

$\mathcal{R}^4\mathcal{C}$: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason

Naoya Inoue^{1,2} Pontus Stenetorp^{2,3} Kentaro Inui^{1,2}
¹Tohoku University ²RIKEN
³University College London
{naoya-i, inui}@ecei.tohoku.ac.jp
p.stenetorp@cs.ucl.ac.uk

Abstract

Recent studies have revealed that reading comprehension (RC) systems learn to exploit annotation artifacts and other biases in current datasets. This prevents the community from reliably measuring the progress of RC systems. To address this issue, we introduce $\mathcal{R}^4\mathcal{C}$, a new task for evaluating RC systems' internal reasoning. $\mathcal{R}^4\mathcal{C}$ requires giving not only answers but also derivations: explanations that justify predicted answers. We present a reliable, crowdsourced framework for scalably annotating RC datasets with derivations. We create and publicly release the $\mathcal{R}^4\mathcal{C}$ dataset, the first, quality-assured dataset consisting of 4.6k questions, each of which is annotated with 3 reference derivations (i.e. 13.8k derivations). Experiments show that our automatic evaluation metrics using multiple reference derivations are reliable, and that $\mathcal{R}^4\mathcal{C}$ assesses different skills from an existing benchmark.

1 Introduction

Reading comprehension (RC) has become a key benchmark for natural language understanding (NLU) systems, and a large number of datasets are now available (Welbl et al., 2018; Kočiský et al., 2018; Yang et al., 2018, i.a.). However, it has been established that these datasets suffer from annotation artifacts and other biases, which may allow systems to “cheat”: Instead of learning to read and comprehend texts in their entirety, systems learn to exploit these biases and find answers via simple heuristics, such as looking for an entity with a particular semantic type (Sugawara et al., 2018; Mudrakarta et al., 2018) (e.g. given a question starting with *Who*, a system finds a person entity found in a document).

To address this issue, the community has introduced increasingly more difficult Question Answering (QA) problems, for example, so that answer-

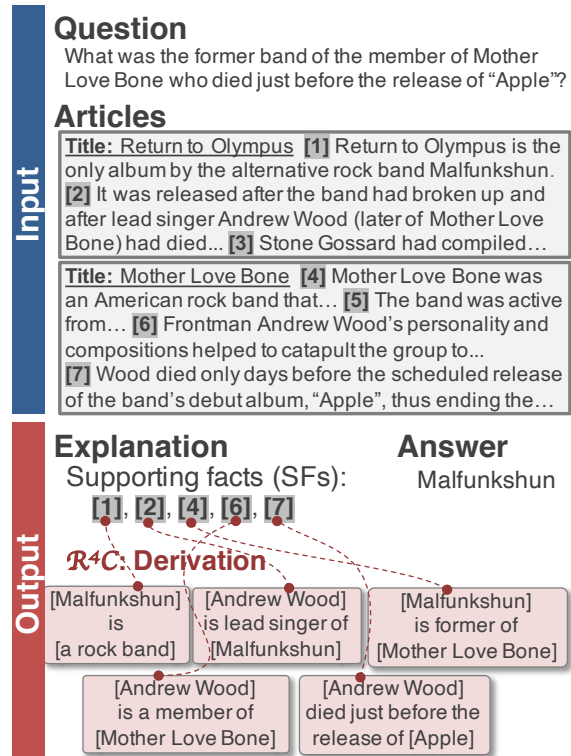


Figure 1: $\mathcal{R}^4\mathcal{C}$, a new RC task extending upon the standard RC setting, requiring systems to provide not only an answer, but also a derivation. The example is taken from HotpotQA (Yang et al., 2018), where sentences [1-2, 4, 6-7] are supporting facts, and [3,5] are not.

related information is scattered across several articles (Welbl et al., 2018; Yang et al., 2018) (i.e. *multi-hop QA*). However, recent studies show that such multi-hop QA also has weaknesses (Chen and Durrett, 2019; Min et al., 2019; Jiang et al., 2019), e.g. combining multiple sources of information is not always necessary to find answers. Another direction, which we follow, includes evaluating a systems’ reasoning (Jansen, 2018; Yang et al., 2018; Thorne and Vlachos, 2018; Camburu et al., 2018; Fan et al., 2019; Rajani et al., 2019). In

the context of RC, Yang et al. (2018) propose HotpotQA, which requires systems not only to give an answer but also to identify *supporting facts* (SFs), sentences containing information that supports the answer. SFs are defined as *sentences* containing information that supports the answer (see “Supporting facts” in Fig. 1 for an example).

As shown in SFs [1], [2], and [7], however, only a subset of SFs may contribute to the necessary reasoning. For example, [1] states two facts: (a) *Return to Olympus is an album by Malfunkshun*; and (b) *Malfunkshun is a rock band*. Among these, only (b) is related to the necessary reasoning. Thus, achieving a high accuracy in the SF detection task does not fully prove a RC systems’s reasoning ability.

This paper proposes $\mathcal{R}^4\mathcal{C}$, a new task of RC that requires systems to provide an answer *and derivation*¹: a minimal explanation that justifies predicted answers in a semi-structured natural language form (see “Derivation” in Fig. 1 for an example). Our main contributions can be summarized as follows:

- We propose $\mathcal{R}^4\mathcal{C}$, which enables us to quantitatively evaluate a systems’ internal reasoning in a finer-grained manner than the SF detection task. We show that $\mathcal{R}^4\mathcal{C}$ assesses different skills from the SF detection task.
- We create and publicly release the first dataset of $\mathcal{R}^4\mathcal{C}$ consisting of 4,588 questions, each of which is annotated with 3 high-quality derivations (i.e. 13,764 derivations), available at <https://naoya-i.github.io/r4c/>.
- We present and publicly release a reliable, crowdsourced framework for scalably annotating existing RC datasets with derivations in order to facilitate large-scale dataset construction of derivations in the RC community.

2 Task description

2.1 Task definition

We build $\mathcal{R}^4\mathcal{C}$ on top of the standard RC task. Given a question q and articles R , the task is (i) to find the answer a from R and (ii) to generate a derivation D that justifies why a is believed to be the answer to q .

There are several design choices for derivations, including whether derivations should be structured, whether the vocabulary should be closed, etc. This

¹ $\mathcal{R}^4\mathcal{C}$ is short for “Right for the Right Reasons RC.”

leads to a trade-off between the expressivity of reasoning and the interpretability of an evaluation metric. To maintain a reasonable trade-off, we choose to represent derivations in a semi-structured natural language form. Specifically, a derivation is defined as a set of *derivation steps*. Each derivation step $d_i \in D$ is defined as a relational fact, i.e. $d_i \equiv \langle d_i^h, d_i^r, d_i^t \rangle$, where d_i^h, d_i^t are entities (noun phrases), and d_i^r is a verb phrase representing a relationship between d_i^h and d_i^t (see Fig. 1 for an example), similar to the Open Information Extraction paradigm (Etzioni et al., 2008). d_i^h, d_i^r, d_i^t may be a phrase not contained in R (e.g. *is lead singer of* in Fig. 1).

2.2 Evaluation metrics

While the output derivations are semi-structured, the linguistic diversity of entities and relations still prevents automatic evaluation. One typical solution is crowdsourced judgement, but it is costly both in terms of time and budget. We thus resort to a reference-based similarity metric.

Specifically, for output derivation D , we assume n sets of golden derivations G_1, G_2, \dots, G_n . For evaluation, we would like to assess how well derivation steps in D can be aligned with those in G_i in the best case. For each golden derivation G_i , we calculate $c(D; G_i)$, an alignment score of D with respect to G_i or a soft version of the number of correct derivation steps in D (i.e. $0 \leq c(D; G_i) \leq \min(|D|, |G_i|)$). We then find a golden derivation G^* that gives the highest $c(D; G^*)$ and define the precision, recall and f_1 as follows:

$$\begin{aligned} \text{pr}(D) &= \frac{c(D; G^*)}{|D|}, \text{rc}(D) = \frac{c(D; G^*)}{|G^*|} \\ f_1(D) &= \frac{2 \cdot \text{pr}(D; G^*) \cdot \text{rc}(D; G^*)}{\text{pr}(D; G^*) + \text{rc}(D; G^*)} \end{aligned}$$

An official evaluation script is available at <https://naoya-i.github.io/r4c/>.

Alignment score To calculate $c(D; G_i)$, we would like to find the best alignment between derivation steps in D and those in G_i . See Fig. 2 for an example, where two possible alignments A_1, A_2 are shown. As derivation steps in D agree with those in G_i with A_2 more than those with A_1 , we would like to consider A_2 when evaluating. We first define $c(D; G_i, A_j)$, the correctness of D given a specific alignment A_j , and then pick the

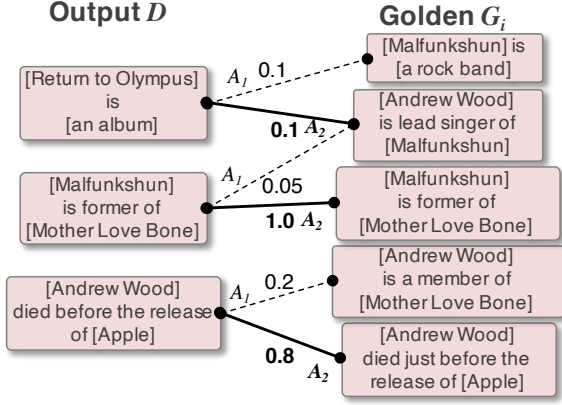


Figure 2: Two possible alignments A_1 and A_2 between D and G_i with their alignment scores $a(\cdot, \cdot)$. The precision and recall of D is $(0.1+1.0+0.8)/3 = 0.633$ and $(0.1+1.0+0.8)/5=0.380$, respectively.

best alignment as follows:

$$c(D; G_i, A_j) = \sum_{(d_j, g_j) \in A_j} a(d_j, g_j)$$

$$c(D; G_i) = \max_{A_j \in \mathcal{A}(D, G_i)} c(D; G_i, A_j),$$

where $a(d_j, g_j)$ is a similarity $[0, 1]$ between two derivation steps d_j, g_j , and $\mathcal{A}(D, G_i)$ denotes all possible one-to-one alignments between derivation steps in D and those in G_i .

For $a(d_j, g_j)$, we consider three variants, depending on the granularity of evaluation. We first introduce two fine-grained scorers, taking only entities or relations into account (henceforth, *entity scorer* and *relation scorer*):

$$a^{\text{ent}}(d_j, g_j) = \frac{1}{2}(s(d_j^h, g_j^h) + s(d_j^t, g_j^t))$$

$$a^{\text{rel}}(d_j, g_j) = s(d_j^r, g_j^r),$$

where $s(\cdot, \cdot)$ denotes an arbitrary similarity measure $[0, 1]$ between two phrases. In this study, we employ a normalized Levenshtein distance. Finally, as a rough indication of overall performance, we also provide a *full scorer* as follows:

$$a^{\text{full}}(d_j, g_j) = \frac{1}{3}(s(d_j^h, g_j^h) + s(d_j^r, g_j^r) + s(d_j^t, g_j^t))$$

3 Data collection

The main purpose of $\mathcal{R}^4\mathcal{C}$ is to *benchmark* an RC systems' internal reasoning. We thus assume a semi-supervised learning scenario where RC systems are trained to answer a given question on a

Question:

The 1988 American comedy film, The Great Outdoors, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

Articles:

Select sentences relevant to your reasoning. Please use highlighted sentences as much as possible (non-highlighted one can be used if necessary).

Article 1: The Great Outdoors (film)

The Great Outdoors is a 1988 American comedy film directed by Howard Deutch, and written and produced by John Hughes.

It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

Question: The 1988 American comedy film, The Great Outdoors, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

Reasoning:

Write your reasoning steps in a simple form *subject-verb-object*. You may rely on words from "suggestions" generated automatically, but *editing may be needed*.

It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

From the above sentence, what information did you infer?

Who/what:

The Great Outdoors (film)

Suggestions: The Great Outdoors (film)

Dan Aykroyd | John Candy

Stephanie Faracy | Annette Bening

her film debut

Did what:

star

Figure 3: Crowdsourcing interface for derivation annotation. Workers click on sentences and create derivation steps in the form of entity-relation triplets.

large-scale RC dataset and *then fine-tuned* to give a correct reasoning on a smaller reasoning-annotated datasets. To acquire a dataset of derivations, we use crowdsourcing (CS).

3.1 Crowdsourcing interface

We design our interface to annotate existing RC datasets with derivations, as a wide variety of high quality RC datasets are already available (Welbl et al., 2018; Yang et al., 2018, etc.). We assume that RC datasets provide (i) a question, (ii) the answer, and (iii) *supporting articles*, articles that support the answer (optionally with SFs).

Initially, in order to encourage crowdworkers (henceforth, *workers*) to read the supporting articles carefully, we ask workers to answer to the question based on the supporting articles (see Appendix A). To reduce the workload, four candidate answers are provided.² We also allow for *neither* as RC datasets may contain erroneous instances.

Second, we ask workers to write derivations for their answer (see Fig. 3). They click on a sentence (either a SF or non-SF) in a supporting article (left) and then input their derivation in the form of triplets (right). They are asked to input entities and relations through free-form textboxes. To reduce the workload and encourage annotation consistency,

²The correct answer and three incorrect answers randomly chosen from the titles of the supporting articles.

Split	# QA	# derivations			
		2 st.	3 st.	≥ 4 st.	Total
train	2,379	4,944	1,553	640	7,137
dev	2,209	4,424	1,599	604	6,627
total	4,588	9,368	3,152	1,244	13,764

Table 1: Statistics of $\mathcal{R}^4\mathcal{C}$ corpus. “st.” denotes the number of derivation steps. Each instance is annotated with 3 golden derivations.

we also provide suggestions. These suggestions include predefined prepositions, noun phrases, and verb phrases automatically extracted from supporting articles.³ We also highlight SFs if they are available for the given RC dataset.

3.2 Workflow

To discourage noisy annotations, we first deploy a qualification test. We provide the same task described in §3.1 in the test and manually identify competent workers in our task. The final annotation is carried out solely by these qualified workers.

We deploy the task on Amazon Mechanical Turk (AMT).⁴ We allow workers with $\geq 5,000$ Human Intelligence Tasks experience and an approval rate of $\geq 95.0\%$ to take the qualification test. For the test, we pay €15 as a reward per instance. For the final annotation task, we assign 3 workers per instance and pay €30 to each worker.

3.3 Dataset

There are a large number of choices of RC datasets that meet the criteria described in §3.1 including SQuAD (Rajpurkar et al., 2016) and WikiHop (Welbl et al., 2018). Our study uses HotpotQA (Yang et al., 2018), one of the most actively used multi-hop QA datasets.⁵ The multi-hop QA setting ensures that derivation steps are spread across documents, thereby posing an interesting unsolved research problem.

For annotation, we sampled 3,000 instances from 90,564 training instances and 3,000 instances from 7,405 development instances. For the qualification test and interface development, we sampled another 300 instances from the training set. We used the annotations of SFs provided by HotpotQA. We assume that the training set is used for *fine-tuning* RC systems’ internal reasoning, and the development set is used for evaluation.

³Spacy: <https://spacy.io/>

⁴<https://requester.mturk.com/>

⁵<https://hotpotqa.github.io/>

3.4 Statistics

In the qualification test, we identified 45 competent workers (out of 256 workers). To avoid noisy annotations, we filter out submissions (i) with a wrong answer and (ii) with a *neither* answer. After the filtering, we retain only instances with exactly three derivations annotated. Finally, we obtained 7,137 derivations for 2,379 instances in the training set and 7,623 derivations for 2,541 instances in the dev set. See Appendix B for annotation examples.

4 Evaluation

4.1 Methodology

To check whether annotated derivations help humans recover answers, we setup another CS task on AMT (*answerability judgement*). Given a HotpotQA question and the annotated derivation, 3 workers are asked whether or not they can answer the question *solely based on* the derivation at three levels. We evaluate all 7,623 derivations from the dev set. For reliability, we targeted only qualified workers and pay €15 as a reward per instance.

To see if each derivation step can actually be derived from its source SF, we asked two expert annotators (non co-authors) to check 50 derivation steps from the dev set (*derivability judgement*).

4.2 Results

For the answerability judgement, we obtained Krippendorff’s α of 0.263 (a fair agreement). With majority voting, we obtained the following results: YES: 95.2%, LIKELY: 2.2%, and NO: 1.3% (split: 1.3%).⁶ For the derivability judgement, 96.0% of the sampled derivation steps (48/50) are judged as derivable from their corresponding SFs by both expert annotators. Despite the complexity of the annotation task, the results indicate that the proposed annotation pipeline can capture competent workers and produce high-quality derivation annotations. For the final dev set, we retain only instances with YES answerability judgement.

The final $\mathcal{R}^4\mathcal{C}$ dataset includes 4,588 questions from HotpotQA (see Table 1), each of which is annotated with 3 reference derivations (i.e. 13,764 derivations). This is the first dataset of RC annotated with semi-structured, multiple reference derivations. The most closest work to our dataset is the WorldTree corpus (Jansen et al., 2018), the largest QA dataset annotated with explanations,

⁶We also evaluated 1,000 training instances: 96.0% with YES judgement with Krippendorff’s α of 0.173.

# rf	Entity P/R/F	Relation P/R/F	Full P/R/F
1	73.3/75.1/73.4	56.9/55.6/55.5	70.1/69.5/69.0
2	79.4/77.6/77.6	66.7/65.4/65.3	74.7/73.2/73.2
3	83.4/81.1/81.4	72.3/69.4/70.0	77.7/75.1/75.6

Table 2: Performance of oracle annotators on $\mathcal{R}^4\mathcal{C}$ as a function of the number of reference derivations.

which contains 1,680 questions. Jansen et al. (2018) use experts for annotation, and the annotated explanations are grounded on a predefined, structured knowledge base. In contrast, our work proposes a non-expert-based annotation framework and grounds explanations using unstructured texts.

5 Analysis

Effect of multiple references Do crowdsourced multiple golden derivations help us to evaluate output derivations more accurately? To verify this, we evaluated oracle derivations using one, two, or all three references. The derivations were written by qualified workers for 100 dev instances.

Table 2 shows that having more references increases the performance, which indicates that references provided by different workers are indeed diverse enough to capture oracle derivations. The peak performance with # rf= 3 establishes the upper bound performance on this dataset.

The larger improvement of the relation-level performance (+14.5) compared to that of the entity-level performance (+8.0) also suggests that relations are linguistically more diverse than entities, as we expected (e.g. *is in*, *is a town in*, and *is located in* are annotated for a locational relation).

Baseline models To analyze the nature of $\mathcal{R}^4\mathcal{C}$, we evaluate the following heuristic models. IE: extracting all entity relations from SFs.⁷ CORE: extracting the core information of SFs. Based on the dependency structure of SFs (with article title t), it extracts a root verb v and the right, first child c_r of v , and outputs $\langle t, v, c_r \rangle$ as a derivation step.

Table 3 shows a large performance gap to the human upper bound, indicating that $\mathcal{R}^4\mathcal{C}$ is different to the HotpotQA’s SF detection task—it does not simply require systems to exhaustively extract information nor to extract core information from SFs. The errors from these baseline models include generating entity relations irrelevant to reasoning (e.g. *Return to Olympus is an album* in Fig. 2) or missing implicit entity relations (e.g. *Andrew Wood is*

⁷We use Stanford OpenIE (Angeli et al., 2015).

Model	Entity P/R/F	Relation P/R/F	Full P/R/F
IE	11.3/53.4/16.6	13.7/62.8/19.9	11.4/52.3/16.5
CORE	66.4/60.1/62.1	51.0/46.0/47.5	59.4/53.6/55.4

Table 3: Performance of baseline models on $\mathcal{R}^4\mathcal{C}$.

a member of Mother Love Bone in Fig. 1). $\mathcal{R}^4\mathcal{C}$ introduces a new research problem for developing RC systems that can explain their answers.

6 Conclusions

Towards evaluating RC systems’ internal reasoning, we have proposed $\mathcal{R}^4\mathcal{C}$ that requires systems not only to output answers but also to give their derivations. For scalability, we have carefully developed a crowdsourced framework for annotating existing RC datasets with derivations. Our experiments have demonstrated that our framework produces high-quality derivations, and that automatic evaluation metrics using multiple reference derivations can reliably capture oracle derivations. The experiments using two simple baseline models highlight the nature of $\mathcal{R}^4\mathcal{C}$, namely that the derivation generation task is not simply the SF detection task. We make the dataset, automatic evaluation script, and baseline systems publicly available at <https://naoya-i.github.io/r4c/>.

One immediate future work is to evaluate state-of-the-art RC systems’ internal reasoning on our dataset. For modeling, we plan to explore recent advances in conditional language models for jointly modeling QA with generating their derivations.

Acknowledgements

This work was supported by the UCL-Tohoku University Strategic Partnership Fund, JSPS KAKENHI Grant Number 19K20332, JST CREST Grant Number JPMJCR1513 (including the AIP challenge program), the European Union’s Horizon 2020 research and innovation programme under grant agreement No 875160, and the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Research Council (EPSRC) under grant EP/R018693/1 (a part of the collaboration between US DOD, UK MOD, and UK EPSRC under the Multidisciplinary University Research Initiative (MURI)). The authors would like to thank Paul Reisert, Keshav Singh, other members of the Tohoku NLP Lab, and the anonymous reviewers for their insightful feedback.

References

- Gabor Angeli, Melvin Johnson Premkumar, and Christopher D. Manning. 2015. [Leveraging linguistic structure for open domain information extraction](#). In *Proc. of ACL-IJCNLP*, pages 344–354.
- Oana-maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. [e-SNLI : Natural Language Inference with Natural Language Explanations](#). In *Proc. of NIPS*, pages 1–13.
- Jifan Chen and Greg Durrett. 2019. [Understanding Dataset Design Choices for Multi-hop Reasoning](#). In *Proc. of NAACL-HLT*, pages 4026–4032.
- Oren Etzioni, Michele Banko, Stephen Soderland, and Daniel S. Weld. 2008. [Open information extraction from the web](#). *Communications of the ACM*, 51(12):68–74.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: Long Form Question Answering](#). In *Proc. of ACL*, pages 3558–3567.
- Peter A. Jansen. 2018. [Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering?](#) In *Proc. of TextGraphs-12*, pages 12–17.
- Peter A. Jansen, Elizabeth Wainwright, Steven Marmorstein, and Clayton T. Morrison. 2018. [WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-Hop Inference](#). In *Proc. of LREC*, pages 2732–2740.
- Yichen Jiang, Nitish Joshi, Yen-Chun Chen, and Mohit Bansal. 2019. [Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension](#). In *Proc. of ACL*, pages 2714–2725.
- Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. [The NarrativeQA Reading Comprehension Challenge](#). *Trans. of ACL*, 6:317–328.
- Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. [Compositional Questions Do Not Necessitate Multi-hop Reasoning](#). In *Proc. of ACL*, pages 4249–4257.
- Pramod K. Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. 2018. [Did the Model Understand the Question?](#) In *Proc. of ACL*, pages 1896–1906.
- Fatema Nanzneen Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. [Explain Yourself ! Leveraging Language Models for Commonsense Reasoning](#). In *Proc. of ACL*, pages 4932–4942.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ Questions for Machine Comprehension of Text](#). In *Proc. of EMNLP*, pages 2383–2392.
- Saku Sugawara, Kentaro Inui, Satoshi Sekine, and Akiko Aizawa. 2018. [What Makes Reading Comprehension Questions Easier?](#) In *Proc. of EMNLP*, pages 4208–4219.
- James Thorne and Andreas Vlachos. 2018. [Automated Fact Checking: Task formulations, methods and future directions](#). In *Proc. of COLING*, pages 3346–3359.
- Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. [Constructing Datasets for Multi-hop Reading Comprehension Across Documents](#). *Trans. of ACL*, 6:287–302.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering](#). In *Proc. of EMNLP*, pages 2369–2380.

A Crowdsourcing interface

Fig. 4 shows the instruction of our annotation task to crowdworkers. Fig. 5 shows the interface of the question-answering task.

Overview

Artificial Intelligence (AI) has become fast and accurate. However, it still falls short when it comes to human-level reasoning. In this task, you are helping us educate an AI how to reason. Specifically, we will ask you to verify whether you can answer a question from articles and describe the steps you used to find the answer.

Instructions

[View instructions \(for first time workers\)](#)

Overview

1. Read a given question and related articles.
2. Answer to the question solely based on the information from each article.
3. Describe your reasoning on how to reach the answer. Each reasoning step needs to be in a simple subject-verb-object form (see example below). Your reasoning must include sentences containing your answer.

You can consult each section for further instructions.

Example

Question

When is Barack Obama's wife's birthday?

Articles

- **Article 1:** Barack Obama is a former US president. His wife is Michelle Obama.
- **Article 2:** Michelle Obama is a former US first lady (born January 17, 1964).

Answer

January 17, 1964

Reasoning

See both a good response and bad response below:

Reasoning	Good/Bad
<ul style="list-style-type: none">• Article 1, second sentence => [Barack Obama] 's wife is [Michelle Obama].• Article 2, first sentence => [Michelle Obama] is born on [January 17, 1964].	GOOD Note that the second reasoning step includes [January 17, 1964], the answer.
<ul style="list-style-type: none">• Article 2, first sentence => [Michelle Obama] is born on [January 17, 1964].	BAD The reasoning does not state the relation between [Michelle Obama] and [Barack Obama]. It is obvious for us, but it is not for AIs. Do not forget write down such things in this task!
<ul style="list-style-type: none">• Article 1, second sentence => [Barack Obama] 's wife is [Michelle Obama].• Article 2, first sentence => [Michelle Obama] is born [on January 17, 1964].	BAD The second reasoning step includes [on January 17, 1964], which is not an object.

Figure 4: Task instruction.

Task

1. Read

Read the following question and related articles carefully. Important sentences are highlighted.

Question

The 1988 American comedy film, The Great Outdoors, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

Articles

Article 1: The Great Outdoors (film)

The Great Outdoors is a 1988 American comedy film directed by Howard Deutch, and written and produced by John Hughes.

It stars Dan Aykroyd, John Candy, Stephanie Faracy and Annette Bening in her film debut.

Article 2: Annette Bening

Annette Carol Bening (born May 29, 1958) is an American actress.

She began her career on stage with the Colorado Shakespeare Festival company in 1980, and played Lady Macbeth in 1984 at the American Conservatory Theatre.

She was nominated for the 1987 Tony Award for Best Featured Actress in a Play for her Broadway debut in "Coastal Disturbances".

She is a four-time Academy Award nominee; for "The Grifters" (1990), "American Beauty" (1999), "Being Julia" (2004) and "The Kids Are All Right" (2010).

In 2006, she received a star on the Hollywood Walk of Fame.

2. Answer

Based upon the related articles, answer to the question.

NOTE:

- If you found multiple possible answers, please choose one of them and write your reasoning for it below.
- If the question contains a typo(s): (i) if you can pretend that the typos are not there, answer the question. (ii) if you cannot (e.g. the question does not make sense), select "NONE OF THE ABOVE".

The 1988 American comedy film, The Great Outdoors, starred a four-time Academy Award nominee, who received a star on the Hollywood Walk of Fame in what year?

Choose your answer:

- ✓ Farrah Fawcett
- 2006
- Shane Stanley
- Beau Bridges
- NONE OF THE ABOVE

Figure 5: Task interface for the first question answering phase. The reasoning annotation interface shown in Fig. 3 follows after this interface.

B Example annotations

Table 4 shows examples of crowdsourced annotations.

Question	Were Scott Derrickson and Ed Wood of the same nationality?
Supporting Art. 1	[1] Scott Derrickson (born July 16, 1966) is an American director, screenwriter and producer.[2] He lives in Los Angeles, California.[3] He is best known for directing horror films such as "Sinister", "The Exorcism of Emily Rose", and "Deliver Us From Evil", as well as the 2016 Marvel Cinematic Universe installment, "Doctor Strange."
Supporting Art. 2	[1] Edward Davis Wood Jr. (October 10, 1924 December 10, 1978) was an American filmmaker, actor, writer, producer, and director.
Derivation step 1	[1, 1] [Scott Derrickson] [is] [an American director]
Derivation step 2	[1, 1] [Ed Wood] [was] [an American filmmaker]
Question	The director of the romantic comedy "Big Stone Gap" is based in what New York city?
Supporting Art. 1	[1] Big Stone Gap is a 2014 American drama romantic comedy film written and directed by Adriana Trigiani and produced by Donna Gigliotti for Altar Identity Studios, a subsidiary of Media Society.[2] Based on Trigiani's 2000 best-selling novel of the same name, the story is set in the actual Virginia town of Big Stone Gap circa 1970s.[3] The film had its world premiere at the Virginia Film Festival on November 6, 2014.
Supporting Art. 2	[1] Adriana Trigiani is an Italian American best-selling author of sixteen books, television writer, film director, and entrepreneur based in Greenwich Village, New York City.[2] Trigiani has published a novel a year since 2000.
Derivation step 1	[1, 1] [Big Stone Gap] [is directed by] [Adriana Trigiani]
Derivation step 2	[2, 1] [Adriana Trigiani] [is from] [Greenwich Village, New York City.]
Question	The arena where the Lewiston Maineiacs played their home games can seat how many people?
Supporting Art. 1	[1] The Lewiston Maineiacs were a junior ice hockey team of the Quebec Major Junior Hockey League based in Lewiston, Maine.[2] The team played its home games at the Androscoggin Bank Colisee.[3] They were the second QMJHL team in the United States, and the only one to play a full season.[4] They won the President's Cup in 2007.
Supporting Art. 2	[1] The Androscoggin Bank Colisee (formerly Central Maine Civic Center and Lewiston Colisee) is a 4,000 capacity (3,677 seated) multi-purpose arena, in Lewiston, Maine, that opened in 1958.[2] In 1965 it was the location of the World Heavyweight Title fight during which one of the most famous sports photographs of the century was taken of Muhammed Ali standing over Sonny Liston.
Derivation step 1	[1,2] [Lewiston Maineiacs] [play in the] [Androscoggin Bank Colisee]
Derivation step 2	[2,1] [Androscoggin Bank Colisee] [is an] [arena]
Derivation step 3	[2,1] [Androscoggin Bank Colisee] [has a seating capacity of] [3,677 seated]

Table 4: Example of annotation results of derivations. Each derivation step is in the following format: [article ID, SF] [Head entity] [Relation] [Tail entity].