Unsupervised Alignment-based Iterative 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

Evidence retrieval is a critical stage of question answering (QA), necessary not only to improve performance, but also to explain the decisions of the corresponding QA method. We introduce a simple, fast, and unsupervised iterative evidence retrieval method, which relies on three ideas: (a) an unsupervised alignment approach to soft-align questions and answers with justification sentences using only GloVe embeddings, (b) an iterative process that reformulates queries focusing on terms that are not covered by existing justifications, which (c) a stopping criterion that terminates retrieval when the terms in the given question and candidate answers are covered by the retrieved justifications. Despite its simplicity, our approach outperforms all the previous methods (including supervised methods) on the evidence selection task on two datasets: MultiRC and QASC. When these evidence sentences are fed into a RoBERTa answer classification component, we achieve state-of-the-art QA performance on these two datasets.

1 Introduction

Explainability in machine learning (ML) remains a critical unsolved challenge that slows the adoption of ML in real-world applications (Biran and Cotton, 2017; Gilpin et al., 2018; Alvarez-Melis and Jaakkola, 2017; Arras et al., 2017).

Question answering (QA) is one of the challenging natural language processing (NLP) tasks that benefits from explainability. In particular, multi-hop QA requires the aggregation of multiple evidence facts in order to answer complex natural language questions (Yang et al., 2018). Several multi-hop QA datasets have been proposed recently (Yang et al., 2018; Khashabi et al., 2018a; Welbl et al., 2018; Dua et al., 2019; Chen and Durrett, 2019; Khot et al., 2019a; Sun et al., 2019b; Jansen and Ustalov, 2019; Rajpurkar et al., 2018). While

several neural methods have achieved state-of-the-art results on these datasets (Devlin et al., 2019; Liu et al., 2019; Yang et al., 2019), we argue that many of these directions lack a human-understandable explanation of their inference process, which is necessary to transition these approaches into real-world applications. This is especially critical for multi-hop, multiple choice QA (MCQA) where: (a) the answer text may not come from an actual knowledge base passage, and (b) reasoning is required to link the candidate answers to the given question (Yadav et al., 2019b). Figure 1 shows one such multi-hop example from a MCQA dataset.

In this paper we introduce a simple *a*lignment-based *i*terative *r*etriever (**AIR**)[1], which retrieves high-quality evidence sentences from unstructured knowledge bases. We demonstrate that these evidence sentences are useful not only to explain the required reasoning steps that answer a question, but they also considerably improve the performance of the QA system itself.

Unlike several previous works that depend on *supervised* methods for the retrieval of justification sentences (deployed mostly in settings that rely on small sets of candidate texts, e.g., HotPotQA, MultiRC), **AIR** is completely *unsupervised* and scales easily from QA tasks that use small sets of candidate evidence texts to ones that rely on large knowledge bases (e.g., QASC (Khot et al., 2019a)). **AIR** retrieves justification sentences through a simple iterative process. In each iteration, **AIR** uses an alignment model to find justification sentences that are closest in embedding space to the current query (Kim et al., 2017; Yadav et al., 2018), which is initialized with the question and candidate answer text. After each iteration, **AIR** adjusts its query to focus on the *missing* information (Khot et al., 2019b) in the current set of justifications.

---

[1] https://github.com/vikas95/AIR-retriever

AIR also conditionally expands the query using the justifications retrieved in the previous steps.

In particular, our key contributions are:

(1) We develop a simple, fast, and unsupervised iterative evidence retrieval method, which achieves state-of-the-art results on justification selection on two multi-hop QA datasets: MultiRC (Khashabi et al., 2018a) and QASC (Khot et al., 2019a). Notably, our simple unsupervised approach that relies solely on GloVe embeddings (Pennington et al., 2014) outperforms three transformer-based supervised state-of-the-art methods: BERT (Devlin et al., 2019), XLnet (Yang et al., 2019) and RoBERTa (Liu et al., 2019) on the justification selection task. Further, when the retrieved justifications are fed into a QA component based on RoBERTa (Liu et al., 2019), we obtain the best QA performance on the development sets of both MultiRC and QASC.[2]

(2) **AIR** can be trivially extended to capture parallel evidence chains by running multiple instances of **AIR** in parallel starting from different initial evidence sentences. We show that aggregating multiple parallel evidences further improves the QA performance over the vanilla **AIR** by 3.7% EM0 on the MultiRC and 5.2% accuracy on QASC datasets (both absolute percentages on development sets). Thus, with 5 parallel evidences from **AIR** we obtain 36.3% EM0 on MultiRC and 81.0% accuracy on QASC **hidden** test sets (on their respective leaderboards). To our knowledge from published works, these results are the new state-of-the-art QA results on these two datasets. These scores are also accompanied by new state-of-the-art performance on evidence retrieval on both the datasets, which emphasizes the interpretability of **AIR**.

(3) We demonstrate that **AIR**'s iterative process that focuses on missing information is more robust to semantic drift. We show that even the supervised RoBERTa-based retriever trained to retrieve evidences iteratively, suffers substantial drops in performance with retrieval from consecutive hops.

2 Related Work

Our work falls under the revitalized direction that focuses on the interpretability of QA systems, where the machine's inference process is explained to the end user in natural language evidence text (Qi et al., 2019; Yang et al., 2018; Wang et al., 2019b; Yadav et al., 2019b; Bauer et al., 2018). Several

---
Question: Exposure to oxygen and water can cause iron to

(A) decrease strength (B) melt (C) uncontrollable burning (D) thermal expansion (E) **turn orange on the surface** (F) vibrate (G) extremes of temperature (H) levitate

---
Gold justification sentences:
1. when a metal rusts , that metal becomes orange on the surface
2. Iron rusts in the presence of oxygen and water.

Parallel evidence chain 1:
1. Dissolved oxygen in water usually causes the oxidation of iron.
2. When iron combines with oxygen it turns orange.

Parallel evidence chain 2:
1. By preventing the exposure of the metal surface to oxygen, oxidation is prevented.
2. When iron oxidizes, it rusts.

Figure 1: An example question that requires multi-hop reasoning, together with its gold justifications from the QASC dataset. The two parallel evidence chains retrieved by **AIR** (see section 3) provide imperfect but relevant explanations for the given question.

datasets in support of interpretable QA have been proposed recently. For example, datasets such as HotPotQA, MultiRC, QASC, Worldtree Corpus, etc., (Yang et al., 2018; Khashabi et al., 2018a; Khot et al., 2019a; Jansen and Ustalov, 2019) provide annotated evidence sentences enabling the automated evaluation of interpretability via evidence text selection.

QA approaches that focus on interpretability can be broadly classified into three main categories: *supervised*, which require annotated justifications at training time, *latent*, which extract justification sentences through latent variable methods driven by answer quality, and, lastly, *unsupervised* ones, which use unsupervised algorithms for evidence extraction.

In the first class of supervised approaches, a supervised classifier is normally trained to identify correct justification sentences driven by a query (Nie et al., 2019; Tu et al., 2019; Banerjee, 2019). Many systems tend to utilize a multi-task learning setting to learn both answer extraction and justification selection with the same network (Min et al., 2018; Gravina et al., 2018). Although these approaches have achieved impressive performance, they rely on annotated justification sentences, which may not be always available. Few approaches have used distant supervision methods (Lin et al., 2018; Wang et al., 2019b) to create noisy training data for evidence retrieval but these usually underperform due to noisy labels.

In the latent approaches for selecting justifica-

---
[2]In settings where external labeled resources are not used.

tions, reinforcement learning (Geva and Berant, 2018; Choi et al., 2017) and PageRank (Surdeanu et al., 2008) have been widely used to select justification sentences without explicit training data. While these directions do not require annotated justifications, they tend to need large amounts of question/correct answer pairs to facilitate the identification of latent justifications.

In unsupervised approaches, many QA systems have relied on structured knowledge base (KB) QA. For example, several previous works have used ConceptNet (Speer et al., 2017) to keep the QA process interpretable (Khashabi et al., 2018b; Sydorova et al., 2019). However, the construction of such structured knowledge bases is expensive, and may need frequent updates. Instead, in this work we focus on justification selection from textual (or unstructured) KBs, which are inexpensive to build and can be applied in several domains. In the same category of unsupervised approaches, conventional information retrieval (IR) methods such as BM25 (Chen et al., 2017) have also been widely used to retrieve independent individual sentences. As shown by (Khot et al., 2019a; Qi et al., 2019), and our table 2, these techniques do not work well for complex multi-hop questions, which require knowledge aggregation from multiple related justifications. Some unsupervised methods extract groups of justification sentences (Chen et al., 2019; Yadav et al., 2019b) but these methods are exponentially expensive in the retrieval step. Contrary to all of these, **AIR** proposes a simpler and more efficient method for chaining justification sentences.

Recently, many supervised iterative justification retrieval approaches for QA have been proposed (Qi et al., 2019; Feldman and El-Yaniv, 2019; Banerjee, 2019; Das et al., 2018). While these were shown to achieve good evidence selection performance for complex questions when compared to earlier approaches that relied on just the original query (Chen et al., 2017; Yang et al., 2018), they all require supervision.

As opposed to all these iterative-retrieval methods and previously discussed directions, our proposed approach **AIR** is completely unsupervised, i.e., it does not require annotated justifications. Further, unlike many of the supervised iterative approaches (Feldman and El-Yaniv, 2019; Sun et al., 2019a) that perform query reformulation in a continuous representation space, **AIR** employs a simpler and more interpretable query reformulation

strategy that relies on explicit terms from the query and the previously retrieved justification. Lastly, none of the previous iterative retrieval approaches address the problem of semantic drift, whereas **AIR** accounts for drift by controlling the query reformulation as explained in section 3.1.

3 Approach

As shown in fig. 2, the proposed QA approach consists of two components: (a) an unsupervised, iterative component that retrieves chains of justification sentences given a query; and (b) an answer classification component that classifies a candidate answer as correct or not, given the original question and the previously retrieved justifications. We detail these components in the next two sub-sections.

3.1 Iterative Justification Retrieval

AIR iteratively builds justification chains given a query. **AIR** starts by initializing the query with the concatenated question and candidate answer text[3]. Then, **AIR** iteratively repeats the following two steps: (a) It retrieves the most salient justification sentence given the current query using an alignment-IR approach(Yadav et al., 2019a). The candidate justification sentences come from dataset-specific KBs. For example, in MultiRC, we use as candidates all the sentences from the paragraph associated with the given question. In QASC, which has a large KB[4] of 17.4 million sentences), similar to Khot et al. (2019a) candidates are retrieved using the Heuristic+IR method which returns 80 candidate sentences for each candidate answer from the provided QASC KB. (b) it adjusts the query to focus on the *missing* information, i.e., the keywords that are not covered by the current evidence chain. **AIR** also dynamically adds new terms to the query from the previously retrieved justifications to nudge multi-hop retrieval. These two iterative steps repeat until a parameter-free termination condition is reached.

We first detail the important components of **AIR**.

Alignment: To compute the similarity score between a given query and a sentence from KB, **AIR**

---

[3]Note that this work can be trivially adapted to reading comprehension tasks. In such tasks (e.g., SQuAD (Rajpurkar et al., 2018)), the initial query would contain just the question text.

[4]In large KB-based QA, **AIR** first uses an off-the-shelf Lucene BM25(Robertson et al., 2009) to retrieve a pool of candidate justification sentences from which the evidence chains are constructed.
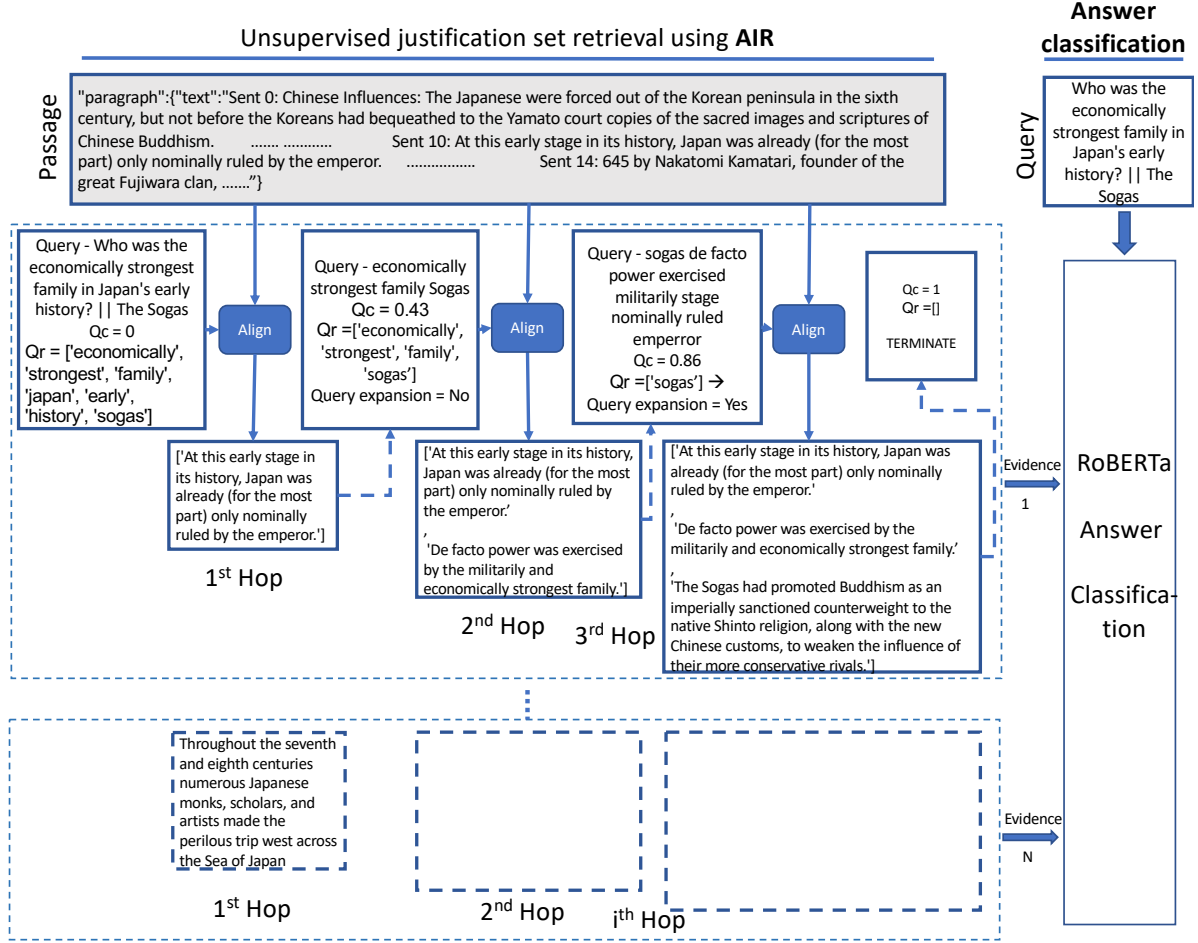
Figure 2: A walkthrough example showing the iterative retrieval of justification sentences by **AIR** on MultiRC. Each current query includes keywords from the original query (which consists of question + candidate answer) that are not covered by previously retrieved justifications (see 2nd hop). If the number of uncovered keywords is too small, the query is expanded with keywords from the most recent justification (3rd hop). The retrieval process terminates when all query terms are covered by existing justifications. Q_c indicates the proportion of query terms covered in the justifications; Q_r indicates the query terms which are still not covered by the justifications. **AIR** can retrieve parallel justification chains by running the retrieval process in parallel, starting from different candidates for the first justification sentence in a chain.

uses a vanilla unsupervised alignment method of Yadav et al. (2019a) which uses only GloVe embeddings (Pennington et al., 2014).[5] The alignment method computes the cosine similarity between the word embeddings of each token in the query and each token in the given KB sentence, resulting in a matrix of cosine similarity scores. For each query token, the algorithm select the most similar token in the evidence text using max-pooling. At the end, the element-wise dot product between this max-pooled vector of cosine-similarity scores and the vector containing the IDF values of the query tokens is calculated to produce the overall alignment score s for the given query Q and the supporting paragraph P_j:

$$s(Q, P_j) = \sum_{i=1}^{|Q|} idf(q_i) \cdot align(q_i, P_j) \quad (1)$$

$$align(q_i, P_j) = \max_{k=1}^{|P_j|} cosSim(q_i, p_k) \quad (2)$$

where q_i and p_k are the ith and k^{th} terms of the query (Q) and evidence sentence (P_j) respectively.

Remainder terms (Q_r): Query reformulation in **AIR** is driven by the remainder terms, which are the set of query terms not yet covered in the justification set of i sentences (retrieved from the first i iterations of the retrieval process):

$$Q_r(i) = t(Q) - \bigcup_{s_k \in S_i} t(s_k) \quad (3)$$

where $t(Q)$ represents the unique set of query terms, $t(s_k)$ represents the unique terms of the kth

4517

justification, and S_i represents the set of i justification sentences. Note that we use soft matching of alignment for the inclusion operation: we consider a query term to be included in the set of terms in the justifications if its cosine similarity with a justification term is larger than a similarity threshold M (we use M=0.95 for all our experiments - see section 5.2), thus ensuring that the two terms are similar in the embedding space.

Coverage (Q_c): measures the coverage of the query keywords by the retrieved chain of justifications S:

$$Q_c(i) = \frac{|\bigcup_{s_k \in S_i} t(Q) \cap t(s_k)|}{|t(Q)|} \qquad (4)$$

where $|t(Q)|$ denotes the size of unique query terms.

The AIR retrieval process

Query reformulation: In each iteration j, **AIR** reformulates the query $Q(j)$ to include only the terms not yet covered by the current justification chain, $Q_r(j-1)$. See, for example, the second hop in fig. 2. To mitigate ambiguous queries, the query is expanded with the terms from all the previously retrieved justification sentences only if the number of uncovered terms is less than T (we used $T = 2$ for MultiRC and $T = 4$ for QASC (see section 5.2). See, for example, the third hop in fig. 2, in which the query is expanded with the terms of all the previously retrieved justification sentences. Formally:

$$Q(j) = \begin{cases} Q_r(j-1), & \text{if } |Q_r(j-1)| > T \\ Q_r(j-1) + (t(s_{j-1}) - t(Q)), & \text{otherwise} \end{cases}$$
$$(5)$$

where j is the current iteration index.

Stopping criteria: **AIR** stops its iterative evidence retrieval process when either of the following conditions is true: (a) no new query terms are discovered in the last justification retrieved, i.e., $Q_r(i-1) == Q_r(i)$, or (b) all query terms are covered by justifications, i.e., $Q_c = 1$.

3.2 Answer Classification

AIR's justification chains can be fed into any supervised answer classification method. For all experiments in this paper, we used RoBERTa (Liu et al., 2019), a state-of-the-art transformer-based method. In particular, for MultiRC, we concatenate the query (composed from question and candidate answer text) with the evidence text, with the

[SEP] token between the two texts. A sigmoid is used over the [CLS] representation to train a binary classification task[6] (correct answer or not).

For QASC, we fine-tune RoBERTa as a multiple-choice QA [7] (MCQA) (Wolf et al., 2019) classifier with 8 choices using a softmax layer(similar to (Khot et al., 2019a)) instead of the sigmoid. The input text consists of eight queries (from eight candidate answers) and their corresponding eight evidence texts. Unlike the case of MultiRC, it is possible to train a MCQA classifier for QASC because every question has only 1 correct answer. We had also tried the binary classification approach for QASC but it resulted in nearly 5% lower performance for majority of the experiments in table 2.

In QA tasks that rely on large KBs there may exist multiple chains of evidence that support a correct answer. This is particularly relevant in QASC, whose KB contains 17.2M facts.[8] Figure 1 shows an example of this situation. To utilize this type of redundancy in answer classification, we extend **AIR** to extract *parallel evidence chains*. That is, to extract N parallel chains, we run **AIR** N times, ensuring that the first justification sentences in each chain are different (in practice, we start a new chain for each justification in the top N retrieved sentences in the first hop). After retrieving N parallel evidence chains, we take the union of all the individual justification sentences to create the supporting evidence text for that candidate answer.

4 Experiments

We evaluated our approach on two datasets:

Multi-sentence reading comprehension (MultiRC), which is a reading comprehension dataset provided in the form of multiple-choice QA task (Khashabi et al., 2018a). Every question is based on a paragraph, which contains the gold justification sentences for each question. We use every sentence of the paragraph as candidate justifications for a given question. Here we use the original

---

#	Computational steps	Supervised selection of justifications?	Method	$F1_m$	$F1_a$	EM0	Evidence selection P	R	F1
			DEVELOPMENT DATASET						
			Baselines						
1	N	No	IR(paragraphs) (Khashabi et al., 2018a)	64.3	60.0	1.4	–		
2	N	No	SurfaceLR (Khashabi et al., 2018a)	66.5	63.2	11.8	–		
3	N	No	Entailment baseline (Trivedi et al., 2019)	51.3	50.4	–	–		
			Previous work						
4	N	Yes	QA+NLI (Pujari and Goldwasser, 2019)	-	-	21.6	–		
5	N	Yes	EER_{DPL} + FT (Wang et al., 2019b)	70.5	67.8	13.3	–		
6	N	Yes	Multee (GloVe) (Trivedi et al., 2019)	71.3	68.3	17.9	–		
7	N	Yes	Multee (ELMo)* (Trivedi et al., 2019)	73.0	69.6	22.8	–		
8	$K \times N$	Yes	RS* (Sun et al., 2019c)	73.1	70.5	21.8	–	–	60.8
9	N	No	BERT + BM25 (Yadav et al., 2019b)	71.1	67.4	23.1	43.8	61.2	51.0
10	$2^N - N - 1$	No	BERT + AutoROCC (Yadav et al., 2019b)	72.9	69.6	24.7	48.2	68.2	56.4
			Alignment + RoBERTa(QA) baselines						
11	-	No	Entire passage + RoBERTa	73.9	71.7	28.7	17.4	100.0	29.6
12	N	No	Alignment ($k = 2$ sentences) + RoBERTa	72.6	69.6	25.9	62.4	55.6	58.8
13	N	No	Alignment ($k = 3$ sentences) + RoBERTa	72.4	69.8	25.1	49.3	65.1	56.1
14	N	No	Alignment ($k = 4$ sentences) + RoBERTa	73.6	71.4	28.0	41.0	72.0	52.3
15	N	No	Alignment ($k = 5$ sentences) + RoBERTa	73.7	70.8	25.0	35.2	77.1	48.4
			RoBERTa retriever + RoBERTa(QA) baselines						
16	N	Yes	RoBERTa-retriever(All passages) + RoBERTa	70.5	68.0	24.9	63.4	61.1	62.3
17	N	Yes	RoBERTa-retriever(Fiction) + RoBERTa	72.8	70.4	24.7	47.8	73.9	58.1
18	N	Yes	RoBERTa-retriever(News) + RoBERTa	69.0	67.3	24.2	60.8	59.2	59.9
19	N	Yes	RoBERTa-retriever(Science-textbook) + RoBERTa	70.3	67.7	25.3	48.1	62.0	54.2
20	N	Yes	RoBERTa-retriever(Society_Law) + RoBERTa	72.8	70.3	25.3	50.4	68.5	58.0
21	$K \times N$	Yes	RoBERTa-**iterative**-retriever + RoBERTa	70.1	67.6	24.0	67.1	58.4	62.5
			RoBERTa + **AIR** (Parallel) Justifications						
22	$K \times N$	No	**AIR** (lexical) top chain + RoBERTa	71.0	68.2	22.9	58.2	49.5	53.5
23	$K \times N$	No	**AIR** top chain + RoBERTa	***74.7***	***72.3***	***29.3***	66.2	63.1	**64.2**
24	$2 \times K \times N$	No	**AIR** Parallel evidence chains ($p = 2$) + RoBERTa	75.5	73.6	32.5	50.4	71.9	59.2
25	$3 \times K \times N$	No	**AIR** Parallel evidence chains ($p = 3$) + RoBERTa	75.8	73.7	30.6	40.8	76.7	53.3
26	$4 \times K \times N$	No	**AIR** Parallel evidence chains ($p = 4$) + RoBERTa	76.3	74.2	31.3	34.8	80.8	48.7
27	$5 \times K \times N$	No	**AIR** Parallel evidence chains ($p = 5$) + RoBERTa	**77.2**	**75.1**	**33.0**	28.6	84.1	44.9
			Ceiling systems with gold justifications						
29	-	Yes	EER_{gt} + FT (Wang et al., 2019b)	72.3	70.1	19.2	–		
30	-	Yes	RoBERTa + Gold knowledge	81.4	80	39	100.0	100.0	100.0
31	-	-	Human	86.4	83.8	56.6	–		
			TEST DATASET						
32	N	No	SurfaceLR (Khashabi et al., 2018a)	66.9	63.5	12.8			
33	N	Yes	Multee (ELMo)* (Trivedi et al., 2019)	73.8	70.4	24.5	–		
34	$2^N - N - 1$	No	BERT + AutoROCC (Yadav et al., 2019b)	73.8	70.6	26.1			
35	$5 \times K \times N$	No	RoBERTa + **AIR** (Parallel evidence = 5)	**79.0**	**76.4**	**36.3**			

Table 1: Results on the MultiRC development and test sets. The first column specifies the runtime overhead required for selection of evidence sentences, where N is the total number of sentences in the passage, and K is the selected number of sentences. The second column specifies if the retrieval system is a supervised method or not. The last three columns indicate evidence selection performance, whereas the previous three indicate overall QA performance. Only the last block of results report performance on the test set. The bold italic font highlights the best performance without using parallel evidences. $\star$ denotes usage of external labeled data for pretraining.

MultiRC dataset,[9] which includes the gold annotations for evidence text, unlike the version available on SuperGlue (Wang et al., 2019a).

Question Answering using Sentence Composition (QASC), a large KB-based multiple-choice QA dataset (Khot et al., 2019a). Each question is provided with 8 answer candidates, out of which 4 candidates are hard adversarial choices. Every question is annotated with a fixed set of two justification sentences for answering the question. The justification sentences are to be retrieved from a KB having 17.2 million facts. As shown in the example of fig. 1 and also highlighted by (Khot et al., 2019a), multiple evidence text are possible for a given question in QASC where the annotated gold justification sentences explain it more precisely.

We report overall question answering performance as well as evidence selection performance in table 1 for MultiRC, and table 2 for QASC[10].

---

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

[10] https://leaderboard.allenai.org/qasc/submissions/public

#	Number of steps used?	Method	Accuracy	Recall@10 *both found*	Recall@10 *atleast one found*
		Baselines			
0	Single	Naive Lucene BM25	35.6	17.2	68.1
1	Two	Naive Lucene BM25	36.3	27.8	65.7
2	Two	Heuristics+IR (Khot et al., 2019a)	32.4	41.6	64.4
3	-	ESIM Q2Choice (Khot et al., 2019a)	21.1	41.6	64.4
		Previous work			
4	Single	BERT-LC (Khot et al., 2019a)	59.8	11.7	54.7
5	Two	BERT-LC (Khot et al., 2019a)	71.0	41.6	64.4
6	Two	BERT-LC[WM]* (Khot et al., 2019a)	78.0	41.6	64.4
		Alignment + RoBERTa baselines			
7	–	No Justifiction + RoBERTa	20.5	0	0
8	Single	Alignment ($K = 1$ sentences) + RoBERTa	54.4	-	-
9	Two	Alignment ($K = 2$ sentences) + RoBERTa	71.5	-	-
10	Two	Alignment ($K = 3$ sentences) + RoBERTa	73.3	-	-
11	Two	Alignment ($K = 4$ sentences) + RoBERTa	73.5	-	-
12	Two	Alignment ($K = 5$ sentences) + RoBERTa	74.1	-	-
		AIR+ RoBERTa			
13	Two	**AIR** (lexical) top chain + RoBERTa	75.8	-	-
14	Two	**AIR** top chain + RoBERTa	*76.2*	-	-
15	Two	**AIR** Parallel evidence chains ($p = 2$) + RoBERTa	79.8	-	-
16	Two	**AIR** Parallel evidence chains ($p = 3$) + RoBERTa	80.9	-	-
17	Two	**AIR** Parallel evidence chains ($p = 4$) + RoBERTa	79.7	-	-
18	Two	**AIR** Parallel evidence chains ($p = 5$) + RoBERTa	**81.4**	**44.8**	**68.6**
		TEST DATASET			
19	Two	BERT-LC (Khot et al., 2019a)	68.5	-	-
20	Two	BERT-LC[WM]* (Khot et al., 2019a)	73.2	-	-
21	Two	**AIR** Parallel evidence chains ($p = 5$) + RoBERTa	**81.4**	-	-

Table 2: QA and evidence selection performance on QASC. We also report recall@10 similar to Khot et al. (2019a). *both found* reports the recall scores when both the gold justifications are found in top 10 ranked sentences and similarly *atleast one found* reports the recall scores when either one or both the gold justifications are found in the top 10 ranked sentences. Recall@10 are not reported (row 8-17) when number of retrieved sentences are lesser than 10. Other notations are same as table 1.

4.1 Baselines

In addition to previously-reported results, we include in the tables several in-house baselines. For MultiRC, we considered three baselines. The first baseline is where we feed *all* passage sentences to the RoBERTa classifier (row 11 in table 1). The second baseline uses the alignment method of (Kim et al., 2017) to retrieve the top k sentences ($k = 2, 5$). Since **AIR** uses the same alignment approach for retrieving justifications in each iteration, the comparison to this second baseline highlights the gains from our iterative process with query reformulation. The third baseline uses a supervised RoBERTa classifier trained to select the gold justifications for every query (rows 16–21 in table 1). Lastly, we also developed a RoBERTa-based iterative retriever by concatenating the query with the retrieved justification in the previous step. We retrain the RoBERTa iterative retriever in every step, using the new query in each step.

We considered two baselines for QASC. The first baseline does not include any justifications (row 7 in table 2). The second baseline uses the top k sentences retrieved by the alignment method (row (8–12 in table 2).

4.2 Evidence Selection Results

For evidence selection, we report precision, recall, and F1 scores on MultiRC (similar to (Wang et al., 2019b; Yadav et al., 2019b)). For QASC, we report Recall@10, similar to the dataset authors (Khot et al., 2019a). We draw several observation from the evidence selection results:

(1) AIR vs. unsupervised methods - **AIR** outperforms all the unsupervised baselines and previous works in both MultiRC (row 9-15 vs. row 23 in table 1) and QASC(rows 0-6 vs. row 18). Thus, highlighting strengths of **AIR** over the standard IR baselines. **AIR** achieves 5.4% better F1 score compared to the best parametric alignment baseline (row 12 in table 1), which highlights the importance of the iterative approach over the vanilla alignment in **AIR**. Similarly, rows (4 and 5) of table 2 also highlight this importance in QASC.

(2) AIR vs. supervised methods - Surprisingly, **AIR** also outperforms the supervised RoBERTa-retriver in every setting(rows 16–21 in table 1). Note that the performance of this supervised re-

trieval method drops considerably when trained on passages from a specific domain (row 19 in table 1), which highlights the domain sensitivity of supervised retrieval methods. In contrast, **AIR** is unsupervised and generalize better as it is not tuned to any specific domain. **AIR** also achieves better performance than supervised RoBERTa-iterative-retriever (row 21 in table 1) which simply concatenates the retrieved justification to the query after every iteration and further trains to retrieve the next justification. The RoBERTa-iterative-retriever achieves similar performance as that of the simple RoBERTa-retriever (row 16 vs. 21) which suggests that supervised iterative retrievers marginally exploit the information from query expansion. On the other hand, controlled query reformulation of **AIR** leads to 5.4% improvement as explained in the previous point. All in all, **AIR** achieves state-of-the-art results for evidence retrieval on both MultiRC (row 23 in table 1) and QASC (row 18 of table 2).

(3) Soft-matching of AIR - the alignment-based **AIR** is 10.7% F1 better than **AIR** that relies on lexical matching (rather than the soft matching) on MultiRC (row 22 vs. 23), which emphasizes the advantage of alignment methods over conventional lexical match approaches.

4.3 Question Answering Results

For overall QA performance, we report the standard performance measures ($F1_a$, $F1_m$, and $EM0$) in MultiRC (Khashabi et al., 2018a), and accuracy for QASC (Khot et al., 2019a).

The results in tables 1 and 2 highlight:

(1) State-of-the-art performance:
Development set - On both MultiRC and QASC, RoBERTa fine-tuned using the **AIR** retrieved evidence chains (row 23 in table 1 and row 14 in table 2) outperforms all the previous approaches and the baseline methods. This indicates that the evidence texts retrieved by **AIR** not only provide better explanations, but also contribute considerably in achieving the best QA performance.

Test set - On the official hidden test set, RoBERTa fine-tuned on 5 parallel evidences from **AIR** achieves new state-of-the-art QA results, outperforming previous state-of-the-art methods by 7.8% accuracy on QASC (row 21 vs. 20), and 10.2% EM0 on MultiRC (row 35 vs. 34).

(2) Knowledge aggregation - The knowledge aggregation from multiple justification sentences

# of hops	BM25	**AIR** (Lexical) *uncontrolled*	Alignment	**AIR** *uncontrolled*
1	38.8	38.8	46.5	46.5
2	48.4	45.9	58.8	54.1
3	48.4	45.8	56.1	52.2
4	47.0	44.0	52.3	49.1
5	44.8	41.1	48.4	46.0

Table 3: Impact of semantic drift across consecutive hops on justification selection F1-performance of MultiRC development set. The *uncontrolled* configuration indicates that the justification sentences retrieved in each hop were appended to the query in each step. Here, **AIR** is forced to retrieve the same number of justifications as indicated by the # of hops.

leads to substantial improvements, particularly in QASC (single justification (row 4 and 8) vs. evidence chains (row 5 and row 9) in table table 2). Overall, the chain of evidence text retrieved by **AIR** enables knowledge aggregation resulting in the improvement of QA performances.

(3) Gains from parallel evidences - Further, knowledge aggregation from parallel evidence chains lead to another 3.7% EM0 improvement on MultiRC (row 27), and 5.6% on QASC over the single **AIR** evidence chain (row 18). To our knowledge, these are new state-of-the-art results in both the datasets.

5 Analysis

To further understand the retrieval process of **AIR** we implemented several analyses.

5.1 Semantic Drift Analysis

To understand the importance of modeling missing information in query reformulation, we analyzed a simple variant of **AIR** in which, rather the focusing on missing information, we simply concatenate the complete justification sentence to the query after each hop. To expose semantic drift, we retrieve a specified number of justification sentences. As seen in table 3, now the **AIR**(lexical)-*uncontrolled* and **AIR**-*uncontrolled* perform worse than both BM25 and the alignment method. This highlights that the focus on missing information during query reformulation is an important deterrent of semantic drift. We repeated the same experiment with the supervised RoBERTa retriever (trained iteratively for 2 steps) and the original parameter-free **AIR**, which decides its number of hops using the stopping conditions. Again, we observe similar performance drops in both: the RoBERTa retriever drops from 62.3% to 57.6% and **AIR** drops to 55.4%.

| Q_r | MultiRC | QASC | |
	F1 score	Both Found	One Found
1	**64.2**	41.7	67.7
2	62.7	42.7	67.7
3	61.8	43.1	68.6
4	60.63	40.6	68.4
5	59.8	39.0	67.5

Table 4: Impact on justification selection F1-performance from the hyper parameter Q_r of **AIR** (eq. (5)).

| M | MultiRC | QASC | |
	F1 score	Both Found	One Found
0.95	**64.2**	43.1	68.6
0.85	63.7	42.3	67.9
0.75	63.4	42.5	68.0

Table 5: Impact on justification selection F1 score from the hyper parameter M in the alignment step (section 3.1).

5.2 Robustness to Hyper Parameters

We evaluate the sensitivity of **AIR** to the 2 hyper parameters: the threshold (Q_r) for query expansion, and the cosine similarity threshold M in computation of alignment. As shown in table 5, evidence selection performance of **AIR** drops with the lower values of M but the drops are small, suggesting that **AIR** is robust to different M values.

Similarly, there is a drop in performance for MultiRC with the increase in the Q_r threshold used for query expansion, hinting to the occurrence of semantic drift for higher values of Q_r (table 4). This is because the candidate justifications are coming from a relatively small numbers of paragraphs in MultiRC; thus even shorter queries (= 2 words) can retrieve relevant justifications. On the other hand, the number of candidate justifications in QASC is much higher, which requires longer queries for disambiguation (>= 4 words).

5.3 Saturation of Supervised Learning

To verify if the MultiRC training data is sufficient to train a supervised justification retrieval method, we trained justification selection classifiers based on BERT, XLNet, and RoBERTa on increasing proportions of the MultiRC training data (table 6). This analysis indicates that all three classifiers approach their best performance at around 5% of the training data. This indicates that, while these supervised methods converge quickly, they are unlikely to outperform **AIR**, an unsupervised method, even if more training data were available.

% of training data	BERT	XLnet	RoBERTa	**AIR**
2	55.2	54.6	62.3	
5	60.0	59.6	60.8	
10	59.9	57.0	59.8	
15	58.3	59.9	59.1	
20	58.5	60.2	60.0	64.2
40	58.5	58.7	58.8	
60	59.1	61.4	59.8	
80	59.3	61.0	60.5	
100	60.9	61.1	62.3	

Table 6: Comparison of **AIR** with XLNet, RoBERTa, and BERT on justification selection task, trained on increasing proportion of the training data on MultiRC.

6 Conclusion

We introduced a simple, unsupervised approach for evidence retrieval for question answering. Our approach combines three ideas: (a) an unsupervised alignment approach to soft-align questions and answers with justification sentences using GloVe embeddings, (b) an iterative process that reformulates queries focusing on terms that are not covered by existing justifications, and (c) a simple stopping condition that concludes the iterative process when all terms in the given question and candidate answers are covered by the retrieved justifications. Overall, despite its simplicity, unsupervised nature, and its sole reliance on GloVe embeddings, our approach outperforms all previous methods (including supervised ones) on the evidence selection task on two datasets: MultiRC and QASC. When these evidence sentences are fed into a RoBERTa answer classification component, we achieve the best QA performance on these two datasets. Further, we show that considerable improvements can be obtained by aggregating knowledge from parallel evidence chains retrieved by our method.

In addition of improving QA, we hypothesize that these simple unsupervised components of **AIR** will benefit future work on supervised neural iterative retrieval approaches by improving their query reformulation algorithms and termination criteria.

Acknowledgments

References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421.

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.

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.

Lisa Bauer, Yicheng Wang, and Mohit Bansal. 2018. Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230.

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

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.

Jifan Chen, Shih-ting Lin, and Greg Durrett. 2019. Multi-hop question answering via reasoning chains. *arXiv preprint arXiv:1910.02610*.

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.

Rajarshi Das, Shehzaad Dhuliawala, Manzil Zaheer, and Andrew McCallum. 2018. Multi-step retriever-reader interaction for scalable open-domain question answering.

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.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. 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 2368–2378.

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 and Jonathan Berant. 2018. Learning to search in long documents using document structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 161–176.

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.

Peter Jansen and Dmitry Ustalov. 2019. Textgraphs 2019 shared task on multi-hop inference for explanation regeneration. In *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*, pages 63–77.

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

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.

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.

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*.

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

Yixin Nie, Songhe Wang, and Mohit Bansal. 2019. Revealing the importance of semantic retrieval for machine reading at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2553–2566.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rajkumar Pujari and Dan Goldwasser. 2019. Using natural language relations between answer choices for machine comprehension. 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 4010–4015.

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.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789.

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

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

Haitian Sun, Tania Bedrax-Weiss, and William Cohen. 2019a. Pullnet: Open domain question answering with iterative retrieval on knowledge bases and text. 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 2380–2390.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019b. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019c. 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.

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.

Alona Sydorova, Nina Poerner, and Benjamin Roth. 2019. Interpretable question answering on knowledge bases and text. *arXiv preprint arXiv:1906.10924*.

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*.

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, 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, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*, pages 5754–5764.

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.