If You Want to Go Far Go Together: Unsupervised Joint Candidate Evidence Retrieval for Multi-hop Question Answering

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

Abstract

Multi-hop reasoning requires aggregation and inference from multiple facts. To retrieve such facts, we propose a simple approach that retrieves and reranks set of evidence facts jointly. Our approach first generates unsupervised clusters of sentences as candidate evidence by accounting links between sentences and coverage with the given query. Then, a RoBERTa-based reranker is trained to bring the most representative evidence cluster to the top. We specifically emphasize on the importance of retrieving evidence jointly by showing several comparative analyses to other methods that retrieve and rerank evidence sentences individually. First, we introduce several attention- and embedding-based analyses, which indicate that jointly retrieving and reranking approaches can learn compositional knowledge required for multi-hop reasoning. Second, our experiments show that jointly retrieving candidate evidence leads to substantially higher evidence retrieval performance when fed to the same supervised reranker. In particular, our joint retrieval and then reranking approach achieves new state-of-the-art evidence retrieval performance on two multi-hop question answering (QA) datasets: 30.5 Recall@2 on QASC, and 67.6% F1 on MultiRC. When the evidence text from our joint retrieval approach is fed to a RoBERTa-based answer selection classifier, we achieve new state-of-the-art QA performance on MultiRC and second best result on QASC.

1 Introduction

Recent advances in question answering (QA) have achieved excellent performance on several benchmark datasets (Wang et al., 2019a), even when relying on partial (Gururangan et al., 2018), incorrect (Jia and Liang, 2017) or no supporting knowledge (Raffel et al., 2019). Specifically, black-box neural QA methods have shown to rely on spurious signals confirming unfaithful or non-explainable behavior (Geva et al., 2019). Thus, justifying the underlying knowledge or evidence text has been deemed very important for faithfulness and explainability of neural QA methods (DeYoung et al., 2019; Yang et al., 2018). Our work is also focused on improving the explainability of QA methods by the means of evidence (or justification) sentence retrieval.

Question: RNA is a small molecule that can squeeze through pores in (A) dermal & vascular tissue (B) space between (C) **eukaryotic cells** (D) jellyfish (E) ··· ··· (H)

Gold evidence sentences:
1. RNA is a small molecule that can squeeze through pores in the nuclear membrane
2. Cells with a nuclear membrane are called eukaryotic.

BM25 sentences:
1. RNA is a small molecule that can squeeze through pores in the nuclear membrane.
2. RNA synthesis in eukaryotic cells is synthesized by three types of RNA polymerases
3. Eukaryotic cells have three different RNA polymerases.
4. the molecule seems to have evolved specifically to parasitize eukaryotic cells

WAIR Step-1 sentences:
1. RNA is a small molecule that can squeeze through pores in the nuclear membrane.
2. RNA synthesis in eukaryotic cells is synthesized by three types of RNA polymerases

WAIR Step-2 sentences:
1. Cells with a nuclear membrane are called eukaryotic
2. Eukaryotic cells have three different RNA polymerases.

Figure 1: An example question from the QASC dataset with evidece sentences retrieved by BM25 and two steps of WAIR. The evidence retrieved in step-2 of WAIR contain information missed by sentences in step-1 and are associated with each other. Both the gold evidence are also found in sentences from step-1 and step-2.

Evidence retrieval for multi-hop QA is a challenging task as it requires compositional inference based aggregation of multiple evidence sentences (Yang et al., 2018; Khashabi et al., 2018; Welbl et al., 2018; Khot et al., 2019a). For such compositional aggregation, we emphasize on the importance of jointly handeling the set of evidence

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4571–4581
June 6–11, 2021.

facts within the QA pipeline. The motivation behind our work is simple: jointly handling evidence sentences gives access to the complete information together and thus enable compositional reasoning. On the other hand, handling evidence sentences individually leads to selection of disconnected evidence that do not support compositional multi-hop reasoning (Jansen, 2018; Chen and Durrett, 2019).

For retrieving compositional evidence, we propose a simple unsupervised retriever - ***w***eighted ***a***lignment-based ***i***nformation ***r***etrieval algorithm (WAIR) that generates candidate evidence chains based on two key heuristics - *coverage* and *associativity*. Coverage denotes the proportion of query covered by the evidence text and associativity denotes links between individual evidence sentences. We show that WAIR evidence candidate chains lead to substantially higher retrieval performance when compared to the other approaches that handle evidence sentences individually. Particularly, we show that just feeding the candidate evidence chain from WAIR to RoBERTa reranker achieves substantially better performance than when the same reranker is instead fed with individual candidate sentences. Further, we present several attention- and embedding-based analyses of the reranker RoBERTa model highlighting that WAIR retrieved chains enable a) learning of compositional reasoning and, b) complementary knowledge aggregation.

Our overall QA approach operates in three steps. We first retrieve candidate evidence chains for a given query using WAIR. In 2 iterations, our unsupervised WAIR approach weighs down query terms that have already been covered by previously retrieved sentences, and increases the weights of reformulated query terms that have not been covered yet. In the second step of our QA framework, we *jointly rerank* clusters of evidence sentences generated by WAIR. The reranking is implemented as a regression task, where the score assigned to each sentence cluster is F1 score computed from the gold annotated evidence sentences. Lastly, the top reranked set of sentences are fed into an answer classification component.

In particular, our key contributions are:

(1) We introduce a simple, unsupervised and fast evidence retrieval approach -WAIR for multi-hop QA that generates complete and associated candidate evidence chains. To show the multi-hop reasoning approximated within WAIR candidate evidence chains, We present several attention weights and embeddings based analyses[1]. Our attention analyses highlights that jointly retrieving candidate evidence chains using WAIR assists the reranker model to learn contextual and compositional knowledge necessary for multi-hop reasoning. Specifically, our transformer based reranker attends more on the linking terms necessary for combining multiple evidence facts. Further, our embedding based analysis shows that the reranking of WAIR evidence chains helps the reranker to project embedding representations of evidence facts differently, thus allowing complementary knowledge aggregation during the QA stage necessary for multi-hop reasoning.

(2) We show that just the simple construction of candidate evidence using WAIR leads to substantial higher (10.2% $Recall$@2 on QASC (Khot et al., 2019a) and 3.6% F1 on MultiRC (Khashabi et al., 2018)) evidence selection performance with the same RoBERTa reranker over the case when it is fed with individual candidate sentences. Specifically, we achieve the new state-of-the-art evidence selection results on two multi-hop QA datasets - (30.5% $Recall$@2 on QASC and 68.0% on MultiRC. Further, our simple candidate chain generation approach can be coupled with any reranker and QA method, and can be applied to different QA settings, e.g., large KB-based QA such as QASC, reading comprehension and passage-based MCQA such as MultiRC, etc. We also show that the QA performance improves by 2.3% EM0 in MultiRC and 5.2% accuracy in QASC when the top reranked WAIR evidence chain is fed to the QA module over the case of feeding individually reranked sentences. By just feeding the top reranked WAIR evidence chain, we achieve state-of-the-art QA performance on MultiRC and second best QA results on QASC.

2 Related Work

Evidence retrieval has been shown to improve explainability of complex inference based QA tasks (Qi et al., 2019). There are two potential ways to retrieve evidence sentences: *individually* or *jointly*.

Retrieving individual evidence sentences: Most unsupervised information retrieval techniques, e.g., BM25 (Robertson et al., 2009), tf-idf (Ramos et al., 2003; Manning et al., 2008), or alignment-based methods (Kim et al., 2017), have

[1]Codes - https://github.com/vikas95/WAIR_interpretability

been widely used to retrieve evidence texts for open-domain QA tasks (Joshi et al., 2017; Dunn et al., 2017). Although these approaches have been strong benchmarks for decades, they usually do not perform well on recent complex reasoning-based QA tasks (Yang et al., 2018; Khot et al., 2019a). More recently, supervised neural network (NN) based retrieval methods have achieved strong results on complex questions (Karpukhin et al., 2020; Nie et al., 2019; Tu et al., 2019). However, these approaches require annotated data for initial retrieval and suffer from the same disadvantages at the reranking stage as the other methods that retrieve+rerank individual evidence sentences, i.e., the retrieval algorithm is not aware of what information has already been retrieved and what is missing, or how individual facts need to be combined for explaining the multi-hop reasoning (Khot et al., 2019b). Our proposed joint retrieval and reranking approach mitigates both these limitations.

Jointly retrieving evidence sentences: Recently, several works have proposed retrieval of evidence chains that has led to stronger evidence retrieval performance (Yadav et al., 2019b; Khot et al., 2019a). Our WAIR approach aligns in the same direction and particularly utilizes coverage and associativity that leads to higher performance. Importantly, our work focuses on highlighting the benefits of feeding evidence *chains* to transformer based reranking methods. First, the evidence retrieval performance of the same reranker is substantially improved resulting in state-of-the-art performance and thus outperforming all the previous approaches. Second, we show that the candidate evidence chain from WAIR assist reranker method to learn compositional and aggregative reasoning.

Other recent works have proposed supervised iterative and multi-task approaches for evidence retrieval (Feldman and El-Yaniv, 2019; Qi et al., 2019; Banerjee, 2019). But, these supervised chain retrieval approaches are expensive in their runtime and do not scale well on large KB based QA datasets. On the contrary, our retrieval approach does not require any labeling data and is faster because of its unsupervised nature. Further, our joint approach is much simpler, performs well and scales on large KB based QA such as QASC.

In this work, we focus on analyzing the multi-hop evidence reasoning via attention (Clark et al., 2019) and learned embeddings (Ethayarajh, 2019) analyses. Several works have shown attention based analysis on pretrained transformer language models (Rogers et al., 2020) on various NLP tasks including QA (van Aken et al., 2019). Our novel analyses are particularly focused on a) evaluating attention scores on linking terms that approximate multi-hop compositionality and, b) complementary knowledge aggregation necessary for multi-hop QA.

Importance of Evidence Retrieval for Question Answering Several neural QA methods have achieved high performance without relying on evidence texts. Many of these approaches utilize external labeled training data (Raffel et al., 2019; Pan et al., 2019), which limits their portability to other domains. Others rely on pretraining, which tends to be computationally expensive but can be used as starting checkpoints (Devlin et al., 2019; Liu et al., 2019). More importantly, many of these directions lack explanation of their selected answers to the end user. In contrast, QA methods that incorporate an evidence retrieval module can provide these evidence texts as human-readable explanations. Further, several works have demonstrated that *retrieve and read* approaches (similar to ours) tend to achieve higher performance than the former QA methods (Chen et al., 2017; Qi et al., 2019). Our work is inspired by these directions but mostly focuses on jointly retrieving+reranking clusters of evidence sentences that leads to substantial QA performance improvements.

3 Proposed Approach

We summarize the overall execution flow of our QA system in Figure 2. The four key components of the system are explained below.

1. Initial evidence sentence retrieval: In the first step, we retrieve candidate evidence sentences (or justification) given a query. We propose a simple unsupervised approach, which, however, has been designed to bridge the "lexical chasm" inherent between multi-hop questions and their answers (Berger et al., 2000). We call our algorithm ***w***eighted ***a***lignment-based ***i***nformation ***r***etrieval (WAIR). WAIR operates in two steps, by combining ideas from embedding based-alignment (Yadav et al., 2019a) and pseudo-relevance feedback (Bernhard, 2010) approaches.

In its first step, WAIR uses a query that consists of the non-stop words of the original ques-

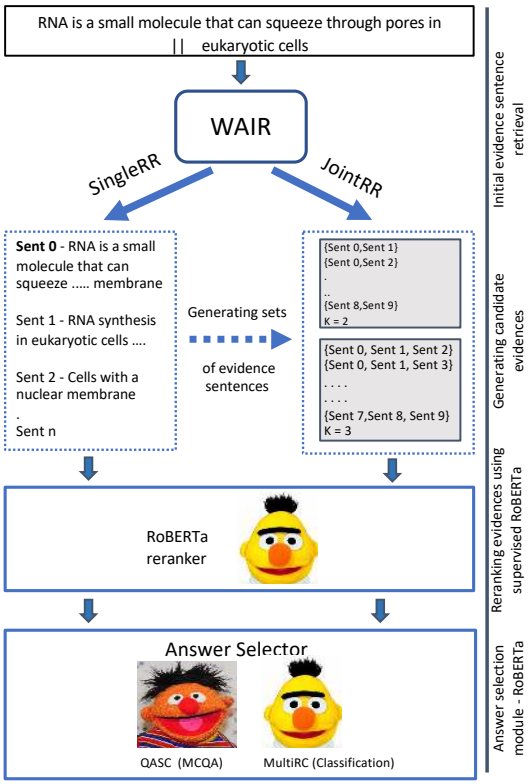

Figure 2: Flow diagram of the overall QA approach. The left branch implements a baseline method, which retrieves and feeds candidate evidence sentences to reranker individually. We denote this method "single sentence retrieval and reranking" (**SingleRR**). The method on the right branch feeds WAIR candidate chains to the RoBERTa reranker which jointly reranks the complete evidence text (referred to as **JointRR**).

tion[2] ($Q = q_1, q_2, ..., q_n$). Using Q, WAIR retrieves k justification sentences ($J_1, J_2, ...J_k$) with the alignment IR method[3] of Yadav et al. (2019a). In the second step, WAIR generates k new queries ($Q_1, Q_2, ...Q_i, ..Q_k$) by concatenating Q with each retrieved justification in the previous step. For each new query Q_i, WAIR assigns a weight[4] of 2 to the original query tokens which are *not* retrieved in the corresponding justification sentence J_i. All the other covered terms in Q_i receive a weight of 1. This simple idea encourages the algorithm to focus on terms that have not yet been retrieved in J_i. Also, weighing uncovered query terms higher encourages the retrieval approach to retrieve the remaining query terms thus yielding higher query

[2]and candidate answer for multiple-choice QA

[3]Please note that for larger KB, BM25 is used to retrieve initial pool of sentences. Then, alignment IR method is applied on this pool to retrieve top k sentences similar to Yadav et al. (2019a)

[4]These term weights were tuned on the training partition

Dataset	WAIR	BM25	Alignment	Gold Evidence
QASC (top 2)	78.85	61.42	63.40	80.81
MultiRC (top 3)	55.92	39.86	52.98	63.95

Table 1: The coverage of question (+candidate answer) terms in the sentences retrieved by various IR techniques. Last column gives the upper-bound of coverage from gold justifications and also suggests the effectiveness of coverage for the retrieval task.

coverage scores as shown in table 1. Further, the concatenation of J_i with Q encourages retrieval of sentences that are *associated or linked* with the previously retrieved sentences. The J_i terms are also weighted 1 to mitigate the semantic drift problem by helping the second retrieval iteration stay close to the original query (see WAIR sentences in fig. 1). In both iterations of WAIR, the score between a given query Q and a justification sentence J is calculated as:

$$s(Q, J) = \sum_{m=1}^{|Q|} idf(q_m) \cdot align(q_m, J) \quad (1)$$

$$align(q_m, J) = \max_{k=1}^{|J|} cosSim(q_m, j_k) \quad (2)$$

where q_m and j_k are the m^{th} and k^{th} terms of the query Q and justification sentence J, respectively. The inverse document frequency values (idf) are computed over the complete knowledge base of QASC (Khot et al., 2019a) and all the paragraphs in MultiRC dataset. The cosine similarity ($cosSim$) is computed over GLoVe embeddings for simiplicity.

2. Generating candidate evidence sets: From the N sentences retrieved in the 2 iterations of previous step, WAIR generates $\binom{N}{p}$ combinations, where p denotes the number of sentences in a candidate evidence chain. To reduce the overhead on the next supervised component, we implemented a beam filter strategy on these sets. We first rank each evidence set E_i by how many query terms are included in the set (referred to as coverage which has been shown as a strong retrieval indicator for multi-hop QA (Wang et al., 2019b) (as also shown in table 1)):

$$C(E_i) = \frac{1}{|t(Q)|} \sum_{w \in t(Q) \cap t(E_i)} idf(w) \quad (3)$$

where $t(Q)$ and $t(E_i)$ denote the unique terms in Q and evidence set E_i, respectively. We then keep the top n sets with the highest coverage score (C). We implement an equivalent process for the **SingleRR**

baseline: we compute the coverage C for individual evidence sentences, and keep the top n.

3. Supervised evidence reranking: This component uses a supervised RoBERTa classifier to rerank evidence sets (for **JointRR**) or classify individual justifications (for **SingleRR**). The latter scenario is modeled as binary classification of individual justification sentences. The former scenario (for **JointRR**) is modeled as a regression task, where the score of each evidence set is the F1 score computed from gold evidence sentences. For example, an evidence set with 3 sentences, out of which 2 are correct has a precision of $2/3$. Assuming 2 gold justifications are not included in the set, its recall is $2/4$, and the F1 score used for regression is 0.57. Please note that we directly use the sets created in the previous step even in the training step i.e., we do not insert gold sentences in the set to keep the consistency between training and test step.

For both classifiers, we used RoBERTa-base with a learning rate of $1e{-}5$, maximum sequence length of 256 [5], batch size of 8, and 4 epochs. For the **SingleRR** approach, all the evidence sentences having probability larger than 0.5 are concatenated to create the final evidence text. For **JointRR** approach, the evidence set with the highest regression score is selected. Similarly, all the sentences in this set are concatenated into a single text.

4. Answer selection: The last component classifies candidate answers given the original question and the evidence text assembled in the previous step. Similar to previous works, we use the multiple-choice question answering (MCQA) architecture of RoBERTa for QASC (Khot et al., 2019a; Wolf et al., 2019) where a softmax is used to discriminate among the eight answer choices. The inputs to RoBERTa-MCQA consist of eight queries (from eight candidate answers) and their corresponding eight evidence texts. The hyperparameters used were: RoBERTa large, maximum sequence length = 128[6] (for each candidate answer), batch size = 8, epoch = 3. For MultiRC, where questions have variable number of candidate answers and multiple correct answers, a RoBERTa binary classifier[7] is used for each candidate seperately.

[5]We tried sequence length of 128 and 512 also but that resulted in 1.5% lower performance

[6]We tried 184 as sequence length (with batch size as 2 to fit on GPU's) but it resulted in 1-2% lower performance for majority of the experiments

[7]hyperparameters same as the RoBERTa retrieval classifier

4 Experimental Results

We focus on complex *non-factoid* and *long answer span* based explainable multi-hop datasets:

Multi-sentence reading comprehension (MultiRC): a reading comprehension dataset provided in the multiple-choice QA format (Khashabi et al., 2018). Every question is supported by one document, from which the answer and justification sentences must be extracted. WAIR retrieves $n = 10$ sentences,[8] which are separately considered as candidates in the downstream components of **SingleRR**. For the **JointRR** approach, we generate combinations of evidence texts with $k \in \{2, 3, 4\}$ sentences, i.e., $\binom{n=10}{k\in\{2,3,4\}}$. We use the original MultiRC dataset[9] which includes the gold annotations for evidence text.

Question Answering using Sentence Composition (QASC): a multiple-choice QA dataset (Khot et al., 2019a), where each question is provided with 8 answer candidates, out of which 4 candidates are hard adversarial choices. The evidence sentences are to be retrieved from a large KB of 17.2 million facts. Similar to Khot et al. (2019a), WAIR first retrieves $n = 10$ sentences[10] for each candidate answer, where the query concatenates the question and candidate answer texts. WAIR uses each of these retrieved sentences to reformulate and reweigh the query, to retrieve an additional 1 sentence in a second iteration. This results in a total of 20 candidate evidence sentences for a given question and candidate answer. We generate evidence chains using the same approach as the one used for MultiRC, except here we focus on $k = 2$, i.e., $\binom{n=20}{k=2}$, because all questions in QASC are annotated with only two gold justification sentences. We report QA and evidence selection performances in both the datasets using standard evaluation measures (Khot et al., 2019a; Khashabi et al., 2018).

4.1 Evidence Retrieval Results

Tables 2 and 4 list the main results for both question answering and evidence retrieval for the two datasets. Table 3 shows a more detailed analysis

[8]The recall for the retrieval of gold evidence sentences is approximately 94% at $n = 10$ in the MultiRC training set.

[9]https://cogcomp.seas.upenn.edu/multirc/

[10]Since QASC is a large KB based dataset, we use BM25 for the retrieval of initial pool of evidence sentences similar to Yadav et al. (2019a) and Khot et al. (2019a).

#	Retrieval steps	Method	Accuracy	Evidence Both found	Evidence At least one found
		Unsupervised Baselines			
1	Single	Lucene BM25	35.6	5.5	56.0
2	Two	Heuristics+IR (Khot et al., 2019a)	32.4	25.2	51.9
		Previous work			
3	-	ESIM Q2Choice (Khot et al., 2019a)	21.1	25.2	51.9
4	Single	BERT-LC (Khot et al., 2019a)	59.8	5.6	54.6
5	Two	BERT-LC (Khot et al., 2019a)	71.0	25.2	51.9
6	Two	BERT-LC[WM]* (Khot et al., 2019a)	78.0	25.2	51.9
7	Two	KF+SIR+2Step* (Banerjee and Baral, 2020)	**82.4**	-	-
8	Two	AIR+RoBERTa (Yadav et al., 2020b)	76.2	25.6	56.6
		Our work			
9	Two	BM25 + RoBERTa	68.0	11.5	51.0
10	Two	Alignment-IR + RoBERTa	71.5	22.8	49.1
11	Two	WAIR + RoBERTa	74.0	23.6	51.1
12	Two	**SingleRR** + RoBERTa	73.4	20.1	**65.3**
13	Two	**JointRR** + RoBERTa	78.6	**30.5**	65.1
14	Two	Pseudo oracle + **JointRR** + RoBERTa	82.4	32.4	69.8
		TEST DATASET			
15	Two	BERT-LC (Khot et al., 2019a)	68.5	-	-
16	Two	BERT-LC[WM]* (Khot et al., 2019a)	73.2	-	-
17	Two	KF+SIR+2Step*(Banerjee and Baral, 2020)	80.0	-	-
18	Two	AIR + RoBERTa* (Yadav et al., 2020b)	**81.0**	-	-
19	Two	**JointRR** + RoBERTa	78.0	-	-

Table 2: Question answering and evidence retrieval results on QASC. The second column indicates if the initial retrieval process is single step (e.g., a single iteration of BM25), or two steps (as in the WAIR approach). $\star$ highlight the methods that use ensembling or external labeled resources. "Both found" reports the recall scores when both the gold justifications are found and "Atleast 1 found" reports the recall when either one or both the gold justifications are found in the top 2 ranked sentences.

for QASC[11] at different levels of recall, i.e., the percentage of gold evidence sentences found in top N reranked evidence sentences ($Recall@N$). We draw following observations from evidence retrieval experiments (answer selection results are discussed in the following subsection):

(1) Unsupervised retrieval: Indicating initial benefits of retrieving evidence chains, our alignment-based evidence retrieval approach (WAIR) outperforms the other IR benchmarks (BM25 and alignment) as shown in rows 10-11 vs. 12-13 in table 4 and rows {1,9,10} vs. 11 in table 2. WAIR also outperforms the two-step IR-based methods for evidence retrieval (row (9, 10 vs. 11) in table 2), highlighting the importance of query reweighing in iterative retrieval methods.

(2) Supervised reranking: Reranking WAIR candidate evidence chains (**JointRR**) leads to absolute 10.4% on QASC (row 12 vs row 13 in table 2) and 3.6% F1 improvement on MultiRC (row 14 vs row 15 in table 4) over the case where the same reranker is fed with individual sentences (**SingleRR**). This highlights the importance of feeding candidate evidence *chains* to the supervised reranker.

(3) Recall comparison: As shown in table 3, just feeding WAIR candidate chains result in higher performance for retrieving complete evidence (the "Both found" columns) than **SingleRR**, especially for low recall scenarios. Notably, **SingleRR** achieves marginally better performance on finding atleast 1 evidence sentence but performs poorly on retrieving both the evidence sentences indicating absence of compositional multi-hop reasoning. We observe similar gains on MultiRC i.e., **JointRR** achieves 6% higher recall compared to **SingleRR** (row 14, row 15 in table 4).

(4) (Pseudo) oracle JointRR: To investigate the ceiling of **JointRR**, we inserted the *gold* justification sentences within the WAIR retrieved sentences and then created candidate evidence chains. These chains were then reranked by the same RoBERTa reranker. As shown in row 18a of table 4 and row 14 of table 2, the performance of **JointRR** approach is substantially improved when gold evidence sentences are retrieved in the initial WAIR pool. The ceiling performance of **JointRR** is much higher than the current actual method (row 13 in table 2 and row 15 in table 4), which suggests there is

[11] We found similar trends for MultiRC but present analysis only on QASC (large KB based QA) because of space constraints.

Recall@N	SingleRR			JointRR		
	Evidence Both found	Evidence Atleast 1 found	QA Accuracy	Evidence Both found	Evidence Atleast 1 found	QA Accuracy
Recall@2	20.1	65.3	73.8	30.5	65.1	78.6
Recall@4	35.0	67.9	74.7	40.5	66.7	80.7
Recall@6	40.2	69.0	77.9	44.1	68.2	80.0
Recall@8	43.3	69.4	76.8	45.2	69.0	79.6
Recall@10	44.4	**69.6**	79.7	**45.3**	69.4	**81.7**

Table 3: Evidence retrieval and QA performance comparison of **SingleRR** and **JointRR** at different recall levels on the QASC development dataset. "Both found" and "Atleast 1 found" notations are same as in table 2 but at top N sentences. Recall@N of "Both found" means when both the gold justifications are found in top N sentences. All the N sentences are concatenated to feed into the answer classifier for QA task.

#	Other-resources /Ensembling	Method	$F1_m$	$F1_a$	EM0	Evidence retrieval P	R	F1
		DEVELOPMENT DATASET						
		Baselines						
1	No	IR(paragraphs) (Khashabi et al., 2018)	64.3	60.0	1.4	–		
2	No	SurfaceLR (Khashabi et al., 2018)	66.5	63.2	11.8	–		
3	No	RoBERTa+ Full passage (Yadav et al., 2020b)	73.9	71.7	28.7	17.4	100.0	29.6
		Previous work						
5	No	EER_{DPL} + FT (Wang et al., 2019b)	70.5	67.8	13.3	–		
6	Yes	Multee (ELMo)* (Trivedi et al., 2019)	73.0	69.6	22.8	–		
7	Yes	RS* (Sun et al., 2019)	73.1	70.5	21.8	–	–	60.8
8	No	AutoROCC (Yadav et al., 2019b)	72.9	69.6	24.7	48.2	68.2	56.4
9	Yes	AIR + RoBERTa (Yadav et al., 2020b)	74.7	72.3	29.3	66.2	63.1	64.2
		Our work						
10	No	3 Evidence sents(BM25) + RoBERTa	70.5	68.0	24.9	42.6	56.1	48.4
11	No	3 Evidence sents(Alignment) + RoBERTa	72.4	69.8	25.1	49.3	65.1	56.1
12	No	3 WAIR sents + RoBERTa	74.3	71.5	24.6	50.9	67.6	58.1
13	No	WAIR max-coverage + RoBERTa	74.2	72.2	27.0	55.0	67.2	60.5
14	No	**SingleRR** + RoBERTa	74.9	72.4	25.9	63.9	64.0	64.0
15	No	**JointRR** + RoBERTa	75.2	72.7	28.2	**65.4**	**69.9**	**67.6**
15a	No	**JointRR** ($\pm$ 1 neighboring sentence) + RoBERTa	**77.0**	**74.5**	**32.9**	65.4	69.9	67.6
		Reranking checkpoints transferred to QA task						
16	No	**SingleRR** transferred	71.7	68.8	21.6	63.9	64.0	64.0
17	No	**JointRR** transferred	75.9	73.1	28.2	65.4	69.9	67.6
		Ceiling systems with gold justifications						
18	No	Oracle knowledge + RoBERTa	81.4	80	39	100.0	100.0	100.0
18a	No	Pseudo oracle + **JointRR** + RoBERTa	77.9	74.8	32.9	87.8	82.9	85.3
19	No	Human	86.4	83.8	56.6	–		
		TEST DATASET						
20	No	SurfaceLR (Khashabi et al., 2018)	66.9	63.5	12.8			
21	Yes	Multee (ELMo)* (Trivedi et al., 2019)	73.8	70.4	24.5	–		
22	No	AutoROCC (Yadav et al., 2019b)	73.8	70.6	26.1			
23	Yes	RoBERTa + AIR (Yadav et al., 2020b)	79.0	76.4	**36.3**			
24	No	**JointRR** ($\pm$ 1 neighboring sentence) + RoBERTa	**79.5**	**76.5**	35.4			

Table 4: Answer selection (column 4-6) and evidence retrieval results (column 7-10) on the MultiRC development and test sets. The second column specifies if any external labeled or ensembling resources were used in the approach. $\pm$ 1 neighboring sentence (row 15a) indicates concatenation of neighboring sentences with the predicted evidence sentences to utilize coreferences in the context.

potential for progress from future works.

(5) State-of-the-art evidence retrieval performance: The top reranked WAIR chain achieves 30.5% Recall@2 on QASC (row 13, table 2) and 67.6% F1 on MultiRC (row 15, table 4). Thus, establishing the new state-of-the-art evidence retrieval performance on both the datasets.

4.2 Answer Selection Results

(1) Impact of two-step evidence retrieval: Unsurprisingly, the two-step evidence retrieval process substantially impacts QA performance (e.g., row 1 vs. row 9 in table 2), which is consistent with the observations of previous works (Khot et al., 2019a; Yadav et al., 2020b). The top reranked WAIR chain

leads to higher QA performance (**+5.2%** on QASC (row 12 vs. 13, table 2), and **2.3% F1** on MultiRC (row 14 vs. 15, table 4)).

(2) Impact of retrieval recall: As shown in table 3, **JointRR** always achieves higher Recall@N score for finding both (or complete) evidence. As a result, it also achieves better QA accuracy when compared to **SingleRR**. On the other hand, **SingleRR** always achieves marginally better performance on finding atleast 1 evidence sentence indicating that retrieval of incomplete information leads to lower QA performance. Further, the best QA performance is also achieved at higher recalls (last row of table 3 and row 15 in table 4).

(3) Ceiling performance: When coupled with the (pseudo) oracle retriever, the QA scores of **JointRR** approaches human performance (row 18, table 4). This emphasizes the importance of evidence retrieval for the QA performance.

(4) Top QA performance: RoBERTa answer classifier that just the uses top reranked evidence of WAIR achieves state-of-the-art QA performance on MultiRC development and test sets. It also achieves the second and third best results on QASC development and test sets. Notably, the approaches that score higher than **JointRR** use ensembling or additional labeled data.

5 Representational Analysis

5.1 Attention Analysis

To better understand the differences in learned features of RoBERTa reranker from WAIR chains (**JointRR**) and individual candidate evidence sentences (**SingleRR**), we performed several analyses of their attention weights. We focus on the attention score on the `[CLS]` token, whose representation is fed into the decision layer of the RoBERTa classifier (Wolf et al., 2019). We compute the attention score from a given token to `[CLS]` by summing up the attention scores from all the 12 heads in each layer (Clark et al., 2019). Similar to Clark et al. (2019); Rogers et al. (2020), we remove the attention scores from $< s >, < /s >$, punctuation and stopword tokens in our analysis.

Attention from semantically matching tokens in query and evidence : Retrieval tasks are often driven by the lexically matching query tokens in the retrieved document(Robertson et al., 2009; Manning et al., 2008). Thus, to understand the focus of the reranker on semantic matching, we compute the attention on `[CLS]` from all the tokens that are not lexically matched between the given question+candidate answer text and the retrieved evidence text (Yadav et al., 2020a). We refer it to as Semantic Matching Attention (*SMA*) score. As shown in table 5, reranker fed with WAIR chain (**JointRR** approach) attends more on the tokens requiring semantic matching when compared to **SingleRR** (50.3% vs 56% on QASC and 60.0 vs. 64.0% on MultiRC) suggesting that it learns how to "bridge the lexical chasm" between question and answers (Berger et al., 2000)

Token type	QASC		MultiRC	
	SingleRR	**JointRR**	**SingleRR**	**JointRR**
SMA	50.3	56.0	60.0	64.0
Linking	50.6	54.8	55.7	64.4

Table 5: Various attention scores of the **SingleRR** and the **JointRR** approaches. These normalized attention values are reported from the average of last 3 layers (10th, 11th and 12th layer) of RoBERTa-base. We observed similar trends with few exceptions in the lower layers as well which are farthest away from the decision layer that uses representation of `[CLS]`.

Attention from linking tokens of evidence: Here, we focus only on the terms that are shared between sentences in the gold evidence texts (referred to as *Linking* terms). As shown in fig. 1, *{nuclear, membrane}* are examples of linking terms that compose the two justification sentences into a complete explanation. The remaining terms in the evidence text, i.e., terms that are uniquely present in any one of the evidence sentences are referred to as *Non linking* terms.As shown in table 5, **JointRR** attends considerably more to the *Linking* terms (50.6 vs. 54.8 and 55.7 vs. 64.4), which suggests that it focuses more on the relevant compositional pieces after the retrieval training.

5.2 Learned Embedding Analysis

We also analyzed the embedding representations of the reranking model (Ethayarajh, 2019). In particular, we computed the embedding based cosine-similarity scores (or alignment scores (Yadav et al., 2019a)) between the two gold evidence sentences to determine their similarity in embedding space. As shown in fig. 3, the inter-justification alignment similarity score of **JointRR** is substantially lower across the majority of the layers after layer

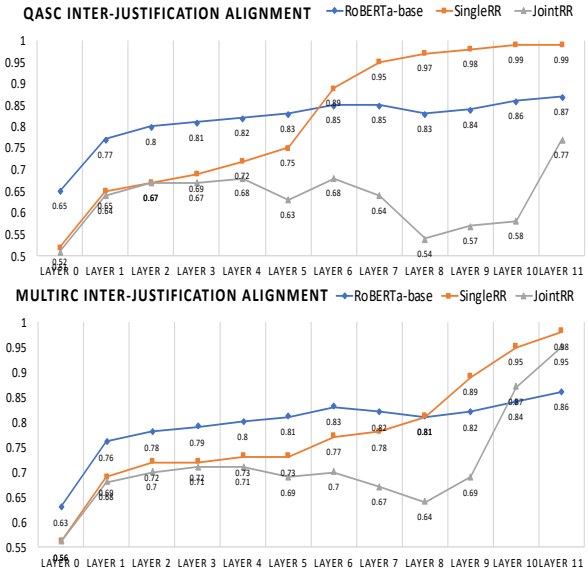

Figure 3: Layer-wise embedding based alignment similarity scores between the two gold justification sentences. In QASC, every question is annotated with just two gold justification sentences; for simplicity, we consider only the subset of MultiRC questions which have two gold justifications(65% of dev set).

3. This indicates that the RoBERTa reranker fed with WAIR chains has learned to differentiate the individual justification sentences (in embedding space) enabling *complementary* and compositional knowledge aggregation. As shown in table 4 (row 17 vs. row 15), this compositionality information is useful when the evidence reranking RoBERTa is transferred to the answer selection component i.e., we see a (small) QA performance improvement. On the other hand, **SingleRR** learns to consider both sentences similar, and this hurts the QA performance by 4.3% EM0 (row 16 vs. row 14, table 4).

Recent works have shown importance of vector normalization (Kobayashi et al., 2020) for analyzing the transformer embeddings. In future works, normalized embedding analysis can be added to further study the behavior of trained retriever's across different layers.

6 Conclusion

We introduced a simple unsupervised approach for retrieving candidate evidence chains that after reranking achieves state-of-the-art evidence retrieval performance on two multi-hop QA datasets: QASC and MultiRC. We highlight the importance of generating and feeding candidate evidence *chains* by showing several benefits over the widely followed approach that retrieves evidence sentences individually. Further, we introduced few attention and embedding analyses demonstrating that jointly retrieving and reranking chains assist in learning compositional information, which is also beneficial to the downstream QA task. Overall, our work highlights the strengths and potential of joint retrieval+reranking approaches for future works.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency (DARPA) under the World Modelers program, grant number W911NF1810014. 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

Pratyay Banerjee. 2019. Asu at textgraphs 2019 shared task: Explanation regeneration using language models and iterative re-ranking. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 78–84.

Pratyay Banerjee and Chitta Baral. 2020. Knowledge fusion and semantic knowledge ranking for open domain question answering. *arXiv preprint arXiv:2004.03101*.

Adam Berger, Rich Caruana, David Cohn, Dayne Freytag, and Vibhu Mittal. 2000. Bridging the lexical chasm: Statistical approaches to answer finding. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval*, Athens, Greece.

Delphine Bernhard. 2010. Query expansion based on pseudo relevance feedback from definition clusters. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 54–62. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879.

Jifan Chen and Greg Durrett. 2019. Understanding dataset design choices for multi-hop reasoning. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4026–4032.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. *arXiv preprint arXiv:1911.03429*.

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*.

Kawin Ethayarajh. 2019. How contextual are contextualized word representations? comparing the geometry of bert, elmo, and gpt-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 55–65.

Yair Feldman and Ran El-Yaniv. 2019. Multi-hop paragraph retrieval for open-domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2296–2309.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. *arXiv preprint arXiv:1803.02324*.

Peter Jansen. 2018. Multi-hop inference for sentence-level textgraphs: How challenging is meaningfully combining information for science question answering? In *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*, pages 12–17.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

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

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. 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.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2019a. Qasc: A dataset for question answering via sentence composition. *arXiv preprint arXiv:1910.11473*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2019b. What's missing: A knowledge gap guided approach for multi-hop question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2807–2821.

Sun Kim, Nicolas Fiorini, W John Wilbur, and Zhiyong Lu. 2017. Bridging the gap: Incorporating a semantic similarity measure for effectively mapping pubmed queries to documents. *Journal of biomedical informatics*, 75:122–127.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge university press.

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International*

Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2553–2566.

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

Peng Qi, Xiaowen Lin, Leo Mehr, Zijian Wang, and Christopher D Manning. 2019. Answering complex open-domain questions through iterative query generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2590–2602.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

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

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *arXiv preprint arXiv:2002.12327*.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019. Improving machine reading comprehension with general reading strategies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2633–2643.

Harsh Trivedi, Heeyoung Kwon, Tushar Khot, Ashish Sabharwal, and Niranjan Balasubramanian. 2019. Repurposing entailment for multi-hop question answering tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2948–2958.

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

Betty van Aken, Benjamin Winter, Alexander Löser, and Felix A Gers. 2019. How does bert answer questions? a layer-wise analysis of transformer representations. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pages 1823–1832.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019a. Superglue: A stickier benchmark for general-purpose language understanding systems. *arXiv preprint arXiv:1905.00537*.

Hai Wang, Dian Yu, Kai Sun, Jianshu Chen, Dong Yu, David McAllester, and Dan Roth. 2019b. Evidence sentence extraction for machine reading comprehension. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 696–707.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2019a. 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, Steven Bethard, and Mihai Surdeanu. 2019b. Quick and (not so) dirty: Unsupervised selection of justification sentences for multi-hop question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2578–2589.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020a. Having your cake and eating it too: Training neural retrieval for language inference without losing lexical match. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1625–1628.

Vikas Yadav, Steven Bethard, and Mihai Surdeanu. 2020b. Unsupervised alignment-based iterative evidence retrieval for multi-hop question answering. *arXiv preprint arXiv:2005.01218*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.