upper-bound less than 100%. For the artificial MNLI-12 dataset, all the true evidences are guaranteed to be included.

System	HotpotQA-50		MNLI-12	
	EM	F1	EM	F1
Baseline Ranker	16.67	41.29	41.61	67.57
DRN + BERT	1.03	35.37	6.20	46.07
Our Method	**20.15**	**49.10**	**53.81**	**73.18**
Upper-Bound	35.49	61.08	100.00	100.00

Table 1: Model Evaluation (%). The upper-bound indicates the amount of true evidences contained by all candidate passages. The baseline ranker is a BERT ranker trained only with relevancy loss.

Table 1 shows that our method achieves significant improvements on both datasets. On HotpotQA-50, all systems have low EM scores, because of the relatively low recall of the BM25 retrieval. Only 35.49% of the samples in the test set contain both ground-truth evidence passages. On MNLI-12, the EM score is around 50%. This is mainly because the segments are usually much shorter than a paragraph, with an average length of 7 words. Therefore it is more challenging in matching the q with the p_is. Specifically, both our method and the BERT baseline surpass the DRN model on all datasets and metrics, which results from our question-conditioned passage encoding approach. Our defined vector space proves beneficial to model the complementation among the evidence with respect to a given question. The ablation study of our loss function further illustrates that the diversity and the compactness terms efficiently bring additional 20%/30% increase in EM score on two datasets and consequently raise the F1 score by about 8/6 absolute points.

Figure 2 gives examples about how our model improves over the baseline. Our method can successfully select complementary passages while the baselines only select passages that look similar to the question. A more interesting example is given at the bottom where the top-50 only covers one supporting passage. The BERT baseline selects two

[3] There is only one positive pair of evidences for each question.

[4] This is the standard setting that starts with BM25 retrieval to make the inference time efficient enough without loss of generality.

incorrect passages that cover identical part of facts required by the question and similarly the DRN baseline select a relevant evidence and an irrelevant evidence, while our method scores lower the second passage that does not bring new information, and reaches a supporting selection. A similar situation contributes to the majority of improvement on one-supporting-evidence data sample in HotpotQA-50.

Inference Speed Our beam search with score function brings slight overheads to the running time. On HotpotQA-50, it takes 1,990 milliseconds (ms) on average to obtain the embeddings of all passages for one data sample whereas our vector-based complementary selection only adds an extra 2 ms which can be negligible compared to the encoding time.

3.4 Future Work

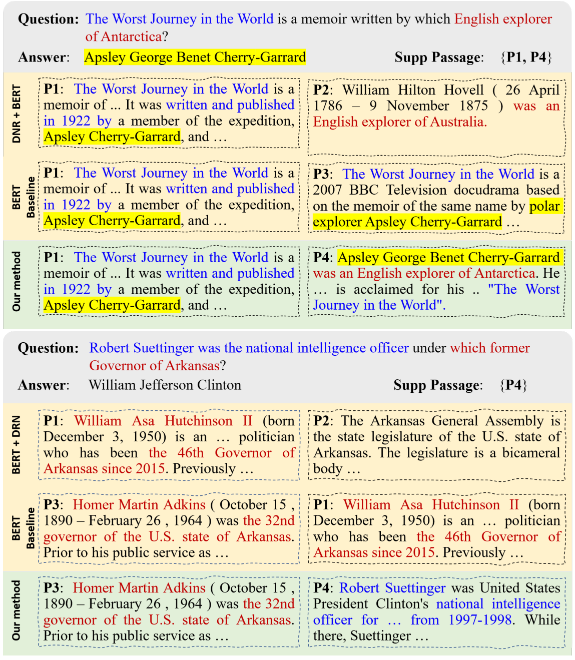
The latest dense retrieval methods (Lee et al., 2019; Karpukhin et al., 2020; Guu et al., 2020) show promising results on efficient inference on the full set of Wikipedia articles, which allows to skip the initial standard BM25 retrieval and avoid the significant loss during the pre-processing step. Our proposed approach is able to directly cooperate with these methods as we all work in the vector space. Therefore, the extension to dense retrieval can be naturally the next step of our work.

4 Conclusion

In the paper, we propose a new problem of complementary evidence identification and define the criterion of complementary evidence in vector space. We further design an algorithm and a loss function to support efficient training and inference for complementary evidence selection. Compared to the baseline, our approach improves more than 20% and remains to scale well to the computationally complex cases.

Acknowledgment

A special thank to Rensselaer-IBM Artificial Intelligence Research Collaboration (RPI-AIRC) for providing externship and other supports. This work is funded by Cognitive and Immersive Systems Lab (CISL), a collaboration between IBM and RPI, and also a center in IBM's AI Horizons Network.



Question: The Worst Journey in the World is a memoir written by which English explorer of Antarctica?
Answer: Apsley George Benet Cherry-Garrard **Supp Passage:** {**P1, P4**}

P1: The Worst Journey in the World is a memoir of ... It was written and published in 1922 by a member of the expedition, Apsley Cherry-Garrard, and …

P2: William Hilton Hovell (26 April 1786 – 9 November 1875) was an English explorer of Australia.

P1: The Worst Journey in the World is a memoir of ... It was written and published in 1922 by a member of the expedition, Apsley Cherry-Garrard, and …

P3: The Worst Journey in the World is a 2007 BBC Television docudrama based on the memoir of the same name by polar explorer Apsley Cherry-Garrard …

P1: The Worst Journey in the World is a memoir of ... It was written and published in 1922 by a member of the expedition, Apsley Cherry-Garrard, and …

P4: Apsley George Benet Cherry-Garrard was an English explorer of Antarctica. He … is acclaimed for his .. "The Worst Journey in the World".

Question: Robert Suettinger was the national intelligence officer under which former Governor of Arkansas?
Answer: William Jefferson Clinton **Supp Passage:** {**P4**}

P1: William Asa Hutchinson II (born December 3, 1950) is an … politician who has been the 46th Governor of Arkansas since 2015. Previously …

P2: The Arkansas General Assembly is the state legislature of the U.S. state of Arkansas. The legislature is a bicameral body …

P3: Homer Martin Adkins (October 15 , 1890 – February 26 , 1964) was the 32nd governor of the U.S. state of Arkansas. Prior to his public service as …

P1: William Asa Hutchinson II (born December 3, 1950) is an … politician who has been the 46th Governor of Arkansas since 2015. Previously …

P3: Homer Martin Adkins (October 15 , 1890 – February 26 , 1964) was the 32nd governor of the U.S. state of Arkansas. Prior to his public service as …

P4: Robert Suettinger was United States President Clinton's national intelligence officer for … from 1997-1998. While there, Suettinger …

Figure 2: Gain from complementary selection. In both examples, the DRN baseline first finds the most relevant evidence to the question (left) and then select a diverse one (right); the BERT baseline model selected the top-2 most relevant passages (**P1**, **P2**) to the question regardless of their complementation; whereas our model made the selection (**P1**, **P3**) with consideration of both relevance and evidence sufficiency. Note that, in the bottom example, one of the ground-truth supporting passages and the answer were excluded when building the dataset.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2019. Learning to retrieve reasoning paths over wikipedia graph for question answering. *arXiv preprint arXiv:1911.10470*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ameya Godbole, Dilip Kavarthapu, Rajarshi Das, Zhiyu Gong, Abhishek Singhal, Hamed Zamani, Mo Yu, Tian Gao, Xiaoxiao Guo, Manzil Zaheer, et al. 2019. Multi-step entity-centric information retrieval for multi-hop question answering. *arXiv preprint arXiv:1909.07598*.

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*.

Shahar Harel, Sefi Albo, Eugene Agichtein, and Kira Radinsky. 2019. Learning novelty-aware ranking of answers to complex questions. In *The World Wide Web Conference*, pages 2799–2805.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. *arXiv preprint arXiv:1906.00300*.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1736–1745.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. *arXiv preprint arXiv:1906.02900*.

Sewon Min, Victor Zhong, Richard Socher, and Caiming Xiong. 2018. Efficient and robust question answering from minimal context over documents. *arXiv preprint arXiv:1805.08092*.

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.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237.

Haoyu Wang, Mo Yu, Xiaoxiao Guo, Rajarshi Das, Wenhan Xiong, and Tian Gao. 2019. Do multi-hop readers dream of reasoning chains? *arXiv preprint arXiv:1910.14520*.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R 3: Reinforced ranker-reader for open-domain question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Shuohang Wang, Mo Yu, Jing Jiang, Wei Zhang, Xiaoxiao Guo, Shiyu Chang, Zhiguo Wang, Tim Klinger, Gerald Tesauro, and Murray Campbell. 2017. Evidence aggregation for answer re-ranking in open-domain question answering. *arXiv preprint arXiv:1711.05116*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. 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 1112–1122. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

A Complementary Evidence Selection via Beam Search

For efficient inference when $L = 2$, we start to select the top-N ($N \ll K$) most relevant passages. Then we score the combinations between each passage pair in the top-N set and another top-M set. This reduces the complexity from $O(K^2)$ to $O(MN)$. M is a hyperparameter corresponding to the beam size. In a more general setting with $L \geq 2$, we have an algorithm with the complexity of $O((L-1)MN)$ instead of $O(K^L)$, which is shown in algorithm 1.

Algorithm 1: Complementary Evidence Selection via Beam Search

```
Data: Vector representation of question ($\boldsymbol{q}$), vector representation of all the $N$ passages $\{p_n\}$ ($\{\boldsymbol{p_n}\}$); the maximum
      number of passage to select ($L$); the beam size ($M$); a vector of weights for all regularization terms $\boldsymbol{\lambda}$.
Result: The top ranked complementary passages.
/* Predict the probability $P(\boldsymbol{p_i})$ of being a supporting passage for each passage
   $\boldsymbol{p_i}$ given $\boldsymbol{q}$                                                    */
for $i \in [1, N]$ do
    $P(\boldsymbol{p_i}) \leftarrow f(\boldsymbol{q}, \boldsymbol{p_i})$;
end
Rank the passages by $P(\boldsymbol{p_i})$;
$\boldsymbol{P_{span}}$ = []
Pick $M$ passages with top $P(\boldsymbol{p_i})$ into $\boldsymbol{P_{span}}$;
for $depth \in [2, L]$ do
    $\boldsymbol{P'_{span}}$ = [] ;
    for $j \in [1, M]$ do
        /* $\boldsymbol{P_j}$ is a selected subset, $s_j$ is the corresponding score            */
        Pop the $j$-th tuple $(\boldsymbol{P_j}, s_j)$ from $\boldsymbol{P_{span}}$;
        for $n \in [1, N]$ do
            if The set $\boldsymbol{P_j} \cup \{\boldsymbol{p_n}\}$ is covered by $\boldsymbol{P'_{span}}$ then
                continue
            end
            /* $\boldsymbol{r}_n$ is the regulation increases by adding $\boldsymbol{p_n}$ to $\boldsymbol{P_j}$       */
            Put $(\boldsymbol{P_j} \cup \{\boldsymbol{p_n}\}, s_j + P(\boldsymbol{p_n}) + \boldsymbol{\lambda r}_n)$ into $\boldsymbol{P'_{span}}$;
            if More than $M$ tuples added based on $\boldsymbol{P_j}$ then
                break
            end
        end
    end
    Rank $\boldsymbol{P'_{span}}$ according to the scores;
    $\boldsymbol{P_{span}} \leftarrow \boldsymbol{P'_{span}}[1:M]$
end
Return $\boldsymbol{P_{span}}[0]$
```

B Data Sampling

MNLI-12 In original MNLI, each premise sentence P corresponds to one entailment E_P, one neutral N_P and one contradiction C_P. We take the premise P as q, and split each of its corresponding hypotheses into two segments with a random cutting point near the middle of the sentence, resulting in a total of 6 segments $\{E_P^1, E_P^2, N_P^1, N_P^2, C_P^1, C_P^2\}$. Mixing them with the 6 segments corresponding to another premise X, we can finally have $\mathcal{P}^+ = \{E_P^1, E_P^2\}$ and $\mathcal{P}^- = \{N_P^1, N_P^2, C_P^1, C_P^2, E_X^1, E_X^2, N_X^1, N_X^2, C_X^1, C_X^2\}$. Consequently, we sample one positive and eight negative pairs respectively from $\mathcal{P}^+$ and $\mathcal{P}^-$. A pair like $\{E_P^1, C_X^2\}$ is considered as negative. To ensure the segments are literally meaningful, each segment is guaranteed to be longer than 5 words.

HotpotQA In HotpotQA, the true supporting paragraphs of each question q are given. Therefore, we can easily form $\mathcal{P}^+$ and $\mathcal{P}^-$ and sample positive and negative pairs of paragraphs respectively from $\mathcal{P}^+$ and $\mathcal{P}^-$. A special pair that contains one true supporting paragraph and one non-supporting paragraph is considered as a negative pair.