An Interface for Annotating Science Questions

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum
College of Information and Computer Sciences
University of Massachusetts, Amherst MA

Maria Chang, Achille Fokoue, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, Michael Witbrock
IBM Research, Yorktown Heights NY

Abstract

Recent work introduces the AI2 Reasoning Challenge (ARC) and the associated ARC dataset that partitions open domain, complex science questions into an Easy Set and a Challenge Set. That work includes an analysis of 100 questions with respect to the types of knowledge and reasoning required to answer them. However, it does not include clear definitions of these types, nor does it offer information about the quality of the labels or the annotation process used. In this paper, we introduce a novel interface for human annotation of science question-answer pairs with their respective knowledge and reasoning types, in order that the classification of new questions may be improved. We build on the classification schema proposed by prior work on the ARC dataset, and evaluate the effectiveness of our interface with a preliminary study involving 10 participants.

1 Introduction

Recent work by Clark et al. (2018) introduces the AI2 Reasoning Challenge (ARC)[1] and the associated ARC dataset. This dataset contains science questions from standardized tests that are separated into an Easy Set and a Challenge Set. The Challenge Set comprises questions that are answered incorrectly by two solvers based on Pointwise Mutual Information (PMI) Information Retrieval (IR). In addition to this division, a survey of the various types of knowledge as well as the types of reasoning that are required to answer various questions in the ARC dataset was presented. This survey was based on an analysis of 100 questions chosen at random from the Challenge Set. However, very little detail is provided about the questions chosen, the annotations provided, or the methodology used. These questions surround the very core of the paper, since the main contribution is a dataset that contains complex questions.

In this work, in order to overcome some of the limitations of Clark et al. (2018) described above, we present a detailed annotation interface for the ARC dataset that allows a distributed set of annotators to label the knowledge and reasoning types (Boratko et al., 2018). Following an annotation round involving over ten people at two institutions, we measure and report statistics such as inter-rater agreement, and the distribution of knowledge and reasoning type labels in the dataset.

2 Annotation Interface

The annotation interface introduced in this paper is shown in Figure 1. The text of the science question is displayed at the top of the left side, followed by the answer options. Each of the answer options is preceded by a radio button: each button is initially transparent, but the annotator can click on a button to check whether the corresponding option is the answer to the question. This facility is to help annotators with extra information if it is needed in labeling the question; however, we leave it blank initially to avoid biasing the annotations.

Clicking on a specific answer option executes a search on the ARC corpus, with the query text of that search set to be the last sentence of the question appended with the entire text of the clicked answer option. The retrieved search results are shown in the bottom left half of the interface. Annotators have the option of labeling retrieved search results as *irrelevant* or *relevant* to answering the question at hand. The query box also accepts free text, and annotators who wish to craft more specific queries are free to do so. We collect all the queries executed, as well as the annotations pertaining to the relevance of the returned results.

[1] http://data.allenai.org/arc/

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (System Demonstrations), pages 102–107
Brussels, Belgium, October 31–November 4, 2018.

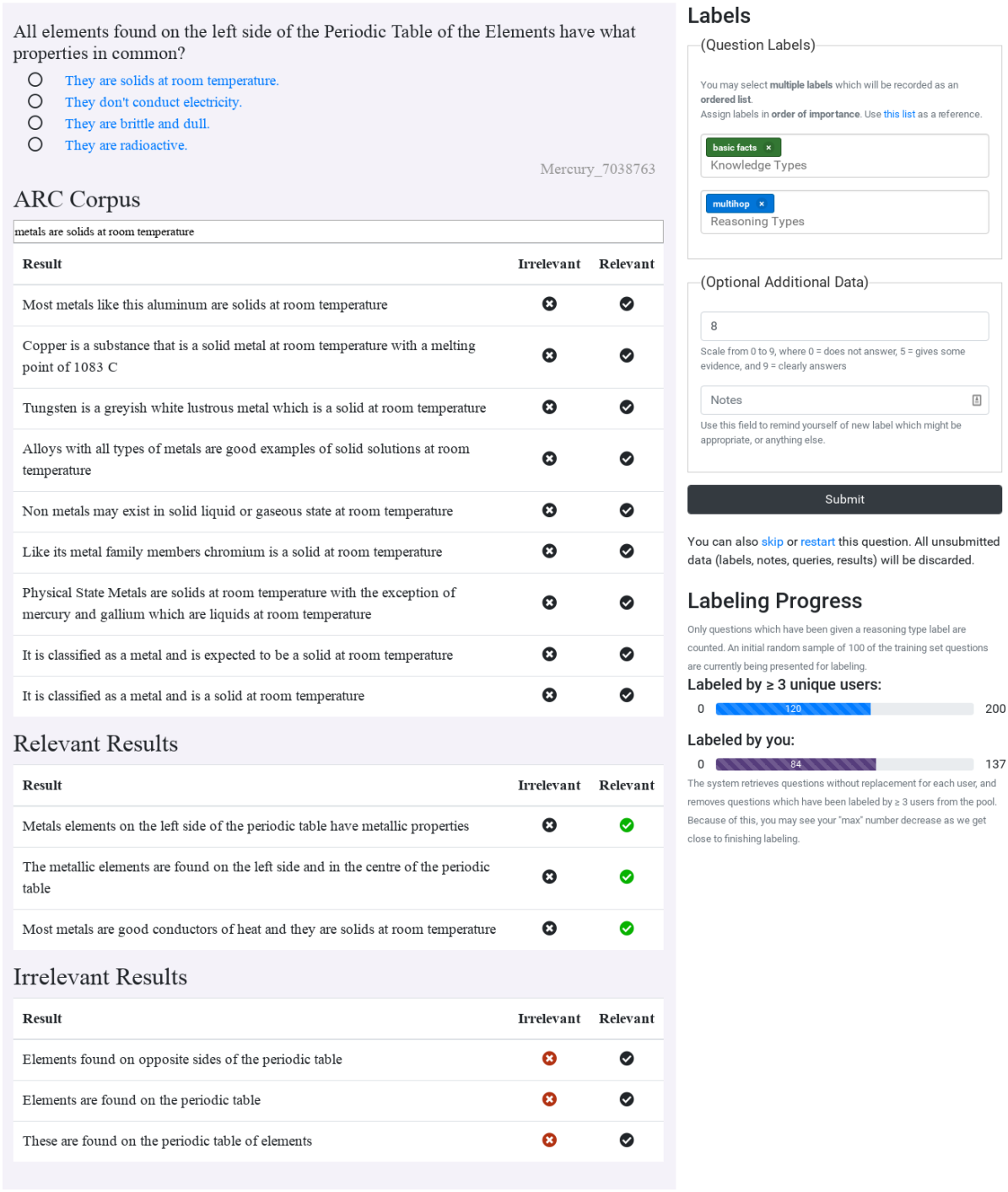

Figure 1: A screenshot of the interface to our annotation system, described in Section 2.

2.1 Question Annotation

The right hand side of the interface deals with the annotation of a given question. There are two boxes for annotating knowledge and reasoning types respectively. The labels are populated from the knowledge and reasoning type tables in Boratko et al. (2018) (more on these types in Section 3). The annotator can also provide optional information on the quality of the retrieved search results if they choose to run a query. Finally, the annotator can use the optional field below quality to enter additional notes about the question; these notes are stored and can be retrieved for subsequent discussion and refinement of the labels.

2.2 Search Result Retrieval & Annotation

In addition to labeling the knowledge and reasoning types systematically, we demonstrate yet an-

other capability of our interface: given a corpus of knowledge, we are able to retrieve and display search results that may be relevant to the question (and its corresponding options) at hand. This is useful because it gives a solution technique an additional signal as it tries to identify the correct answer to a given question. In open-domain question answering, the retriever plays as important a role as the machine reader (Chen et al., 2017). In the past few years, there has been a lot of effort in designing sophisticated neural architectures for reading a small piece of text (e.g. paragraph) (Wang and Jiang, 2016; Xiong et al., 2016; Seo et al., 2016; Lee et al., 2016, inter alia). However, most work in open domain settings (Chen et al., 2017; Clark and Gardner, 2017; Wang et al., 2018) only uses simple retrievers (such as TF-IDF based ones). As a result, there is a notable decrease in the performance of the QA system. One roadblock for training a sophisticated retriever is the lack of available training data which annotates the relevance of a retrieved context with respect to the question. We believe our annotated retrieval data can be used to train a better ranker/retriever without obliging annotators to explicitly connect the supporting passages (Jansen et al., 2018).

The underlying retriever in our interface is a simple Elasticsearch, similar to the one used by Clark et al. (2018). The interface is populated by default with the top ranked sentences that are retrieved with the given question as the input query. However, we noticed that results thus retrieved were often irrelevant to answering the question. To address this, our labeling interface also allows annotators to input their own custom queries. We found that reformulating the initial query significantly improved the quality of the retrieved context (results). We encouraged the annotators to mark the contexts (results) that they thought were relevant to answering the question at hand. For example, in Figure 1, the annotator came up with a novel query – `'metals are solid at room temperatures'` – and also marked the relevant sentences which are needed to answer this question. Note that sometimes we need to reason over multiple sentences to arrive at the answer. For example, the question in Figure 1 can be answered by combining the first and third sentences in the '`Relevant Results`' tab.

3 Knowledge & Reasoning Types

In previous work (Clark et al., 2018), the standardized test questions under consideration were split into various categories based on the kinds of *knowledge* and *reasoning* that are needed to answer those questions. The idea of classifying questions by these two types is central to the notion of standardized testing, which endeavors to test students on various kinds of knowledge, as well as various problem types and solution techniques. These categories allow for the classification of questions, which makes it easier to partition them into subsets to measure performance and improve solution strategies.

3.1 Knowledge Types

In most question-answering (QA) scenarios, the knowledge that is present with the system (or the agent) determines whether a given question can be answered. The full list of the revised knowledge labels (types) – along with the instructions given to annotators and respective exemplars from the ARC question set – can be found in our complementary work (Boratko et al., 2018). For the annotation of knowledge types using our interface, annotators were given the following instructions:

> *You are to answer the question, "In a perfect world given an ideal knowledge source, what types of knowledge would you as a human need to answer this question?" You are allowed to select* ***multiple labels*** *for this type which will be recorded as an ordered list. You are to assign labels* ***in the order of importance*** *to answering the questions at hand.*

In order to level the field among annotators, we included phrasing about an *ideal knowledge source*. Additionally, displaying the retrieved search results in the interface provides another way for the annotators to share some common ground with respect to the typical kind of knowledge that is likely to be available. We also provide instruction-based definitions for each class, as opposed to the single exemplars provided previously. We believe this greatly simplifies the annotation task for new annotators, since they no longer need to perform a preliminary manual analysis of the QA set in order to understand the distinctions between the classes.

3.2 Reasoning Types

The annotation instructions for reasoning types follow a similar pattern to the knowledge types described in the previous section. The annotators

were given the following instructions when annotating the reasoning types:

> *You are to answer the question, "What types of reasoning or problem solving would a competent student with access to Wikipedia need to answer this question?" You are allowed to select* ***multiple labels*** *for this type which will be recorded as an ordered list. You are to assign labels* ***in the order of importance*** *to answering the questions at hand.*
>
> *You may use the search results to help differentiate between the linguistic and multi-hop reasoning types. Any label other than these should take precedence if they apply. For example, a question that requires using a mathematical formula along with linguistic matching should be labeled* **algebraic, linguistic**.

Notice that the instructions in this case refer to being able to access a specific knowledge corpus, and allow for the selection of multiple labels in decreasing order of applicability. We also provide specific instructions on the order of precedence as relates to *linguistic* and *multi-hop* reasoning types: this is based on our empirical observation that many questions can be classified trivially into these reasoning categories, and we would prefer (for downstream application use) a clean split into as many distinct categories as possible.

4 Results

Members of the annotation group were given access to the annotation interface (which includes the question, answers, query search results and more information as described above). Each annotator was shown the questions in a random order, and was allowed to skip or pass any question.

Statistics. We collected labels from at least 3 unique annotators (out of the possible 10) for 192 distinct questions. This annotation process produced 1.42 knowledge type labels and 1.7 reasoning type labels per question. Figure 2 and Figure 3 shows the distribution of annotation labels by all raters at any position. While *Basic Facts* dominates the knowledge type labels, there is no clear cut consensus for the reasoning type. Indeed, *qn logic*, *linguistic*, and *explanation* occur most frequently.

4.1 Inter-Rater Agreement

A comprehensive look at the labels and inter-rater agreement can be found in Table 1 and Table 2. Fleiss' κ is often used to measure inter-rater agreement (Cohen, 1995). Informally, this measures the

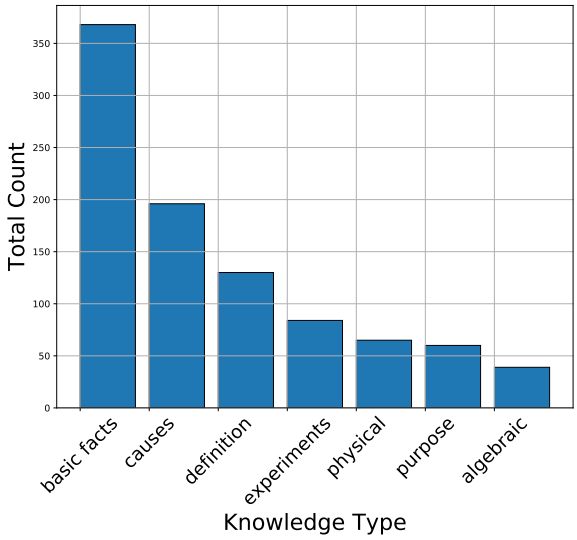

Figure 2: Histogram of the first (most important) knowledge label for each question; the Y-axis refers to annotations.

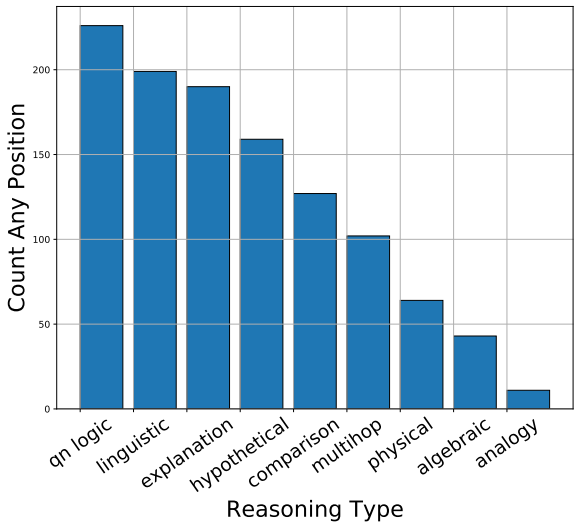

Figure 3: Histogram of the first (most important) reasoning label for each question; the Y-axis refers to annotations.

amount of agreement, beyond chance, based on the number of raters, objects and classes. $\kappa > 0.2$ is typically taken to denote good agreement between raters, while a negative value means that there was little to no agreement. Since Fleiss' κ is only defined for a single set of labels, we consider only the first (most important) label for each question in the statistic we report.

In addition to Fleiss' κ we also use the Kemeny voting rule (Kemeny, 1959) to measure the consensus by the annotators. The Kemeny voting rule minimizes the Kendall Tau (Kendall, 1938) (flip) distance between the output ordering and the ordering of all annotators. One theory of voting (aggregation) is that there is a true or correct ordering and all voters provide a noisy observation of the

ground truth. This method of thinking is largely credited to Condorcet (de Caritat, 1785; Young, 1988) and there is recent work in characterizing other voting rules as maximum likelihood estimators (MLEs) (Conitzer et al., 2009). The Kemeny voting rule is the MLE of the Condorcet Noise Model, in which pairwise inversions of the preference order happen uniformly at random (Young, 1988, 1995). Hence, if we assume all annotators make pairwise errors uniformly at random then Kemeny is the MLE of label orders they report.

Label	Appears	Majority	Consensus
basic facts	125	69	28
algebraic	13	5	2
definition	52	16	5
causes	78	33	15
experiments	35	19	13
purpose	30	13	0
physical	21	3	1
Fleiss' $\kappa = 0.342$			

Table 1: Pairwise inter-rater agreement for Knowledge Labels, along with the mean and Fleiss' κ for survey responses.

Label	Appears	Majority	Consensus
linguistic	66	31	8
algebraic	15	8	3
explanation	80	22	4
hypothetical	62	21	6
multihop	45	6	0
comparison	46	13	3
qn logic	78	33	2
physical	18	3	0
analogy	4	1	1
Fleiss' $\kappa = -0.683$			

Table 2: Pairwise inter-rater agreement for Reasoning Labels, along with the mean and Fleiss' κ for survey responses.

4.1.1 Knowledge Labels

We achieve $\kappa = 0.342$, which means that our raters did a reasonable job of independently agreeing on the types of knowledge required to answer the questions. The mean Kemeny score of the consensus ranking for each question is 2.57, meaning that on average there are less than three flips required to get from the consensus ranking to each of the annotators' rankings. The most frequent label in the first position was *basic facts*, followed by *causes*. Overall, there was a reasonable amount of consensus between the raters for knowledge type: $64/192$ questions had a consensus amongst all the raters. Taken together, our results on knowledge type indicate that most questions deal with *basic facts*, *causes*, and *definitions*; and that labeling can be done reliably.

4.1.2 Reasoning Labels

The inter-rater agreement score for the reasoning labels tells a very different story from the knowledge labels. The agreement was $\kappa = -0.683$, which indicates that raters did not agree above chance on their labels. Strong evidence for this comes from the fact that only $27/192$ questions had a consensus label. This may be due to the fact that we allow multiple labels, and the annotators simply disagree on the *order* of the labels. However, the score of the consensus ranking for each question is 6.57, which indicates that on average the ordering of the labels is quite far apart.

Considering the histogram in Figure 3, we see that *qn logic*, *linguistic*, and *explanation* are the most frequent label types; this may indicate that getting better at understanding the questions themselves could lead to a big boost for reasoners. For Figure 4, we have merged the first and second label (if present) for all annotators. Now, the set of all possible labels is all singletons as well as all pairs of labels. Comparing this histogram to the one in Figure 3, we see that while *linguistic* and *explanation* remain somewhat unchanged, the *qn logic* label becomes very spread out across the types. This is more support for our hypothesis that annotators may be disagreeing on the ordering of the labels, rather than the content itself.

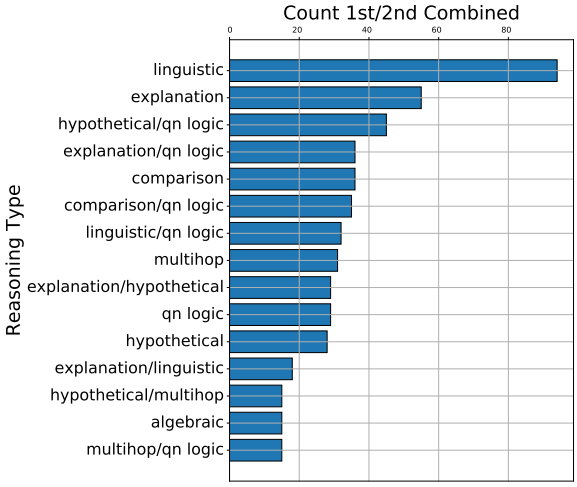

Figure 4: Histogram of the reasoning labels when we combine the first and (if present) second label of every annotator. The count refers to annotations.

4.2 Search Results

To quantitatively measure the efficacy of the annotated context (search results) from the interface, we evaluated 47 questions and their respective human-annotated relevant sentences with a pretrained `DrQA` model (Chen et al., 2017). We compared this to a baseline which only returned the sentences retrieved by using the text of the

question plus given options as input queries. Since DrQA returns a span from the input sentences, we picked the multiple choice option that maximally overlapped with the returned answer span. Our baseline results are 7 correct out of 47 questions. With the annotated context, the performance increased to 27 correctly answered questions - a 42% increase in accuracy. Encouraged by these results, we posit that the community should focus a lot of attention on improving the retrieval portions of the various QA systems available; we think that annotated context will certainly help in training a better ranker. We conclude that the community should focus on improving the retrieval portion of their QA system and we think that the annotated context would help in training a better ranker.

5 Conclusion & Future Work

In this paper, we introduce a novel annotation interface and define annotation instructions for the knowledge and reasoning type labels that are used for question analysis for standardized tests. We annotate approximately 200 questions from the ARC Challenge Set shared by AI2 with the types of knowledge and reasoning required to answer the respective questions. Each question has at least 3 annotators, with high agreement on the requirements for knowledge type. We will leverage the knowledge and reasoning type annotations, as well as the search annotations, to improve the performance of QA systems. We will also release these annotations to the community to complement the ARC Dataset, and make our annotation interface available to interested researchers for use with other question-answering (QA) tasks.

References

Michael Boratko, Harshit Padigela, Divyendra Mikkilineni, Pritish Yuvraj, Rajarshi Das, Andrew McCallum, Maria Chang, Achille Fokoue-Nkoutche, Pavan Kapanipathi, Nicholas Mattei, Ryan Musa, Kartik Talamadupula, and Michael Witbrock. 2018. A Systematic Classification of the Knowledge, Reasoning, and Context within the ARC Dataset. In *Proceedings of the ACL 2018 Workshop on Machine Reading for Question Answering (MRQA)*. Association for Computational Linguistics.

M. J. A. N. de Caritat. 1785. *Essai sur l'application de l'analyse à la probabilité des décisions: rendues à la pluralité des voix*. Paris: L'Imprimerie Royale.

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

Christopher Clark and Matt Gardner. 2017. Simple and effective multi-paragraph reading comprehension. *arXiv preprint arXiv:1710.10723*.

P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. 2018. Think you have solved question answering? Try ARC, the AI2 Reasning Challenge. In *ArXiv e-prints 1803.05457*.

P. R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. MIT Press.

V. Conitzer, M. Rognlie, and L. Xia. 2009. Preference functions that score rankings and maximum likelihood estimation. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI)*, pages 109–115.

Peter Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton Morrison. 2018. WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

J. G. Kemeny. 1959. Mathematics without numbers. *Daedalus*, 88(4):577–591.

M. G. Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kenton Lee, Shimi Salant, Tom Kwiatkowski, Ankur Parikh, Dipanjan Das, and Jonathan Berant. 2016. Learning recurrent span representations for extractive question answering. *arXiv preprint arXiv:1611.01436*.

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

Shuohang Wang and Jing Jiang. 2016. Machine comprehension using match-lstm and answer pointer. *arXiv preprint arXiv:1608.07905*.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018. R3: Reinforced ranker-reader for open-domain question answering.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

H. P. Young. 1988. Condorcet's theory of voting. *The American Political Science Review*, 82(4):1231 – 1244.

H. P. Young. 1995. Optimal voting rules. *The Journal of Economic Perspectives*, 9(1):51–64.