# EVIDENCE AGGREGATION FOR ANSWER RE-RANKING IN OPEN-DOMAIN QUESTION ANSWERING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Very recently, it comes to be a popular approach for answering open-domain questions by first searching question-related passages, then applying reading comprehension models to extract answers. Existing works usually extract answers from single passages independently, thus not fully make use of the multiple searched passages, especially for the some questions requiring several evidences, which can appear in different passages, to be answered. The above observations raise the problem of evidence aggregation from multiple passages. In this paper, we deal with this problem as answer re-ranking. Specifically, based on the answer candidates generated from the existing state-of-the-art QA model, we propose two different re-ranking methods, strength-based and coverage-based re-rankers, which make use of the aggregated evidences from different passages to help entail the ground-truth answer for the question. Our model achieved state-of-the-arts on three public open-domain QA datasets, Quasar-T, SearchQA and the open-domain version of TriviaQA, with about 8% improvement on the former two datasets.

## 1 INTRODUCTION

Open-domain question answering (QA) aims to answer questions from a broad range of domains by effectively retrieving evidence from one or more huge open-domain knowledge sources and by comprehending the evidence to infer the correct answers. Such resources can be Wikipedia (Chen et al., 2017), the whole web (Ferrucci et al., 2010), structured knowledge bases (Berant et al., 2013; Yu et al., 2017) or combinations of the above (Baudiš & Šedivỳ, 2015).

Recent work on open-domain QA focuses on using unstructured text from the web as sources, with the help from machine comprehension models (Chen et al., 2017; Dhingra et al., 2017b; Wang et al., 2017). These studies adopt a two-step-process: an information retrieval (IR) model to coarsely select relevant passages to a question, followed by a reading comprehension (RC) model (Wang & Jiang, 2017; Seo et al., 2017; Chen et al., 2017) to infer an answer from the passages. These works use each passage independently to train RC model, thus fail to fully explore the benefits of using multiple passages.

Information from different passages can provide enhanced or complementary evidence. We propose to improve open-domain QA by explicitly aggregating evidence from multiple passages. Specifically, we consider two important evidence aggregation problems from previous open-domain QA result analysis:

- First, the correct answer is suggested by multiple retrieved passages. Even if the best single passage can only provide weaker evidence for the correct answer than the other answer candidates (in terms of low prediction scores), aggregating the weaker evidence from multiple passages can enhance the confidence of predicting the correct answer. For example, in Figure 1(a), the correct answer "*danny boy*" has more passages providing evidence relevant to the question compared to the incorrect one. This observation can be seen as multiple passages strengthening each other to provide evidence for the correct answer.
- Second, sometimes the question may cover multiple aspects of the answer, and these aspects may spread over multiple passages. In order to infer the correct answer, one has to take these multiple passages into consideration. In Figure 1(b), for example, the correct answer "*Galileo Galilei*" at

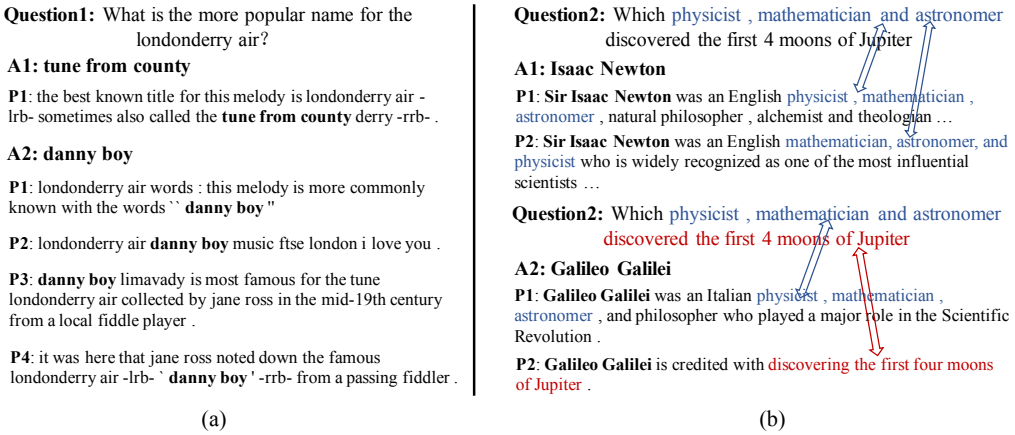


Figure 1: Two examples of questions and candidate answers. (a) A question benefiting from the coherence among evidence. Bottom: the correct answer has supporting evidence from multiple passages. Top: a wrong answer with only a single supporting passage. (b) A question benefiting from the union of multiple pieces of evidence for each answer. Bottom: the correct answer has evidence passages that could match both the first half and the second half of the question. Top: a wrong answer's evidence passages might be able to cover a small portion of the question (the first part), but not as much as the correct answer does.

the bottom has passages P1, "*Galileo was a physicist ...*" and P2, "*Galileo discovered the first 4 moons of Jupiter*", mentioning two pieces of evidence to match the question. In this case, the aggregation of these two pieces of evidence from the two passages can help entail the ground-truth answer "*Galileo Galilei*" for the question. In comparison, the incorrect answer "*Isaac Newton*" has passages providing partial evidence on only "*physicist , mathematician and astronomer*". This observation can be seen as multiple passages providing ***complementary*** evidence to the correct answer; and a method to take the union of the evidence for question answering will be helpful. Such a method is important especially for cases where the question is complex and a single passage is not sufficient to answer the question. In this case it is necessary to turn to complementary evidence from other passages.

To provide more accurate answers for open-domain QA, we hope to make better use of multiple passages for the same question by aggregating both the strengthened and the complementary evidence from all the passages. We formulate the above evidence aggregation as an answer re-ranking problem. Re-ranking has been commonly used in NLP problems, such as in parsing and translation, in order to make use of high-order or global features that are too expensive for decoding algorithms (Collins & Koo, 2005; Shen et al., 2004; Huang, 2008; Dyer et al., 2016). Here we borrow the idea to efficiently make use of global information from multiple pieces of textual evidence for each answer candidate, without significantly increase the complexity of the prediction of the RC model: We first collect the top-$K$ candidate answers based on their probabilities computed by a standard QA system. Then we propose two approaches of re-ranking the answer candidates with different ways of aggregating each candidate's evidence.

- The first re-ranker, named ***strength-based re-ranker***, ranks the answer candidates according to how often their evidence occurs in different passages. The re-ranker is based on a simple fact that if one answer candidate has multiple evidence, then each passage with its evidence will tend to predict it as the answer with a relatively high score (although it may not be the top score). As a result, the answer appearing more times in the top-$K$ list [1] is the one with more evidence scattered in different passages. Thus the count of each answer reflects how strong its evidence is.
- The second one, named ***coverage-based re-ranker***, aims to rank a answer candidate higher if the union of all its contexts in different passages could cover more aspects asked in the question. To achieve this, for each answer we concatenate all the passages that contain the answer together.

[1]The answers can be duplicated in this list, as the same answer can be extracted from different passages by the RC model.

The result is a new context that aggregates all the evidence necessary to entail the answer for the question. We then treat the new context as one sequence to represent the answer, and build an attention-based match-LSTM model (Wang & Jiang, 2017) between the sequence and the question to measure how well the new aggregated context could entail the question.

Overall, our contributions are as follows: 1) We propose a reranking-based framework to make use of the evidence from multiple passages in open-domain QA, and two re-ranking methods to address the common cases of coherent and complementary evidence aggregation in existing open-domain QA datasets. We find that the latter one performs better on two of the three public datasets. 2) Our proposed approach leads to state-of-the-art results on three different datasets (**Quasar-T** (Dhingra et al., 2017b), **SearchQA** (Dunn et al., 2017) and **TriviaQA** (Joshi et al., 2017)) and outperforms previous state-of-the-arts by large margins. Especially, we achieved up to 8% improvement on F1 on both Quasar-T and SearchQA compared to the previous best results.

## 2 METHOD

Given an open-domain question $\mathbf{q}$, the problem we are solving is to find the correct answer $\mathbf{a}^g$ to $\mathbf{q}$ based on all the information from the web. To retrieve all the contexts most related to the question, an IR model (with the help of a search engine such as Google or Bing) is used to provide the top-$N$ candidate passages, $\mathbf{p}_1$, $\mathbf{p}_2$, ..., $\mathbf{p}_N$, from the web. Then a reading comprehension (RC) model is used to extract the answer from these passages.

This setting is different from closed-domain QA (Rajpurkar et al., 2016), where a single fixed passage is given, from which the answer is to be extracted. When developing a closed-domain QA system, we can use the specific positions of the answer sequence in the given passage for training. In comparison, in the open-domain setting, the RC models are usually trained under distant supervision (Chen et al., 2017; Dhingra et al., 2017b; Joshi et al., 2017). Specifically, as the training data doesn't label the positions of the answer spans in passages, during the training stage, the RC model will match all passages that contain the ground-truth answer with the question one by one. Throughout the paper, we will first directly apply an existing RC model called $R^3$ (Wang et al., 2017) to extract the candidate answers.

After the candidate answers are extracted, we start to aggregate evidence from multiple passages for better answer prediction by adopting a re-ranking approach over answer candidates. Given a question $\mathbf{q}$, suppose we have a baseline open-domain QA system that could generate the top-$K$ answer candidates $\mathbf{a}_1, \ldots, \mathbf{a}_K$, each being a text span in some passage $\mathbf{p}_i$. The goal of the re-ranker is to rank this list of candidates such that the top-ranked candidates have more chance to be the correct answer $\mathbf{a}^g$.

This re-ranking step considers more powerful features and more context information that cannot be easily handled by the baseline system. In this paper, we investigate two types of such more powerful information based on the multiple pieces of evidence for each answer: *evidence strength* and *evidence coverage*, as aggregated from multiple passages. An overview of our method is shown in Figure 2.

### 2.1 EVIDENCE AGGREGATION FOR STRENGTH-BASED RE-RANKER

For open-domain QA, unlike the closed-domain setting, we have more passages retrieved by the IR model and the ground-truth answer may appear in different passages, which means different answer spans may be referring to the same answer. Considering this property, we provide two features to further re-rank the Top-$K$ answers generated by the RC model.

**Measuring Strength by Count** For this re-ranking method, we hypothesize that the more passages we have that could entail a particular answer for the question, the stronger the evidence is for that answer. In the top-$K$ answer spans generated by the RC model, each time an answer appears in the top-$K$ answer list, we increase the counter for that answer by one. In this way, in the end the answer with the highest count will be predicted to be the final answer.

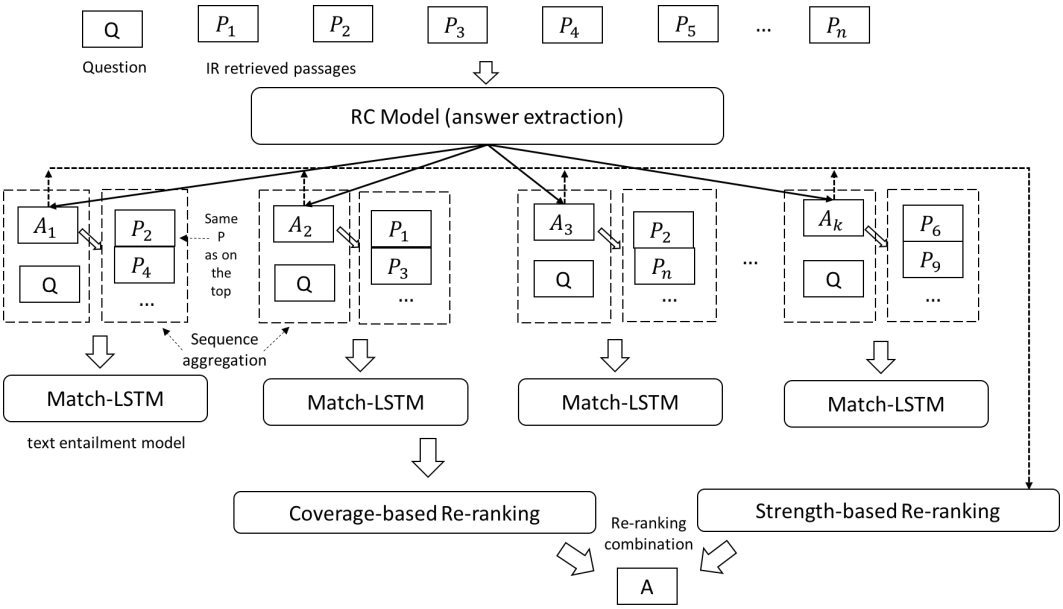


Figure 2: An overview of our model. It consists of strength-based and coverage-based re-ranking.

**Measuring Strength by Probability** Since we can get the probability of each answer span in a passages based on the RC model, we can also sum up the probabilities of the answer spans that are referring to the same answer. Then the answer with the highest probability would be the final prediction [2]. In the re-ranking scenario, it is not necessary to exhaustively consider all the probabilities of all the spans in the passages, as there may be a large number of different answer spans and most of them are irrelevant to the ground-truth answer.

Remark: Note that neither method above requires new model training. Both methods just take the candidate predictions from the baseline QA system and then perform counting or probability calculation. During testing the time complexity of strength-based re-ranking is negligible.

### 2.2 Evidence Aggregation for Coverage-based Re-ranker

This section focuses on another type of evidence aggregation which needs to take the union of complementary information from different evidence passages. As shown in Figure 1, the two answer candidates both have evidence matching the first half of the question. However, only the correct answer has evidence that could also match the second half. In this case, the strength-based re-ranker will treat both answer candidates the same due to the equal amount of supporting evidence, while actually the second answer has complementary evidence satisfying all aspects of the question. To handle such case, we propose a *coverage-based re-ranker* that could rank the answer candidates according to whether the unions of their evidence from different passages could well cover the question.

In order to take the union of evidence into consideration, we first concatenate the passages containing the answer into a single pseudo passage. Then we measure how well the new passage could entail the answer for the question. As examples shown in Figure 1(b), we hope the textual entailment model could reflect (i) how each aspect of the question is matched by the union of multiple passages, where we simply define the aspect to be one hidden state of bi-directional LSTM (Hochreiter & Schmidhuber, 1997); (ii) whether all the aspects of the question can be matched by the union of multiple passages. The match-LSTM (Wang & Jiang, 2016) model is one way to achieve the above effect in entailment. Therefore we build our coverage-based re-ranker on top of the concatenated pseudo passages through Match-LSTM. The detailed method is described below.

[2]This is an extension of the Attention Sum method in (Kadlec et al., 2016) from single-token answers to phrase answers.

**Passage Aggregation** We consider the top-$K$ answers, $\mathbf{a}_1, \ldots, \mathbf{a}_K$, provided by the baseline QA system. For each answer $\mathbf{a}_k, k \in [1, K]$, we concatenate all the passages that contain $\mathbf{a}_k$, $\{\mathbf{p}_n | \mathbf{a}_k \in \mathbf{p}_n, n \in [1, N]\}$, to form ***the union passage*** $\hat{\mathbf{p}}_k$. Our further model is to identify which union passage, e.g. $\hat{\mathbf{p}}_k$, could better entail its answer, e.g. $\mathbf{a}_k$, for the question.

**Measuring Aspect(Word)-Level Matching** As discussed earlier, the first mission of the coverage-based re-ranker is to measure how each aspect of the question is matched by the union of multiple passages. We achieve this with word-by-word attention followed by a comparison module.

First, we write the answer candidate $\mathbf{a}$, question $\mathbf{q}$ and the union passage $\hat{\mathbf{p}}$ of $\mathbf{a}$ as matrices $\mathbf{A}, \mathbf{Q}, \hat{\mathbf{P}}$, with each column being the embedding of a word in the sequence. We then feed them to the bi-directional LSTM as follows:

$$\mathbf{H}^{\mathrm{a}} = \mathrm{BiLSTM}(\mathbf{A}), \quad \mathbf{H}^{\mathrm{q}} = \mathrm{BiLSTM}(\mathbf{Q}), \quad \mathbf{H}^{\mathrm{p}} = \mathrm{BiLSTM}(\hat{\mathbf{P}}), \tag{1}$$

where $\mathbf{H}^{\mathrm{a}} \in \mathbb{R}^{l \times A}$ , $\mathbf{H}^{\mathrm{q}} \in \mathbb{R}^{l \times Q}$ and $\mathbf{H}^{\mathrm{p}} \in \mathbb{R}^{l \times P}$ are the hidden states for the answer candidate, question and passage respectively; $l$ is the dimension of the hidden states, and $A$,$Q$ and $P$ are the length of the three sequences, respectively.

Next, we enhance the question representation $\mathbf{H}^{\mathrm{q}}$ with $\mathbf{H}^{\mathrm{a}}$:

$$\mathbf{H}^{\mathrm{aq}} = [\mathbf{H}^{\mathrm{a}}; \mathbf{H}^{\mathrm{q}}], \tag{2}$$

where $[\cdot;\cdot]$ is the concatenation of two matrices in row and $\mathbf{H}^{\mathrm{aq}} \in \mathbb{R}^{l \times (A+Q)}$. As most of the answer candidates do not appear in the question, this is for better matching with the passage and finding more answer-related information from the passage. Now we can view each ***aspect*** of the question as a column vector (i.e. a hidden state at each word position in the answer-question concatenation) in the enhanced question representation $\mathbf{H}^{\mathrm{aq}}$. Then the task becomes to measure how well each column vector can be matched by the union passage; and we achieve this by computing the attention vector Parikh et al. (2016) for each hidden state of sequences $\mathbf{a}$ and $\mathbf{q}$ as follows:

$$\alpha = \mathrm{SoftMax}\left((\mathbf{H}^{\mathrm{p}})^{\mathrm{T}} \mathbf{H}^{\mathrm{aq}}\right), \quad \overline{\mathbf{H}}^{\mathrm{aq}} = \mathbf{H}^{\mathrm{p}} \alpha, \tag{3}$$

where $\alpha \in \mathbb{R}^{P \times (A+Q)}$ is the attention weight matrix which is normalized in column through softmax. $\overline{\mathbf{H}}^{\mathrm{aq}} \in \mathbb{R}^{l \times (A+Q)}$ are the attention vectors for each word of the answer and the question by weighted summing all the hidden states of the passage $\hat{\mathbf{p}}$. Now in order to see whether the aspects in the question can be matched by the union passage, we use the following matching function:

$$\mathbf{M} = \mathrm{ReLU}\left(\mathbf{W}^{\mathrm{m}} \begin{bmatrix} \mathbf{H}^{\mathrm{aq}} \bigodot \overline{\mathbf{H}}^{\mathrm{aq}} \\ \mathbf{H}^{\mathrm{aq}} - \overline{\mathbf{H}}^{\mathrm{aq}} \\ \mathbf{H}^{\mathrm{aq}} \\ \overline{\mathbf{H}}^{\mathrm{aq}} \end{bmatrix} + \mathbf{b}^{\mathrm{m}} \otimes \mathbf{e}_{(A+Q)}\right), \tag{4}$$

where $\cdot \otimes \mathbf{e}_{(A+Q)}$ is to repeat the vector (or scalar) on the left $A+Q$ times; $(\cdot \bigodot \cdot)$ and $(\cdot - \cdot)$ are the element-wise operations for checking whether the word in the answer and question can be matched by the evidence in the passage. We also concatenate these matching representations with the hidden state representations $\mathbf{H}^{\mathrm{aq}}$ and $\overline{\mathbf{H}}^{\mathrm{aq}}$, so that some stop words or punctuation representations may not affect the final aspect-level matching representations $\mathbf{M} \in \mathbb{R}^{2l \times (A+Q)}$, which is computed through the non-linear transformation on four different representations with parameters $\mathbf{W}^{\mathrm{m}} \in \mathbb{R}^{2l \times 4l}$ and $b^{\mathrm{m}} \in \mathbb{R}^{l}$.

**Measuring the Entire Question Matching** Next, in order to measure how the entire question is matched by the union passage $\hat{\mathbf{p}}$ by taking into consideration of the matching result at each aspect, we add another bi-directional LSTM on top of it to aggregate the aspect-level matching information [3]:

$$\mathbf{H}^{\mathrm{m}} = \mathrm{BiLSTM}(\mathbf{M}), \quad \mathbf{h}^{\mathrm{s}} = \mathrm{MaxPooling}(\mathbf{H}^{\mathrm{m}}), \tag{5}$$

[3]Note that we use LSTM here to capture the conjunction information (the dependency) among aspects, i.e. how all the aspects are jointly matched. In comparison simple pooling methods will treat the aspects independently. Low-rank tensor inspired neural architectures like in (Lei et al., 2017) could be another choice and we will investigate them in future work.

where $\mathbf{H}^{\mathrm{m}} \in \mathbb{R}^{l \times (A+Q)}$ is to denote all the hidden states and $\mathbf{h}^{\mathrm{s}} \in \mathbb{R}^{l}$, the max of pooling of each dimension of $\mathbf{H}^{\mathrm{m}}$, is the entire matching representation which reflects how well the evidences in questions could be matched by the union passage.

**Re-ranking Objective Function** Our re-ranking is based on the entire matching representation. For each candidate answer $\mathbf{a}_k, k \in [1, K]$, we can get a matching representation $\mathbf{h}^{\mathrm{s}}_k$ between the answer $\mathbf{a}_k$, question $\mathbf{q}$ and the union passage $\hat{\mathbf{p}}_k$ through Eqn. (1-5). Then we transform all representations into scalar values followed by a normalization process for ranking as follows:

$$\mathbf{R} = \mathrm{Tanh}\left(\mathbf{W}^{\mathrm{r}}[\mathbf{h}^{\mathrm{s}}_1; \mathbf{h}^{\mathrm{s}}_2; ...; \mathbf{h}^{\mathrm{s}}_K] + \mathbf{b}^{\mathrm{r}} \otimes \mathbf{e}_K\right), \quad \mathbf{o} = \mathrm{Softmax}(\mathbf{w}^o \mathbf{R} + b^o \otimes \mathbf{e}_K), \tag{6}$$

where we concatenate the match representations for each answer in row through $[\cdot;\cdot]$, and do a non-linear transformation by parameters $\mathbf{W}^{\mathrm{r}} \in \mathbb{R}^{l \times l}$ and $\mathbf{b}^{\mathrm{r}} \in \mathbb{R}^{l}$ to get hidden representation $\mathbf{R} \in \mathbb{R}^{l \times K}$. Finally, we map the transformed matching representations into scalar values through parameters $\mathbf{w}^o \in \mathbb{R}^{l}$ and $\mathbf{w}^o \in \mathbb{R}$. $\mathbf{o} \in \mathbb{R}^{K}$ is the normalized probability for the candidate answers to be ground-truth. Due to the aliases of the ground-truth answer, there may be multiple answers in the candidates are ground-truth, we use KL distance as our objective function as follows:

$$\sum_{k=1}^{K} y_k \left(\log(y_k) - \log(o_k)\right), \tag{7}$$

where $y_k$ indicates whether $\mathbf{a}_k$ the ground-truth answer or not and is normalized by $\sum_{k=1}^{K} y_k$ and $o_k$ is the ranking output of our model for $\mathbf{a}_k$.

### 2.3 Combination of Different Types of Aggregations

Although the coverage-based re-ranker tries to deal with more difficult cases compared to the strength-based re-ranker, the strength-based re-ranker works on more common cases according to the distributions of most open-domain QA datasets. As a result we hope to make the best of both methods.

We achieve our ***full re-ranker*** by weighted combination of the outputs of the above different re-rankers without further training. Specifically, we first use softmax to re-normalize the top-5 answer scores provided by the two strength-based rankers and the one coverage-based re-ranker; we then weighted sum up the scores for the same answer and select the answer with the largest score as the final prediction.

## 3 Experimental Settings

We have experiments on three public open-domain QA datasets, **Quasar-T** (Dhingra et al., 2017b), **SearchQA** (Dunn et al., 2017) and **TriviaQA** (Joshi et al., 2017), to evaluate our model. The passages retrieved for all the answers based on search engine, such as Google and Bing, are given. And we will not retrieve more passages for the experiment.

### 3.1 Datasets

The statistics of the three datasets are shown in Table 1.

**Quasar-T** [4] (Dhingra et al., 2017b) is based on a trivia question set and makes use of "lucene index" as IR model to collect 100 sentence-level passages for each question from ClueWeb09 data source. The human performance is evaluated in an open-book setting, i.e. the people had access to the passages retrieved by the same IR model and tried to find the answers from the passages.

**SearchQA**[5] (Dunn et al., 2017) is based on a Jeopardy! questions and use search engine Google to collect about 50 web page snippets as passages for each question. The human performance is evaluated in a similar way to Quasar-T dataset.

[4] https://github.com/bdhingra/quasar

[5] https://github.com/nyu-dl/SearchQA

| | #q(train) | #q(dev) | #q(test) | #p | #p(truth) | #p(aggregated) |
|---|---|---|---|---|---|---|
| Quasar-T | 28,496 | 3,000 | 3,000 | 100 | 14.8 | 5.2 |
| SearchQA | 99,811 | 13,893 | 27,247 | 50 | 16.5 | 5.4 |
| TriviaQA | 66,828 | 11,313 | 10,832 | 100 | 16.0 | 5.6 |

Table 1: Statistics of the datasets. #q represents the number of questions for training (not counting the questions that don't have ground-truth answer in the corresponding passages for training set), development, and testing datasets. #p is the number of passages for each question. For TriviaQA, we split the raw documents into sentence level passages and select the top 100 passages based on the its overlaps with the corresponding question. #p(golden) means the number of passages that contain the ground-truth answer in average. #p(aggregated) is the number of passages we aggregated in average for top 10 candidate answers provided by RC model.

**TriviaQA (Open-Domain Setting)**[6] (Joshi et al., 2017) collects trivia questions coming from 14 trivia and quiz-league websites, and makes use of the Bing Web search API to collect the top 50 webs most related to the questions. We focus on the open domain setting of the dataset [7] and our model will use all the IR retrieved information.

### 3.2 Baselines

Our baseline models [8] include: GA (Dhingra et al., 2017a;b), a reading comprehension model with gated-attention; BiDAF (Seo et al., 2017), a RC model with bidirectional attention flow; AQA (Buck et al., 2017), a reinforced system learning to aggregate the answers generated by the re-written questions; $R^3$ (Wang et al., 2017), a reinforced model making use of a ranker for selecting passages to train RC model. As $R^3$ is the first step of our system for generating candidate answers, the improvement our re-ranking methods can be directly compared to this baseline.

### 3.3 Implementation Details

We first use a pre-trained $R^3$ model (Wang et al., 2017), which gets the stat-of-the-art performance on the 3 public datasets we consider, to generate top 50 candidate spans for both training, dev and test datasets, and we use them for further ranking. During training, if the ground-truth answer doesn't appear in the answer candidates, we will manually add it into the answer candidate list.

For the details of the Coverage-based Re-Ranker, we use Adamax (Kingma & Ba, 2015) to optimize the model. Word embeddings are initialized by GloVe (Pennington et al., 2014) and are not updated during training. We set all the words beyond Glove as zero vector. We set $l$ to 300, batch size to 30, learning rate to 0.002. We tune the dropout probability fro 0 to 0.5 and the number of candidate answers for re-ranking ($K$) in [3, 5, 10] [9].

## 4 Results and Analysis

In this section, we will show the results of our different re-ranking methods on three different public and the further analysis about our models.

### 4.1 Overall Results

The performance of our models are shown in Table 2. We use F1 score and Exact Match (EM) as our evaluation metrics[10]. From the results, we can clearly see that our full re-ranker, the combination of

[6] `http://nlp.cs.washington.edu/triviaqa/data/triviaqa-unfiltered.tar.gz`

[7] The dataset also includes the closed-domain setting which is a single-passage RC task.

[8] Most of the results of different models come from the public paper. While we re-run model $R^3$ (Wang et al., 2017) based on the authors' source code and extend the model to the datasets of SearchQA and TriviaQA datasets.

[9] Our code will be released.

[10] Our evaluation is based on the tool from SQuAD (Rajpurkar et al., 2016).

| | Quasar-T | | SearchQA | | TriviaQA (open) | |
|---|---|---|---|---|---|---|
| | EM | F1 | EM | F1 | EM | F1 |
| GA (Dhingra et al., 2017a) | 26.4 | 26.4 | - | - | - | - |
| BiDAF (Seo et al., 2017) | 25.9 | 28.5 | 28.6 | 34.6 | - | - |
| AQA (Buck et al., 2017) | - | - | 40.5 | 47.4 | - | - |
| $R^3$ (Wang et al., 2017) | 35.3 | 41.7 | 49.0 | 55.3 | 47.3 | 53.7 |
| Full Re-Ranker | **42.3** | **49.6** | **57.0** | **63.2** | **50.6** | **57.3** |
| Strength-Based Re-Ranker (Probability) | 36.1 | 42.4 | 50.4 | 56.5 | 49.2 | 55.1 |
| Strength-Based Re-Ranker (Counting) | 37.1 | 46.7 | 54.2 | 61.6 | 46.1 | 55.8 |
| Coverage-Based Re-Ranker | 40.6 | 49.1 | 53.6 | 60.6 | 50.0 | 57.0 |
| Human Performance | 51.5 | 60.6 | 43.9 | - | - | - |

Table 2: Experiment results on three open-domain QA test datasets: Quasar-T, SearchQA and Triviaqa (open domain setting). EM: Exact Match. Full Re-ranker is the combination of three different re-rankers.

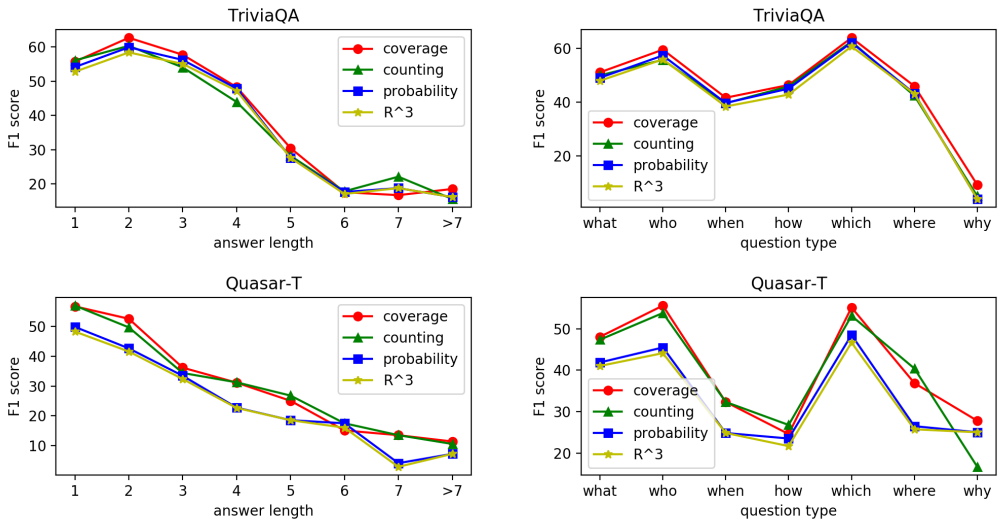


Figure 3: Performance decomposition according to the length of answers and the question types.

different re-rankers, can significantly outperform the previous best performance in a large margin, especially on the datasets of Quasar-T and SearchQA. Moreover, our model is much better than the human performance on the SearchQA dataset. Besides, we can see that our coverage-based re-ranker performs a generally well performance on the three datasets, even though gets 1% lower than strength-based re-ranker on SearchQA dataset.

### 4.2 Analysis

In this subsection, we will analyze the benefits of re-ranking models.

**Re-ranking performance versus answer lengths and question types** Figure 3 decomposes the performance according to the length of golden answers and the types of questions on TriviaQA and Quasar-T. We ignore the analysis on SearchQA because the Jeopardy! style questions are more difficult to distinguish the questions types, and the range of answer length is smaller.

According to the results, the coverage-based re-ranker could outperform the baseline in different length of answers and different types of questions. The strength-based re-ranker (counting) also gives improvement but is less stable across different datasets. While the strength-based re-ranker

| | Quasar-T | | SearchQA | | TriviaQA (open) | |
|---|---|---|---|---|---|---|
| **Top-K** | **EM** | **F1** | **EM** | **F1** | **EM** | **F1** |
| 1 | 35.1 | 41.6 | 51.2 | 57.3 | 47.6 | 53.5 |
| 3 | 46.2 | 53.5 | 63.9 | 68.9 | 54.1 | 60.4 |
| 5 | 51.0 | 58.9 | 69.1 | 73.9 | 58.0 | 64.5 |
| 10 | 56.1 | 64.8 | 75.5 | 79.6 | 62.1 | 69.0 |

Table 3: The upper bound (recall) of the Top-K answer candidates generated by the baseline $R^3$ system (on dev set), which indicates the potential of the coverage-based re-ranker.

| | Re-Ranker Results | | Upper Bound | |
|---|---|---|---|---|
| **Candidate Set** | **EM** | **F1** | **EM** | **F1** |
| top-3 | 40.5 | 47.8 | 46.2 | 53.5 |
| top-5 | **41.8** | 50.1 | 51.0 | 58.9 |
| top-10 | 41.3 | **50.8** | 56.1 | 64.8 |

Table 4: Results of running coverage-based re-ranker on different number of top-$K$ answer candidates on Quasar-T (dev set).

(probability) tends to have results and trends that are close to the baseline curves, which is probably because the method is dominated by the probabilities predicted by the baseline.

The coverage-based re-ranker and the strength-based re-ranker (counting) have similar trends on most of the question types. The only exception is that the strength-based re-ranker performs significantly worse compared to the coverage-based re-ranker on the "*why*" questions. This is possibly because those questions usually have non-factoid answers, which are less likely to have exactly the same text spans predicted on different passages by the baseline.

**Potential improvement of re-rankers** Table 3 shows the percentage of time when the correct answer is included in the top-$K$ answer predictions of the baseline $R^3$ method. More concretely, the scores are computed by selecting the answer from the top-$K$ predictions with the best EM/F1 score. Therefore the final top-$K$ EM and F1 can be viewed as recall or upper bound of the top-$K$ prediction. From the results, although the top-1 prediction of $R^3$ is not very accurate, there is high chance that a moderate top-$K$ list could cover the correct answer. This explains why our re-ranking approach could achieve large improvement. Also by comparing the upper bound performance of top-5 and our re-ranking performance in Table 2, we can see there is still a clear gap of about 10% on both datasets and on both F1 and EM, showing the great potential improvement for the re-ranking model in future work.

**Effect of the selection of $K$ for the coverage-based re-ranker** As shown in Table 3, with the increase of $K$ from 1 to 10, the recall of top-$K$ predictions from the baseline $R^3$ system increase significantly. Ideally, if we use a larger $K$, then the candidate lists will be more likely to contain good answers. At the same time, the lists to be ranked are longer thus the re-ranking problem is harder. Therefore, there is a trade-off between the coverage of rank lists and the difficulty of re-ranking; and selecting an appropriate $K$ becomes important. Table 4 shows the effects of $K$ on the performance of coverage-based re-ranker. We train and test the coverage-based re-ranker on top-$K$ predictions from the baseline, here $K \in \{3, 5, 10\}$. The upper bound results are the same ones from Table 3. The results show that when $K$ is small, like $K$=3, the performance is not very good due to the low coverage (thus low upper bound) of the candidate list. With the increase of $K$, the performance becomes better, but the top-5 and top-10 results are on par with each other. This is because the higher upper bound of top-10 results counteracts the harder problem of re-ranking longer lists. Since there is no significant advantage of the usage of $K$=10 while the computation is slower, we report all testing results with $K$=5.

**Effect of the selection of $K$ for the strength-based re-ranker** Similar to Table 4, we conduct experiments to show the effects of $K$ on the performance of strength-based re-ranker. We

| Candidate Set | Re-Ranker Results EM | Re-Ranker Results F1 | Upper Bound EM | Upper Bound F1 |
|---|---|---|---|---|
| top-10 | **37.9** | 46.1 | 56.1 | 64.8 |
| top-50 | 37.8 | **47.8** | 64.1 | 74.1 |
| top-100 | 36.4 | 47.3 | 66.5 | 77.1 |
| top-200 | 33.7 | 45.8 | 68.7 | 79.5 |

Table 5: Results of running strength-based re-ranker (counting) on different number of top-$M$ answer candidates on Quasar-T (dev set).

| | | | |
|---|---|---|---|
| Q: | Which children 's TV programme , which first appeared in November 1969 , has won a record 122 Emmy Awards in all categories ? | | |
| A1: | **Great Dane** | A2: | **Sesame Street** |
| P1 | The world 's most famous Great Dane first appeared on television screens on Sept. 13 , 1969 . | P1: | In its long history , Sesame Street has received more *Emmy Awards* than any other program , ... |
| P2 | premiered on broadcast television ( CBS ) Saturday morning , Sept. 13 , 1969 , ... yet beloved great Dane . | P2: | Sesame Street ... is recognized as a pioneer of the contemporary standard which combines education and entertainment in *children 's television shows* . |

Table 6: An example from Quasar-T dataset. The ground-truth answer is "*Sesame Street*". Q: question, A: answer, P: passages containing corresponding answer.

run the strength-based re-ranker (counting) on top-$K$ predictions from the baseline, here $K \in \{10, 50, 100, 200\}$, as well as evaluate the upper bound results for these $K$s. Note that the strength-based re-ranker runs very fast and the different number of $K$ does not affect the computation speed very much compared to the other QA components. The results are shown in Table 5, where we achieve the best results when $K$=50. The performance drops significantly when $K$ increases to 200. This is because the ratio of incorrect answers increases a lot, making incorrect answers also likely to have high counts. When $K$ is smaller, such incorrect answers appear less because statistically they have lower prediction scores. We report all testing results with $K$=50.

**Examples** Table 6 shows an example from Quasar-T where the re-ranker successfully correct the wrong answer predicted by the baseline. This is a case where the coverage-based re-ranker helps: the correct answer "*Sesame Street*" has evidence from different passages that covers the aspects "*Emmy Award*" and "*children 's television shows*". Although it still doesn't fully match all the facts in the question, it still helps to rank the correct answer higher than the top-1 prediction "*Great Dane*" from the $R^3$ baseline, which only has evidence covering "*TV*" and "*1969*" in the question.

## 5 Related Work

**Open Domain Question Answering** The task of **Open domain question answering** dates back to as early as (Green Jr et al., 1961) and was popularized by TREC-8 (Voorhees, 1999). The task is to produce the answer to a question by exploiting resources such as documents (Voorhees, 1999), webpages (Kwok et al., 2001) or structured knowledge bases (Berant et al., 2013; Bordes et al., 2015; Yu et al., 2017).

Recent efforts (Chen et al., 2017; Dunn et al., 2017; Dhingra et al., 2017b; Wang et al., 2017) benefit from the advance of machine reading comprehension (RC) and follow the search-and-read QA direction. These deep learning based methods usually rely on a document retrieval module to retrieve a list of passages for RC models to extract answers. As there is no passage-level annotation about which passages entails the answer, the model has to find proper ways to handle the noise introduced in the IR step. Chen et al. (2017) uses bi-gram passage index to improve the retrieval step; and Dunn et al. (2017); Dhingra et al. (2017b) propose to reduce the length of the retrieved passages. Wang et al. (2017) focus more on noise reduction in the passage ranking step, in which a ranker module is jointly trained with the RC model with reinforcement learning.

To the best of our knowledge, our work is the first to improve neural open-domain QA systems by benefiting from multiple passages for evidence aggregation.

**Multi-Step Approaches for Reading Comprehension** We are the first to introduce re-ranking methods to neural open-domain QA and multi-passage RC. In the meanwhile, our two-step fashion approach shares some similarities to the previous multi-step approaches proposed for standard single-passage RC, in terms of the purposes of either using additional information or re-fining answer predictions that are not easily handled by the standard answer extraction models for RC.

On cloze-test tasks (Hermann et al., 2015), Epireader Trischler et al. (2016) relates to our work in the sense that it is a two-step extractor-reasoner model, which first extracts $K$ most probable single-token answer candidates, then constructs a hypothesis by combining each answer candidate to the question and compares the hypothesis with all the sentences in the passage. Their model differs from ours in several aspects: (i) Epireader matches a hypothesis to *every* single sentence, including all the "noisy" ones that does not contain the answer, that makes the model inappropriate for open-domain QA setting; (ii) The sentence matching is based on the sentence embedding vectors computed by a convolutional neural network, which makes it hard to distinguish redundant and complementary evidence in aggregation; and (iii) Epireader passes the probabilities predicted by the extractor to the reasoner directly to sustain differentiability, which can not be easily adapted to our problem to handle phrases as answers or to use part of passages.

Similarly, (Cui et al., 2017) also combined answer candidates to the question to form hypotheses, and then explicitly use language models trained on documents to re-rank the hypotheses. This method benefits from the consistency between the documents and gold hypotheses (which are titles of the documents) in cloze-test datasets, but does not handle multiple evidence aggregation like our work.

S-Net (Tan et al., 2017) proposes a two-step approach to achieve generative QA. The model first extracts an text span as the answer clue and then generates the answer according to the question, passage and the text span. Besides the different goal on answer generation instead of re-ranking like this work, their approach also differ from ours on that it extracts only one text span from a single selected passage.

## 6 Conclusion

We propose to improve open-domain QA with evidence aggregation from multiple passages. The aggregation process is formulated as answer re-ranking; and two re-rankers are proposed to deal with two types of important evidence aggregation in open-domain QA, aggregation on pieces of evidence strengthening each other and on pieces of evidence complementary to each other. Both re-rankers helped to significant improvement our results and greatly advances the state-of-the-arts on three open-domain QA datasets.

## References

Petr Baudiš and Jan Šedivỳ. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 222–228. Springer, 2015.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2013.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. Large-scale simple question answering with memory networks. *Proceedings of the International Conference on Learning Representations*, 2015.

Christian Buck, Jannis Bulian, Massimiliano Ciaramita, Andrea Gesmundo, Neil Houlsby, Wojciech Gajewski, and Wei Wang. Ask the right questions: Active question reformulation with reinforcement learning. *arXiv preprint arXiv:1705.07830*, 2017.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading Wikipedia to answer open-domain questions. In *Proceedings of the Conference on Association for Computational Linguistics*, 2017.

Michael Collins and Terry Koo. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25–70, 2005.

Yiming Cui, Zhipeng Chen, Si Wei, Shijin Wang, Ting Liu, and Guoping Hu. Attention-over-attention neural networks for reading comprehension. *Proceedings of the Conference on Association for Computational Linguistics*, 2017.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. *Proceedings of the Conference on Association for Computational Linguistics*, 2017a.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. QUASAR: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*, 2017b.

Matthew Dunn, Levent Sagun, Mike Higgins, Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A Smith. Recurrent neural network grammars. In *Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics*, 2016.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

Bert F Green Jr, Alice K Wolf, Carol Chomsky, and Kenneth Laughery. Baseball: an automatic question-answerer. In *Papers presented at the May 9-11, 1961, western joint IRE-AIEE-ACM computer conference*, pp. 219–224. ACM, 1961.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*, pp. 1693–1701, 2015.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Liang Huang. Forest reranking: Discriminative parsing with non-local features. In *Proceedings of the Conference on Association for Computational Linguistics*, 2008.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2017.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2016.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*, 2015.

Cody Kwok, Oren Etzioni, and Daniel S Weld. Scaling question answering to the web. *ACM Transactions on Information Systems (TOIS)*, 19(3):242–262, 2001.

Tao Lei, Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Deriving neural architectures from sequence and graph kernels. *International Conference on Machine Learning*, 2017.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2014.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *Proceedings of the International Conference on Learning Representations*, 2017.

Libin Shen, Anoop Sarkar, and Franz Josef Och. Discriminative reranking for machine translation. In *Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics*, 2004.

Chuanqi Tan, Furu Wei, Nan Yang, Weifeng Lv, and Ming Zhou. S-net: From answer extraction to answer generation for machine reading comprehension. *arXiv preprint arXiv:1706.04815*, 2017.

Adam Trischler, Zheng Ye, Xingdi Yuan, and Kaheer Suleman. Natural language comprehension with the epireader. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

Ellen M. Voorhees. The trec-8 question answering track report. In *Trec*, volume 99, pp. 77–82, 1999.

Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In *Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics*, 2016.

Shuohang Wang and Jing Jiang. Machine comprehension using match-LSTM and answer pointer. In *Proceedings of the International Conference on Learning Representations*, 2017.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. R3: Reinforced reader-ranker for open-domain question answering. *arXiv preprint arXiv:1709.00023*, 2017.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. Improved neural relation detection for knowledge base question answering. *Proceedings of the Conference on Association for Computational Linguistics*, 2017.