Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences

Daniel Khashabi[1], **Snigdha Chaturvedi**[2], **Michael Roth**[3], **Shyam Upadhyay**[1], **Dan Roth**[1]
[1]University of Pennsylvania, [2]University of California, Santa Cruz, [3]Saarland University
{danielkh,shyamupa,danroth}@cis.upenn.edu, snigdha@ucsc.edu, mroth@coli.uni-sb.de

Abstract

We present a reading comprehension challenge in which questions can only be answered by taking into account information from multiple sentences. We solicit and verify questions and answers for this challenge through a 4-step crowdsourcing experiment. Our challenge dataset contains $\sim 6k$ questions for $+800$ paragraphs across 7 different domains (elementary school science, news, travel guides, fiction stories, etc) bringing in linguistic diversity to the texts and to the questions wordings. On a subset of our dataset, we found human solvers to achieve an F1-score of 86.4%. We analyze a range of baselines, including a recent state-of-art reading comprehension system, and demonstrate the difficulty of this challenge, despite a high human performance. The dataset is the first to study multi-sentence inference at scale, with an open-ended set of question types that requires reasoning skills.

1 Introduction

Machine Comprehension of natural language text is a fundamental challenge in AI and it has received significant attention throughout the history of AI (Greene, 1959; McCarthy, 1976; Reiter, 1976; Winograd, 1980). In particular, in natural language processing (NLP) it has been studied under various settings, such as multiple-choice Question-Answering (QA) (Green Jr. et al., 1961), Reading Comprehension (RC) (Hirschman et al., 1999), Recognizing Textual Entailment (RTE) (Dagan et al., 2013) etc. The area has seen rapidly increasing interest, thanks to the existence of sizable datasets and standard benchmarks. CNN/Daily Mail (Hermann et al., 2015), SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2016) to name a few, are some of the datasets that were released recently with the goal of facilitating research in machine comprehension. Despite all the excitement fueled by that large data sets and the ability to directly train statistical learning models, current QA systems do not have capabilities comparable to elementary school or younger children (Clark and Etzioni, 2016). For many of these datasets, researchers point out that models neither need to 'comprehend' in order to correctly predict an answer, nor do they learn to 'reason' in a way that generalizes across datasets. For example, Khashabi et al. (2016) showed that adversarial perturbation in candidate answers results in a significant drop in performance of a few state-of-art science QA systems. Similarly, Jia and Liang (2017) show that adding an adversarially selected sentence to the instances in the SQuAD datasets drastically reduces the performance of many of the existing baselines. Chen et al. (2016) show that in the CNN/Daily Mail datasets, "the required reasoning and inference level ... is quite simple" and that a relatively simple algorithm can get almost close to the upper-bound. We believe that one key reason that simple algorithms can deal with the existing large datasets but, nevertheless, fail at generalization, is that the datasets do not actually require a deep understanding.

We propose to address this shortcoming by developing a reading comprehension challenge in which answering each of the questions requires reasoning over multiple sentences.

There is evidence that answering 'single-sentence questions', i.e. questions that can be answered from a single sentence of the given paragraph, is easier than answering multi-sentence questions', which require multiple sentences to answer a given question. For example, Richardson et al. (2013) released a reading comprehension dataset that contained both single-sentence and multi-sentence questions; models proposed for this task yielded considerably better performance on the single-sentence questions than on the multi-

Proceedings of NAACL-HLT 2018, pages 252–262
New Orleans, Louisiana, June 1 - 6, 2018.

sentence questions (according to Narasimhan and Barzilay (2015) accuracy of about 83% and 60% on these two types of questions, respectively).

There could be multiple reasons for this. First, multi-sentence reasoning seems to be inherently a difficult task. Research has shown that while complete-sentence construction emerges as early as first grade for many children, their ability to integrate sentences emerges only in fourth grade (Berninger et al., 2011). Answering multi-sentence questions might be more challenging for an automated system because it involves more than just processing individual sentences but rather combining linguistic, semantic and background knowledge across sentences—a computational challenges in itself. Despite these challenges, multi-sentence questions can be answered by humans and hence present an interesting yet reasonable goal for AI systems (Davis, 2014).

In this work, we propose a multi-sentence QA challenge in which questions can be answered only using information from multiple sentences. Specifically, we present MultiRC (Multi-Sentence Reading Comprehension)[1]—a dataset of short paragraphs and multi-sentence questions that can be answered from the content of the paragraph. Each question is associated with several choices for answer-options, out of which *one or more* correctly answer the question. Figure 1 shows two examples from our dataset. Each instance consists of a multi-sentence paragraph, a question, and answer-options. All instances were constructed such that it is not possible to answer a question correctly without gathering information from multiple sentences. Due to space constraints, the figure shows only the relevant sentences from the original paragraph. The entire corpus consists of 871 paragraphs and about $\sim$ 6k multi-sentence questions.

The goal of this dataset is to encourage the research community to explore approaches that can do more than sophisticated lexical-level matching. To accomplish this, we designed the dataset with three key challenges in mind. (i) The number of correct answer-options for each question is not pre-specified. This removes the over-reliance of current approaches on answer-options and forces them to decide on the correctness of each candidate answer independently of others. In other words, unlike previous work, the task here is not to simply identify the best answer-option, but to evaluate the correctness of each answer-option individually. For example, the first question in Figure 1 can be answered by combining information from sentences 3, 5, 10, 13 and 15. It requires not only understanding that the stalker's name is Timothy but also that he is the man who Mary had hit. (ii) The correct answer(s) is not required to be a span in the text. For example, the correct answer, A, of the second question in Figure 1 is not present in the paragraph verbatim. It is instead a combination of two spans from 2 sentences: 12 and 13. Such answer-options force models to process and understand not only the paragraph and the question but also the answer-options. (iii) The paragraphs in our dataset have diverse provenance by being extracted from 7 different domains such as news, fiction, historical text etc., and hence are expected to be more diverse in their contents as compared to single-domain datasets. We also expect this to lead to diversity in the types of questions that can be constructed from the passage.

Overall, we introduce a reading comprehension dataset that significantly differs from most other datasets available today in the following ways:

S3: Hearing noises in the garage, Mary Murdock finds a bleeding man, mangled and impaled on her jeep's bumper.
S5: Panicked, she hits him with a golf club.
S10: Later the news reveals the missing man is kindergarten teacher, Timothy Emser.
S12: It transpires that Rick, her boyfriend, gets involved in the cover up and goes to retrieve incriminatory evidence off the corpse, but is killed, replaced in Emser's grave.
S13: It becomes clear Emser survived.
S15: He stalks Mary many ways.

Who is stalking Mary?

A)* Timothy	D) Rick
B) Timothy's girlfriend	E) Murdock
C)* The man she hit	F) Her Boyfriend

S1: Most young mammals, including humans, play.
S2: Play is how they learn the skills that they will need as adults.
S6: Big cats also play.
S8: At the same time, they also practice their hunting skills.
S11: Human children learn by playing as well.
S12: For example, playing games and sports can help them learn to follow rules.
S13: They also learn to work together.

What do human children learn by playing games and sports?
A)* They learn to follow rules and work together
B) hunting skills
C)* skills that they will need as adult

Figure 1: Examples from our MultiRCcorpus. Each example shows relevant excerpts from a paragraph; multi-sentence question that can be answered by combining information from multiple sentences of the paragraph; and corresponding answer-options. The correct answer(s) is indicated by a *. Note that there can be multiple correct answers per question.

[1] http://cogcomp.org/multirc/

- $\sim 6k$ high-quality multiple-choice RC questions that are generated (and manually verified via crowdsourcing) to require integrating information from multiple sentences.
- The questions are not constrained to have a *single* correct answer, generalizing existing paradigms for representing answer-options.
- Our dataset is constructed using 7 different sources, allowing more diversity in content, style, and possible question types.
- We show a significant performance gap between current solvers and human performance, indicating an opportunity for developing sophistical reasoning systems.

2 Relevant Work

Automated reasoning is arguably one of the major problems in contemporary AI research. Brachman et al. (2005) suggest challenges for developing AI program that can pass the SAT exams. In similar spirit Clark and Etzioni (2016) advocate elementary-school tests as a new test for AI. Davis (2014) proposes hand-construction of multiple-choice challenge sets that are easy for children but difficult for computers. Despite Davis' claim on simplicity of his target questions, it is not clear how easy it is to generate such questions, as he doesn't provide any reasonably-sized dataset matching his proposal. Weston et al. (2015) present a relatively small dataset of 10 reasoning categories, and propose to build a system that uses a world model and a linguistic model. The fundamental limitation of the dataset is that it is generated according to a restricted set of reasoning categories, which possibly limits the complexity and diversity of questions.

Some other recent datasets proposed for machine comprehension also pay attention to type of questions and reasoning required. For example, RACE (Lai et al., 2017) attempts to incorporate different types of reasoning phenomena, and MCTest (Richardson et al., 2013) attempted to contain at least 50% multi-sentence reasoning questions. However, since the crowdsourced workers who created the dataset were only encouraged, and not required, to write such questions, it is not clear how many of these questions actually require multi-sentence reasoning (see Sec. 3.5). Similarly, only about 25% of question in the RACE dataset require multi-sentence reasoning as reported in their paper. Remedia (Hirschman et al., 1999) also contains 5 different types of questions (based on question words) but is a much smaller dataset. Other datasets which do not deliberately attempt to include multi-sentence reasoning, like SQuAD (Rajpurkar et al., 2016) and the CNN/Daily Mail dataset (Hermann et al., 2015), suffer from even lower percentage of such questions (12% and 2% respectively (Lai et al., 2017)). There are several other corpora which do not guarantee specific reasoning types, including MS MARCO (Nguyen et al., 2016), WikiQA (Yang et al., 2015), and TriviaQA (Joshi et al., 2017).

The complexity of reasoning required for a reading comprehension dataset would depend on several factors such as the source of questions or paragraphs; the way they are generated; and the order in which they are generated (i.e. questions from paragraphs, or the reverse). Specifically, paragraphs' source could influence the complexity and diversity of the language of the paragraphs and questions, and hence the required level of reasoning capabilities. Unlike most current datasets which rely on only one or two sources for their paragraphs (e.g. CNN/Daily Mail and SQuAD rely only on news and Wikipedia articles respectively) our dataset uses 7 different domains.

Another factor that distinguishes our dataset from previously proposed corpora is the way answers are represented. Several datasets represent answers as multiple-choices with a single correct answer. While multiple-choice questions are easy to grade, coming up with non-trivial correct and incorrect answers can be challenging. Also, assuming exactly one correct answer (e.g., as in MCTest and RACE) inadvertently changes the task from choosing the correct answer to choosing the most likely answer. Other datasets (e.g MS-MARCO and SQuAD) represent answers as a contiguous substring within the passage. This assumption of the answer being a span of the paragraph, limits the questions to those whose answer is contained verbatim in the paragraph. Unfortunately, it rules out more complicated questions whose answers are only implied by the text and hence require a deeper understanding. Because of these limitations, we designed our dataset to use multiple-choice representations, but without specifying the number of correct answers for each question.

3 Construction of MultiRC

In this section we describe our principles and methodology of dataset collection. This includes automatically collecting paragraphs, composing questions and answer-options through crowdsourcing platform, and manually curating the collected data. We also summarize a pilot study that helped us design this process, and end with a summary of statistics of the collected corpus.

3.1 Principles of design

Questions and answers in our dataset are designed based on the following key principles:

Multi-sentenceness. Questions in our challenge require models to use information from multiple sentences of a paragraph. This is ensured through explicit validation. We exclude any question that can be answered based on a single sentence from a paragraph.

Open-endedness. Our dataset is not restricted to questions whose answer can be found verbatim in a paragraph. Instead, we provide a set of handcrafted answer-options for each question. Notably, they can represent information that is not explicitly stated in the text but is only inferable from it (e.g. implied counts, sentiments, and relationships).

Answers to be judged independently. The total number of answer options per question is variable in our data and we explicitly allow multiple correct and incorrect answer options (e.g. 2 correct and 1 incorrect options). As a consequence, correct answers cannot be guessed solely by a process of elimination or by simply choosing the best candidates out of the given options.

Through these principles, we encourage users to explicitly model the semantics of text beyond individual words and sentences, to incorporate extra-linguistic reasoning mechanisms, and to handle answer options independently of one another.

Variability. We encourage variability on different levels. Our dataset is based on paragraphs from multiple domains, leading to linguistically diverse questions and answers. Also, we do not impose any restrictions on the questions, to encourage different forms of reasoning.

3.2 Sources of documents

The paragraphs used in our dataset are extracted from various sources. Here is the complete list of the text types and sources used in our dataset, and the number of paragraphs extracted from each category (indicated in square brackets on the right):

1. News: [121]
   - CNN (Hermann et al., 2015)
   - WSJ (Ide et al., 2008)
   - NYT (Ide et al., 2008)
2. Wikipedia articles [92]
3. Articles on society, law and justice (Ide and Suderman, 2006) [91]
4. Articles on history and anthropology (Ide et al., 2008) [65]
5. Elementary school science textbooks [2] [153]
6. 9/11 reports (Ide and Suderman, 2006) [72]
7. Fiction: [277]
   - Stories from the Gutenberg project
   - Children stories from MCTest (Richardson et al., 2013)
   - Movie plots from CMU Movie Summary corpus (Bamman et al., 2013)

From each of the above-mentioned sources we extracted paragraphs that had enough content. To ensure this we followed a 3-step process. In the first step we selected top few sentences from paragraphs such that they contained 1k-1.5k characters. To ensure coherence, all sentences were contiguous and extracted from the same paragraph. In this process we also discarded paragraphs that seemed to deviate too much from third person narrative style. For example, while processing Gutenberg corpus we considered files that had at least 5k lines because we found that most of them were short poetic texts. In the second step, we annotated (Khashabi et al., 2018b) the paragraphs and automatically filtered texts using conditions such as the average number of words per sentence; number of named entities; number of discourse connectives in the paragraph. These were designed by the authors of this paper after reviewing a small sample of paragraphs. A complete set of conditions is listed in Table 1. Finally in the last step, we manually verified each paragraph and filtered out the ones that had formatting issues or other concerns that seemed to compromise their usability.

[2] https://www.ck12.org

Condition	bound
Number of sentences	≥ 6 & ≤ 18
Number of NER(CoNLL) mentions	≥ 2
Avg. number of NER(CoNLL) mentions	≥ 0.2
Number of NER(Ontonotes) mentions	≥ 4
Avg. number of NER(Ontonotes) mentions	≥ 0.25
Avg. number of words per sentence	≥ 5
Number of coreference mentions	≥ 3
Avg. number of coreference mentions	≥ 0.1
Number of coreference relations	≥ 3
Avg. number of coreference relations	≥ 0.08
Number of coreference chains	≥ 2
Avg. number of coreference chains	≥ 0.1
Number of discourse markers	≥ 2

Table 1: Bounds used to select paragraphs for dataset creation.

3.3 Pipeline of question extraction

In this section, we delineate details of the process for collecting questions and answers. Figure 2 gives a high-level idea of the process. The first two steps deal with creating multi-sentence questions, followed by two steps for construction of candidate answers. Interested readers can find more details on set-ups of each step in Appendix I.

Step 1: Generating questions. The goal of the first step of our pipeline is to collect multi-sentence questions. We show each paragraph to 5 turkers and ask them to write 3-5 questions such that: (1) the question is answerable from the passage, and (2) only those questions are allowed whose answer cannot be determined from a single sentence. We clarify this point by providing example paragraphs and questions. In order to encourage turkers to write meaningful questions that fit our criteria, we additionally ask them for a correct answer and for the sentence indices required to answer the question. To ensure the grammatical quality of the questions collected in this step, we limit the turkers to the countries with English as their major language. After the acquisition of questions in this step, we filter out questions which required less than 2 or more than 4 sentences to be answered; we also run them through an automatic spell-checker[3] and manually correct questions regarding typos and unusual wordings.

Step 2: Verifying multi-sentenceness of questions. In a second step, we verify that each question can only be answered using more than one sentence. For each question collected in the previous step, we create question-sentence pairs by pairing it with each of the sentences necessary for answering it as indicated in the previous step. For a given question-sentence pair, we then ask turkers to annotate if they could answer the question from the sentence it is paired with (binary annotation). The underlying idea of this step is that a multi-sentence question would not be answerable from a single sentence, hence turkers should not be able to give a correct answer for any of the question-sentence pair. Accordingly, we determine a question as requiring multiple sentences only if the correct answer cannot be guessed from any single question-sentence pair. We collected at least 3 annotations per pair, and to avoid sharing of information across sentences, no two pairs shown to a turker came from the same paragraph. We aggregate the above annotations for each question-answer pair and retain only those questions for which no pair was judged as answerable by a majority of turkers.

Step 3: Generating answer-options. In this step, we collect answer-options that will be shown with each question. Specifically, for each verified question from the previous steps, we ask 3 turkers to write as many correct and incorrect answer options as they can think of. In order to not curb creativity, we do not place a restriction on the number of options they have to write. We explicitly ask turkers to design difficult and non-trivial incorrect answer-options (e.g. if the question is about a person, a non-trivial incorrect answer-option would be other people mentioned in the paragraph).

After this step, we perform a light clean up of the candidate answers by manually correcting minor errors (such as typos), completing incomplete sentences and rephrasing any ambiguous sentences. We further make sure there is not much repetition in the answer-options, to prevent potential exploitation of correlation between some candidate answers in order to find the correct answer. For example, we drop obviously duplicate answer-options (i.e. identical options after lower-casing, lemmatization, and removing stop-words).

Step 4: Verifying quality of the dataset. This step serves as the final quality check for both questions and the answer-options generated in the previous steps. We show each paragraph, its questions, and the corresponding answer-options to 3 turkers, and ask them to indicate if they find any errors (grammatical or otherwise), in the questions and/or answer-options. We then manually review,

[3]Grammarly: www.grammarly.com

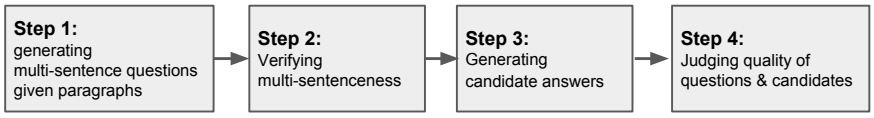

Figure 2: Pipeline of our dataset construction.

and correct if needed, all erroneous questions and answer-options. This ensures that we have meaningful questions and answer-options. In this step, we also want to verify that the correct (or incorrect) options obtained from Step 3 were indeed correct (or incorrect). For this, we additionally ask the annotators to select all correct answer-options for the question. If their annotations did not agree with the ones we had after Step 3 (e.g. if they unanimously selected an 'incorrect' option as the answer), we manually reviewed and corrected (if needed) the annotation.

3.4 Pilot experiments

The 4-step process described above was a result of detailed analysis and substantial refinement after two small pilot studies.

In the first pilot study, we ran a set of 10 paragraphs extracted from the CMU Movie Summary Corpus through our pipeline. Our then pipeline looked considerably different from the one described above. We found the steps that required turkers to write questions and answer-options to often have grammatical errors, possibly because a large majority of turkers were non-native speakers of English. This probslem was more prominent in questions than in answer-options. Because of this, we decided to limit the task to native speakers. Also, based on the results of this pilot, we overhauled the instructions of these steps by including examples of grammatically correct—but undesirable (not multi-sentence)—questions and answer-options, in addition to several minor changes.

Thereafter, we decided to perform a manual validation of the verification steps (current Steps 2 and 4). For this, we (the authors of this paper) performed additional annotations ourselves on the data shown to turkers, and compared our results with those provided by the turkers. We found that in the verification of answer-options, our annotations were in high agreement (98%) with those obtained from mechanical turk. However, that was not the case for the verification of multi-sentence questions. We made several further changes to the first two steps. Among other things, we clarified in the instructions that turkers should not use their background knowledge when writing and verifying questions, and also included negative examples of such questions. Additionally, when turkers judged a question to be answerable using a single sentence, we decided to encourage (but not require) them to guess the answer to the question. This improved our results considerably, possibly because it forced annotators to think more carefully about what the answer might be, and whether they *actually* knew the answer or they just *thought* that they knew it (possibly because of background knowledge or because the sentence contained a lot of information relevant to the question). Guessed answers in this step were only used to verify the validity of multi-sentence questions. They were not used in the dataset or subsequent steps.

After revision, we ran a second pilot study in which we processed a set of 50 paragraphs through our updated pipeline. This second pilot confirmed that our revisions were helpful, but thanks to its larger size, also allowed us to identify a couple of borderline cases for which additional clarifications were required. Based on the results of the second pilot, we made some additional minor changes and then decided to apply the pipeline for creating the final dataset.

3.5 Verifying multi-sentenceness

While collecting our dataset, we found that, even though Step 1 instructed turkers to write multi-sentence questions, not all generated questions indeed required multi-sentence reasoning. This happened even after clarifications and revisions to the corresponding instructions, and we attribute it to honest mistakes. Therefore, we designed the subsequent verification step (Step 2).

There are other datasets which aim to include multi-sentence reasoning questions, especially MCTest. Using our verification step, we systematically verify their multi-sentenceness. For this, we conducted a small pilot study on about 60 multi-sentence questions from MCTest. As for our own verification, we created question-sentence pairs for each question and asked annotators to judge whether they can answer a question from the single sentence shown. Because we did not know

which sentences contain information relevant to a question, we created question-sentence pairs using all sentences from a paragraph. After aggregation of turker annotations, we found that about half of the questions annotated as multi-sentence could be answered from a single sentence of the paragraph. This study, though performed on a subset of the data, underscores the necessity of rigorous verification step for multi-sentence reasoning when studying this phenomenon.

3.6 Statistics on the dataset

We now provide a brief summary of MultiRC. Overall, it contains roughly $\sim 6k$ multi-sentence questions collected for about $+800$ paragraphs.[4] The median number of correct and total answer options for each question is 2 and 5, respectively. Additional statistics are given in Table 2.

In Step 1, we also asked annotators to identify sentences required to answer a given question. We found that answering each question required 2.4 sentences on average. Also, required sentences are often not contiguous, and the average distance between sentences is 2.4. Next, we analyze the types of questions in our dataset. Figure 4 shows the count of first word(s) for our questions. We can see that while the popular question words (*What*, *Who*, etc.) are very common, there is a wide variety in the first word(s) indicating a diversity in question types. About 28% of our questions require binary decisions (true/false or yes/no).

We randomly selected 60 multi-sentence questions from our corpus and asked two independent annotators to label them with the type of reasoning phenomenon required to answer them.[5] During this process, the annotators were shown a list of common reasoning phenomena (shown below), and they had to identify one or more of the phenomena relevant to a given question. The list of phenomena shown to the annotators included the following categories: mathematical and logical reasoning, spatio-temporal reasoning, list/enumeration, coreference resolution (including implicit references, abstract pronouns, event coreference, etc.), causal relations, paraphrases and contrasts (including lexical relations such as synonyms, antonyms), commonsense knowledge, and 'other'. The categories were selected after a manual inspection of a subset of questions by two of the authors. The annotation process revealed that answering questions in our corpus requires a broad variety of reasoning phenomena. The left plot in Figure 3 provides detailed results.

The figure shows that a large fraction of questions require coreference resolution, and a more careful inspection revealed that there were different types of coreference phenomena at play here. To investigate these further, we conducted a follow-up experiment in which manually annotated all questions that required coreference resolution into finer categories. Specifically, each question was shown to two annotators who were asked to select one or more of the following categories: entity coreference (between two entities), event coreference (between two events), set inclusion coreference (one item is part of or included in the other) and 'other'. Figure 3 (right) shows the results of this experiment. We can see that, as expected, entity coreference is the most common type of coreference resolution needed in our corpus. However, a significant number of questions also require other types of coreference resolution. We provide some examples of questions along with the required reasoning phenomena in Appendix II.

Parameter	Value
# of paragraphs	871
# of questions	9,872
# of multi-sentence questions	5,825
avg # of candidates (per question)	5.44
avg # of correct answers (per question)	2.58
avg paragraph length (in sentences)	14.3 (4.1)
avg paragraph length (in tokens)	263.1 (92.4)
avg question length (in tokens)	10.9 (4.8)
avg answer length (in tokens)	4.7 (5.5)
% of yes/no/true/false questions	27.57%
avg # of sent. used for questions	2.37 (0.63)
avg distance between the sent.'s used	2.4 (2.58)
% of correct answers verbatim in paragraph	34.96%
% of incorrect answers verbatim in paragraph	25.84%

Table 2: Various statistics of our dataset. Figures in parentheses represent standard deviation.

4 Analysis

In this section, we provide a quantitative analysis of several baselines for our challenge.

Evaluation Metrics. We define precision and recall for a question q as: $\text{Pre}(q) = \frac{|A(q) \cap \hat{A}(q)|}{|\hat{A}(q)|}$ and $\text{Rec}(q) = \frac{|A(q) \cap \hat{A}(q)|}{|A(q)|}$, where $A(q)$ and $\hat{A}(q)$ are the sets of correct and selected answer-options.

[4] We will also release the $3.7k$ questions that did not pass Step 2. Though not multi-sentence questions, they could be a valuable resource on their own.

[5] The annotations were adjudicated by two authors of this paper.

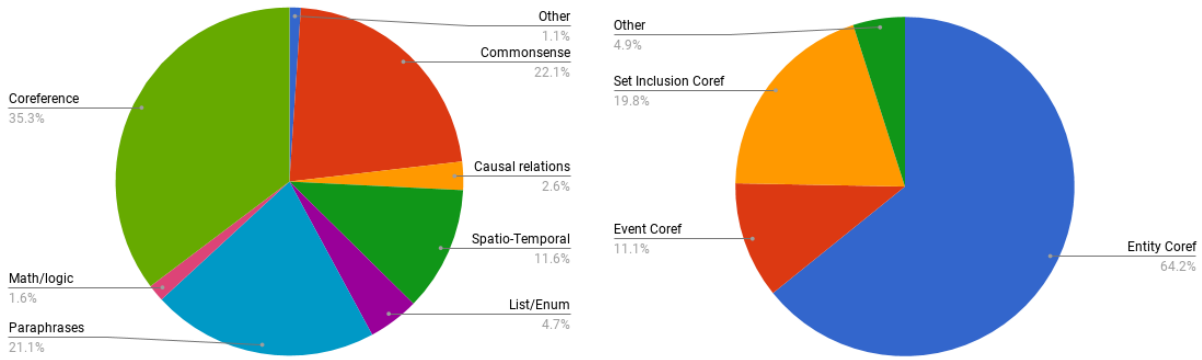

Figure 3: Distribution of (left) general phenomena; (right) variations of the "coreference" phenomena.

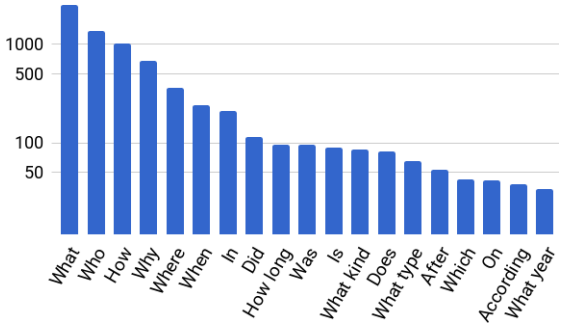

Figure 4: Most frequent first chunks of the questions (counts in log scale).

We define (macro-average) F1_m as the harmonic mean of average-precision $avg_{q \in Q}(\text{Pre}(q))$ and average-recall $avg_{q \in Q}(\text{Rec}(q))$ with Q as the set of all questions.

Since by design, each answer-option can be judged independently, we consider another metric, F1_a, evaluating binary decisions on all the answer-options in the dataset. We define F1_a to be the harmonic mean of $\text{Pre}(Q)$ and $\text{Rec}(Q)$, with $\text{Pre}(Q) = \frac{|A(Q) \cap \hat{A}(Q)|}{|\hat{A}(Q)|}$; $A(Q) = \bigcup_{q \in Q} A(q)$; and similar definitions for $\hat{A}(Q)$ and $\text{Rec}(Q)$.

4.1 Baselines

Human. Human performance provides us with an estimate of the best achievable results on datasets. Using mechanical turk, we ask 4 people (limited to native speakers) to solve our data. We evaluate score of each label by averaging the decision of the individuals.

Random. To get an estimate on the lower-bound we consider a random baseline, where each answer option is selected as correct with a probability of 50% (an unbiased coin toss). The numbers reported for this baseline represent the expected outcome (statistical expectation).

IR (information retrieval baseline). This baseline selects answer-options that best match sentences in a text corpus (Clark et al., 2016). Specifically, for each question q and answer option a_i, the IR solver sends $q + a_i$ as a query to a search engine (we use Lucene) on a corpus, and returns the search engine's score for the top retrieved sentence s, where s must have at least one non-stopword overlap with q, and at least one with a_i.

We create two versions of this system. In the first variation IR(paragraphs) we create a corpus of sentences extracted from all the paragraphs in the dataset. In the second variation, IR(web) in addition to the knowledge of the paragraphs, we use extensive external knowledge extracted from the web (Wikipedia, science textbooks and study guidelines, and other webpages), with 5×10^{10} tokens (280GB of plain text).

SurfaceLR (logistic regression baseline). As a simple baseline that makes use of our small training set, we reimplemented and trained a logistic regression model using word-based overlap features. As described in (Merkhofer et al., 2018), this baseline takes into account the lengths of a text, question and each answer candidate, as well as indicator features regarding the (co-)occurrences of any words in them.

SemanticILP (semi-structured baseline). This state-of-the-art solver, originally proposed for science questions and biology tests, uses a semi-structured representation to formalize the scoring problem as a subgraph optimization problem over multiple layers of semantic abstrac-

	Dev		Test	
	$F1_m$	$F1_a$	$F1_m$	$F1_a$
Random	44.3	43.8	47.1	47.6
IR(paragraphs)	64.3	60.0	54.8	53.9
SurfaceLR	66.1	63.7	66.7	63.5
Human	86.4	83.8	84.3	81.8

Table 3: Performance comparison for different baselines tested on a subset of our dataset (in percentage). There is a significant gap between the human performance and current statistical methods.

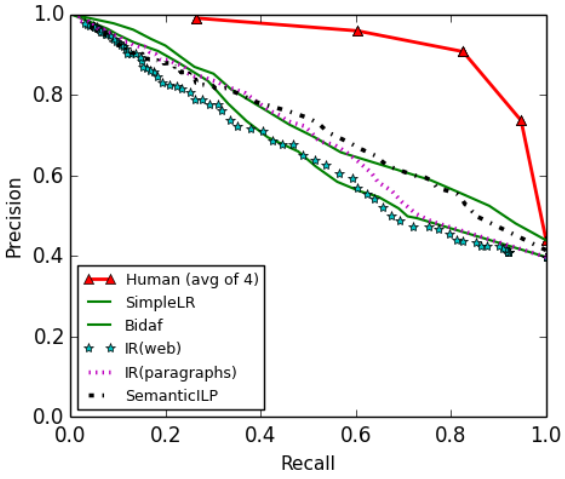

Figure 5: PR curve for each of the baselines. There is a considerable gap with the baselines and human.

tions (Khashabi et al., 2018a). Since the solver is designed for multiple-choice with single-correct answer, we adapt it to our setting by running it for each answer-option. Specifically for each answer-option, we create a single-candidate question, and retrieve a real-valued score from the solver.

BiDAF (neural network baseline). As a neural baseline, we apply this solver by Seo et al. (2017), which was originally proposed for SQuAD but has been shown to generalize well to another domain (Min et al., 2017). Since BiDAF was designed for cloze style questions, we apply it to our multiple-choice setting following the procedure by Kembhavi et al. (2017): Specifically, we score each answer-option by computing the similarity value of it's output span with each of the candidate answers, computed by phrasal similarity tool of Wieting et al. (2015).

4.2 Results

To get a sense of our dataset's hardness, we evaluate both human performance and multiple computational baselines. Each baseline scores an answer-option with a real-valued score, which we threshold to decide whether an answer option is selected or not, where the threshold is tuned on the development set. Table 3 shows performance results for different baselines. The significantly high human performance shows that humans do not have much difficulties in answering the questions. Similar observations can be made in Figure 5 where we plot $avg_{q \in Q}(\text{Pre}(q))$ vs. $avg_{q \in Q}(\text{Rec}(q))$, for different threshold values.

5 Conclusion

In this paper we have presented MultiRC, a reading comprehension dataset in which questions require reasoning over multiple sentences to be answered. Our dataset contains $\sim 6k$ questions extracted from about +800 paragraphs. For each question, it contains multiple answer-options out of which one or more can be correct. The paragraphs (and questions) originate from different domains and hence are amenable to a wide variety and complexity of required reasoning phenomena. We found human performance on this corpus to be about 88% while state-of-the-art machine comprehension models do not exceed a F1-score of 60%. We hope that this significant difference in performance will encourage the community to work towards more sophisticated reasoning systems.

6 Acknowledgement

The authors would like to thank all the contributors to the project. This work was supported by Contract HR0011-15-2-0025 with the US Defense Advanced Research Projects Agency (DARPA). Approved for Public Release, Distribution Unlimited. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This work was partly funded by grants from the German Research Foundation (DFG EXC 284 and RO 4848/1-1), by the Allen Institute for Artificial Intelligence (allenai.org); by Google; and by IBM-ILLINOIS Center for Cognitive Computing Systems Research (C3SR)—a research collaboration as part of the IBM AI Horizons Network.

References

David Bamman, Brendan O'Connor, and Noah A. Smith. 2013. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, Volume 1: Long Papers*. pages 352–361. http://aclweb.org/anthology/P/P13/P13-1035.pdf.

Virginia W. Berninger, William Nagy, and Scott Beers. 2011. Child writers construction and reconstruction of single sentences and construction of multi-sentence texts: Contributions of syntax and transcription to translation. *Reading and writing* 24(2):151–182.

Ronald Brachman, David Gunning, Selmer Bringsjord, Michael Genesereth, Lynette Hirschman, and Lisa Ferro. 2005. Selected grand challenges in cognitive science. Technical report, MITRE Technical Report 05-1218.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*. http://aclweb.org/anthology/P/P16/P16-1223.pdf.

Peter Clark and Oren Etzioni. 2016. My computer is an honor student - but how intelligent is it? standardized tests as a measure of AI. *AI Magazine* 37(1):5–12. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2636.

Peter Clark, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter D. Turney, and Daniel Khashabi. 2016. Combining retrieval, statistics, and inference to answer elementary science questions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. pages 2580–2586. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11963.

Ido Dagan, Dan Roth, Mark Sammons, and Fabio Massimo Zanzoto. 2013. *Recognizing Textual Entailment: Models and Applications*. Morgan and Claypool.

Ernest Davis. 2014. The limitations of standardized science tests as benchmarks for artificial intelligence research: Position paper. *CoRR* abs/1411.1629. http://arxiv.org/abs/1411.1629.

Bert F. Green Jr., Alice K. Wolf, Carol Chomsky, and Kenneth Laughery. 1961. Baseball: An automatic question-answerer. In *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*. ACM, IRE-AIEE-ACM '61 (Western), pages 219–224. https://doi.org/10.1145/1460690.1460714.

Peter H. Greene. 1959. An approach to computers that perceive, learn, and reason. In *Papers Presented at the the March 3-5, 1959, Western Joint Computer Conference*. ACM, IRE-AIEE-ACM '59 (Western), pages 181–186. https://doi.org/10.1145/1457838.1457870.

Karl Moritz Hermann, Tomás Kociský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015*. pages 1693–1701.

Lynette Hirschman, Marc Light, Eric Breck, and John D. Burger. 1999. Deep read: A reading comprehension system. In *27th Annual Meeting of the Association for Computational Linguistics, ACL 1999*. http://www.aclweb.org/anthology/P99-1042.

Nancy Ide, Collin F. Baker, Christiane Fellbaum, Charles J. Fillmore, and Rebecca J. Passonneau. 2008. MASC: the manually annotated sub-corpus of american english. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008*. http://www.lrec-conf.org/proceedings/lrec2008/summaries/617.html.

Nancy Ide and Keith Suderman. 2006. Integrating linguistic resources: The american national corpus model. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC 2006*. pages 621–624. http://www.lrec-conf.org/proceedings/lrec2006/pdf/560_pdf.pdf.

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, EMNLP 2017*. pages 2021–2031. https://aclanthology.info/papers/D17-1215/d17-1215.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*. pages 1601–1611. https://doi.org/10.18653/v1/P17-1147.

Aniruddha Kembhavi, Min Joon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*. pages 5376–5384. https://doi.org/10.1109/CVPR.2017.571.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni, and Dan Roth. 2016. Question answering via integer programming over semi-structured knowledge. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*. pages 1145–1152. http://www.ijcai.org/Abstract/16/166.

Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2018a. Question answering as global reasoning over semantic abstractions. In *Proceedings of The Thirty-Second AAAI Conference on Artificial Intelligence, AAAI 2018 (to appear)*. http://cogcomp.org/page/publication_view/824.

Daniel Khashabi, Mark Sammons, Ben Zhou, Tom Redman, Christos Christodoulopoulos, Vivek Srikumar, Nicholas Rizzolo, Lev Ratinov, Guanheng Luo, Quang Do, Chen-Tse Tsai, Subhro Roy, Stephen Mayhew, Zhilli Feng, John Wieting, Xiaodong Yu, Yangqiu Song, Shashank Gupta, Shyam Upadhyay, Naveen Arivazhagan, Qiang Ning, Shaoshi Ling, and Dan Roth. 2018b. CogCompNLP: your swiss army knife for nlp. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard H. Hovy. 2017. RACE: large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*. pages 785–794. https://aclanthology.info/papers/D17-1082/d17-1082.

J. McCarthy. 1976. An example for natural language understanding and the AI problems it raises. Available at http://jmc.stanford.edu/articles/mrhug/mrhug.pdf.

Elizabeth Merkhofer, John Henderson, David Bloom, Laura Strickhart, and Guido Zarrella. 2018. Mitre at semeval-2018 task 11: Commonsense reasoning without commonsense knowledge. In *Proceedings of the International Workshop on Semantic Evaluation (SemEval-2018)*. New Orleans, LA, USA.

Sewon Min, Min Joon Seo, and Hannaneh Hajishirzi. 2017. Question answering through transfer learning from large fine-grained supervision data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 2: Short Papers*. pages 510–517. https://doi.org/10.18653/v1/P17-2081.

Karthik Narasimhan and Regina Barzilay. 2015. Machine comprehension with discourse relations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, Volume 1: Long Papers*. pages 1253–1262. http://aclweb.org/anthology/P/P15/P15-1121.pdf.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. *CoRR* abs/1611.09268. http://arxiv.org/abs/1611.09268.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*. pages 2383–2392. http://aclweb.org/anthology/D/D16/D16-1264.pdf.

Raymond Reiter. 1976. A semantically guided deductive system for automatic theorem proving. *IEEE Transactions on Computers* 25:328–334. https://doi.org/10.1109/TC.1976.1674613.

Matthew Richardson, Christopher J. C. Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013*. pages 193–203. http://aclweb.org/anthology/D/D13/D13-1020.pdf.

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

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR* abs/1611.09830. http://arxiv.org/abs/1611.09830.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *CoRR* abs/1502.05698. http://arxiv.org/abs/1502.05698.

J. Wieting, M. Bansal, K. Gimpel, K. Livescu, and D. Roth. 2015. From paraphrase database to compositional paraphrase model and back. *TACL* 3:345–358.

Terry Winograd. 1980. Extended inference modes in reasoning by computer systems. *Artificial Intelligence* 13:5–26.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. pages 2013–2018. http://aclweb.org/anthology/D/D15/D15-1237.pdf.