Getting Closer to AI Complete Question Answering: A Set of Prerequisite Real Tasks

Anna Rogers, Olga Kovaleva, Matthew Downey, Anna Rumshisky
Department of Computer Science, University of Massachusetts Lowell
Lowell, MA 01854
{arogers, okovalev, mdowney, arum}@cs.uml.edu

Abstract

The recent explosion in question answering research produced a wealth of both factoid reading comprehension (RC) and commonsense reasoning datasets. Combining them presents a different kind of task: deciding not simply whether information is present in the text, but also whether a confident guess could be made for the missing information. We present QuAIL, the first RC dataset to combine text-based, world knowledge and unanswerable questions, and to provide question type annotation that would enable diagnostics of the reasoning strategies by a given QA system. QuAIL contains 15K multi-choice questions for 800 texts in 4 domains. Crucially, it offers both general and text-specific questions, unlikely to be found in pretraining data. We show that QuAIL poses substantial challenges to the current state-of-the-art systems, with a 30% drop in accuracy compared to the most similar existing dataset.

1 Introduction

Evaluation of NLP systems relies heavily on high-level reasoning tasks such as natural language inference and question answering (QA). They are used as de-facto Turing test proxies: if a model can perform such reasoning, then we know that it does indeed capture a lot of language knowledge. However, for that to work, the benchmarks need to cover a large (if not complete) set of reasoning strategies employed by humans.

It is clear by now that (1) the existing datasets typically explore a specific type of reasoning and are unbalanced for question types, and (2) even the most successful NLP models rely on annotation artifacts. Furthermore, much of reading comprehension (RC) datasets target common knowledge (e.g. "What is the birth place of Dante"?), which could have directly occurred in pretraining data of large models, such as BERT or XLNet.

QuAIL (Question answering for Artificial IntelLigence) is a new benchmark that aims to achieve the following:

- explore *the full spectrum of uncertainty* in QA: information directly present in the text, a combination of text-based information with world knowledge, and deciding that neither source provides sufficient information;
- attempt to *balance the types of reasoning and domains* in the dataset and provide *reasoning type annotation*, which would be used for model diagnostics;
- collect a new corpus with *information that is unlikely to have occurred in typical pretraining data*.

QuAIL contains 15K questions across 4 domains and is publicly available[1]. Experiments with 7 baseline systems show QuAIL to be challenging, with the system that achieves state-of-the-art results on the most similar dataset experiencing up to 30% drop in accuracy.

2 Related Work

The QA field is exploding: at the time of submission of this paper there were already over 80 datasets, with at least 40 published or announced in 2018-2019. It comprises two subfields: open-world QA and reading comprehension (RC).

In open-world QA there are multiple possible sources of information (typically web snippets) that may or may not contain the correct answer. The questions are collected independently of texts that contain the answers, and tend to focus on generic factoid information, such as trivia (e.g. TriviaQA, (Joshi et al. 2017)) and search engine queries (e.g. NaturalQuestions (Kwiatkowski et al. 2019)).

RC datasets come in several flavors: cloze/span-selection task, multiple-choice questions, and (the most rare) freeform answers. Most RC datasets are in the cloze and span-selection categories, as they are easy to generate. These include the popular SQuAD (Rajpurkar et al. 2016), CBT (Hill et al. 2015), and CNN/Daily Mail (Hermann et al. 2015) datasets that have long dominated QA research. The obvious limitation of extractive datasets is that they can only target information explicitly mentioned in the text, and often get solved with shallow lexical matching. Interesting recent attempts to force more complex reasoning include unanswerable questions (Rajpurkar, Jia, and Liang 2018), reasoning over long texts (Kocisky et al. 2018), and multiple documents ((Yang et al. 2018)).

[1] http://text-machine.cs.uml.edu/projects/quail/

Another way to enforce more complex reasoning and simultaneously enable questions that require both context and external knowledge is to switch from extractive to multiple-choice questions. This format is found in collections of school tests that are accompanied with "answer evidence" texts (e.g. RACE (Lai et al. 2017) and ARC (Clark et al. 2018)). It also lends itself naturally to commonsense reasoning tasks, where most datasets are based on crowdsourced narratives (MCScript (Ostermann et al. 2018), RocStories (Mostafazadeh et al. 2017)). Datasets like SWAG (Zellers et al. 2018) and Winograd (Levesque, Davis, and Morgenstern 2012) could also be considered in this category, except that their text consists of a single sentence.

Motivation for the Present Work

All of over 80 current QA and RC datasets that we are aware of have one or more of the following problems.

- **Single domain.** The absolute majority of the available datasets target only one domain, with rare exceptions such as CoQA (Reddy, Chen, and Manning 2018).
- **Non text-specific knowledge.** What exacerbates the single-domain problem is that this one domain is most often encyclopedia, which means that the facts targeted by the questions are common knowledge. Given that the leaderboards are currently dominated by large models like BERT (Devlin et al. 2018) that had a lot of pretraining, the target facts could have occurred in the pretraining data.
- **Built-in assumptions about the source of information.** Extractive datasets assume that the knowledge is either found in the text or not (for unanswerable questions). Likewise, world knowledge datasets assume that external knowledge should be required. A more difficult challenge (indeed, a new task) would be to make the model decide *whether* it has enough evidence, and when it should stop trying to find it.
- **Lack of question type annotation.** The common strategy is to collect the dataset and then to manually analyze a small random sample, which typically shows that a few types of reasoning are much more frequent than the others. For example, NarrativeQA has 30.54% questions about people, 24.5% descriptions, 9.73% locations, 9.4% reasons, and the rest concerns entities, objects, numbers, duration, and relations (Kocisky et al. 2018).

The last problem is one of the most pervasive and the least discussed, and it would be a major contribution of the current work if it managed to draw attention to it. So far question type annotation is available only in in synthetic datasets like the original bAbI (Weston et al. 2015). The lack of this feature is particularly surprising in the datasets that could have provided it easily because they relied on question templates (e.g. emrQA (Pampari et al. 2018)) or simulated worlds (Labutov et al. 2018).

Here is why the general lack of interest in annotating reasoning types is problematic.

- *There is no way to diagnose the model for what it gets right and wrong.* When developing QA models, we keep trying different architectures until we find something that does overall better than the previous best result – and that is likely to come at the expense of some types of reasoning that are less frequent in the dataset. Combine this with the high conference preference for publishing models achieving SOTA on some benchmark, and we get an incentive for developing highly specialized systems for given tasks rather than pursuing general language understanding.
- *Crowdsourced datasets are susceptible to annotation artifacts* (Gururangan et al. 2018), and, intuitively, *balancing data subcategories such as different question types should help to encourage diversity and balance*. It is now clear that Turkers do not "naturally" yield much diversity (Geva, Goldberg, and Berant 2019), and even the top models like BERT are susceptible to annotation artifacts (Niven and Kao 2019). Most disturbingly, datasets can be exploited to find a "trigger" phrase that will cause the model to output the same prediction if added to any input (Wallace et al. 2019).
- Crowdsourced datasets are often suffering from poor lexical diversity and can be solved with simple heuristics (Jia and Liang 2017; Chen, Bolton, and Manning 2016). The obvious solution is adversarial authoring, but combining text-based and world knowledge questions and providing question type annotation would also help to see where the models are having suspiciously high success rates.

A recent step towards robustness of verbal reasoning is training on a combination of different datasets (Talmor and Berant 2019). MRQA shared task (Fisch et al. 2019) and Open Reading Benchmark [2] offer a curated collection of extractive question answering datasets with different domains. However, note that these collections still cannot offer diagnostics of reasoning types that can be performed by the system (beyond wh-word distribution analysis), as the component datasets did not have it in the first place.

To the best of our knowledge, QuAIL is the first multi-domain, human-written QA dataset to be balanced and annotated for question types, the first to combine unanswerable questions with the text-based questions and questions requiring external knowledge, and also one of the few datasets to combine the domains with openly available knowledge and domains where the contexts can be assumed to be unique.

3 Constructing QuAIL

Corpus Collection

QuAIL comes with a new balanced multi-domain corpus of 800 texts, drawing from resources which either have a Creative Common (CC) license or for which the usage permission has been obtained. Table 1 shows the domains, each of which is represented by 200 texts. The texts were hand-picked from beginning of texts or chapters/sections, so that they would be comprehensible without previous context.

Please note that two of the domains (news and, to some degree, blogs) contain information that could be found elsewhere in a large pretraining corpus. However, fiction (by recent and not well-known writers) and personal stories shared

[2] https://leaderboard.allenai.org/orb/submissions/public

on Quora could be assumed to describe unique combinations of events and characters. This means that performance on these two splits of QuAIL data could be used to diagnose the performance of a QA model and pre-training effect.

Last but not the least, the length of QuAIL texts makes it more complex than most current datasets that are limited to a single paragraph or a web snippet. Each QuAIL text is 300-350 words of text (250 is a standard essay page).

Domain	Description
Fiction	fiction published under CC license[1]
News	political news from Voice of America[2]
Blogs	a collection of blogs, assembled manually
User stories	User stories published on Quora[3]

[1] http://manybooks.net/categories/CCL
[2] https://www.voanews.com/
[3] https://www.quora.com/about/tos

Table 1: Domains in QuAIL

Question Types

One takeaway from surveying the large body of existing RC datasets is that all the authors use completely different sets of reasoning types to analyze random samples of their data, depending on what kind of corpus they had. This highlights an important limitation on multi-domain datasets: for them to be balanced, the questions have to be such that can be asked by crowd workers in all the domains.

For QuAIL, we found a set of 9 questions types that fit our 4-domain corpus after much experimentation. They include:

- **Text-based questions**: (1) reasoning about *factual information* in the text, (2) *temporal order of events*, (3) *character identity*.
- **Questions that require world knowledge** cannot be answered based on the text alone, but world knowledge makes one of the answer options more likely. These include: (4) *causality*[3], (5) *inference about properties and qualities* of characters, (6) their *belief states*, (7) the most likely *subsequent state after the narrated events*, and (8) the likely *duration of the narrated events*.
- **Unanswerable questions** (9) cannot be answered with the information in the text, and the world knowledge does not make one of the options more likely.

A sample of QuAIL data is provided in Fig. 1.

Ensuring Question Type Correctness

The biggest problem that we needed to solve was ensuring the correct types of the generated questions. This proved to be difficult, which is consistent with the fact that when the authors of other published datasets performed qualitative analysis of small data samples, there were usually several highly prevalent types of questions (usually factoid wh-questions) (Kocisky et al. 2018; Dua et al. 2019).

We experimented with 3 setups:

- crowd-assisted writing by semi-skilled annotators;
- crowdsourcing questions with keyword-based validation;
- expert writing.

Crowdsourcing with manual validation was conducted in 2 phases. First, we set up a task on Amazon Mechanical Turk in which the crowd workers were asked to read a text and to compose 9 questions with multi-choice answers, of which only the first one must be correct. Each question had 3 separate input fields for the answers, one of them explicitly labeled as the "correct answer", and the others as "plausible answer". For the unanswerable questions, the form emphasized that they all needed to be incorrect but plausible (Levy et al. 2017). The form contained both descriptions an examples for different question types, adapted for the different domains. The forms can be found in the project repository.

Following Rajpurkar et al., we hire workers from US and Canada who have a 97% HIT acceptance rate, a minimum of 1000 HITs, and are located in the United States, Canada, or the UK. For each text, 9 questions are generated by 2 Turkers, totaling 18 questions per text. We paid 2.5$ per HIT.

In the second phase, the questions were manually edited by students (CS undergrad interns). The students were asked to check both the question type and the correctness of the answers. We provided several rounds of training for the task.

Crowdsourcing with automatic validation. In this setup, the crowd workers were presented with the same form as above, except that now all the input fields contained validation by 10-20 keywords and phrases that we collected after much analysis of unconstrained input by crowdworkers. For example, temporal order questions had to contain words like "before", "after", "when", etc. The code for this HIT can be found in the project repository.

Crowdsourcing with expert adversarial validation also relies on correcting Turker-generated output, but is performed by an expert linguist whose task is to not only check the question types and correct answers, but also paraphrase questions that are too obvious. The results of this experiment will be described in Section 5.

Data Quality Analysis

To compare manual and automatic validation procedures, we manually annotated 180 questions for each procedure. We checked (a) whether the questions were of the right type, and (b) whether there was a single correct answer.

We found that the students successfully corrected typos and grammar errors, but did not fully eliminate the noise: 11.1% problematic questions for crowd-assisted writing, 16.7% for automatic validation. Most problems involved more than one correct answer, and in a few rare cases all answers were incorrect.

In terms of question types, the Turkers actually fared batter: 6.7% errors, as opposed to 11.7% for the CS students

[3] Causality borders on text-based questions: it can involve either causal links that follow from temporal order of events, or those inferrable with some external knowledge.

Genre: fiction

The air exploded in a flash of bone and steel and blood. The clash of metal rang through the forest. An arrow pierced through the darkness, its barbed head tearing through flesh and muscle. A roar echoed off of the mountains far to the west. A cry broke through soon after. Then silence.
Char stood over a pile of black fur and red blood. He held a curved sword, jagged half way down the wide blade and hilted in bone. He held a large thick bow in the other. Lorfel and Ranur stood behind him, panting. Lorfel, a short man of twenty six held a large axe in both hands and still prepared to swing it hard. Ranur, the largest of the three held a pike in one hand, its tip hanging low towards the ground. He buried his other hand in his gray tunic.
"Did it get either of you?" Char's voice rasped low in the silence of the night.
"No" Lorfel said. He planted his axe head on the ground with a thud and leaned on the tall handle. There was a pause. Char turned towards Ranur.
"Are you hurt?"
"Mm...My hand." Ranur took his hand out of his tunic. Moonlight gleamed red off of the ragged wound. Char thought he saw a glimmer of bone.
"Did he claw you or bite you?" Char's voice held an urgency that set both Lorfel and Ranur on edge.
Ranur paused and then spoke low. "He bit me."
Char picked Lorfel and Ranur as his hunting partners for their speed and sharpness in battle. They had hunted beasts of the deep woods all of their lives. They hunted the beasts that hunted men. They all knew the risks of battling such creatures. The old man dropped his curved sword, drew his bow, and fired. The arrow hammered into Ranur's chest, burying itself in his heart. Lorfel saw the gleaming arrow head sticking almost a foot out of his companion's back. Ranur fell face first to the ground.

Source: *The Bear* by Michael E. Shea

T *Temporal order: the order in which events happened*

When did the roar happen?

(a) After the cry
(b) before the silence
(c) not enough information to answer this question
(d) when Char was speaking

T *Coreference: linking a pronoun with its coreferent noun*

Who bit Ranur?

(a) the beast
(b) Lorfel
(c) Char
(d) not enough information to answer this question

T *Factual questions*

What was the color of the beast's fur?

(a) brown
(b) not enough information to answer this question
(c) black
(d) red

T *Causality: causal links between events*

Why was there blood?

(a) because Char shot something
(b) not enough information to answer this question
(c) because Lorfel had an axe
(d) because Char had a sword

W *Subsequent state after a narrated event*

After the end of this text, Ranur is:

(a) standing up
(b) not enough information to answer this question
(c) on the ground
(d) in the sky

W *The most likely event duration*

Ranur probably died:

(a) a month later
(b) instantly
(c) not enough information to answer this question
(d) a year later

W *Entity properties inferrable from the text*

What is probably true about the beast's bite?

(a) it is harmless
(b) it is extremely dangerous
(c) not enough information to answer this question
(d) it helps people

W *Belief states of characters*

Who was concerned about his companions' injuries?

(a) not enough information to answer this question
(b) Char
(c) Lorfel
(d) Ranur

U *Unanswerable questions: equally likely answer options*

What was done with Ranur's body?

(a) burned to avoid spreading disease
(b) left abandoned along with the beasts' corpse
(c) buried in the ground
(d) not enough information to answer this question

Figure 1: Sample of QuAIL data

(even after several training sessions). In both conditions, there were outliers: the texts containing a third of all errors.

In both conditions the bulk of the errors was due to confusion of unanswerable and character identity questions with factual questions. Our instructions originally solicited questions to pronouns that would require coreference resolution, but about 25% of the resulting questions were simply factual who-questions (e.g. "Who are the cake?" if the text said "John ate the cake".) It would thus be more correct to say that the resulting questions are about "character identity".

In some cases questions could be attributed to more than one category (e.g. a question character's beliefs at a certain point in time involves both temporal reasoning and inference about belief states). We considered the original writer's question type correct, even if more types were applicable.

We also found that crowd-assisted writing introduced a new problem. The number of trained writers would necessarily be limited (we had 3), and while processing large volumes of data they could develop their own strategies, thus introducing additional biases. In this sample, we found that 25% that the annotator labeled as entity properties were in fact on causality: the type of error none of the Turkers made.

We developed the automatic validation after we generated the data for 200 fiction texts. Since the results were comparable in terms of noise, but validation was faster and yielded more diversity (although with more language errors), the remaining domains were processed with this procedure.

4 Human Evaluation

Human evaluation was performed on a random sample of 180 questions. We measured agreement (Krippendorff's alpha) between the numbers of the answer options labeled as correct question authors, and those selected by two volunteers (native English speakers, non-linguists). We relied on volunteers rather than MTurk workers because the QuAIL texts were rather long, and our pilot experiments showed that Turkers are not necessarily motivated to spend sufficient time to look for non-obvious answers. This would make their performance on the task an inaccurate estimate of "human performance".

Let us reiterate that a big new challenge in QuAIL is the combination of text-based, world knowledge and unanswerable questions. This covers the full range of uncertainty we face in real life. For example, suppose you are considering a purchase of a new laptop. You know where to look for technical characteristics, and from the photos you will deduce what you would look like on video calls with this laptop. But the level of comfort of the keyboard will be an unknown until you spend some time with it.

The problem is, however, how to test this ability in a way that would be fair to both humans and NLP systems. We experimented with the following setups.

- **All questions**. The straightforward formulation of the task would be "Please read the text and choose the answer option you think is correct". Each question has the option "not enough information to answer the question", and that option is the correct one for unanswerable questions.

Question type	All questions	Text+ Unanswerable	World knowledge
Temp. order	0.66	0.67	–
Coreference	0.70	0.79	–
Factual	0.75	0.82	–
Causality	0.76	–	0.86
Subsequent	0.53	–	0.62
Duration	0.32	–	0.37
Properties	0.67	–	0.78
Beliefs	0.62	–	0.85
Unanswerable	0.25	0.83	–
Total	0.60	0.78	0.70

Table 2: Human evaluation per question type

- **Text-based + unanswerable questions.** The task is formulated as above, but world knowledge questions are excluded (task similar to SquAD 2.0 (Rajpurkar, Jia, and Liang 2018).)
- **World knowledge**. This setup is the same as above, but with only world knowledge questions. Unanswerable questions are not present, but we do keep the "not enough information" answer option, which is normally not present in the existing commonsense datasets such as SWAG (Zellers et al. 2018). In this condition, we formulated it as *Given these options, I would rather not guess*.

The results of these experiments, broken down by question types, are shown in Table 2. The big takeaway here is that presenting people with the full spectrum of uncertainty results in poor agreement. It seems that in this formulation our volunteers treated world knowledge questions as unanswerable questions, and were not willing to make any guesses. However, the same data that had poor agreement in the all-questions condition fared much better when the data is split into world knowledge and text-based+unanswerable questions. This is consistent with the reported agreement on datasets like SQuAD and SWAG, which have about the same split, but it reveals two major challenge for QA research:

- The current multi-choice QA task does not reveal the ful spectrum of human capability for uncertainty estimation;
- Given the possible variations due to different data splits and different instructions for the task, how do we ensure fair comparison with NLP system performance?

5 Model Evaluation

Simple heuristics. The simplest baseline of selecting a random choice of 4 options provides 25% accuracy. ***LongChoice*** is a baseline that selects the longest ansert option with a probability of 8/9 (the probability of an answerable question in the train dataset) and the "Not enough information" choice otherwise.

Similarity-based approaches. The first baseline represents a given text, a question, and each of the choices as the average of 300-dimensional CommonCrawl FastText

word embeddings (Bojanowski et al. 2016) of its constituent words. We compute pairwise cosine similarities, and then select the choice with the highest similarity to the input text/question as the prediction. We experimented with selecting the answer based on its similarity to (a) the given text only, (b) the question only, and (c) the average representation of the text and the question, achieving the accuracy of (a) 24.6%, (b) 19.7%, and (c) 22.1%, respectively. In subsequent sections we we report the performance of the third configuration, referred to as ***AvgCos***, as it requires good representations of both the text and the question.

Another baseline used the last hidden state of the bi-directional ***LSTM*** encoding (Graves and Schmidhuber 2005) to represent the texts, questions, and answer options.

$$\begin{aligned}\overleftarrow{\mathbf{h}}(\text{input}) &= \overleftarrow{LSTM}(\text{input}) \\ \overrightarrow{\mathbf{h}}(\text{input}) &= \overrightarrow{LSTM}(\text{input}) \\ \mathbf{h}(\text{input}) &= [\overleftarrow{\mathbf{h}}(\text{input}); \overrightarrow{\mathbf{h}}(\text{input})]\end{aligned} \tag{1}$$

where $\mathbf{h}$ is the hidden state for the last timestep, input $\in \{p, q, c\}$, and p, q, and c denote token-level representations of the passage, the question, and each of the choices, respectively. Pre-trained word embeddings were the input to the model. We used 300-dimensional word embeddings initialized from FastText and 128-dimensional hidden state LSTM vectors. Like before, the final model prediction was produced with cosine similarity using one of the three options:

$$\begin{aligned}&\underset{i}{\text{argmax}}\,[\mathbf{h}(p) \cdot \mathbf{h}(c_i)] \\ &\underset{i}{\text{argmax}}\,[\mathbf{h}(q) \cdot \mathbf{h}(c_i)] \\ &\underset{i}{\text{argmax}}\left[\frac{\mathbf{h}(p) + \mathbf{h}(q)}{2} \cdot \mathbf{h}(c_i)\right]\end{aligned} \tag{2}$$

These configurations achieved 35.0%, 39.4% and 37.2% accuracy, respectively. Further sections report the latter option.

The Pointwise Mutual Information Solver (***PMI***, as described by Clark et al., is a classic similarity-based filter. For a corpus C (the text), and n-grams x and y from the question and answer respectively, the PMI is defined as:

$$\text{PMI}(x, y) = \log \frac{P(x, y)}{P(x)P(y)} \tag{3}$$

$P(x, y)$ is the probability that x and y are found together within a 10-word window in the corpus C. $P(x)$ and $P(y)$ are the probabilities that x, and y independently appear in the corpus, and $P(x)P(y)$ is the expected probability of their co-occurrence. PMI is the ratio of the observed to the expected co-occurrence. The higher it is, the larger the association between x and y. The rank for each answer option is the average of the PMI values for all possible x, y combinations in the question and answer option, and the highest-ranking option is chosen as the prediction. If all options yield zeros, the question is considered unanswerable.

We also experimented with an information retrieval (***IR***) solver, as described by (Clark et al. 2018). IR solver ranks the answer options for each question by using a search engine to find if the question along with the answer are explicitly stated in some sentence in the text. We used elasticsearch as the search engine. The input query is the question concatenated with an answer option, and the result is the score of the top retrieved sentence returned from the search engine. The answer with the highest result is chosen as the prediction. If no results are returned for any option, the question is considered unanswerable.

Integrating external knowledge. We tested the ***TriAN*** model (Wang et al. 2018), which is currently the best-performing system on MCScript dataset and the winner of SemEval 2018 Task 11. At the moment, this dataset is the closest to QuAIL in terms of question format and required reasoning abilities, which makes TriAN a strong baseline for QuAIL. The TriAN architecture incorporates knowledge from the ConceptNet knowledge base (Speer and Havasi 2012). It models the interactions between the question, the text span, and the answer choices, and uses linguistic and handcrafted features. We used the same hyperparameters as reported in the original study (initial learning rate of $2 \cdot 10^{-3}$, batch size of 32, dropout rate of 0.4, and hidden vectors of size 96) trained the TriAN model on our dataset for 50 epochs and recorded test set accuracy for every question type. Experiments with different hyperparameter configurations did not yield significant changes in the output accuracy.

Transformer-based approach. Following the recent state-of-the-art trends of fine-tuning pre-trained language models, we build a baseline that leveraged the base-uncased ***BERT*** model (Devlin et al. 2018) adjusted for the multiple choice question answering task. The particular model we used was originally developed for SWAG (Zellers et al. 2018), a multiple-choice commonsense inference task. This baseline was selected because BERT-based systems currently achieve state-of-the-art on a number of QA datasets, including SQuAD(Rajpurkar et al. 2016), the most popular reading comprehension dataset that includes unanswerable questions, and CoQA (Reddy, Chen, and Manning 2018), the question answering challenge involving answering conversational questions from diverse domains.

For fine-tuning BERT on our dataset, we used a PyTorch implementation[4]. For each answer option, the context, question, and choice are joined and used as input (Radford et al. 2018), and the output is its probability. The most likely option is selected as the answer.

The input had to be large for this task to fit the context, question, and answer. We used 400d input, and as a result, the largest possible batch size was 1. To help with the low batch size, there were 25 gradient accumulation steps. The learning rate was 1e-5, and the vocab size was 30522. We fine-tuned the model on our dataset for 2 epochs.

For TriAn, we did not make any adjustments for the unanswerable questions, i.e. the model would need to learn the meaning of the option "not enough information to answer

[4]https://github.com/huggingface/pytorch-pretrained-BERT

Model	*Text-based questions*				*World knowledge questions*				Unanswe-rable questions	All
	Temp. order	Coref.	Causa-lity	Factual	Subseq. state	Event duration	Entity properties	Belief states		
LongChoice	36.3	32.3	46.8	35.9	29.5	33.6	35.0	30.9	12.2	35.6
AvgCos	13.6	5.9	28.2	6.3	20.0	7.7	27.7	21.8	**65.9**	22.1
LSTM	37.0	32.4	38.5	20.2	36.8	43.6	30.8	34.7	51.8	37.2
PMI	42.5	48.3	57.8	**57.5**	32.9	37.0	33.7	37.5	23.3	41.8
IR	27.9	30.0	42.5	30.8	29.6	35.4	27.5	32.0	28.8	32.4
TriAN	**55.5**	**53.1**	60.1	55.0	47.5	56.9	45.8	43.3	65.0	54.7
BERT	52.9	46.2	**67.1**	55.8	**56.7**	**63.8**	**48.8**	**55.0**	54.2	**55.9**

Table 3: Accuracy of baseline models on QuAIL by question types.

the question". For BERT, we used the strategy that was developed for SQuAD 2.0 (first making a binary prediction for whether or not a given question is answerable, and then trying to answer it or not).

Results

We tested all baselines on 15% of the full QuAIL data (the same amount used for development). The results are shown in Table 3. The top system achieves only 55.9%, confirming that it is a challenging dataset that we hope would considerably raise the bar for the field. Note that TriAN achieves 84% accuracy on MCScript, the most similar existing dataset, and it experiences nearly 30% drop on QuAIL. Similarly, BERT that scores 86.3% on SWAG yields only 55.9% on QuAIL.

The overall pattern of evaluation results is predictable: the BERT-based comes the first, in line with its current domination on the SQUAD leaderboard. TriAN was the close second (Table 3). However, the analysis of behavior of these models enabled by QuAIL question type annotation shows something unexpected:

- TriAN could be expected to work best on the world-knowledge questions, due to ConceptNet. However, it actually worked best for text-based questions, in particular temporal order and coreference questions.
- BERT does the best on the world-knowledge questions and also causality questions, which, as we mentioned in Table 3, often border on world knowledge. That could only be explained by the knowledge it accumulated in pre-training on a large external corpus – which must then be at least as effective as ConceptNet on this task.
- Unanswerable questions were the most reliably detected by the system that simply averaged word embeddings (AvgCos). Recall that QuAIL texts are relatively long, up to 350 words. It seems that the average of all words in a long text becomes meaningless enough to be the most similar to the average of words not found there (*not enough information to answer the question*). This would explain why AvgCos did much worse than chance in all other question types.

Both PMI and IR baselines could be expected to do better on text-based than on world knowledge question, and that is indeed the case. Surprisingly, PMI outperforms both word-embedding-based approaches, and on factoid questions it beats both BERT and TriAN.

Paraphrased data results A big trend in the recent QA research is adversarial attacks on the models that learned to exploit some annotation artifacts or shallow cues. For instance, DROP (Dua et al. 2019) was produced with a model-in-the-loop, with Turkers required to formulate a question that the model could not answer.

Qtype	TriAN	BERT	PMI	IR
Temporal order	0.51 (0.06)	0.24 (-0.25)	0.49 (-0.01)	0.42 (0.09)
Coreference	0.42 (-0.11)	0.31 (-0.13)	0.49 (0.07)	0.24 (-0.11)
Factual	0.32 (-0.21)	0.4 (-0.12)	0.54 (-0.01)	0.26 (-0.09)
Causality	0.33 (-0.2)	0.3 (-0.27)	0.51 (-0.11)	0.39 (0.06)
Subsequent state	0.23 (-0.12)	0.3 (-0.18)	0.33 (-0.04)	0.3 (0.0)
Event duration	0.62 (0.0)	0.48 (-0.13)	0.4 (-0.1)	0.25 (-0.15)
Entity properties	0.31 (-0.06)	0.42 (0.02)	0.38 (-0.03)	0.37 (0.14)
Belief states	0.26 (-0.27)	0.31 (-0.39)	0.39 (-0.08)	0.36 (0.06)
Unanswerable	0.55 (-0.1)	0.48 (-0.13)	0.22 (0.05)	0.29 (-0.15)
All questions	0.4 (-0.11)	0.36 (-0.17)	0.42 (-0.03)	0.32 (-0.02)

Table 4: Evaluation on paraphrased fiction data per question type. The numbers in brackets indicate the difference with respect to non-paraphrased data for the same texts.

In scope of this work, as mentioned in Section 3, we additionally produced an diagnostic dataset of 556 questions for 30 fiction texts. The questions were Turker questions edited by an expert linguist, who developed specific strategies for adversarial rewriting of different question types. The goals were to (a) create distractors likely to be picked up by a model relying on shallow cues, and (b) paraphrase the questions and/or answers so that the span potentially containing the answer would be harder to find. Detailed description of the strategies that were used in the generation of this dataset can be found in the project repository.

The first finding from this experiment is that nothing changed in terms of human evaluation: in the setting where the task is simply to pick one correct option, and all question types are present, Kripperndorff's alpha for a sample of 180 questions was 0.61. As before, the participants had lower scores for world knowledge questions, with the chance-level scores for unanswerable and duration questions. We interpret this as additional evidence of high variability in human world knowledge, which will be a problem for any tests going beyond stereotypical scripts.

However, the models did find the new dataset more challenging. To test the effect, we trained our baselines on training data for all four domains, and tested on either original

fiction data, or the paraphrased versions of the same texts. The results are shown in Table 4.

Both BERT and TriAN worked significantly worse on new data, with BERT losing 17% in accuracy on average. The most significant losses were in questions on temporal order and belief states (nearly 40% drop). Both baselines were at about the same level as IR and PMI, which were less impacted by the paraphrasing (presumably because they did not have as much to lose).

The results of this experiments confirm the known trend of QA models being non-robust to adversarial data, and we hope that the strategies we developed for attacking models trained on QuAIL would be useful in subsequent studies. The fact that human performance did not change much while the model performance dropped significantly also confirm that the current models and humans do not reason in the same way when performing question answering.

6 Discussion

Question Types

Overall Table 3 suggests that the core idea of QuAIL - providing a reading comprehension dataset that is annotated for reasoning types - is working. We can see where the different baselines yield different performance patterns, and this information can provide insights for development and fine-tuning of both QA models and the dataset itself.

For example, analysis by question types suggests that ConceptNet does not provide an edge over BERT pretraining corpus as a source of world knowledge data. It is also clear that PMI is a strong baseline for factoid questions and should not be omitted in studies targeting this kind of RC.

As a quick example of how question type annotation is useful for diagnosing the dataset itself, consider that LongChoice achieves 35.6% accuracy (vs 25% for random choice). It works suspiciously well on causality questions, which suggests an annotation artifact: the Turkers were spending more time on a plausible correct than incorrect answers. Since the "correct option" input field was always presented the first, it may help to randomise the order of input fields (although at the risk that the workers would mix correct and incorrect answers).

Domains

Let us consider the results of BERT on answerable, world-knowledge and unanswerable questions for each genre. We should see fiction and user stories as more difficult than news and blogs, since their contexts can be assumed to be unique.

Question types	Fiction	News	Blogs	User stories
Text-based	45.5	38.8	60.5	**61.6**
World knowledge	**61.0**	58.0	58.3	55.6
Unanswerable	58.3	**68.3**	40.0	50.0
All questions	55.5	52.7	**57.0**	**57.0**

Table 5: BERT accuracy by domains and question types

Table 5 suggests that, in fact, news is a little easier than either blogs and user stories, but BERT is doing suspiciously well on the unanswerable questions here. This points to an annotation artifact: news texts contain more names and words that are hard to paraphrase (e.g. "president", "congresswoman"), and the Turkers were more likely to formulate questions with direct lexical overlaps with the text. This would make unanswerable questions easy to detect as ones that had the least matches with the text.

Are the Models Making the Same Errors?

Finally, let us consider whether our baselines are making the same or different errors on QuAIL. Fig. 2 shows BERT compared to TriAN, PMI, LSTM and IR baselines in terms of shared correct and incorrect answers.

Fig. 2 suggests that there is not much potential for ensembles: no model answers correctly many questions that BERT fails. Furthermore, BERT and TriAN, make the most shared choices (54%, including both correct and incorrect), although they are supposed to use very different reasoning strategies. This is an extra piece of evidence that the ConceptNet knowledge in TriAN is somehow similar to the knowledge BERT learned from the corpus in pre-training.

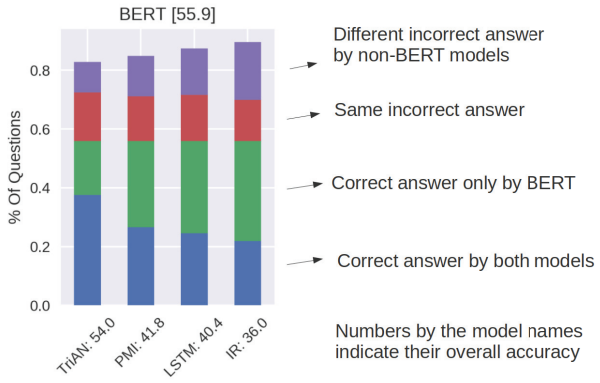

Figure 2: Answer overlap: BERT and other baselines

Challenges

Multi-choice format . Like RocStories, MCScript, and SWAG, QuAIL may be criticized for being a multiple-choice dataset, which may seem "artificial". However, the extractive approach only works for text-based factoid questions. Freeform answers would be the best, but evaluating generated text is arguably the biggest NLP challenge at the moment, and until it is solved, we believe multiple-choice QA is the best compromise.

The participants in our human evaluation rounds reported that they felt the task was unpleasant in forcing them to choose one correct option when they felt that more than one was applicable. Until there is a solution for evaluating generated text, a straightforward extension of the current task would be to simply allow multiple correct answers, with both the models and the people able to select as many as they think were plausible.

Writers seeing the text. QuAIL could also be criticized for the setup in which the writers of the questions see the evidence text, which is "unnatural" (Kwiatkowski et al. 2019). The argument would be that (a) such questions tend to target information from just one part of text rather than aggregated information from several parts of text, and (b) there is danger of copy-pasting or the questions being too similar to the evidence in the text.

However, "natural questions" from search engines are limited to open-world questions with factoid answers. This approach would not work for questions combining context and world knowledge, and especially for texts with unique events (as opposed to Wikipedia). The approach of generating questions based on summaries (e.g. NarrativeQA) precludes any questions about small details or non-major plot events. The most interesting alternative was suggested in QuAC: one Turker tries to learn the information from a text that only the other Turker can see (Choi et al. 2018). However, the authors had to let the "teacher" provide hints about what information could be targeted in the questions, which arguably gives the game away.

We found that the problem with the questions targeting only one small segment of text actually varies by question type: simple factual questions such as *What school did John go to?* do not normally require a large segment of text to answer. On the other hand, questions about causality or temporal order naturally tend to involve longer segments.

Reasoning types. An obvious limitation is that QuAIL does not cover all possible types of verbal reasoning, and, as discussed above, the choice of question types was guided by the considerations of cross-domain balance rather than needs of any particular industrial application. It could be argued that in practice it is desirable to have a training dataset with question type distribution that matches the actual use. QuAIL is aimed at development of general AI rather than industry QA, but we would like to argue that even for downstream applications it is beneficial to have datasets with question type annotation, because they enable tuning the system for a particular use case. For instance, the text-based part of QuAIL could be used for training specifically factoid QA systems.

Question type balance. Our experience with QuAIL confirms the known problem of little variety in Turk data. In particular, we found many cases when two Turkers working on the same text asked essentially the same question (typically something obvious, e.g. a question about the cause of the most salient event in the text). It is not clear how to counteract this, as the saliency of different events and facts comes from the text and will be similar for most Turkers.

Another aspect of this problem became clear from analysis of QuAIL data for unanswerable questions. In theory, unanswerable questions can have all the same types as the answerable ones (temporal order, causality, etc.), and to teach a system to handle them well we would need a representative sample from all these subcategories. Our forms did not prompt for any specific type of unanswerable questions. The result was that the Turkers were mainly asking unanswerable questions about character identity and different facts in the text (presumably because they were the easiest to formulate). This is problematic not only because it creates unbalanced data, but also because the system is getting a large hint: it could simply learn that, say, questions about event duration are mostly answerable, and ignore the "not enough information" option for all of them.

7 Conclusion

We presented QuAIL[5], the first multi-domain text comprehension challenge that is balanced and annotated for 9 types of verbal reasoning. QuAIL aims to show the extent to which current models can generalize over different domains and reasoning strategies and handle questions that can be answered with the information in a given text, unanswerable questions and questions that require extra world knowledge. We hope that QuAIL will stimulate efforts to develop generalist systems tackling different kinds of verbal reasoning, and that it will be useful in diagnostics and qualitative analysis for new QA systems.

8 Acknowledgements

This project is funded in part by an NSF CAREER award to Anna Rumshisky (IIS-1652742).

We would like to express our gratitude to our interns and volunteers in human evaluation rounds: Samuel Dodson, David Donahue, Aleksandr Drozd, Kevin Huang, Marzena Karpinska, Viktoria Khaichuk, Vladislav Lialin, Joseph Rogers, and Sergiusz Rzepkowski.

References

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Chen, D.; Bolton, J.; and Manning, C. D. 2016. A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. In *Proceedings of ACL (Volume 1: Long Papers)*, 2358–2367. ACL.

Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.-t.; Choi, Y.; Liang, P.; and Zettlemoyer, L. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2174–2184. ACL.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv:1803.05457 [cs]*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv:1810.04805 [cs]*.

Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; and Gardner, M. 2019. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. In *Proceedings of NAACL*, 2368–2378.

Fisch, A.; Talmor, A.; Jia, R.; Seo, M.; Choi, E.; and Chen, D. 2019. MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. In *Proc. of the 2nd Workshop on Machine Reading for Question Answering*, 1–13.

[5] http://text-machine.cs.uml.edu/projects/quail/

Geva, M.; Goldberg, Y.; and Berant, J. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *Proc. of EMNLP*.

Graves, A., and Schmidhuber, J. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks* 18(5-6):602–610.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proc. of NAACL*, 107–112.

Hermann, K. M.; Kočiský, T.; Grefenstette, E.; Espeholt, L.; Kay, W.; Suleyman, M.; and Blunsom, P. 2015. Teaching Machines to Read and Comprehend. In *Proceedings of NIPS - Volume 1*, NIPS'15, 1693–1701. MIT Press.

Hill, F.; Bordes, A.; Chopra, S.; and Weston, J. 2015. The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv:1511.02301 [cs]*.

Jia, R., and Liang, P. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2021–2031. ACL.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 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)*, 1601–1611. ACL.

Kocisky, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K. M.; Melis, G.; and Grefenstette, E. 2018. The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics* 6:317–328.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Kelcey, M.; Devlin, J.; Lee, K.; Toutanova, K. N.; Jones, L.; Chang, M.-W.; Dai, A.; Uszkoreit, J.; Le, Q.; and Petrov, S. 2019. Natural Questions: A Benchmark for Question Answering Research. *TACL*.

Labutov, I.; Yang, B.; Prakash, A.; and Azaria, A. 2018. Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 833–844. ACL.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 785–794. ACL.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, 552–561.

Levy, O.; Seo, M.; Choi, E.; and Zettlemoyer, L. 2017. Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, 333–342. ACL.

Mostafazadeh, N.; Roth, M.; Chambers, N.; and Louis, A. 2017. LSDSem 2017 Shared Task: The Story Cloze Test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-Level Semantics*, 46–51. ACL.

Niven, T., and Kao, H.-Y. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4658–4664. ACL.

Ostermann, S.; Roth, M.; Modi, A.; Thater, S.; and Pinkal, M. 2018. SemEval-2018 Task 11: Machine Comprehension Using Commonsense Knowledge. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 747–757. ACL.

Pampari, A.; Raghavan, P.; Liang, J.; and Peng, J. 2018. emrQA: A Large Corpus for Question Answering on Electronic Medical Records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2357–2368. ACL.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. *Preprint*.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2383–2392. ACL.

Rajpurkar, P.; Jia, R.; and Liang, P. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of ACL*, 784–789.

Reddy, S.; Chen, D.; and Manning, C. D. 2018. CoQA: A Conversational Question Answering Challenge. *arXiv:1808.07042 [cs]*.

Speer, R., and Havasi, C. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, 3679–3686.

Talmor, A., and Berant, J. 2019. MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. *arXiv:1905.13453 [cs]*.

Wallace, E.; Feng, S.; Kandpal, N.; Gardner, M.; and Singh, S. 2019. Universal Adversarial Triggers for Attacking and Analyzing NLP. *arXiv:1908.07125 [cs]*.

Wang, L.; Sun, M.; Zhao, W.; Shen, K.; and Liu, J. 2018. Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension. *arXiv preprint arXiv:1803.00191*.

Weston, J.; Bordes, A.; Chopra, S.; Rush, A. M.; van Merriënboer, B.; Joulin, A.; and Mikolov, T. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of EMNLP*, 2369–2380. ACL.

Zellers, R.; Bisk, Y.; Schwartz, R.; and Choi, Y. 2018. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 93–104. ACL.