ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension

Sheng Zhang[†][*], **Xiaodong Liu**[‡], **Jingjing Liu**[‡], **Jianfeng Gao**[‡],
Kevin Duh[†] and **Benjamin Van Durme**[†]
[†]Johns Hopkins University
[‡]Microsoft Research

Abstract

We present a large-scale dataset, ReCoRD, for machine reading comprehension requiring commonsense reasoning. Experiments on this dataset demonstrate that the performance of state-of-the-art MRC systems fall far behind human performance. ReCoRD represents a challenge for future research to bridge the gap between human and machine commonsense reading comprehension. ReCoRD is available at `http://nlp.jhu.edu/record`.

1 Introduction

Machine reading comprehension (MRC) is a central task in natural language understanding, with techniques lately driven by a surge of large-scale datasets (Hermann et al., 2015; Hill et al., 2015; Rajpurkar et al., 2016; Trischler et al., 2017; Nguyen et al., 2016), usually formalized as a task of answering questions given a passage. An increasing number of analyses (Jia and Liang, 2017; Rajpurkar et al., 2018; Kaushik and Lipton, 2018) have revealed that a large portion of questions in these datasets can be answered by simply matching the patterns between the question and the answer sentence in the passage. While systems may match or even outperform humans on these datasets, our intuition suggests that there are at least some instances in human reading comprehension that require more than what existing challenge tasks are emphasizing. One primary type of questions these datasets lack are the ones that require reasoning over common sense or understanding across multiple sentences in the passage (Rajpurkar et al., 2016; Trischler et al., 2017).

Passage
(CNN) -- A lawsuit has been filed claiming that the iconic Led Zeppelin song "Stairway to Heaven" was far from original. The suit, filed on May 31 in the United States District Court Eastern District of Pennsylvania, was brought by the estate of the late musician Randy California against the surviving members of Led Zeppelin and their record label. The copyright infringement case alleges that the Zeppelin song was taken from the single "Taurus" by the 1960s band Spirit, for whom California served as lead guitarist. "Late in 1968, a then new band named Led Zeppelin began touring in the United States, opening for Spirit," the suit states. "It was during this time that Jimmy Page, Led Zeppelin's guitarist, grew familiar with 'Taurus' and the rest of Spirit's catalog. Page stated in interviews that he found Spirit to be 'very good' and that the band's performances struck him 'on an emotional level.' "

- Suit claims similarities between two songs
- Randy California was guitarist for the group Spirit
- Jimmy Page has called the accusation "ridiculous"

(Cloze-style) Query
According to claims in the suit, "Parts of 'Stairway to Heaven,' instantly recognizable to the music fans across the world, sound almost identical to significant portions of '**X**.'"

Reference Answers
Taurus

Figure 1: An example from ReCoRD. The **passage** is a snippet from a news article followed by some bullet points which summarize the news event. Named entities highlighted in the passage are possible answers to the query. The **query** is a statement that is factually supported by the passage. **X** in the statement indicates a missing named entity. The goal is to find the correct entity in the passage that best fits **X**.

To overcome this limitation, we introduce a large-scale dataset for reading comprehension, ReCoRD ([ˈrɛkərd]), which consists of over 120,000 examples, most of which require deep commonsense reasoning. ReCoRD is an acronym for the **Re**ading **Co**mprehension with **Co**mmonsense **R**easoning **D**ataset.

Figure 1 shows a ReCoRD example: the passage describes a lawsuit claiming that the band "*Led Zeppelin*" had plagiarized the song "*Taurus*"

[*] Work done when Sheng Zhang was visiting Microsoft.

to their most iconic song, "*Stairway to Heaven*". The cloze-style query asks what does "*Stairway to Heaven*" sound similar to. To find the correct answer, we need to understand from the passage that "*a copyright infringement case alleges that 'Stairway to Heaven' was taken from 'Taurus'*", and from the bullet point that "*these two songs are claimed similar*". Then based on the commonsense knowledge that "*if two songs are claimed similar, it is likely that (parts of) these songs sound almost identical*", we can reasonably infer that the answer is "*Taurus*".

Differing from most of the existing MRC datasets, all queries and passages in ReCoRD are automatically mined from news articles, which maximally reduces the human elicitation bias (Gordon and Van Durme, 2013; Misra et al., 2016; Zhang et al., 2017), and the data collection method we propose is cost-efficient. Further analysis shows that a large portion of ReCoRD requires commonsense reasoning.

Experiments on ReCoRD demonstrate that human readers are able to achieve a high performance at 91.69 F1, whereas the state-of-the-art MRC models fall far behind at 46.65 F1. Thus, ReCoRD presents a real challenge for future research to bridge the gap between human and machine commonsense reading comprehension.

2 Task Motivation

> *A program has common sense if it automatically deduces for itself a sufficiently wide class of immediate consequences of anything it is told and what it already knows.* – McCarthy (1959)

Commonsense Reasoning in MRC As illustrated by the example in Figure 1, the commonsense knowledge "*if two songs are claimed similar, it is likely that (parts of) these songs sound almost identica*" is not explicitly described in the passage, but is necessary to acquire in order to generate the answer. Human is able to infer the answer because the commonsense knowledge is commonly known by nearly all people. Our goal is to evaluate whether a machine is able to learn such knowledge. However, since commonsense knowledge is massive and mostly implicit, defining an explicit free-form evaluation is challenging (Levesque et al., 2011). Motivated by McCarthy (1959), we instead evaluate a machine's ability of commonsense reasoning – a reasoning process requiring commonsense knowledge; that is, if a machine has common sense, it can deduce for itself the likely consequences or details of anything it is told and what it already knows rather than the unlikely ones. To formalize it in MRC, given a passage $\mathbf{p}$ (i.e., "*anything it is told*" and "*what it already knows*"), and a set of consequences or details $\mathcal{C}$ which are factually supported by the passage $\mathbf{p}$ with different likelihood, if a machine $\mathbf{M}$ has common sense, it can choose the most likely consequence or detail $\mathbf{c}^*$ from $\mathcal{C}$, i.e.,

$$\mathbf{c}^* = \arg\max_{\mathbf{c} \in \mathcal{C}} P(\mathbf{c} \mid \mathbf{p}, \mathbf{M}). \tag{1}$$

Task Definition With the above discussion, we propose a specific task to evaluate a machine's ability of commonsense reasoning in MRC: as shown in Figure 1, given a passage $\mathbf{p}$ describing an event, a set of text spans $\mathbf{E}$ marked in $\mathbf{p}$, and a cloze-style query $Q(\mathbf{X})$ with a missing text span indicated by $\mathbf{X}$, a machine $\mathbf{M}$ is expected to act like human, reading the passage $\mathbf{p}$ and then using its hidden commonsense knowledge to choose a text span $\mathbf{e} \in \mathbf{E}$ that best fits $\mathbf{X}$, i.e.,

$$\mathbf{e}^* = \arg\max_{\mathbf{e} \in \mathbf{E}} P(Q(\mathbf{e}) \mid \mathbf{p}, \mathbf{M}). \tag{2}$$

Once the cloze-style query $Q(\mathbf{X})$ is filled in by a text span $\mathbf{e}$, the resulted statement $Q(\mathbf{e})$ becomes a consequence or detail $\mathbf{c}$ as described in Equation (1), which is factually supported by the passage with certain likelihood.

3 Data Collection

We describe the framework for automatically generating the dataset, ReCoRD, for our task defined in Equation (2), which consists of passages with text spans marked, cloze-style queries, and reference answers. We collect ReCoRD in four stages as shown in Figure 2: (1) curating CNN/Daily Mail news articles, (2) generating passage-query-answers triples based on the news articles, (3) filtering out the queries that can be easily answered by state-of-the-art MRC models, and (4) filtering out the queries ambiguous to human readers.

3.1 News Article Curation

We choose to create ReCoRD by exploiting news articles, because the structure of news makes it a good source for our task: normally, the first few paragraphs of a news article summarize the news

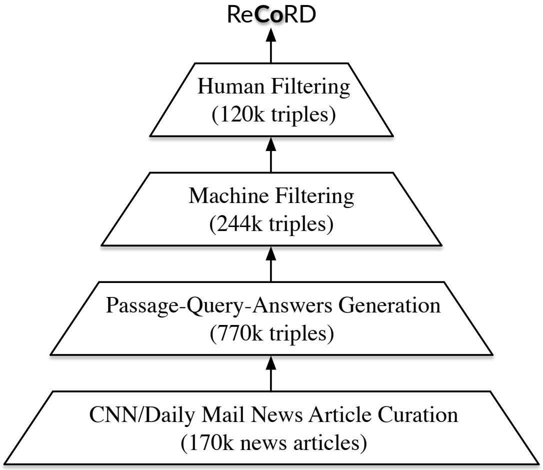


Figure 2: The overview of data collection stages.

event, which can be used to generate passages of the task; and the rest of the news article provides consequences or details of the news event, which can be used to generate queries of the task. In addition, news providers such as CNN and Daily Mail supplement their articles with a number of bullet points (Svore et al., 2007; Woodsend and Lapata, 2010; Hermann et al., 2015), which outline the highlights of the news and hence form a supplemental source for generating passages.

We first downloaded CNN and Daily Mail news articles using the script[1] provided by Hermann et al. (2015), and then sampled 148K articles from CNN and Daily Mail. In these articles, named entities and their coreference information have been annotated by a Google NLP pipeline, and will be used in the second stage of our data collection. Since these articles can be easily downloaded using the public script, we are concerned about potential cheating if using them as the source for generating the dev./test datasets. Therefore, we crawled additional 22K news articles from the CNN and Daily Mail websites. These crawled articles have no overlap with the articles used in Hermann et al. (2015). We then ran the state-of-the-art named entity recognition model (Peters et al., 2018) and the end-to-end coreference resolution model (Lee et al., 2017) provided by AllenNLP (Gardner et al., 2018) to annotate the crawled articles. Overall, we have collected 170K CNN/Daily Mail news articles with their named entities and coreference information annotated.

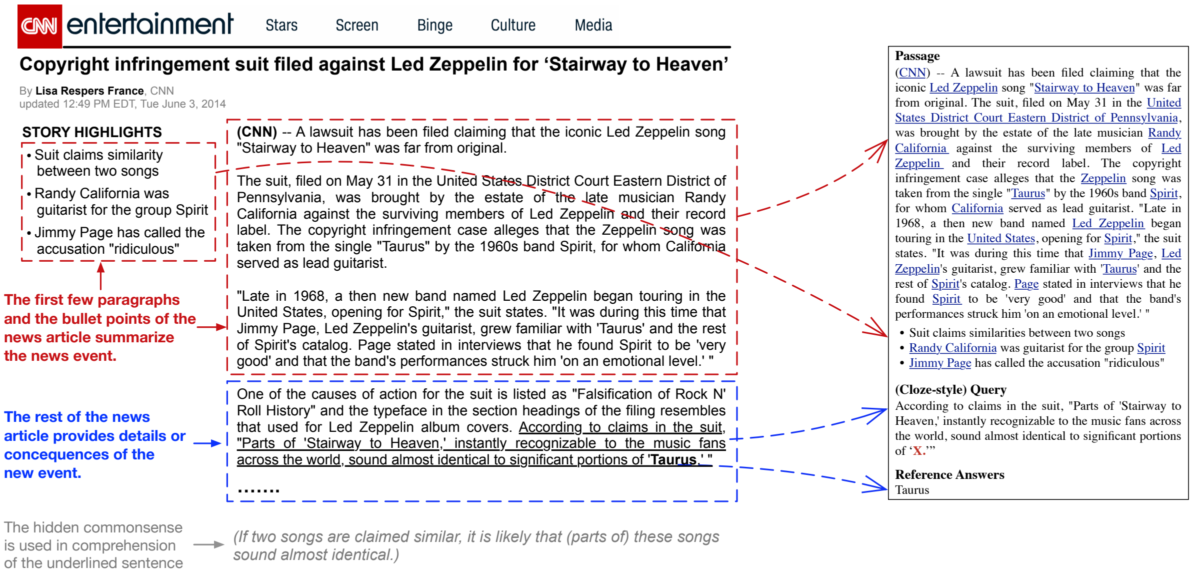
3.2 Passage-Query-Answers Generation

All passages, queries and answers in ReCoRD were automatically generated from the curated news articles. Figure 3 illustrates the generation process. (1) we split each news article into two parts as described in Section 3.1: the first few paragraphs which summarize the news event, and the rest of the news which provides the details or consequences of the news event. These two parts make a good source for generating passages and queries of our task respectively. (2) we enriched the first part of news article with the bullet points provided by the news editors. The first part of news article, together with the bullet points, is considered as a candidate passage. To ensure that the candidate passages are informative enough, we required the first part of news article to have at least 100 tokens and contain at least four different entities. (3) for each candidate passage, the second part of its corresponding news article was split into sentences by Stanford CoreNLP (Manning et al., 2014). Then we selected the sentences that satisfy the following conditions as potential details or consequences of the news event described by the passage:

- Sentences should have at least 10 tokens, as longer sentences contain more information and thus are more likely to be inferrable details or consequences.
- Sentences should not be questions, as we only consider details or consequences of a news event, not questions.
- Sentences should not have 3-gram overlap with the corresponding passage, so they are less likely to be paraphrase of sentences in the passage.
- Sentences should have at least one named entity, so that we can replace it with **X** to generate a cloze-style query.
- All named entities in sentences should have precedents in the passage according to coreference, so that the sentences are not too disconnected from the passage, and the correct entity can be found in the passage to fill in **X**.

Finally, we generated queries by replacing entities in the selected sentences with **X**. We only replaced one entity in the selected sentence each time, and generated one cloze-style query. Based on coreference, the precedents of the replaced en-

[1] https://github.com/deepmind/rc-data



CNN entertainment Stars Screen Binge Culture Media

Copyright infringement suit filed against Led Zeppelin for 'Stairway to Heaven'

By **Lisa Respers France**, CNN
updated 12:49 PM EDT, Tue June 3, 2014

STORY HIGHLIGHTS

- Suit claims similarity between two songs
- Randy California was guitarist for the group Spirit
- Jimmy Page has called the accusation "ridiculous"

(CNN) -- A lawsuit has been filed claiming that the iconic Led Zeppelin song "Stairway to Heaven" was far from original.

The suit, filed on May 31 in the United States District Court Eastern District of Pennsylvania, was brought by the estate of the late musician Randy California against the surviving members of Led Zeppelin and their record label. The copyright infringement case alleges that the Zeppelin song was taken from the single "Taurus" by the 1960s band Spirit, for whom California served as lead guitarist.

"Late in 1968, a then new band named Led Zeppelin began touring in the United States, opening for Spirit," the suit states. "It was during this time that Jimmy Page, Led Zeppelin's guitarist, grew familiar with 'Taurus' and the rest of Spirit's catalog. Page stated in interviews that he found Spirit to be 'very good' and that the band's performances struck him 'on an emotional level.' "

One of the causes of action for the suit is listed as "Falsification of Rock N' Roll History" and the typeface in the section headings of the filing resembles that used for Led Zeppelin album covers. According to claims in the suit, "Parts of 'Stairway to Heaven,' instantly recognizable to the music fans across the world, sound almost identical to significant portions of '**Taurus**.' "

....

The first few paragraphs and the bullet points of the news article summarize the news event.
The rest of the news article provides details or concequences of the new event.
The hidden commonsense is used in comprehension of the underlined sentence
(If two songs are claimed similar, it is likely that (parts of) these songs sound almost identical.)

Passage

(CNN) -- A lawsuit has been filed claiming that the iconic Led Zeppelin song "Stairway to Heaven" was far from original. The suit, filed on May 31 in the United States District Court Eastern District of Pennsylvania, was brought by the estate of the late musician Randy California against the surviving members of Led Zeppelin and their record label. The copyright infringement case alleges that the Zeppelin song was taken from the single "Taurus" by the 1960s band Spirit, for whom California served as lead guitarist. "Late in 1968, a then new band named Led Zeppelin began touring in the United States, opening for Spirit," the suit states. "It was during this time that Jimmy Page, Led Zeppelin's guitarist, grew familiar with 'Taurus' and the rest of Spirit's catalog. Page stated in interviews that he found Spirit to be 'very good' and that the band's performances struck him 'on an emotional level.' "

- Suit claims similarities between two songs
- Randy California was guitarist for the group Spirit
- Jimmy Page has called the accusation "ridiculous"

(Cloze-style) Query

According to claims in the suit, "Parts of 'Stairway to Heaven,' instantly recognizable to the music fans across the world, sound almost identical to significant portions of '**X.**'"

Reference Answers

Taurus

Figure 3: Passage-query-answers generation from a CNN news article.

tity in the passage became reference answers to the query. The passage-query-answers generation process matched our task definition in Section 2, and therefore created queries that require some aspect of reasoning beyond immediate pattern matching. In total, we generated 770k (passage, query, answers) triples.

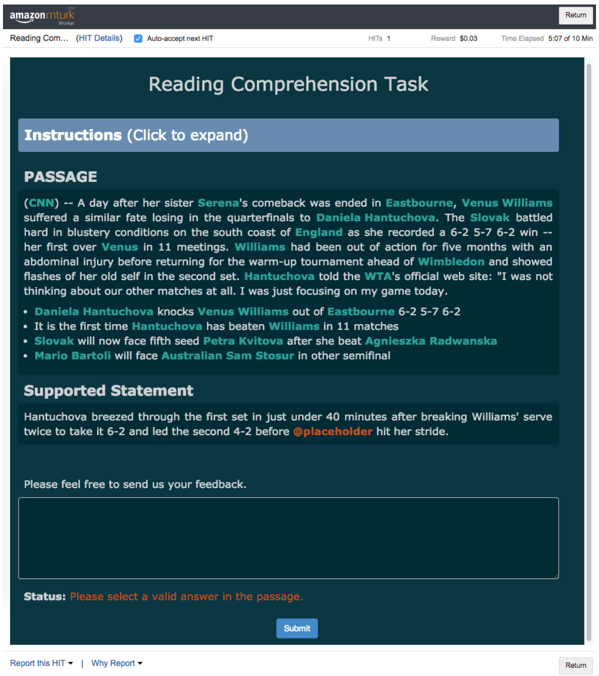
3.3 Machine Filtering

As discussed in Jia and Liang (2017); Rajpurkar et al. (2018); Wang and Bansal (2018); Kaushik and Lipton (2018), existing MRC models mostly learn to predict the answer by simply paraphrasing questions into declarative forms, and then matching them with the sentences in the passages. To overcome this limitation, we filtered out triples whose queries can be easily answered by the state-of-the-art MRC architecture, Stochastic Answer Networks (SAN) (Liu et al., 2018). We choose SAN because it is competitive on existing MRC datasets, and it has components widely used in many MRC architectures such that low bias was anticipated in the filtering (which is confirmed by evaluation in Section 5). We used SAN to perform a five-fold cross validation on all 770k triples. The SAN models correctly answered 68% of these triples. We excluded those triples, and only kept 244k triples that could not be answered by SAN. These triples contain queries which could not be answered by simple paraphrasing, and other types of reasoning such as commonsense reasoning and multi-sentence reasoning are needed.

3.4 Human Filtering

Since the first three stages of data collection were fully automated, the resulted triples could be noisy and ambiguous to human readers. Therefore, we employed crowdworkers to validate these triples. We used Amazon Mechanical Turk for validation. Crowdworkers were required to: 1) have a 95% HIT acceptance rate, 2) a minimum of 50 HITs, 3) be located in the United States, Canada, or Great Britain, and 4) not be granted the qualification of poor quality (which we will explain later in this section). Workers were asked to spend at least 30 seconds on each assignment, and paid $3.6 per hour on average.

Figure 4 shows the crowdsourcing web interface. Each HIT corresponds to a triple in our data collection. In each HIT assignment, we first showed the expandable instructions for first-time workers, to help them better understand our task (see the Appendix A.2). Then we presented workers with a passage in which the named entities are highlighted and clickable. After reading the passage, workers were given a supported statement with a placeholder (i.e., a cloze-style query) indicating a missing entity. Based on their understanding of the events that might be inferred from the passage, workers were asked to find the correct entity in the passage that best fits the placeholder. If workers thought the answer is not obvious, they were allowed to guess one, and were required to report that case in the feedback box. Workers were also encouraged to write other feedback.


amazon mturk Worker
Return
Reading Com... (HIT Details)
☑ Auto-accept next HIT
HITs 1
Reward $0.03
Time Elapsed 5:07 of 10 Min
Instructions (Click to expand)
Submit
Report this HIT ▾ | Why Report ▾
Return

Reading Comprehension Task

PASSAGE

(CNN) -- A day after her sister Serena's comeback was ended in Eastbourne, Venus Williams suffered a similar fate losing in the quarterfinals to Daniela Hantuchova. The Slovak battled hard in blustery conditions on the south coast of England as she recorded a 6-2 5-7 6-2 win -- her first over Venus in 11 meetings. Williams had been out of action for five months with an abdominal injury before returning for the warm-up tournament ahead of Wimbledon and showed flashes of her old self in the second set. Hantuchova told the WTA's official web site: "I was not thinking about our other matches at all. I was just focusing on my game today.

- Daniela Hantuchova knocks Venus Williams out of Eastbourne 6-2 5-7 6-2
- It is the first time Hantuchova has beaten Williams in 11 matches
- Slovak will now face fifth seed Petra Kvitova after she beat Agnieszka Radwanska
- Mario Bartoli will face Australian Sam Stosur in other semifinal

Supported Statement

Hantuchova breezed through the first set in just under 40 minutes after breaking Williams' serve twice to take it 6-2 and led the second 4-2 before @placeholder hit her stride.

Please feel free to send us your feedback.

Status: Please select a valid answer in the passage.

Figure 4: The crowdsourcing web interface.

To ensure quality and prevent spamming, we used the reference answers in the triples to compute workers' average performance after every 1000 submissions. While there might be coreference or named entity recognition errors in the reference answers, as reported in Chen et al. (2016) (also confirmed by our analysis in Section 4), they only accounted for a very small portion of all the reference answers. Thus, the reference answers could be used for comparing workers' performance. Specifically, if a worker's performance was significantly lower than the average performance of all workers, we blocked the worker by granting the qualification of poor quality. In practice, workers were able to correctly answer about 50% of all queries. We blocked workers if their average accuracy was lower than 20%, and then republished their HIT assignments. Overall, 2,257 crowdworkers have participated in our task, and 51 of them have been granted the qualification of poor quality.

Train / Dev. / Test Splits Among all the 244k triples collected from the third stage, we first obtained one worker answer for each triple. Compared to the reference answers, workers correctly answered queries in 122k triples. We then selected around 100k correctly-answered triples as the training set, restricting the origins of these triples to the news articles used in Hermann et al. (2015). As for the development and test sets, we solicited another worker answer to further ensure their quality. Therefore, each of the rest 22k triples has been validated by two workers. We only kept 20k triples that were correctly answered by both workers. The origins of these triples are either articles used in Hermann et al. (2015) or articles crawled by us (as described in Section 3.1), with a ratio of 3:7. Finally, we randomly split the 20k triples into development and test sets, with 10k triples for each set. Table 1 summarizes the statistics of our dataset, ReCoRD.

	Train	Dev.	Test	Overall
queries	100,730	10,000	10,000	120,730
unique passages	65,709	7,133	7,279	80,121
passage vocab.	352,491	93,171	94,386	395,356
query vocab.	119,069	30,844	31,028	134,397
tokens / passage	169.5	168.6	168.1	169.3
entities / passage	17.8	17.5	17.3	17.8
tokens / query	21.3	22.1	22.2	21.4

Table 1: Statistics of ReCoRD

4 Data Analysis

ReCoRD differs from other reading comprehension datasets due to its unique requirement for reasoning more than just paraphrasing. In this section, we provide a qualitative analysis of ReCoRD which highlights its unique features.

Reasoning Types We sampled 100 examples from the development set, and then manually categorized them into types shown in table 2. The results show that significantly different from existing datasets such as SQuAD (Rajpurkar et al., 2016), and NewsQA (Trischler et al., 2017), ReCoRD requires commonsense reasoning to answer 75% of queries. Owing to the machine filtering stage, only 3% queries could be answered by paraphrasing. The small percentage (6%) of ambiguous queries demonstrate the benefit of the human filtering stage. We also noticed that 10% queries can be answered through partial clues. As the example shows, some of partial clues were caused by the incompleteness of named entity recognition in the stage of news article curation.

Types of Commonsense Reasoning Formalizing the commonsense knowledge needed for even simple reasoning problems is a huge undertaking. Based on the observation of the sampled queries that required commonsense reasoning, we roughly categorized them into the following four coarse-gained types:

Reasoning	Description	Example	%
Paraphrasing	The answer sentence can be found by paraphrasing the query with some syntactic or lexical variation.	**P:** . . . Ralph Roberts. . . then acquired other cable systems, changed the name of the company to Comcast and ran the company until he was aged 82 **Q:** X began acquiring smaller cable systems and built the company into the nation's fifth-largest by 1988. **A:** [Ralph Roberts]	3%
Partial Clue	Although a complete semantic match cannot be found between the query and the passage, the answer can be inferred through partial clues, such as some word/concept overlap.	**P:**. . . Hani Al-Sibai says he has 'severe mobility problems' to get disability cash. . . **Q:** However the photographs caught X-Sibai walking with apparent ease in the sunshine. **A:** [Hani Al]	10%
Multi-sentence Reasoning	It requires anaphora, or higher-level fusion of multiple sentences to find the answer.	**P:** Donald Trump is officially a $10 billion man. . . HIs campaign won't release a copy of the financial disclosure even though the FEC says it can do so on its own. . . **Q:** The X campaign did provide a one-page summary of the billionaire's investment portfolio, which is remarkably modest for a man of his means. **A:** [Donald Trump]	6%
Commonsense Reasoning	It requires inference drew on common sense as well as multi-sentence reasoning to find the answer.	**P:** . . . Daniela Hantuchova knocks Venus Williams out of Eastbourne 6-2 5-7 6-2 . . . **Q:** Hantuchova breezed through the first set in just under 40 minutes after breaking Williams' serve twice to take it 6-2 and led the second 4-2 before X hit her stride. **A:** [Venus Williams]	75%
Ambiguous	The passage is not informative enough, or the query does not have a unique answer.	**P:** The supermarket wars have heated up with the chief executive of Wesfarmers suggesting successful rival Aldi may not be paying its fair share of tax in Australia. . . **Q:** X's average corporate tax rate for the last three years was almost 31 per cent of net profit, and in 2013 it paid $81.6 million in income tax. **A:** [Aldi]	6%

Table 2: An analysis of types of reasoning needed in 100 random samples from the dev. set of ReCoRD.

Conceptual Knowledge: the presumed knowledge of properties of concepts (Miller, 1995; Liu and Singh, 2004; Paşca and Van Durme, 2008; Zhang et al., 2017).

Causal Reasoning: the causal bridging inference invoked between two events, which is validated against common sense (Singer et al., 1992; Roemmele et al., 2011).

Naïve Psychology: the predictable human mental states in reaction to events (Stich and Ravenscroft, 1994).

Other: Other types of common sense, such as social norms, planning, spatial reasoning, etc.

We annotated one or more types to each of these queries, and computed the percentage of them in these queries as shown in Table 3.

5 Evaluation

We are interested in the performance of existing MRC architectures on ReCoRD. According to the task definition in Section 2, ReCoRD can be formalized as two types of machine reading comprehension (MRC) datasets: passages with cloze-style queries, or passages with queries whose answers are spans in the passage. Therefore, we can evaluate two types of MRC models on ReCoRD, and compare them with human performance. All the evaluation is carried out based on the train /dev. /test split as illustrated in Table 1.

5.1 Methods

DocQA[2] (Clark and Gardner, 2018) is a strong baseline model for queries with extractive answers. It consists of components such as bi-directional attention flow (Seo et al., 2016) and self attention which are widely used in MRC models. We also evaluate DocQA with ELMo (Peters et al., 2018) to analyze the impact of largely pre-trained encoder on our dataset.

[2] https://github.com/allenai/document-qa

Reasoning	Example	%
Conceptual Knowledge	**P:** Suspended hundreds of feet in the air amid glistening pillars of ice illuminated with ghostly lights from below, this could easily be a computer-generated scene from the latest sci-fi blockbuster movie. But in fact these ethereal photographs were taken in real life…captured by photographer Thomas Senf as climber Stephan Siegrist, 43, scaled frozen waterfall… **Q:** With bright lights illuminating his efforts from below, Mr**X** appears to be on the set of a sci-fi movie. **A:** [Stephan Siegrist] **Commonsense knowledge:** *Scenes such as "a person suspended hundreds of feet in the air amid glistening pillars of ice illuminated with ghostly lights from below" tend to be found in sci-fi movies.*	49.3%
Causal Reasoning	**P:** …Jamie Lee Sharp, 25, stole keys to £40,000 Porsche Boxster during raid…He filmed himself boasting about the car before getting behind the wheel **Q:** **X** was jailed for four years after pleading guilty to burglary, aggravated vehicle taking, driving whilst disqualified, drink-driving and driving without insurance. **A:** [Jamie Lee Sharp] **Commonsense knowledge:** *If a person steals a car, the person may be arrested and jailed.*	32.0%
Naïve Psychology	**P:** Uruguay star Diego Forlan said Monday that he is leaving Atletico Madrid and is set to join Serie A Inter Milan…Forlan said "…At the age of 33, going to a club like Inter is not an opportunity that comes up often…" **Q:** "I am happy with the decision that I have taken, it is normal that some players come and others go," **X** added. **A:** [Diego Forlan, Forlan] **Commonsense knowledge:** *If a person has seized an valuable opportunity, the person will feel happy for it.*	28.0%
Other	**P:** A British backpacker who wrote a romantic note to locate a handsome stranger after spotting him on a New Zealand beach has finally met her Romeo for the first time. Sarah Milne, from Glasgow, left a handmade poster for the man, who she saw in Picton on Friday…She said she would return to the same spot in Picton, New Zealand, on Tuesday in search for him…William Scott Chalmers revealed himself as the man and went to meet her… **Q:** Mr Chalmers, who brought a bottle of champagne with him, walked over to where Milne was sitting and said "Hello, I'm **X**, you know you could have just asked for my number." **A:** [William Scott Chalmers] **Commonsense knowledge:** *When two people meet each other for the first time, they will likely first introduce themselves.*	12.0%

Table 3: An analysis of specific types of commonsense reasoning in 75 random sampled queries illustrated in Table 2 which requires common sense reasoning. A query may require multiple types of commonsense reasoning.
.

QANet[3] (Yu et al., 2018) is one of the top MRC models for SQuAD-style datasets. It is different from many other MRC models due to the use of transformer (Vaswani et al., 2017). Through QANet, we can evaluate the reasoning ability of transformer on our dataset.

SAN[4] (Liu et al., 2018) is also a top-rank MRC model. It shares many components with DocQA, and employs a stochastic answer module. Since we used SAN to filter out easy queries in our data collection, it is necessary to verify that the queries we collect is hard for not only SAN but also other MRC architectures.

ASReader[5] (Kadlec et al., 2016) is a strong baseline model for cloze-style datasets such as (Hermann et al., 2015; Hill et al., 2015). Unlike other baseline models which search among all text spans in the passage, ASReader directly predicts answers from the candidate named entities.

Language Models[6] (LMs) (Trinh and Le, 2018) trained on large corpora recently achieved the state-of-the-art scores on the Winograd Schema Challenge (Levesque et al., 2011). Following in the same manner, we first concatenate the passage and the query together as a long sequence, and substitute **X** in the long sequence with each candidate entity; we use LMs to compute the probability of each resultant sequence and the substitution that results in the most probable sequence will be the predicted answer.

Random Guess acts as the lower bound of the evaluated models. It considers the queries in our

[3] The official implementation of QANet is not released. We use the implementation at `https://github.com/NLPLearn/QANet`.

[4] `https://github.com/kevinduh/san_mrc`

[5] `https://github.com/rkadlec/asreader`

[6] `https://github.com/tensorflow/models/tree/master/research/lm_commonsense`

dataset as cloze style, and randomly picks a candidate entity from the passage as the answer.

5.2 Human Performance

As described in Section 3.4, we obtained two worker answers for each query in the development and test sets, and confirmed that each query has been correctly answered by two different workers. To get human performance, we obtained an additional worker answer for each query, and compare it with the reference answers.

5.3 Metrics

We use two evaluation metrics similar to those used by SQuAD (Rajpurkar et al., 2016). Both ignore punctuations and articles (e.g., *a, an, the*).

Exact Match (EM) measures the percentage of predictions that match any one of the reference answers exactly.

(Macro-averaged) **F1** measures the average overlap between the prediction and the reference answers. We treat the prediction and the reference answer as bags of tokens, and compute their F1. We take the maximum F1 over all of the reference answers for a given query, and then average over all of the queries.

5.4 Results

We show the evaluation results in Table 4. Humans are able to get 91.31 EM and 91.69 F1 on the set, with similar results on the development set. In contrast, the best automatic method – DocQA with ELMo – achieves 45.44 EM and 46.65 F1 on the test set, illustrating a significant gap between human and machine reading comprehension on ReCoRD. All other methods without ELMo get EM/F1 scores significantly lower than DocQA with ELMo, which shows the positive impact of ELMo (see in Section 5.5). We also note that SAN leads to a result comparable with other strong baseline methods. This confirms that since SAN shares general components with many MRC models, using it to do machine filtering does help us filter out queries that are relatively easy to all the methods we evaluate. Finally, to our surprise, the unsupervised method (i.e., LM) which achieved the state-of-the-art scores on the Winograd Schema Challenge only leads to a result similar to the random guess baseline: a potential explanation is the lack of domain knowledge on our dataset. We leave this question for future work.

	Exact Match		F1	
	Dev.	Test	Dev.	Test
Human	**91.28**	**91.31**	**91.64**	**91.69**
DocQA w/ ELMo	44.13	45.44	45.39	46.65
DocQA w/o ELMo	36.59	38.52	37.89	39.76
SAN	38.14	39.77	39.09	40.72
QANet	35.38	36.51	36.75	37.79
ASReader	29.24	29.80	29.80	30.35
LM	16.73	17.57	17.41	18.15
Random Guess	18.41	18.55	19.06	19.12

Table 4: Performance of various methods and human.

5.5 Analysis

Human Errors About 8% dev./test queries have not been correctly answered in the human evaluation. We analyzed samples from these queries, and found that in most queries human was able to narrow down the set of possible candidate entities, but not able to find a unique answer. In many cases, two candidate entities equally fit **X** unless human has the specific background knowledge. We show an example in the Appendix A.1.

For the method analysis, we mainly analyzed the results of three representative methods: DocQA w/ ELMo, DocQA, and QANet.

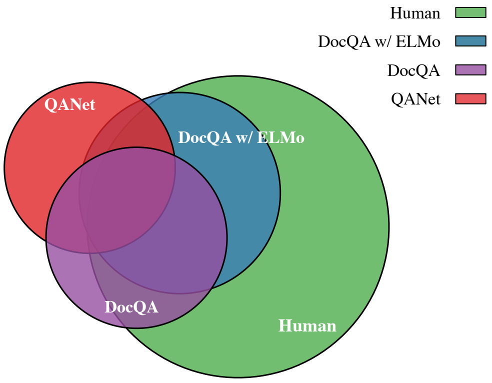


Figure 5: The Venn diagram of correct predictions from various methods and human on the development set.

Impact of ELMo As shown in Figure 5, among all three methods the correct predictions of DocQA w/ ELMo have the largest overlap (92.6%) with the human predictions. As an ablation study, we analyzed queries which were only correctly answered after ELMo was added. We found that in some cases ELMo helped the prediction by incorporating the knowledge of language models. We show an example in the Appendix A.1.

Predictions of QANet Figure 5 shows that QANet correctly answered some ambiguous queries, which we think was due to the randomness of parameter initialization and did not reflect the true reasoning ability. Since QANet uses the transformer-based encoder and DocQA uses the LSTM-based encoder, we see a significant difference of predictions between QANet and DocQA.

Method	OOC Rate
DocQA w/ ELMo	6.27%
DocQA	6.37%
QANet	6.41%

Table 5: The out-of-candidate-entities (OOC) rate of three analyzed methods.

Impact of Cloze-style Setting Except ASReader, all the MRC models were evaluated under the extractive setting, which means the information of candidate named entities was not used. Instead, extractive models searched answers from all possible text spans in passages. To show the potential benefit of using the candidate entities in these models, we computed the percentage of model predictions that could not be found in the candidate entities. As shown in Table 5, all three methods have about 6% OOC predictions. Making use of the candidate entities would potentially help them increase the performance by 6%.

In Section 4, we manually labeled 100 randomly sampled queries with different types of reasoning. In Figure 6 and 7, we show the performance of three analyzed methods on these queries.

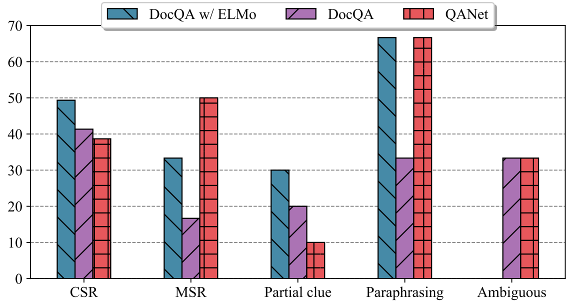


Figure 6: Performance of three analyzed methods on the 100 random samples with reasoning types labeled.(CSR stands for commonsense reasoning, and MSR stands for multi-sentence reasoning.)

Figure 6 shows that three methods performed poorly on queries requiring commonsense reasoning, multi-sentence reasoning and partial clue. Compared to DocQA, QANet performed better on multi-sentence reasoning queries probably due to the use of transformer. Also, QANet outperformed DocQA on paraphrased queries probably because we used SAN to filtering queries and SAN has an architecture similar to DocQA. As we expect, ELMo improved the performance of DocQA on paraphrased queries.

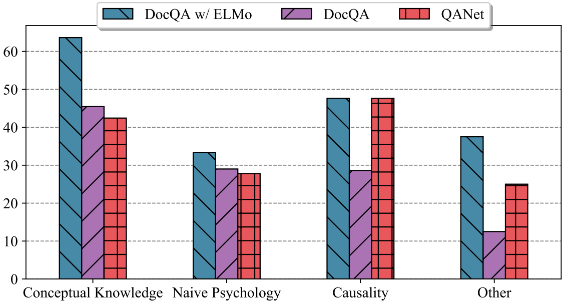


Figure 7: Performance of three analyzed methods on 75% of the random samples with specific commonsense reasoning types labeled.

Among the 75% sampled queries that require commonsense reasoning, we see that ELMo significantly improved the performance of commonsense reasoning with presumed knowledge. For all other types of commonsense reasoning, all three methods have relatively poor performance.

6 Related Datasets

ReCoRD relates to two strands of research in datasets: data for reading comprehension, and that for commonsense reasoning.

Reading Comprehension *The CNN/Daily Mail Corpus* (Hermann et al., 2015), *The Children's Book Test* (CBT) (Hill et al., 2015), and LAMBADA (Paperno et al., 2016) are closely related to ReCoRD: (1) *The CNN/Daily Mail Corpus* constructed queries from the bullet points, most of which required limited reasoning ability (Chen et al., 2016). (2) CBT is a collection of 21 consecutive sentences from book excerpts, with one word randomly removed from the last sentence. Since CBT has no machine or human filtering to ensure quality, only a small portion of the CBT examples really probes machines' ability to understand the context. (3) Built in a similar manner to CBT, LAMBADA was filtered to be human-guessable in the broader context only. Differing from ReCoRD, LAMBADA was designed to be a language modeling problem where contexts were not required to be event summaries, and answers were not necessarily in the context.

Since all candidate answers were extracted from in the passage, ReCoRD can also be formalized as a extractive MRC dataset, similar to SQuAD (Rajpurkar et al., 2016) and NewsQA (Trischler et al., 2017). The difference is that questions in these datasets were curated from crowdworkers. Since it is hard to control the quality of crowdsourced questions, a large portion of questions in these datasets can be answered by word matching or paraphrasing (Jia and Liang, 2017; Rajpurkar et al., 2018; Wang and Bansal, 2018). There are other large-scale datasets (Nguyen et al., 2016; Joshi et al., 2017; Lai et al., 2017; Dunn et al., 2017; Kocisky et al., 2018; Reddy et al., 2018; Choi et al., 2018; Yang et al., 2018) targeting different aspects of reading comprehension. See (Gao et al., 2018) for a recent survey.

Commonsense Reasoning ROCStories Corpus (Mostafazadeh et al., 2016), SWAG (Zellers et al., 2018), and *The Winograd Schema Challenge* (WSC) (Levesque et al., 2011) are related ReCoRD: (1) ROCStories assesses commonsense reasoning in story understanding by choosing the correct story ending from only two candidates. Stories in the corpus were all curated from crowdworkers, which could suffer from human elicitation bias (Gordon and Van Durme, 2013; Misra et al., 2016; Zhang et al., 2017). (2) SWAG unifies commonsense reasoning and natural language inference. It selects an ending from multiple choices which is most likely to be anticipated from the situation describe in the premise. The counterfactual endings in SWAG were generated using language models with adversarial filtering. (3) WSC foucses on intra-sentential pronoun disambiguation problems that require commonsense reasoning. There are other datasets (Roemmele et al., 2011; Zhang et al., 2017; Rashkin et al., 2018a,b) targeting different aspects of commonsense reasoning.

7 Conclusion

We introduced ReCoRD, a large-scale reading comprehension dataset requiring commonsense reasoning. Unlike existing machine reading comprehension (MRC) datasets, ReCoRD contains a large portion of queries that require commonsense reasoning to be answered. Our baselines, including top performers on existing MRC datasets, are no match for human competence on ReCoRD. We hope that ReCoRD will spur more research in MRC with commonsense reasoning.

References

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 (Volume 1: Long Papers)*, pages 2358–2367, Berlin, Germany. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855. Association for Computational Linguistics.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

Jianfeng Gao, Michel Galley, and Lihong Li. 2018. Neural approaches to conversational ai. *arXiv preprint arXiv:1809.08267*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. Allennlp: A deep semantic natural language processing platform. *arXiv preprint arXiv:1803.07640*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Karl Moritz Hermann, Tomáš Kočiský, 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*, pages 1693–1701.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

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

2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel 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 (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. 2016. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 908–918. Association for Computational Linguistics.

Divyansh Kaushik and Zachary C. Lipton. 2018. How much reading does reading comprehension require? a critical investigation of popular benchmarks.

Tomas Kocisky, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gabor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2011. The winograd schema challenge. In *Aaai spring symposium: Logical formalizations of commonsense reasoning*.

H. Liu and P. Singh. 2004. Conceptnet &mdash; a practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4):211–226.

Xiaodong Liu, Yelong Shen, Kevin Duh, and Jianfeng Gao. 2018. Stochastic answer networks for machine reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1694–1704. Association for Computational Linguistics.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60.

John McCarthy. 1959. Programs with common sense. In *Proceedings of the Teddington Conference on the Mechanization of Thought Processes*, London: Her Majesty's Stationery Office.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Ishan Misra, C Lawrence Zitnick, Margaret Mitchell, and Ross Girshick. 2016. Seeing through the human reporting bias: Visual classifiers from noisy human-centric labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2930–2939.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

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. *arXiv preprint arXiv:1611.09268*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernandez. 2016. The lambada dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534. Association for Computational Linguistics.

Marius Paşca and Benjamin Van Durme. 2008. Weakly-supervised acquisition of open-domain classes and class attributes from web documents and query logs. In *Proceedings of ACL-08: HLT*, pages 19–27. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

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

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*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Hannah Rashkin, Antoine Bosselut, Maarten Sap, Kevin Knight, and Yejin Choi. 2018a. Modeling naive psychology of characters in simple commonsense stories. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2289–2299. Association for Computational Linguistics.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018b. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*, pages 90–95.

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

Murray Singer, Michael Halldorson, Jeffrey C Lear, and Peter Andrusiak. 1992. Validation of causal bridging inferences in discourse understanding. *Journal of Memory and Language*, 31(4):507 – 524.

Stephen Stich and Ian Ravenscroft. 1994. What is folk psychology? *Cognition*, 50(1-3):447–468.

Krysta Svore, Lucy Vanderwende, and Christopher Burges. 2007. Enhancing single-document summarization by combining RankNet and third-party sources. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 448–457, Prague, Czech Republic. Association for Computational Linguistics.

Trieu H Trinh and Quoc V Le. 2018. A simple method for commonsense reasoning. *arXiv preprint arXiv:1806.02847*.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2017. Newsqa: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Yicheng Wang and Mohit Bansal. 2018. Robust machine comprehension models via adversarial training. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 575–581. Association for Computational Linguistics.

Kristian Woodsend and Mirella Lapata. 2010. Automatic generation of story highlights. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 565–574, Uppsala, Sweden. Association for Computational Linguistics.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

A Appendices

A.1 Case Study

Human Error Table 6 shows an example where the ambiguous query caused human error. The passage in this example describes "*ambiverts*", and there are two experts studying it: "*Vanessa Van Edwards*" and "*Adam Grant*". Both of them fit in the query asking who gave advice to ambiverts. There is no further information to help human choose a unique answer for this query.

Passage: Your colleagues think you're quiet, but your friends think you're a party animal. If that sounds like you, then you may be what psychologists describe as an 'ambivert'. Scientists believe around two-thirds of people are ambiverts; a personality category that has, up until now, been given relatively little attention. 'Most people who are ambiverts have been told the wrong category their whole life,' Vanessa Van Edwards, an Orgeon-based behavioural expert, told DailyMail.com 'You hear extrovert and you hear introvert, and you think 'ugh, that's not me'.' Ambiversion is a label that has been around for some time, but gained popularity in 2013 with a paper in the journal Psychological Science, by Adam Grant the University of Pennsylvania.

- Most ambiverts have been labelled incorrectly their whole life
- They slide up and down personality spectrum depending on the situation
- Ambiverts are good at gaining people's trust and making their point heard
- They often feel pressure to mirror personality of the person they are with

Query: 'Read each situation more carefully,' X advised ambiverts, 'and ask yourself, 'What do I need to do right now to be most happy or successful?"
Reference answers: Adam Grant

Table 6: An example illustrating a ambiguous query.

Impact of ELMo Table 7 shows an example where DocQA w/ ELMo correctly answered but DocQA failed. The passage in this example describes a woman artist "*Sarah Milne*" who launched a public appeal to find a handsome stranger "*William Scott Chalmers*", and invited him to meet her. The query asks the missing information in the greetings from "*William Scott Chalmers*" when he went to meet "*Sarah Milne*". Our common sense about social norms tells us when two people meet each other for the first time, they are very likely to first introduce themselves. In the query of this example, when Mr. Chalmers said "*Hello, I'm . . .*", it is very likely that he was introducing himself. Therefore, the name of Mr Chalmer fit **X** best.

In this example, the prediction of DocQA without ELMo is "*New Zealand*" which is not even close to the reference answer. The benefit of using ELMo in this example is that its language model will help exclude "*New Zealand*" from the likely candidate answers, because "*I'm . . .*" is usually followed by a person name rather than a location name. Such a pattern learnt by ELMo is useful in narrowing down candidiate answers in ReCoRD.

Passage: A British backpacker who wrote a romantic note to locate a handsome stranger after spotting him on a New Zealand beach has finally met her Romeo for the first time. Sarah Milne, from Glasgow, left a handmade poster for the man, who she saw in Picton on Friday and described as 'shirtless, wearing black shorts with stars tattooed on his torso and running with a curly, bouncy and blonde dog'. In her note, entitled 'Is this you? ', she invited the mystery stranger to meet her on the same beach on Tuesday. But the message soon became a source of huge online interest with the identity of both the author and its intended target generating unexpected publicity.

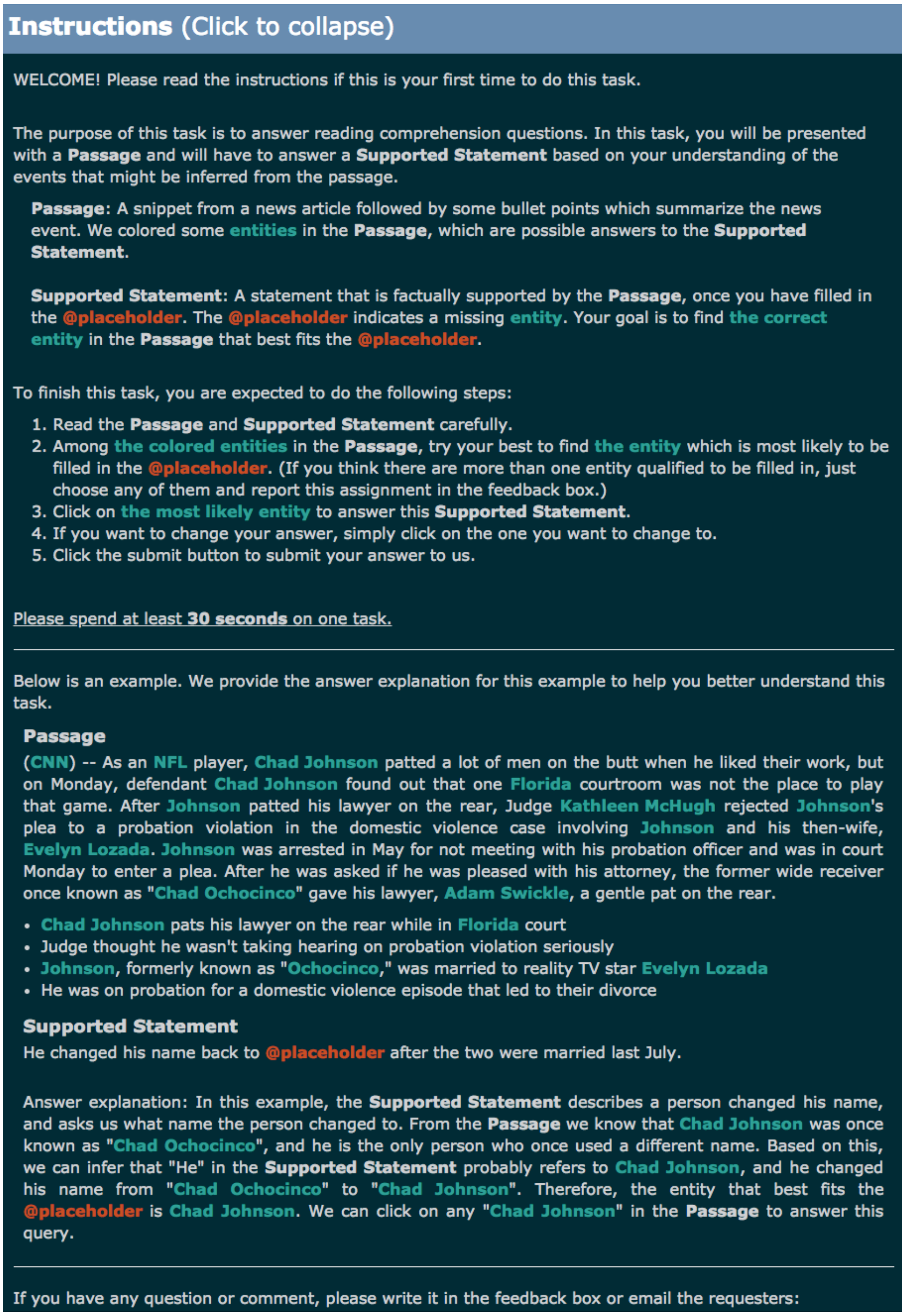
- Sarah Milne, a Glasgow artist, launched a public appeal to find the mystery man
- She wrote a heart-warming message and drew a picture of him with his dog
- She said she would return to the same spot in Picton, New Zealand, on Tuesday in search for him
- William Scott Chalmers revealed himself as the man and went to meet her
- He told Daily Mail Australia that he would ask her out for dinner

Query: Mr Chalmers, who brought a bottle of champagne with him, walked over to where Milne was sitting and said 'Hello, I'm X, you know you could have just asked for my number.'
Reference answers: William Scott Chalmers

Table 7: An example illustrating the impact of ELMo.

A.2 HIT Instructions

We show the instructions for Amazon Mechanical Turk HITs in Figure 8.



Instructions (Click to collapse)

WELCOME! Please read the instructions if this is your first time to do this task.

The purpose of this task is to answer reading comprehension questions. In this task, you will be presented with a **Passage** and will have to answer a **Supported Statement** based on your understanding of the events that might be inferred from the passage.

Passage: A snippet from a news article followed by some bullet points which summarize the news event. We colored some **entities** in the **Passage**, which are possible answers to the **Supported Statement**.

Supported Statement: A statement that is factually supported by the **Passage**, once you have filled in the **@placeholder**. The **@placeholder** indicates a missing **entity**. Your goal is to find **the correct entity** in the **Passage** that best fits the **@placeholder**.

To finish this task, you are expected to do the following steps:

1. Read the **Passage** and **Supported Statement** carefully.
2. Among **the colored entities** in the **Passage**, try your best to find **the entity** which is most likely to be filled in the **@placeholder**. (If you think there are more than one entity qualified to be filled in, just choose any of them and report this assignment in the feedback box.)
3. Click on **the most likely entity** to answer this **Supported Statement**.
4. If you want to change your answer, simply click on the one you want to change to.
5. Click the submit button to submit your answer to us.

Please spend at least **30 seconds** on one task.

Below is an example. We provide the answer explanation for this example to help you better understand this task.

Passage

(**CNN**) -- As an **NFL** player, **Chad Johnson** patted a lot of men on the butt when he liked their work, but on Monday, defendant **Chad Johnson** found out that one **Florida** courtroom was not the place to play that game. After **Johnson** patted his lawyer on the rear, Judge **Kathleen McHugh** rejected **Johnson**'s plea to a probation violation in the domestic violence case involving **Johnson** and his then-wife, **Evelyn Lozada**. **Johnson** was arrested in May for not meeting with his probation officer and was in court Monday to enter a plea. After he was asked if he was pleased with his attorney, the former wide receiver once known as "**Chad Ochocinco**" gave his lawyer, **Adam Swickle**, a gentle pat on the rear.

- **Chad Johnson** pats his lawyer on the rear while in **Florida** court
- Judge thought he wasn't taking hearing on probation violation seriously
- **Johnson**, formerly known as "**Ochocinco**," was married to reality TV star **Evelyn Lozada**
- He was on probation for a domestic violence episode that led to their divorce

Supported Statement

He changed his name back to **@placeholder** after the two were married last July.

Answer explanation: In this example, the **Supported Statement** describes a person changed his name, and asks us what name the person changed to. From the **Passage** we know that **Chad Johnson** was once known as "**Chad Ochocinco**", and he is the only person who once used a different name. Based on this, we can infer that "He" in the **Supported Statement** probably refers to **Chad Johnson**, and he changed his name from "**Chad Ochocinco**" to "**Chad Johnson**". Therefore, the entity that best fits the **@placeholder** is **Chad Johnson**. We can click on any "**Chad Johnson**" in the **Passage** to answer this query.

If you have any question or comment, please write it in the feedback box or email the requesters:

Figure 8: Amazon Mechanical Turk HIT Instructions.