Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering

Vikas Yadav, Steven Bethard, Mihai Surdeanu
University of Arizona, Tucson, AZ, USA
{vikasy, bethard, msurdeanu}@email.arizona.edu

Abstract

We propose an unsupervised strategy for the selection of justification sentences for multi-hop question answering (QA) that (a) maximizes the relevance of the selected sentences, (b) minimizes the overlap between the selected facts, and (c) maximizes the coverage of both question and answer. This unsupervised sentence selection method can be coupled with any supervised QA approach. We show that the sentences selected by our method improve the performance of a state-of-the-art supervised QA model on two multi-hop QA datasets: AI2's Reasoning Challenge (ARC) and Multi-Sentence Reading Comprehension (MultiRC). We obtain new state-of-the-art performance on both datasets among approaches that do not use external resources for training the QA system: 56.82% F1 on ARC (41.24% on Challenge and 64.49% on Easy) and 26.1% EM0 on MultiRC. Our justification sentences have higher quality than the justifications selected by a strong information retrieval baseline, e.g., by 5.4% F1 in MultiRC. We also show that our unsupervised selection of justification sentences is more stable across domains than a state-of-the-art supervised sentence selection method.

1 Introduction

Interpretable machine learning (ML) models, where the end user can understand how a decision was reached, are a critical requirement for the wide adoption of ML solutions in many fields such as healthcare, finance, and law (Samek et al., 2017; Alvarez-Melis and Jaakkola, 2017; Arras et al., 2017; Gilpin et al., 2018; Biran and Cotton, 2017)

For complex natural language processing (NLP) such as question answering (QA), human readable explanations of the inference process have been proposed as a way to interpret QA models (Zhou et al., 2018).

To which organ system do the esophagus, liver, pancreas, small intestine, and colon belong?
(A) reproductive system (B) excretory system
(C) **digestive system** (D) endocrine system

ROCC-selected justification sentences:
1. vertebrate **digestive system** has oral cavity, teeth and pharynx, *esophagus* and stomach, *small intestine, pancreas, liver* and the large intestine
2. **digestive system** consists liver, stomach, large intestine, *small intestine, colon*, rectum and anus

BM25-selected justification sentences:
1. their **digestive system** consists of a stomach, *liver*, *pancreas*, *small intestine*, and a large intestine
2. the *liver pancreas* and gallbladder are the solid organ of the **digestive system**

Figure 1: A multiple-choice question from the ARC dataset with the correct answer in bold, followed by justification sentences selected by our approach (ROCC) vs. sentences selected by a strong IR baseline (BM25). ROCC justification sentences fully cover the five key terms in the question (shown in italic), whereas BM25 misses two: *esophagus* and *colon*. Further, the second BM25 sentence is largely redundant with the first, not covering other query terms.

Recently, multiple datasets have been proposed for *multi-hop* QA, in which questions can only be answered when considering information from multiple sentences and/or documents (Clark et al., 2018; Khashabi et al., 2018a; Yang et al., 2018; Welbl et al., 2018; Mihaylov et al., 2018; Bauer et al., 2018; Dunn et al., 2017; Dhingra et al., 2017; Lai et al., 2017; Rajpurkar et al., 2018; Sun et al., 2019). The task of selecting justification sentences is complex for multi-hop QA, because of the additional knowledge aggregation requirement (examples of such questions and answers are shown in Figures 1 and 2). Although various neural QA methods have achieved high performance on some of these datasets (Sun et al., 2018; Trivedi et al., 2019; Tymoshenko et al., 2017; Seo et al., 2016; Wang and Jiang, 2016; De Cao et al., 2018; Back et al., 2018), we argue that more effort must be dedicated to explaining their inference process.

In this work we propose an *unsupervised* algorithm for the selection of *multi-hop* justifications from *unstructured* knowledge bases (KB). Unlike other supervised selection methods (Dehghani et al., 2019; Bao et al., 2016; Lin et al., 2018; Wang et al., 2018b,a; Tran and Niedereée, 2018; Trivedi et al., 2019), our approach does not require any training data for justification selection. Unlike approaches that rely on structured KBs, which are expensive to create, (Khashabi et al., 2016; Khot et al., 2017; Zhang et al., 2018; Khashabi et al., 2018b; Cui et al., 2017; Bao et al., 2016), our method operates over KBs of only unstructured texts. We demonstrate that our approach has a bigger impact on downstream QA approaches that use these justification sentences as additional signal than a strong baseline that relies on information retrieval (IR). In particular, the contributions of this work are:

(1) We propose an unsupervised, non-parametric strategy for the selection of justification sentences for multi-hop question answering (QA) that (a) maximizes the ***R***elevance of the selected sentences; (b) minimizes the lexical ***O***verlap between the selected facts; and (c) maximizes the lexical ***C***overage of both question and answer. We call our approach ROCC. ROCC operates by first creating $\binom{n}{k}$ justification sets from the top n sentences selected by the BM25 information retrieval model (Robertson et al., 2009), where k ranges from 2 to n, and then ranking them all by a formula that combines the three criteria above. The set with the top score becomes the set of justifications output by ROCC for a given question and candidate answer. As shown in Figure 1, the justification sentences selected by ROCC perform more meaningful knowledge aggregation than a strong IR baseline (BM25), which does not account for overlap (or complementarity) and coverage.

(2) ROCC can be coupled with any supervised QA approach that can use the selected justification sentences as additional signal. To demonstrate its effectiveness, we combine ROCC with a state-of-the-art QA method that relies on BERT (Devlin et al., 2018) to classify correct answers, using the text of the question, the answer, and (now) the justification sentences as input. On the Multi-Sentence Reading Comprehension (MultiRC) dataset (Khashabi et al., 2018a), we achieved a gain of 8.3% EM0 with ROCC justifications when compared to the case where the complete comprehension passage was provided to the BERT classifier. On AI2's Reasoning Challenge (ARC) dataset (Clark et al., 2018), the QA approach enhanced with ROCC justifications outperforms the QA method without justifications by 9.15% accuracy, and the approach that uses top sentences provided by BM25 by 2.88%. Further, we show that the justification sentences selected by ROCC are considerably more correct on their own than justifications selected by BM25 (e.g., the justification score in MultiRC was increased by 11.58% when compared to the best performing BM25 justifications), which indicates that the interpretability of the overall QA system was also increased.

(3) Lastly, our analysis indicates that ROCC is more stable across the different domains in the MultiRC dataset than a supervised strategy for the selection of justification sentences that relies on a dedicated BERT-based classifier, with a difference of over 10% F1 score in some configurations.

The ROCC system and the codes for generating all the analysis are provided here - `https://github.com/vikas95/AutoROCC`.

2 Related Work

The body of QA work that addresses the selection of justification sentences can be classified into roughly four categories: (a) supervised approaches that require training data to learn how to select justification sentences (i.e., questions and answers coupled with correct justifications); (b) methods that treat justifications as latent variables and learn jointly how to answer questions and how to select justifications from questions and answers alone; (c) approaches that rely on information retrieval to select justification sentences; and, lastly, (d) methods that do not use justification sentences at all.

In the first category, previous works (e.g., (Trivedi et al., 2019)) have used entailment resources including labeled trained datasets such as SNLI (Bowman et al., 2015) and MultiNLI (Williams et al., 2017) to train components for selecting justification sentences for QA. Other works have explicitly focused on training sentence selection components for QA models (Min et al., 2018; Lin et al., 2018; Wang et al., 2019). In datasets where gold justification sentences are not provided, researchers have trained such components by retrieving justifications from structured KBs (Cui et al., 2017; Bao et al., 2016; Zhang et al., 2016; Hao et al., 2017) such as ConceptNet (Speer et al., 2017), or from IR systems coupled

with denoising components (Wang et al., 2019). While these works offer exciting directions, they all rely on training data for justifications, which is expensive to generate and may not be available in real-world use cases.

The second group of methods tend to rely on reinforcement learning (Choi et al., 2017; Lai et al., 2018; Geva and Berant, 2018) or PageRank (Surdeanu et al., 2008) to learn how to select justification sentences without explicit training data. Other works have used end-to-end (mostly RNNs with attention mechanisms) QA architectures for learning to pay more attention on better justification sentences (Min et al., 2018; Seo et al., 2016; Yu et al., 2014; Gravina et al., 2018). While these approaches do not require annotated justifications, they need large amounts of question/answer pairs during training so they can discover the latent justifications. In contrast to these two directions, our approach requires no training data at all for the justification selection process.

The third category of methods utilize IR techniques to retrieve justifications from both unstructured (Yadav et al., 2019) and structured (Khashabi et al., 2016) KBs. Our approach is closer in spirit to this direction, but it is adjusted to account for more intentional knowledge aggregation. As we show in Section 4, this is important for both the quality of the justification sentences and the performance of the downstream QA system.

The last group of QA approaches learn how to classify answers without any justification sentences (Mihaylov et al., 2018; Sun et al., 2018; Devlin et al., 2018). While this has been shown to obtain good performance for answer classification, we do not focus on it in this work because these methods cannot easily explain their inference.

Note that some of the works discussed here transfer knowledge from external datasets into the QA task they address (Chung et al., 2017; Sun et al., 2018; Pan et al., 2019; Min et al., 2017; Qiu et al., 2018; Chen et al., 2017). In this work, we focus solely on the resources provided in the task itself because such compatible external resources may not be available in real-world applications of QA.

3 Approach

ROCC, coupled with a QA system, operates in the following steps (illustrated in Figure 2):

(1) Retrieval of candidate justification sentences: For datasets that rely on huge supporting KBs (e.g., ARC), we retrieve the top n sentences[1] from this KB using an IR query that concatenates the question and the candidate answer, similar to Clark et al. (2018); Yadav et al. (2019). We implemented this using the BM25 IR model with the default parameters in Lucene[2]. For reading comprehension datasets where the question is associated with a text passage (e.g., MultiRC), all the sentences in this passage become candidates.

(2) Generation of candidate justification sets: Since its focus is on knowledge aggregation, ROCC ranks *sets* of justification sentences (see below) rather than individual sentences. In this step we create candidate justification sets by generating $\binom{n}{k}$ groups of sentences from the previous n sentences, using multiple values of k.

(3) Ranking of candidate justification sets: For every candidate justification set, we calculate its ROCC score (see Section 3.1), which estimates the likelihood that this group of justifications explains the given answer. We then rank the justification sets in descending order of ROCC score, and choose the top set as the group of justifications that is the output of ROCC for the given question and answer. In MultiRC, we rearrange the justification sentences according to their original indexes in the given passage to bring coherence in the selected sequence of sentences.

(4) Answer classification: ROCC can be coupled with any supervised QA component for answer classification. In this work, we feed in the question, answer, and justification texts into a state-of-the-art classifier that relies on BERT (see Section 3.2). Because the justification sentences in the reading comprehension use case (e.g., MultiRC) come from the same passage and their sequence is likely to be coherent, we concatenate them into a single passage, and use a single BERT instance for classification. This approach is shown on the left side of the answer classification component in Figure 2. On the other hand, the justification sentences retrieved from an external KB (e.g., ARC) may not form a coherent passage when aggregated. For this reason, in the ARC use case, we classify each justification sentence separately (together with the question and candidate answer), and then average all these scores to produce a single score for the candidate answer (right-hand side of the figure).

[1] In this work we used $n = 20$ as in Yadav et al. (2019)

[2] https://lucene.apache.org

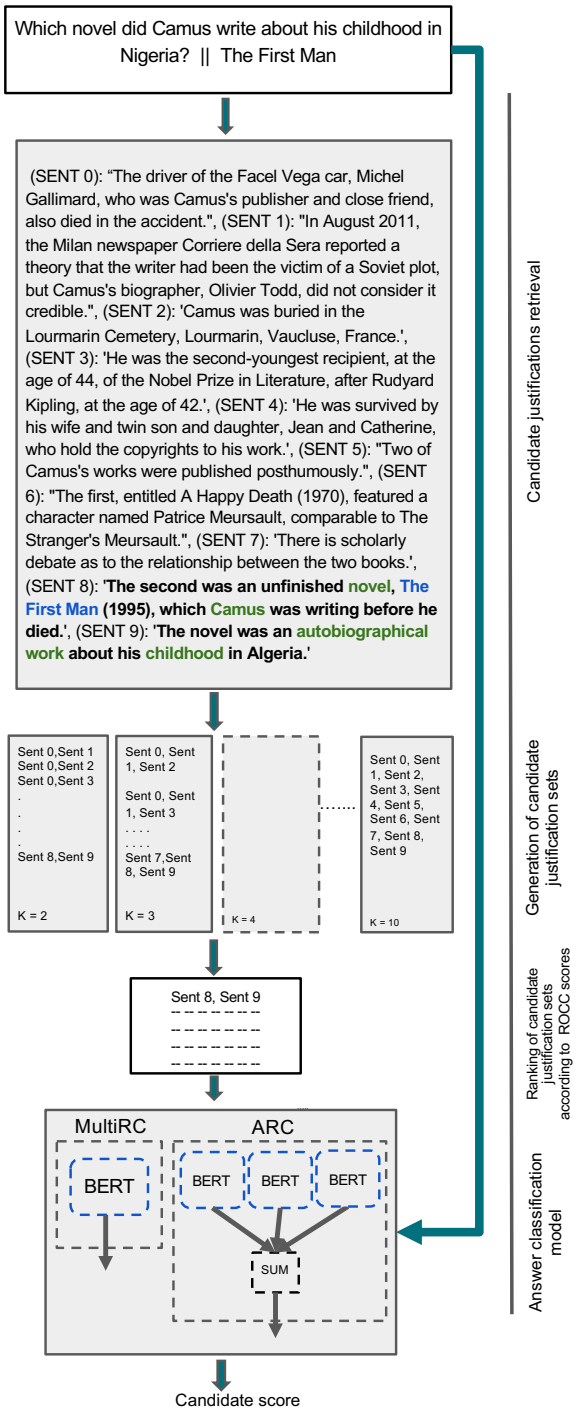

Figure 2: An example of the ROCC process for a question from the MultiRC dataset. Here, ROCC correctly extracts the two justification sentences necessary to explain the correct answer.

3.1 Ranking of Candidate Justification Sets

Each set of justifications is ranked based on its ROCC score, which: (a) maximizes the *R*elevance of the selected sentences; (b) minimizes the lexical *O*verlap between the selected facts; and (c) maximizes the lexical *C*overage of both question and answer (C_{ques}, C_{ans}). The overall score for a given justification set P_i is calculated as:

$$S(P_i) = \frac{R}{\epsilon + O(P_i)} \cdot (\epsilon + C(A)) \cdot (\epsilon + C(Q)) \quad (1)$$

To avoid zeros, we add a small constant ($\epsilon = 1$ here) to each component that can have a value of 0.[3] We detail the components of this formula below.

Relevance (R) We use the Lucene implementation[4] of the BM25 IR model (Robertson et al., 2009) to estimate the relevance of each justification sentence to a given question and candidate answer. In particular, we form a query that concatenates the question and candidate answer, and use as underlying document collection (necessary to compute document statistics such as inverse document frequencies (IDF)) either: sentences in the entire KB (for ARC), or all sentences in the corresponding passage in the case of reading comprehension (MultiRC). The arithmetic mean of BM25 scores over all sentences in a given justification set gives the value of R for the entire set.

Overlap (O) To ensure diversity and complementarity between justification sentences, we compute the overlap between all sentence pairs in a given group. Thus, minimizing this score reduces redundancy and encourages the aggregated sentences to address *different* parts of the question and answer:

$$O(S) = \frac{\sum_{s_i \in S} \sum_{s_j \in S - s_i} \frac{|t(s_i) \cap t(s_j)|}{\max(|t(s_i)|, |t(s_j)|)}}{\binom{|S|}{2}} \quad (2)$$

where S is the given set of justification sentences; s_i is the i^{th} sentence in S; and $t(s_i)$ denotes the set of unique terms in sentence s_i. Note that we divide by $\binom{|S|}{2}$ to normalize across different sizes of justification sets.

Coverage (C) Complementing the overlap score, this component measures the lexical coverage of the question and the answer texts by the given set of justifications S. This coverage is weighted by the IDF of question and answer terms. Thus, maximizing this value encourages the justifications to address more of the meaningful content mentioned

[3] Our R score relies on BM25, which is larger than 0 on the top n sentences.

[4] https://lucene.apache.org/core/7_0_1/core/org/apache/lucene/search/similarities/BM25Similarity.html

in the question ($X = Q$) and the answer ($X = A$):

$$C_t(X) = \bigcup_{s_i \in S} t(X) \cap t(s_i) \quad (3)$$

$$C(X) = \frac{\sum_{t=1}^{|C_t(X)|} IDF[C_t(X)[t]]}{|t(X)|} \quad (4)$$

where $t(X)$ denotes the unique terms in X, and $C_t(X)$ represents the set of all unique terms in X that are present in any of the sentences of the given justification set. $C(X)$ gives the IDF weighted average of $C_t(X)$ terms.

3.2 Answer Classification

As indicated earlier, we propose two flavors for the answer classification component: if the sentences in a justification group come from the same passage and, thus, are likely to be coherent, they are concatenated into a single text before classification, and handled by a single answer classifier. If the sentences come from different texts, they are handled by separate instances of the answer classifier. In the latter case, all scores are averaged to produce a single score for a candidate answer. In all situations we used BERT (Devlin et al., 2018) for answer classification. In particular, we employed BERT as a binary classifier operating over two texts. The first text consists of the concatenated question and answer, and the second text consists of the justification text. The classifier operates over the hidden states of the two texts, i.e., the state corresponding to the `[CLS]` token (Devlin et al., 2018).[5]

We observed empirically that pre-training the BERT classifier on *all* n sentences retrieved by BM25, and then fine tuning on the ROCC justifications improves performance on all datasets we experimented with. This resembles the transfer learning discussed by Howard and Ruder (2018), where the source domain would be the BM25 sentences, and the target domain the ROCC justifications. However, one important distinction is that, in our case, all this knowledge comes solely from the resources provided within each dataset, and is retrieved using unsupervised method (BM25). We conjecture that this helped mainly because the pre-training step exposed BERT to more data which, even if imperfect, is topically related to the corresponding question and answer.

[5]We used the following hyper parameters with BERT Large: learning rate of 1e-5, maximum sequence length of 128, batch size = 16, number of epochs = 6.

Question + answer text	Justification set
Animal cells obtain energy by || absorbing nutrients	1) obtain water and nutrient by absorbing them directly into plant cell 2) the animal obtain nourishment by absorbing nutrient released by symbiotic bacteria

Table 1: Example of a justification set in ARC which was scored by annotator with a precision of $\frac{1}{2}$ because the first justification sentence is not relevant, and a coverage of $\frac{1}{2}$ because the link between *nourishment* and *energy* is not covered.

4 Empirical Evaluation

We evaluated ROCC coupled with the proposed QA approach on two QA datasets. We use the standard train/development/test partitions for each dataset, as well as the standard evaluation measures: accuracy for ARC (Clark et al., 2018), and $F1_m$ (macro-F1 score), $F1_a$ (micro-F1 score), and EM0 (exact match) for MultiRC (Khashabi et al., 2018a).

Multi-sentence reading comprehension (MultiRC): this is a reading comprehension dataset implemented as multiple-choice QA (Khashabi et al., 2018a). Each question is accompanied by a supporting passage, which contains the correct answer. We use all sentences from such paragraphs as candidate justifications for the corresponding questions.

AI2's Reasoning Challenge (ARC): this is a multiple-choice question dataset, containing questions from science exams from grade 3 to grade 9 (Clark et al., 2018). The dataset is split in two partitions: Easy and Challenge, where the latter partition contains the more difficult questions that require reasoning. Most of the questions have 4 answer choices, with <1% of all the questions having either 3 or 5 answer choices. Importantly, ARC includes a supporting KB of 14.3M unstructured text passages. We use BM25 over this entire KB to retrieve candidate justification sentences for ROCC.

4.1 Justification Results

To demonstrate that ROCC has the capacity to select better justification sentences, we also report the quality of the extracted justification sentences. For MultiRC, we report precision/recall/F1 justification scores, computed against the gold justification sentences provided by the dataset.[6] For ARC, where gold justifications are not provided, we used an

[6]We use these gold justifications only for evaluation, *not* for training, since ROCC is an unsupervised algorithm.

#	External resource?	Supervised selection of justifications?	Method	$F1_m$	$F1_a$	EM0	Justification		
							P	R	F1
			DEVELOPMENT DATASET						
			Baselines						
0	No	No	Predict 1 (Khashabi et al., 2018a)	61.0	59.9	0.8	–		
1	No	No	IR(paragraphs) (Khashabi et al., 2018a)	64.3	60.0	1.4	–		
2	No	No	SurfaceLR (Khashabi et al., 2018a)	66.5	63.2	11.8	–		
3	No	No	Entailment baseline (Trivedi et al., 2019)	51.3	50.4	–	–		
			Previous work						
4	Yes	Yes	EER_{DPL} + FT Wang et al. (2019)	70.5	67.8	13.3	–		
5	Yes	Yes	Multee (GloVe) (Trivedi et al., 2019)	71.3	68.3	17.9	–		
6	No	Yes	Multee (ELMo) (Trivedi et al., 2019)	***70.3***	67.3	***22.8***	–		
7	Yes	Yes	Multee (ELMo) (Trivedi et al., 2019)	73.0	69.6	22.8	–		
8	No	Yes	RS (Sun et al., 2018)	69.7	***67.9***	16.9	–		
9	Yes	Yes	RS (Sun et al., 2018)	**73.1***	**70.5***	21.8	–		
			BERT + IR baselines						
10	No	No	BERT + entire passage	65.7	62.7	17.0	17.4	100.0	29.6
11	No	No	BERT + BM25 ($k = 1$ sentence)	66.2	62.8	17.9	**61.0**	27.1	37.5
12	No	No	BERT + BM25 ($k = 2$ sentences)	68.1	64.8	21.0	51.6	45.6	48.4
13	No	No	BERT + BM25 ($k = 3$ sentences)	69.1	65.7	21.6	42.6	56.1	48.4
14	No	No	BERT + BM25 ($k = 4$ sentences)	70.05	66.7	22.3	36.9	64.6	47.0
15	No	No	BERT + BM25 ($k = 5$ sentences)	71.2	67.7	23.4	32.7	71.1	44.8
			BERT + parametric ROCC						
16	No	No	BERT + ROCC ($k = 2$ sentences)	69.8	66.8	22.7	54.7	48.5	51.4
17	No	No	BERT + ROCC ($k = 3$ sentences)	72.7	**69.7**	25.2	48.0	63.5	54.7
18	No	No	BERT + ROCC ($k = 4$ sentences)	72.2	69.0	25.0	40.6	71.0	51.6
19	No	No	BERT + ROCC ($k = 5$ sentences)	71.6	68.7	22.7	35.0	**76.5**	48.1
			BERT + non-parametric ROCC						
20	No	No	BERT + AutoROCC ($k \in \{2, 3, 4\}$)	72.0	69.0	21.9	48.9	66.5	56.3
21	No	No	BERT + AutoROCC ($k \in \{2, 3, 4, 5\}$)	72.0	68.8	23.5	48.3	67.7	56.4
22	No	No	BERT + AutoROCC ($k \in \{2, 3, 4, 5, 6\}$)	72.1	69.2	**25.3**	48.2	68.2	**56.4**
23	No	No	BERT + BM25 (k from best AutoROCC)	71.1	67.4	23.1	43.8	61.2	51.0
24	No	No	BERT + AutoROCC ($k \in \{2, 3, 4, 5, 6\}$, pre-trained)	**72.9**	69.6	24.7	48.2	68.2	**56.4**
			Ceiling systems with gold justifications						
25	Yes	Yes	EER_{gt} + FT (Wang et al., 2019)	72.3	70.1	19.2	- -		
26	No	Yes	BERT + Gold knowledge	79.1	75.4	37.6	100.0	100.0	100.0
27	-	-	Human	86.4	83.8	56.6	–		
			TEST DATASET						
28	No	No	SurfaceLR (Khashabi et al., 2018a)	66.9	63.5	12.8			
29	Yes	Yes	Multee (ELMo) (Trivedi et al., 2019)	73.8	70.4	24.5	–		
30	No	No	BERT + AutoROCC ($k \in \{2, 3, 4, 5, 6\}$, pre-trained)	**73.8**	**70.6**	**26.1**			

Table 2: Performance on the MultiRC dataset, under various configurations. k indicates the size(s) of the sets of justification sentences. In parametric ROCC, k is a hyper parameter; in AutoROCC, k is selected automatically. The pre-trained ROCC configurations pre-train BERT on the entire passage corresponding to the question, before fine tuning it on the ROCC sentences. Bold values with * indicate state-of-the-art results that used external labeled resources or other supervised methods for the selection of justification sentences. Italicized bold values show state-of-the-art results from experiments that do not use any external labeled resources.

external annotator to annotate the justifications for a random stratified sample of 70 questions, with 10 questions selected from each grade (3 – 9).The annotator reported two scores: precision, and coverage. Precision was defined as the fraction of justification sentences that are relevant for the inference necessary to connect the corresponding question and candidate answer. Coverage was defined as 1 if the justification set completely covers the inference process for the given question and answer, $1/2$ if the set of justifications partially addresses the inference, and 0 if the justification set is completely irrelevant. Table 1 illustrates these scores with an actual output from ARC.

4.2 Question answering results

In addition to comparing ROCC with previously reported results, we include multiple baselines: (a) the BERT answer classifier trained on the entire passage of the given question (MultiRC), to demonstrate that ROCC has the capacity to filter out irrelevant content from these paragraphs; (b) BERT trained without any justification sentences (ARC), to show that ROCC has the capacity to aggregate useful information from large unstructured KBs, and (c) BERT trained on sentences retrieved using BM25, to demonstrate that ROCC performs better than other unsupervised approaches. Note that the

#	External resources used?	Supervised selection of justifications?	Method	Challenge	Easy	All	Justification P, Coverage
			Baselines				
0	No	No	AI2 IR Solver (Clark et al., 2018)	59.99	23.98		> 0
1	No	No	Sanity Check (Yadav et al., 2018)	58.36	26.56		> 0
2	Yes	No	Tuple-Inf (Clark et al., 2018)	60.71	23.83		> 0
3	Yes	No	DGEM (Clark et al., 2018)	58.97	27.11		> 0
			Previous work				
4	Yes	–	Bi-LSTM max-out (Mihaylov et al., 2018)	33.87	34.26		$= 0$
8	No	No	AHE (Yadav et al., 2019)	33.28	***63.22***	53.31	
9	No	–	Reading Strategies (Sun et al., 2018)	***35.40***	63.10	***53.94***	$= 0$
10	Yes	–	Reading Strategies (Sun et al., 2018)	**42.30***	**68.90***	**60.19***	$= 0$
			BERT + IR baselines				
11	No	–	BERT	35.11	52.75	46.94	
12	No	No	BERT + BM25 ($k = 1$ sentence)	33.87	56.23	48.85	
13	No	No	BERT + BM25 ($k = 2$ sentences)	38.65	60.50	53.29	
14	No	No	BERT + BM25 ($k = 3$ sentences)	41.04	63.19	55.89	
15	No	No	BERT + BM25 ($k = 4$ sentences)	37.9	63.49	53.90	
16	No	No	BERT + BM25 ($k = 5$ sentences)	38.01	61.28	53.60	
			BERT + parametric ROCC				
17	No	No	BERT + ROCC ($k = 2$ sentences)	36.65	60.59	52.69	
18	No	No	BERT + ROCC ($k = 3$ sentences)	39.29	62.97	55.16	
19	No	No	BERT + ROCC ($k = 4$ sentences)	40.39	61.13	54.29	
20	No	No	BERT + ROCC ($k = 5$ sentences)	40.62	59.96	53.58	
			BERT + non-parametric ROCC				
21	No	No	BERT + AutoROCC ($k \in \{2, 3, ...20\}$)	40.73	63.64	56.09	48.04, 62.50
22	No	No	BERT + BM25 (k from best AutoROCC)	39.24	61.01	53.83	42.55, 55.88
23	No	No	BERT + AutoROCC ($k \in \{2, 3, ...20\}$, pre-trained)	**41.24**	**64.49**	**56.82**	**48.04, 62.50**

Table 3: Performance on the ARC dataset, under various configurations. Notations are the same as in Table 2.

BM25 baseline has an additional hyper parameter: the number of sentences to be considered (k).

Table 2 reports comprehensive results on MultiRC, including both overall QA performance, measured using $F1_m$, $F1_a$, and EM0, as well as justification quality, measured using standard precision (P), recall (R), and F1. Note that the bulk of the results are reported on the development partition. The last row in the table reports results on the test partition, computed using the official submission portal which can be accessed only once per model (including its variants). To understand ROCC's behavior, the table includes both the parametric form of ROCC, where the size of the justification sets (k) is manually tuned as well as the non-parametric ROCC, where k is automatically selected in the third step of the ROCC algorithm (see Figure 2) by sorting across all sizes of justification sets together, instead of sorting within each value of k. Table 3 lists equivalent results on ARC.

We draw several observations from these tables:

(1) Despite its simplicity, ROCC combined with the BERT classifier obtains new state-of-the-art performance on both MultiRC and ARC for the class of approaches that do not use external resources to either train the justification sentence selection or the answer classifier. For example, ROCC outperforms the previous best result in MultiRC by 2.5 EM0 points on the development partition (row 24 vs. row 6), and 1.6 EM0 points on test (row 30 vs. row 29). In ARC, ROCC outperforms the previous best approach by 5.8% accuracy on the Challenge partition, and 2.9% overall (row 23 vs. row 9).

(2) On both datasets, the non-parametric form of ROCC (AutoROCC) slightly outperforms the parametric variant. Importantly, it always achieves higher justification scores compared to the parametric ROCC. In MultiRC, AutoROCC outperforms our baseline of BERT + entire passage (row 10 vs 22) by 8.3% EM0, indicating that AutoROCC can filter out irrelevant content. In ARC, AutoROCC outperforms the baseline with no justification sentences by 9.1% (row 21 vs row 11), demonstrating that ROCC aggregates useful knowledge.

(3) The results of the parametric forms of ROCC (rows 16 – 19 in Table 2 and rows 17 – 20 in Table 3) indicate that performance continues to increase until $k = 4$ in MultiRC and $k = 3$ in ARC. This indicates that: (a) knowledge aggregation is beneficial for these tasks; (b) ROCC can robustly handle non-trivial cases of aggregation with larger values of k; and (c) similar to other QA methods (Chen and Durrett, 2019), performance

train/test	Science textbook	Fiction	News	Wiki articles	wikiMovie Summaries	Society, Law and Justice	All
AutoROCC	54.57	53.88	54.32	60.49	57.10	61.06	56.44
BERT+All passages	55.15	55.46	68.77	65.14	57.39	58.79	60.90
BERT+Science textbook	55.67	41.01	51.45	50.06	54.96	48.84	50.79
BERT+Fiction	45.16	57.60	63.05	63.13	59.98	50.94	58.31
BERT+News	44.11	50.77	68.82	65.45	57.01	58.30	59.30
GPT-2 (Wang et al., 2019)	-	-	-	-	-	-	60.7

Table 4: Domain robustness of the non-parametric ROCC vs. a supervised sentence selection model, evaluated on the gold justification sentences from MultiRC. Each column represents a section of the MultiRC development set. Each row after AutoROCC represents a justification sentence selection component trained only on the specified section of MultiRC (these sections are listed in descending order of the number of passages in the training data).

decreases for large values of k, suggesting that knowledge aggregation remains an open research challenge.

(4) The justification scores in both datasets are considerably higher than the equivalent configuration that uses BM25 instead of ROCC (i.e., row 24 vs. row 23 in Table 2, and row 23 vs. row 22 in Table 3). This confirms that the *joint* scoring of sets of justifications that ROCC performs is better than the individual ranking of justification sentences performed by standard IR models such as BM25.

4.3 Domain Robustness Analysis

To understand ROCC's domain robustness, we compared it against a supervised BERT-based classifier for the selection of justification sentences, as well as against GPT-2 (Wang et al., 2019). For this experiment, we used MultiRC, where gold justifications are provided. We used this data to train a classifier for the selection of justification sentences on various domain-specific sections of MultiRC. The results of this experiment are shown in Table 4. Unsurprisingly, training and testing in the same domain (e.g., Fiction) leads to the best performance on sentence selection. However, ROCC is more stable across domains than the supervised sentence selection component, with a difference of over 10 F1 points in some configurations. This suggests that ROCC is a better solution for real-world use cases where the distribution of the test data may be very different from the training data.

Compared to BERT, the unsupervised AutoROCC achieves almost the same or better performance in the majority of the domains except Wiki articles and News. We conjecture this happens because the BERT language model was trained on a large text corpus that comes from these two domains. However, importantly, AutoROCC is more robust across domains that are different from these two, since it is an unsupervised approach that is not tuned for any specific domain.

#	Ablations	ARC	MultiRC EM0	MultiRC Justification F1
0	Full AutoROCC	**56.09**	**25.29**	**56.44**
1	– IDF	54.11	24.65	54.19
2	– $C(A)$	54.90	21.82	52.93
3	– $C(Q)$	54.66	23.61	52.09
4	– O	55.88	24.03	55.97
5	R*	53.90	23.40	44.81

Table 5: Ablation study, removing different components of ROCC. The scores are reported on the ARC test set and MultiRC dev set. R* denotes the best approach that relies just on the R score. The hyper parameter k in R*, was tuned on the development partition of the respective dataset.

The ARC dataset does not provide justification sentences, so we instead ask how well our question-answering models do on a related inference task, the SciTail entailment dataset (Khot et al., 2018). We trained three QA classifiers on the ARC dataset: BERT with no justification, BERT with BM25 ($k = 4$) justifications, and BERT with AutoROCC justifications. We tested these on SciTail, and achieved 64.49%, 69.70%, and 73.46% accuracy, respectively, indicating that AutoROCC's knowledge aggregation is a valid proxy for entailment.

4.4 Ablation Analysis

Table 5 shows an ablation of the different components of ROCC. Row 0 reports the score from the full AutoROCC model. In row 1, we remove IDF weights from coverage calculations (see eq. (4))

Question type	Precision	Recall	F-1 score
True/False/Yes/No	54.1	68.9	60.6
Verbatim	49.7	71.2	58.5
Non-verbatim	47.3	68.7	56.0

Table 6: Justification selection performance of AutoROCC on different types of questions, in the MultiRC development dataset.

of both question and answer text. In row 2, 3 and 4, we remove the coverage of answer, coverage of question, and overlap from the ROCC formula (see eq. (1)) respectively. In all the cases, we found small drops in both performance and justification scores across both the datasets, with the removal of either $C(A)$ or $C(Q)$ having the largest impact.

4.5 Error Analysis

We analyzed ROCC's justification selection performance on three different types of questions in MultiRC: True/False/Yes/No, Verbatim, and Non-verbatim (Khashabi et al., 2018b). As shown in Table 6, AutoROCC achieves higher recall scores on Verbatim questions, where the answer text is likely to appear within the given justification passage, and worse recall on question types where such overlap does not exist, e.g., Non-verbatim and True/False. This suggests that the $C(A)$ component of ROCC is important for the extraction of meaningful justifications.

4.6 Alignment ROCC

To understand the dependence between ROCC and exact lexical match, we compare the justification selection performance of ROCC when its score components are computed based on lexical match (the approach used throughout the paper up to this point) vs. the semantic alignment match of Yadav et al. (2018). The latter approach relaxes the requirement for lexical match, i.e., two tokens are considered to be matched when the cosine similarity of their embedding vectors is larger than 0.95.[7] As shown in Table 7, the alignment-based ROCC indeed performs better than the ROCC that relies on lexical match. However, the improvements are not large, e.g., the maximum improvement is 1.6% (when $k = 4$), which indicates that ROCC is robust to a certain extent to lexical variation.

[7]This threshold was tuned on the MultiRC development set. We used 100-dimensional GloVe embeddings for this experiment, which performed similarly to larger embedding vectors (300), but allowed for faster experiments.

ROCC (k sentences)	Lexical ROCC	Align. ROCC
ROCC ($k = 2$ sentences)	51.4	51.4
ROCC ($k = 3$ sentences)	54.7	55.5
ROCC ($k = 4$ sentences)	51.6	53.2
ROCC ($k = 5$ sentences)	48.1	49.2

Table 7: Justification selection performance of the ROCC configuration that uses lexical match (BM25) to retrieve candidate justifications (Lexical ROCC), compared against a ROCC variant that uses the semantic alignment approach of Yadav et al. (2018) to retrieve candidates (Align. ROCC). This experiment used the MultiRC development dataset.

5 Conclusion

We introduced ROCC, a simple unsupervised approach for selecting justification sentences for question answering, which balances relevance, overlap of selected sentences, and coverage of the question and answer. We coupled this method with a state-of-the-art BERT-based supervised question answering system, and achieved a new state-of-the-art on the MultiRC and ARC datasets among approaches that do not use external resources during training. We showed that ROCC-based QA approaches are more robust across domains, and generalize better to other related tasks like entailment. In the future, we envision that ROCC scores can be used as distant supervision signal to train supervised justification selection methods.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the World Modelers program, grant number W911NF1810014, and by the National Science Foundation (NSF) under grant IIS-1815948.

Mihai Surdeanu declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

David Alvarez-Melis and Tommi S Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *arXiv preprint arXiv:1707.01943*.

Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017.

" what is relevant in a text document?": An interpretable machine learning approach. *PloS one*, 12(8):e0181142.

Seohyun Back, Seunghak Yu, Sathish Reddy Indurthi, Jihie Kim, and Jaegul Choo. 2018. Memoreader: Large-scale reading comprehension through neural memory controller. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2131–2140.

Junwei Bao, Nan Duan, Zhao Yan, Ming Zhou, and Tiejun Zhao. 2016. Constraint-based question answering with knowledge graph. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2503–2514.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. *arXiv preprint arXiv:1809.06309*.

Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *IJCAI-17 workshop on explainable AI (XAI)*, volume 8.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

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

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. *arXiv preprint arXiv:1904.12106*.

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–220.

Yu-An Chung, Hung-Yi Lee, and James Glass. 2017. Supervised and unsupervised transfer learning for question answering. *arXiv preprint arXiv:1711.05345*.

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*.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. Kbqa: learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10(5):565–576.

Nicola De Cao, Wilker Aziz, and Ivan Titov. 2018. Question answering by reasoning across documents with graph convolutional networks. *arXiv preprint arXiv:1808.09920*.

Mostafa Dehghani, Hosein Azarbonyad, Jaap Kamps, and Maarten de Rijke. 2019. Learning to transform, combine, and reason in open-domain question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 681–689. ACM.

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*.

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

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

Mor Geva and Jonathan Berant. 2018. Learning to search in long documents using document structure. *arXiv preprint arXiv:1806.03529*.

Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 80–89. IEEE.

Alessio Gravina, Federico Rossetto, Silvia Severini, and Giuseppe Attardi. 2018. Cross attention for selection-based question answering. In *2nd Workshop on Natural Language for Artificial Intelligence*. Aachen: R. Piskac.

Yanchao Hao, Yuanzhe Zhang, Kang Liu, Shizhu He, Zhanyi Liu, Hua Wu, and Jun Zhao. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 221–231.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018a. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. 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 252–262.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. *arXiv preprint arXiv:1604.06076*.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018b. Question answering as global reasoning over semantic abstractions. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2017. Answering complex questions using open information extraction. *arXiv preprint arXiv:1704.05572*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. Scitail: A textual entailment dataset from science question answering. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Tuan Manh Lai, Trung Bui, and Sheng Li. 2018. A review on deep learning techniques applied to answer selection. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2132–2144.

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.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Sewon Min, Minjoon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. *arXiv preprint arXiv:1702.02171*.

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*.

Xiaoman Pan, Kai Sun, Dian Yu, Heng Ji, and Dong Yu. 2019. Improving question answering with external knowledge. *arXiv preprint arXiv:1902.00993*.

Minghui Qiu, Liu Yang, Feng Ji, Weipeng Zhao, Wei Zhou, Jun Huang, Haiqing Chen, W Bruce Croft, and Wei Lin. 2018. Transfer learning for context-aware question matching in information-seeking conversations in e-commerce. *arXiv preprint arXiv:1806.05434*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Robert Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, pages 4444–4451.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge dataset and models for dialogue-based reading comprehension. *arXiv preprint arXiv:1902.00164*.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2018. Improving machine reading comprehension with general reading strategies. *arXiv preprint arXiv:1810.13441*.

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.

Nam Khanh Tran and Claudia Niedereée. 2018. Multihop attention networks for question answer matching. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 325–334. ACM.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. *arXiv preprint arXiv:1904.09380*.

Kateryna Tymoshenko, Daniele Bonadiman, and Alessandro Moschitti. 2017. Ranking kernels for structures and embeddings: A hybrid preference and classification model. In *EMNLP*.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, Dan Roth, and David McAllester. 2019. Evidence sentence extraction for machine reading comprehension. *arXiv preprint arXiv:1902.08852*.

Shuohang Wang and Jing Jiang. 2016. A compare-aggregate model for matching text sequences. *arXiv preprint arXiv:1611.01747*.

Shuohang Wang, Mo YU, Jing JIANG, Wei ZHANG, Xiaoxiao GUO, Shiyu CHANG, Zhiguo WANG, Tim KLINGER, Gerald TESAURO, and Murray CAMPBELL. 2018a. Evidence aggregation for answer re-ranking in open-domain question answering.

Yizhong Wang, Kai Liu, Jing Liu, Wei He, Yajuan Lyu, Hua Wu, Sujian Li, and Haifeng Wang. 2018b. Multi-passage machine reading comprehension with cross-passage answer verification. *arXiv preprint arXiv:1805.02220*.

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

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019. Alignment over heterogeneous embeddings for question answering. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long Papers)*, Minneapolis, USA. Association for Computational Linguistics.

Vikas Yadav, Rebecca Sharp, and Mihai Surdeanu. 2018. Sanity check: A strong alignment and information retrieval baseline for question answering. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1217–1220. ACM.

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*.

Lei Yu, Karl Moritz Hermann, Phil Blunsom, and Stephen Pulman. 2014. Deep learning for answer sentence selection. *arXiv preprint arXiv:1412.1632*.

Yuanzhe Zhang, Shizhu He, Kang Liu, and Jun Zhao. 2016. A joint model for question answering over multiple knowledge bases. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Yuyu Zhang, Hanjun Dai, Kamil Toraman, and Le Song. 2018. Kg$\hat{2}$: Learning to reason science exam questions with contextual knowledge graph embeddings. *CoRR*, abs/1805.12393.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2018. An interpretable reasoning network for multi-relation question answering. *arXiv preprint arXiv:1801.04726*.