BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions

Christopher Clark[*1], **Kenton Lee**[†], **Ming-Wei Chang**[†], **Tom Kwiatkowski**[†]

Michael Collins [†2], **Kristina Toutanova**[†]

[*]Paul G. Allen School of CSE, University of Washington
csquared@cs.uw.edu

[†]Google AI Language
{kentonl, mingweichang, tomkwiat, mjcollins, kristout}@google.com

Abstract

In this paper we study yes/no questions that are *naturally occurring* — meaning that they are generated in unprompted and unconstrained settings. We build a reading comprehension dataset, BoolQ, of such questions, and show that they are unexpectedly challenging. They often query for complex, non-factoid information, and require difficult entailment-like inference to solve. We also explore the effectiveness of a range of transfer learning baselines. We find that transferring from entailment data is more effective than transferring from paraphrase or extractive QA data, and that it, surprisingly, continues to be very beneficial even when starting from massive pre-trained language models such as BERT. Our best method trains BERT on MultiNLI and then re-trains it on our train set. It achieves 80.4% accuracy compared to 90% accuracy of human annotators (and 62% majority-baseline), leaving a significant gap for future work.

1 Introduction

Understanding what facts can be inferred to be true or false from text is an essential part of natural language understanding. In many cases, these inferences can go well beyond what is immediately stated in the text. For example, a simple sentence like "Hanna Huyskova won the gold medal for Belarus in freestyle skiing." implies that (1) Belarus is a country, (2) Hanna Huyskova is an athlete, (3) Belarus won at least one Olympic event, (4) the USA did *not* win the freestyle skiing event, and so on.

To test a model's ability to make these kinds of inferences, previous work in natural language inference (NLI) proposed the task of labeling candidate statements as being entailed or contradicted by a given passage. However, in practice, generating candidate statements that test for complex inferential abilities is challenging. For instance, evidence suggests (Gururangan et al., 2018; Jia and Liang, 2017; McCoy et al., 2019) that simply asking human annotators to write candidate statements will result in examples that typically only require surface-level reasoning.

In this paper we propose an alternative: we test models on their ability to answer naturally occurring yes/no questions. That is, questions that were authored by people who were not prompted to write particular kinds of questions, including even being required to write yes/no questions, and who did not know the answer to the question they were asking. Figure 1 contains some examples from our dataset. We find such questions often query for

Q:	Has the UK been hit by a hurricane?
P:	The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands ...
A:	Yes. [An example event is given.]
Q:	Does France have a Prime Minister and a President?
P:	... The extent to which those decisions lie with the Prime Minister or President depends upon ...
A:	Yes. [Both are mentioned, so it can be inferred both exist.]
Q:	Have the San Jose Sharks won a Stanley Cup?
P:	... The Sharks have advanced to the Stanley Cup finals once, losing to the Pittsburgh Penguins in 2016 ...
A:	No. [They were in the finals once, and lost.]

Figure 1: Example yes/no questions from the BoolQ dataset. Each example consists of a question (**Q**), an excerpt from a passage (**P**), and an answer (**A**) with an explanation added for clarity.

[1]Work completed while interning at Google.

[2]Also affiliated with Columbia University, work done at Google.

non-factoid information, and that human annotators need to apply a wide range of inferential abilities when answering them. As a result, they can be used to construct highly inferential reading comprehension datasets that have the added benefit of being directly related to the practical end-task of answering user yes/no questions.

Yes/No questions do appear as a subset of some existing datasets (Reddy et al., 2018; Choi et al., 2018; Yang et al., 2018). However, these datasets are primarily intended to test other aspects of question answering (QA), such as conversational QA or multi-step reasoning, and do not contain naturally occurring questions.

We follow the data collection method used by Natural Questions (NQ) (Kwiatkowski et al., 2019) to gather 16,000 naturally occurring yes/no questions into a dataset we call BoolQ (for Boolean Questions). Each question is paired with a paragraph from Wikipedia that an independent annotator has marked as containing the answer. The task is then to take a question and passage as input, and to return "yes" or "no" as output. Figure 1 contains some examples, and Appendix A.1 contains additional randomly selected examples.

Following recent work (Wang et al., 2018), we focus on using transfer learning to establish baselines for our dataset. Yes/No QA is closely related to many other NLP tasks, including other forms of question answering, entailment, and paraphrasing. Therefore, it is not clear what the best data sources to transfer from are, or if it will be sufficient to just transfer from powerful pre-trained language models such as BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018). We experiment with state-of-the-art unsupervised approaches, using existing entailment datasets, three methods of leveraging extractive QA data, and using a few other supervised datasets.

We found that transferring from MultiNLI, and the unsupervised pre-training in BERT, gave us the best results. Notably, we found these approaches are surprisingly complementary and can be combined to achieve a large gain in performance. Overall, our best model reaches 80.43% accuracy, compared to 62.31% for the majority baseline and 90% human accuracy. In light of the fact BERT on its own has achieved human-like performance on several NLP tasks, this demonstrates the high degree of difficulty of our dataset. We present our data and code at `https://goo.gl/boolq`.

2 Related Work

Yes/No questions make up a subset of the reading comprehension datasets CoQA (Reddy et al., 2018), QuAC (Choi et al., 2018), and HotPotQA (Yang et al., 2018), and are present in the ShARC (Saeidi et al., 2018) dataset. These datasets were built to challenge models to understand conversational QA (for CoQA, ShARC and QuAC) or multi-step reasoning (for HotPotQA), which complicates our goal of using yes/no questions to test inferential abilities. Of the four, QuAC is the only one where the question authors were not allowed to view the text being used to answer their questions, making it the best candidate to contain naturally occurring questions. However, QuAC still heavily prompts users, including limiting their questions to be about pre-selected Wikipedia articles, and is highly class imbalanced with 80% "yes" answers.

The MS Marco dataset (Nguyen et al., 2016), which contains questions with free-form text answers, also includes some yes/no questions. We experiment with heuristically identifying them in Section 4, but this process can be noisy and the quality of the resulting annotations is unknown. We also found the resulting dataset is class imbalanced, with 80% "yes" answers.

Yes/No QA has been used in other contexts, such as the templated bAbI stories (Weston et al., 2015) or some Visual QA datasets (Antol et al., 2015; Wu et al., 2017). We focus on answering yes/no questions using natural language text.

Question answering for reading comprehension in general has seen a great deal of recent work (Rajpurkar et al., 2016; Joshi et al., 2017), and there have been many recent attempts to construct QA datasets that require advanced reasoning abilities (Yang et al., 2018; Welbl et al., 2018; Mihaylov et al., 2018; Zellers et al., 2018; Zhang et al., 2018). However, these attempts typically involve engineering data to be more difficult by, for example, explicitly prompting users to write multi-step questions (Yang et al., 2018; Mihaylov et al., 2018), or filtering out easy questions (Zellers et al., 2018). This risks resulting in models that do not have obvious end-use applications since they are optimized to perform in an artificial setting. In this paper, we show that yes/no questions have the benefit of being very challenging even when they are gathered from natural sources.

Natural language inference is also a well

studied area of research, particularly on the MultiNLI (Williams et al., 2018) and SNLI (Bowman et al., 2015) datasets. Other sources of entailment data include the PASCAL RTE challenges (Bentivogli et al., 2009, 2011) or SciTail (Khot et al., 2018). We note that, although SciTail, RTE-6 and RTE-7 did not use crowd workers to generate candidate statements, they still use sources (multiple choices questions or document summaries) that were written by humans with knowledge of the premise text. Using naturally occurring yes/no questions ensures even greater independence between the questions and premise text, and ties our dataset to a clear end-task. BoolQ also requires detecting entailment in paragraphs instead of sentence pairs.

Transfer learning for entailment has been studied in GLUE (Wang et al., 2018) and SentEval (Conneau and Kiela, 2018). Unsupervised pre-training in general has recently shown excellent results on many datasets, including entailment data (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018).

Converting short-answer or multiple choice questions into entailment examples, as we do when experimenting with transfer learning, has been proposed in several prior works (Demszky et al., 2018; Poliak et al., 2018; Khot et al., 2018). In this paper we found some evidence suggesting that these approaches are less effective than using crowd-sourced entailment examples when it comes to transferring to natural yes/no questions.

Contemporaneously with our work, Phang et al. (2018) showed that pre-training on supervised tasks could be beneficial even when using pre-trained language models, especially for a textual entailment task. Our work confirms these results for yes/no question answering.

This work builds upon the Natural Questions (NQ) (Kwiatkowski et al., 2019), which contains some natural yes/no questions. However, there are too few (about 1% of the corpus) to make yes/no QA a very important aspect of that task. In this paper, we gather a large number of additional yes/no questions in order to construct a dedicated yes/no QA dataset.

3 The BoolQ Dataset

An example in our dataset consists of a question, a paragraph from a Wikipedia article, the title of the article, and an answer, which is either "yes" or "no". We include the article title since it can potentially help resolve ambiguities (e.g., coreferent phrases) in the passage, although none of the models presented in this paper make use of them.

3.1 Data Collection

We gather data using the pipeline from NQ (Kwiatkowski et al., 2019), but with an additional filtering step to focus on yes/no questions. We summarize the complete pipeline here, but refer to their paper for a more detailed description.

Questions are gathered from anonymized, aggregated queries to the Google search engine. Queries that are likely to be yes/no questions are heuristically identified: we found selecting queries where the first word is in a manually constructed set of indicator words[3] and are of sufficient length, to be effective.

Questions are only kept if a Wikipedia page is returned as one of the first five results, in which case the question and Wikipedia page are given to a human annotator for further processing.

Annotators label question/article pairs in a three-step process. First, they decide if the question is *good*, meaning it is comprehensible, unambiguous, and requesting factual information. This judgment is made before the annotator sees the Wikipedia page. Next, for good questions, annotators find a passage within the document that contains enough information to answer the question. Annotators can mark questions as "not answerable" if the Wikipedia article does not contain the requested information. Finally, annotators mark whether the question's answer is "yes" or "no". Annotating data in this manner is quite expensive since annotators need to search entire Wikipedia documents for relevant evidence and read the text carefully.

Note that, unlike in NQ, we only use questions that were marked as having a yes/no answer, and pair each question with the selected passage instead of the entire document. This helps reduce ambiguity (ex., avoiding cases where the document supplies conflicting answers in different paragraphs), and keeps the input small enough so that existing entailment models can easily be applied to our dataset.

We combine 13k questions gathered from this

[3]The full set is: {"did", "do", "does", "is", "are", "was", "were", "have", "has", "can", "could", "will", "would"}.

Question Topic			
Category	Example	Percent	Yes%
Entertainment Media	Is You and I by Lady Gaga a cover?	22.0	65.9
Nature/Science	Are there blue whales in the Atlantic Ocean?	22.0	56.8
Sports	Has the US men's team ever won the World Cup?	11.0	54.5
Law/Government	Is there a seat belt law in New Hampshire?	10.0	70.0
History	Were submarines used in the American Civil War?	5.0	70.0
Fictional Events	Is the Incredible Hulk part of the avengers?	4.0	87.5
Other	Is GDP per capita same as per capita income?	26.0	65.4
Question Type			
Category	Example	Percent	Yes%
Definitional	Is thread seal tape the same as Teflon tape?	14.5	55.2
Existence	Is there any dollar bill higher than a 100?	14.5	69.0
Event Occurrence	Did the great fire of London destroy St. Paul's Cathedral?	11.5	73.9
Other General Fact	Is there such thing as a dominant eye?	29.5	62.7
Other Entity Fact	Is the Arch in St. Louis a national park?	30.0	63.3

Table 1: Question categorization of BoolQ. Question topics are shown in the top half and question types are shown in the bottom half.

pipeline with an additional 3k questions with yes/no answers from the NQ training set to reach a total of 16k questions. We split these questions into a 3.2k dev set, 3.2k test set, and 9.4k train set, ensuring questions from NQ are always in the train set. "Yes" answers are slightly more common (62.31% in the train set). The queries are typically short (average length 8.9 tokens) with longer passages (average length 108 tokens).

3.2 Analysis

In the following section we analyze our dataset to better understand the nature of the questions, the annotation quality, and the kinds of reasoning abilities required to answer them.

3.3 Annotation Quality

First, in order to assess annotation quality, three of the authors labelled 110 randomly chosen examples. If there was a disagreement, the authors conferred and selected a single answer by mutual agreement. We call the resulting labels "gold-standard" labels. On the 110 selected examples, the answer annotations reached 90% accuracy compared to the gold-standard labels. Of the cases where the answer annotation differed from the gold-standard, six were ambiguous or debatable cases, and five were errors where the annotator misunderstood the passage. Since the agreement was sufficiently high, we elected to use singly-annotated examples in the training/dev/test sets in order to be able to gather a larger dataset.

3.4 Question Types

Part of the value of this dataset is that it contains questions that people genuinely want to answer. To explore this further, we manually define a set of topics that questions can be about. An author categorized 200 questions into these topics. The results can be found in the upper half of Table 1.

Questions were often about entertainment media (including T.V., movies, and music), along with other popular topics like sports. However, there are still a good portion of questions asking for more general factual knowledge, including ones about historical events or the natural world.

We also broke the questions into categories based on what kind of information they were requesting, shown in the lower half of Table 1. Roughly one-sixth of the questions are about whether anything with a particular property exists (Existence), another sixth are about whether a particular event occurred (Event Occurrence), and another sixth ask whether an object is known by a particular name, or belongs to a particular category (Definitional). The questions that do not fall into these three categories were split between requesting facts about a specific entity, or requesting more general factual information.

We do find a correlation between the nature of the question and the likelihood of a "yes" answer. However, this correlation is too weak to help outperform the majority baseline because, even if the topic or type is known, it is never best to guess the minority class. We also found that question-only models perform very poorly on this task (see Section 5.3), which helps confirm that the questions

Reasoning Types		Yes/No Question Answering Examples
Paraphrasing (38.7%) The passage explicitly asserts or refutes what is stated in the question.	**Q**: **P**: **A**:	Is Tim Brown in the Hall of Fame? Brown has also played for the Tampa Bay Buccaneers. In 2015, he was inducted into the Pro Football Hall of Fame. Yes. ["inducted into" directly implies he is in Hall of Fame.]
By Example (11.8%) The passage provides an example or counter-example to what is asserted by the question.	**Q**: **P**: **A**:	Are there any nuclear power plants in Michigan? . . . three nuclear power plants supply Michigan with about 30% of its electricity. Yes. [Since there must be at least three.]
Factual Reasoning (8.5%) Answering the question requires using world-knowledge to connect what is stated in the passage to the question.	**Q**: **P**: **A**:	Was designated survivor filmed in the White House? The series is. . . filmed in Toronto, Ontario. No. [The White House is not located in Toronto.]
Implicit (8.5%) The passage mentions or describes entities in the question in way that would not make sense if the answer was not yes/no.	**Q**: **P**: **A**:	Is static pressure the same as atmospheric pressure? The aircraft designer's objective is to ensure the pressure in the aircraft's static pressure system is as close as possible to the atmospheric pressure. . . No. [It would not make sense to bring them "as close as possible" if those terms referred to the same thing.]
Missing Mention (6.6%) We can conclude the answer is yes or no because, if this was not the case, it would have been mentioned in the passage.	**Q**: **P**: **A**:	Did Bonnie Blair's daughter make the Olympic team? Blair and Cruikshank have two children: a son, Grant, and daughter, Blair.... Blair Cruikshank competed at the 2018 United States Olympic speed skating trials at the 500 meter distance. No. [The passage describes Blair Cruikshank's daughter's skating accomplishments, so it would have mentioned it if she had qualified.]
Other Inference (25.9%) The passage states a fact that can be used to infer whether the answer is true or false, and does not fall into any of the other categories.	**Q**: **P**: **A**:	Is the sea snake the most venomous snake? . . . the venom of the inland taipan, drop by drop, is the most toxic among all snakes No. [If inland taipan is the most venomous snake, the sea snake must not be.]

Table 2: Kinds of reasoning needed in the BoolQ dataset.

do not contain sufficient information to predict the answer on their own.

3.5 Types of Inference

Finally, we categorize the kinds of inference required to answer the questions in BoolQ[4]. The definitions and results are shown in Table 2.

Less than 40% of the examples can be solved by detecting paraphrases. Instead, many questions require making additional inferences (categories "Factual Reasoning", "By Example", and "Other Inference") to connect what is stated in the passage to the question. There is also a significant class of questions (categories "Implicit" and "Missing Mention") that require a subtler kind of inference based on how the passage is written.

3.6 Discussion

Why do natural yes/no questions require inference so often? We hypothesize that there are several factors. First, we notice factoid questions that ask about simple properties of entities, such as "Was Obama born in 1962?", are rare. We suspect this is because people will almost always prefer to phrase such questions as short-answer questions (e.g., "When was Obama born?"). Thus, there is a natural filtering effect where people tend to use yes/no questions exactly when they want more complex kinds of information.

Second, both the passages and questions rarely include negation. As a result, detecting a "no" answer typically requires understanding that a positive assertion in the text excludes, or makes unlikely, a positive assertion in the question. This requires reasoning that goes beyond paraphrasing (see the "Other-Inference" or "Implicit" examples).

We also think it was important that annotators only had to answer questions, rather than generate them. For example, imagine trying to construct questions that fall into the categories of "Missing Mention" or "Implicit". While possible, it would require a great deal of thought and creativity. On the other hand, detecting when a yes/no question can be answered using these strategies seems much easier and more intuitive. Thus, having annotators answer pre-existing questions opens the door to building datasets that contain more inference and have higher quality labels.

[4]Note the dataset has been updated since we carried out this analysis, so it might be slighly out-of-date.

4 Training Yes/No QA Models

Models on this dataset need to predict an output class given two pieces of input text, which is a well studied paradigm (Wang et al., 2018). We find training models on our train set alone to be relatively ineffective. Our best model reaches 69.6% accuracy, only 8% better than the majority baseline. Therefore, we follow the recent trend in NLP of using transfer learning. In particular, we experiment with *pre-training* models on related tasks that have larger datasets, and then *fine-tuning* them on our training data. We list the sources we consider for pre-training below.

Entailment: We consider two entailment datasets, *MultiNLI* (Williams et al., 2018) and *SNLI* (Bowman et al., 2015). We choose these datasets since they are widely-used and large enough to use for pre-training. We also experiment with ablating classes from MultiNLI. During fine-tuning we use the probability the model assigns to the "entailment" class as the probability of predicting a "yes" answer.

Multiple-Choice QA: We use a multiple choice reading comprehension dataset, *RACE* (Lai et al., 2017), which contains stories or short essays paired with questions built to test the reader's comprehension of the text. Following what was done in SciTail (Khot et al., 2018), we convert questions and answer-options to statements by either substituting the answer-option for the blanks in fill-in-the-blank questions, or appending a separator token and the answer-option to the question. During training, we have models independently assign a score to each statement, and then apply the softmax operator between all statements per each question to get statement probabilities. We use the negative log probability of the correct statement as a loss function. To fine-tune on BoolQ, we apply the sigmoid operator to the score of the question given its passage to get the probability of a "yes" answer.

Extractive QA: We consider several methods of leveraging extractive QA datasets, where the model must answer questions by selecting text from a relevant passage. Preliminary experiments found that simply transferring the lower-level weights of extractive QA models was ineffective, so we instead consider three methods of constructing entailment-like data from extractive QA data.

First, we use the *QNLI* task from GLUE (Wang et al., 2018), where the model must determine if a sentence from SQuAD 1.1 (Rajpurkar et al., 2016) contains the answer to an input question or not. Following previous work (Hu et al., 2018), we also try building entailment-like training data from *SQuAD 2.0* (Rajpurkar et al., 2018). We concatenate questions with either the correct answer, or with the incorrect "distractor" answer candidate provided by the dataset, and train the model to classify which is which given the question's supporting text.

Finally, we also experiment with leveraging the long-answer portion of NQ, where models must select a paragraph containing the answer to a question from a document. Following our method for Multiple-Choice QA, we train a model to assign a score to (question, paragraph) pairs, apply the softmax operator on paragraphs from the same document to get a probability distribution over the paragraphs, and train the model on the negative log probability of selecting an answer-containing paragraph. We only train on questions that were marked as having an answer, and select an answer-containing paragraph and up to 15 randomly chosen non-answer-containing paragraphs for each question. On BoolQ, we compute the probability of a "yes" answer by applying the sigmoid operator to the score the model gives to the input question and passage.

Paraphrasing: We use the Quora Question Paraphrasing (*QQP*) dataset, which consists of pairs of questions labelled as being paraphrases or not.[5] Paraphrasing is related to entailment since we expect, at least in some cases, passages will contain a paraphrase of the question.

Heuristic Yes/No: We attempt to heuristically construct a corpus of yes/no questions from the MS Marco corpus (Nguyen et al., 2016). MS Marco has free-form answers paired with snippets of related web documents. We search for answers starting with "yes" or "no", and then pair the corresponding questions with snippets marked as being related to the question. We call this task *Y/N MS Marco*; in total we gather 38k examples,

[5] data.quora.com/First-Quora-Dataset-Release-Question-Pairs

80% of which are "yes" answers.

Unsupervised: It is well known that unsupervised pre-training using language-modeling objectives (Peters et al., 2018; Devlin et al., 2018; Radford et al., 2018), can improve performance on many tasks. We experiment with these methods by using the pre-trained models from *ELMo*, *BERT*, and OpenAI's Generative Pre-trained Transformer (*OpenAI GPT*) (see Section 5.2).

5 Results

5.1 Shallow Models

First, we experiment with using a linear classifier on our task. In general, we found features such as word overlap or TF-IDF statistics were not sufficient to achieve better than the majority-class baseline accuracy (62.17% on the dev set). We did find there was a correlation between the number of times question words occurred in the passage and the answer being "yes", but the correlation was not strong enough to build an effective classifier. "Yes" is the most common answer even among questions with zero shared words between the question and passage (with a 51% majority), and more common in other cases.

5.2 Neural Models

For our experiments that do not use unsupervised pre-training (except the use of pre-trained word vectors), we use a standard recurrent model with attention. Our experiments using unsupervised pre-training use the models provided by the authors. In more detail:

Our *Recurrent* model follows a standard recurrent plus attention architecture for text-pair classification (Wang et al., 2018). It embeds the premise/hypothesis text using fasttext word vectors (Mikolov et al., 2018) and learned character vectors, applies a shared bidirectional LSTM to both parts, applies co-attention (Parikh et al., 2016) to share information between the two parts, applies another bi-LSTM to both parts, pools the result, and uses the pooled representation to predict the final class. See Appendix A.2 for details.

Our *Recurrent +ELMo* model uses the language model from Peters et al. (2018) to provide contextualized embeddings to the baseline model outlined above, as recommended by the authors.

Our *OpenAI GPT* model fine-tunes the 12 layer 768 dimensional uni-directional transformer from Radford et al. (2018), which has been pre-trained as a language model on the Books corpus (Zhu et al., 2015).

Our $BERT_L$ model fine-tunes the 24 layer 1024 dimensional transformer from Devlin et al. (2018), which has been trained on next-sentence-selection and masked language modelling on the Book Corpus and Wikipedia.

We fine-tune the BERT_L and the OpenAI GPT models using the optimizers recommended by the authors, but found it important to tune the optimization parameters to achieve the best results. We use a batch size of 24, learning rate of 1e-5, and 5 training epochs for BERT and a learning rate of 6.25e-5, batch size of 6, language model loss of 0.5, and 3 training epochs for OpenAI GPT.

5.3 Question/Passage Only Results

Following the recommendation of Gururangan et al. (2018), we first experiment with models that are only allowed to observe the question or the passage. The pre-trained BERT_L model reached 64.48% dev set accuracy using just the question and 66.74% using just the passage. Given that the majority baseline is 62.17%, this suggests there is little signal in the question by itself, but that some language patterns in the passage correlate with the answer. Possibly, passages that present more straightforward factual information (like Wikipedia introduction paragraphs) correlate with "yes" answers.

5.4 Transfer Learning Results

The results of our transfer learning methods are shown in Table 3. All results are averaged over five runs. For models pre-trained on supervised datasets, both the pre-training and the fine-tuning stages were repeated. For unsupervised pre-training, we use the pre-trained models provided by the authors, but continue to average over five runs of fine-tuning.

QA Results: We were unable to transfer from RACE or SQuAD 2.0. For RACE, the problem might be domain mismatch. In RACE the passages are stories, and the questions often query for passage-specific information such as the author's intent or the state of a particular entity from the passage, instead of general knowledge.

We would expect SQuAD 2.0 to be a better match for BoolQ since it is also Wikipedia-based, but its possible detecting the adversarially-

Transfer Task	Model	Transfer Data	#Examples	Source Acc.	BoolQ Acc.
N/A	Majority	-	-	-	62.17
N/A	Recurrent	-	-	-	69.60
Extractive QA	Recurrent	QNLI	108k	79.66	71.36
		SQuAD 2.0	130k	69.45	69.83
		NQ Long Answer	93k	71.78	72.78
Paraphrasing	Recurrent	QQP	364k	89.58	71.30
Heuristic Y/N	Recurrent	Y/N MS Marco	39k	87.26	71.40
Entailment	Recurrent	MultiNLI	392k	78.23	75.57
		- w/o Entail	262k	84.26	72.95
		- w/o Contradict	262k	81.16	72.85
		- w/o Neutral	262k	89.72	74.83
		SNLI	351k	88.17	73.16
MC QA	Recurrent	RACE	549k	42.30	68.40
Unsupervised	Recurrent +ELMo	Billion Word	1000M	-	71.41
	OpenAI GPT	Books	800M	-	72.87
	BERT_L	Books/Wikipedia	3,300M	-	76.90

Table 3: Transfer learning results on the BoolQ dev set after fine-tuning on the BoolQ training set. Results are averaged over five runs. In all cases directly using the pre-trained model without fine-tuning did not achieve results better than the majority baseline, so we do not include them here.

constructed distractors used for negative examples does not relate well to yes/no QA.

We got better results using QNLI, and even better results using NQ. This shows the task of selecting text relevant to a question is partially transferable to yes/no QA, although we are only able to gain a few points over the baseline.

Entailment Results: The MultiNLI dataset out-performed all other supervised methods by a large margin. Remarkably, this approach is only a few points behind BERT despite using orders of magnitude less training data and a much more light-weight model, showing high-quality pre-training data can help compensate for these deficiencies.

Our ablation results show that removing the neutral class from MultiNLI hurt transfer slightly, and removing either of the other classes was very harmful, suggesting the neutral examples had limited value. SNLI transferred better than other datasets, but worse than MultiNLI. We suspect this is due to limitations of the photo-caption domain it was constructed from.

Other Supervised Results: We obtained a small amount of transfer using QQP and Y/N MS Marco. Although Y/N MS Marco is a yes/no QA dataset, its small size and class imbalance likely contributed to its limited effectiveness. The web snippets it uses as passages also present a large domain shift from the Wikipedia passages in BoolQ.

Model	Dev Acc.	Test Acc.
Majority Class	62.17	62.31
Recurrent	70.28	67.52
+MultiNLI	76.15	74.24
Pre-trained BERT_L	78.09	76.70
+MultiNLI	82.20	80.43

Table 4: Test set results on BoolQ, "+MultiNLI" indicates models that were additionally pre-trained on MultiNLI before being fine-tuned on the train set.

Unsupervised Results: Following results on other datasets (Wang et al., 2018), we found BERT_L to be the most effective unsupervised method, surpassing all other methods of pre-training.

5.5 Multi-Step Transfer Results

Our best single-step transfer learning results were from using the pre-trained BERT_L model and

MultiNLI. We also experiment with combining these approaches using a two-step pre-training regime. In particular, we fine-tune the pre-trained BERT_L on MultiNLI, and then fine-tune the resulting model again on the BoolQ train set. We found decreasing the number of training epochs to 3 resulted in a slight improvement when using the model pre-trained on MultiNLI.

We show the test set results for this model, and some other pre-training variations, in Table 4. For these results we train five versions of each model using different training seeds, and show the model that had the best dev-set performance.

Given how extensively the BERT_L model has been pre-trained, and how successful it has been across many NLP tasks, the additional gain of 3.5 points due to using MultiNLI is remarkable. This suggests MultiNLI contains signal orthogonal to what is found in BERT's unsupervised objectives.

5.6 Sample Efficiency

In Figure 2, we graph model accuracy as more of the training data is used for fine-tuning, both with and without initially pre-training on MultiNLI. Pre-training on MultiNLI gives at least a 5-6 point gain, and nearly a 10 point gain for BERT_L when only using 1000 examples. For small numbers of examples, the recurrent model with MultiNLI pre-training actually out-performs BERT_L.

5.7 Discussion

A surprising result from our work is that the datasets that more closely resemble the format of BoolQ, meaning they contain questions and multi-sentence passages, such as SQuAD 2.0, RACE, or Y/N MS Marco, were not very useful for transfer. The entailment datasets were stronger despite consisting of sentence pairs. This suggests that adapting from sentence-pair input to question/passage input was not a large obstacle to achieving transfer. Preliminary work found attempting to convert the yes/no questions in BoolQ into declarative statements did not improve transfer from MultiNLI, which supports this hypothesis.

The success of MultiNLI might also be surprising given recent concerns about the generalization abilities of models trained on it (Glockner et al., 2018), particularly related to "annotation artifacts" caused by using crowd workers to write the hypothesis statements (Gururangan et al., 2018). We have shown that, despite these weaknesses, it can still be an important starting point for models being used on natural data.

We hypothesize that a key advantage of MultiNLI is that it contains examples of contradictions. The other sources of transfer we consider, including the next-sentence-selection objective in BERT, are closer to providing examples of entailed text vs. neutral/unrelated text. Indeed, we found that our two step transfer procedure only reaches 78.43% dev set accuracy if we remove the contradiction class from MultiNLI, regressing its performance close to the level of BERT_L when just using unsupervised pre-training.

Note that it is possible to pre-train a model on several of the suggested datasets, either in succession or in a multi-task setup. We leave these experiments to future work. Our results also suggest pre-training on MultiNLI would be helpful for other corpora that contain yes/no questions.

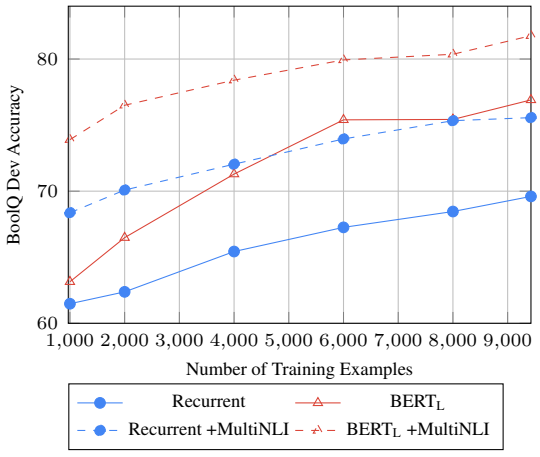

Figure 2: Accuracy for various models on the BoolQ dev set as the number of training examples varies.

6 Conclusion

We have introduced BoolQ, a new reading comprehension dataset of naturally occurring yes/no questions. We have shown these questions are challenging and require a wide range of inference abilities to solve. We have also studied how transfer learning performs on this task, and found crowd-sourced entailment datasets can be leveraged to boost performance even on top of language model pre-training. Future work could include building a document-level version of this task, which would increase its difficulty and its correspondence to an end-user application.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *Proceedings of the IEEE international conference on computer vision*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The Sixth PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2011. The Seventh PASCAL Recognizing Textual Entailment Challenge. In *TAC*.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A Large Annotated Corpus for Learning Natural Language Inference. In *EMNLP*.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced LSTM for Natural Language Inference. In *ACL*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuaC: Question Answering in Context. In *EMNLP*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An Evaluation Toolkit for Universal Sentence Representations. In *LREC*.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming Question Answering Datasets Into Natural Language Inference Datasets. *Computing Research Repository*, arXiv:1809.02922. Version 2.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Computing Research Repository*, arXiv:1810.04805. Version 1.

Max Glockner, Vered Shwartz, and Yoav Goldberg. 2018. Breaking NLI Systems with Sentences that Require Simple Lexical Inferences. In *ACL*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *NAACL*.

Minghao Hu, Yuxing Peng, Zhen Huang, Nan Yang, Ming Zhou, et al. 2018. Read+ Verify: Machine Reading Comprehension with Unanswerable Questions. In *CoRR*.

Robin Jia and Percy Liang. 2017. Adversarial Examples for Evaluating Reading Comprehension Systems. In *EMNLP*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *ACL*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A Textual Entailment Dataset from Science Question Answering. In *AAAI*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *ICLR*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. In *TACL*.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-Scale Reading Comprehension Dataset from Examinations. In *EMNLP*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *Computing Research Repository*, arXiv:1902.01007. Version 1.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *EMNLP*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in Pre-Training Distributed Word Representations. In *LREC*.

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. *Computing Research Repository*, arXiv:1611.09268. Version 3.

Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A Decomposable Attention Model for Natural Language Inference. In *EMNLP*.

Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *NAACL*.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence Encoders on STILTs: Supplementary Training on Intermediate Labeled-data Tasks. *Computing Research Repository*, arXiv:1811.01088. Version 2.

Adam Poliak, Aparajita Haldar, Rachel Rudinger, J Edward Hu, Ellie Pavlick, Aaron Steven White, and Benjamin Van Durme. 2018. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *EMNLP*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-training.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *ACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ Questions for Machine Comprehension of Text. In *EMNLP*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. CoQA: A Conversational Question Answering Challenge. In *TACL*.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *EMNLP*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional Attention Flow for Machine Comprehension. In *ICLR*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *EMNLP*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing Datasets for Multi-hop Reading Comprehension Across Documents. In *ACL*.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In *ICLR*.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In *NAACL*.

Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual Question Answering: A Survey of Methods and Datasets. In *Computer Vision and Image Understanding*. Elsevier.

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 *EMNLP*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. Swag: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *EMNLP*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018. ReCoRD: Bridging the Gap between Human and Machine Commonsense Reading Comprehension. *Computing Research Repository*, arXiv:1810.12885. Version 1.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

A Appendices

A.1 Randomly Selected Examples

We include a number of randomly selected examples from the BoolQ train set in Figure 3. For each example we show the question in bold, followed by the answer in parentheses, and then the passage below.

A.2 Recurrent Model

Our recurrent model is a standard model from the text pair classification literature, similar to the one used in the GLUE baseline (Wang et al., 2018) and the model from Chen et al. (2017). Our model has the following stages:

Embed: Embed the words using a character CNN following what was done by Seo et al. (2017), and the fasttext crawl word embeddings (Mikolov et al., 2018). Then run a BiLSTM over the results to get context-aware word hypothesis embeddings $\langle u_1, u_2, u_3, ... \rangle$ and premise embeddings $\langle v_1, v_2, v_3, ... \rangle$.

Co-Attention: Compute a co-attention matrix, A, between the hypothesis and premise where $A_{ij} = w_1 \cdot u_i + w_2 \cdot v_j + w_3 \cdot (u_i \circ v_j)$, $\circ$ is elementwise multiplication, and w_1, w_2, and w_3 are weights to be learned.

Attend: For each row in A, apply the softmax operator and use the results to compute a weighed sum of the hypothesis embeddings, resulting in attended vectors $\langle \tilde{u}_1, \tilde{u}_2, ... \rangle$. We use the transpose of A to compute vectors $\langle \tilde{v}_1, \tilde{v}_2, ... \rangle$ from the premise embeddings in a similar manner.

Pool: Run another BiLSTM over $\langle [v_1; \tilde{v}_1; \tilde{v}_1 \circ v_1], [v_2; \tilde{v}_2; \tilde{v}_2 \circ v_2], ... \rangle$ to get embeddings $\langle h_1, h_2, ... \rangle$. Then pool these embeddings by computing attention scores $a_i = w \cdot h_i$, $p = softmax(a)$, and then the sum $v^* = \sum_i p_i h_i$. Likewise we compute p^* from the premise.

Classify: Finally we feed $[v^*; p^*]$ into a fully

connected layer, and then through a softmax layer to predict the output class.

We apply dropout at a rate of 0.2 between all layers, and train the model using the Adam optimizer (Kingma and Ba, 2014). The learning rate is decayed by 0.999 every 100 steps. We use 200 dimensional LSTMs and a 100 dimensional fully connected layer.

Is there a catalytic converter on a diesel? (Y)
A catalytic converter is an exhaust emission control device that converts toxic gases and pollutants in exhaust gas from an internal combustion engine into less-toxic pollutants by catalyzing a redox reaction (an oxidation and a reduction reaction). Catalytic converters are usually used with internal combustion engines fueled by either gasoline or diesel–including lean-burn engines as well as kerosene heaters and stoves.

Is there a season 2 of Pride and Prejudice? (N)
Pride and Prejudice is a six-episode 1995 British television drama, adapted by Andrew Davies from Jane Austen's 1813 novel of the same name. Jennifer Ehle and Colin Firth starred as Elizabeth Bennet and Mr. Darcy. Produced by Sue Birtwistle and directed by Simon Langton, the serial was a BBC production with additional funding from the American A&E Network. BBC1 originally broadcast the 55-minute episodes from 24 September to 29 October 1995. The A&E Network aired the series in double episodes on three consecutive nights beginning 14 January 1996. There are six episodes in the series.

Is Saving Private Ryan based on a book? (N)
In 1994, Robert Rodat wrote the script for the film. Rodat's script was submitted to producer Mark Gordon, who liked it and in turn passed it along to Spielberg to direct. The film is loosely based on the World War II life stories of the Niland brothers. A shooting date was set for June 27, 1997.

Is The Talk the same as The View? (N)
In November 2008, the show's post-election day telecast garnered the biggest audience in the show's history at 6.2 million in total viewers, becoming the week's most-watched program in daytime television. It was surpassed on July 29, 2010, during which former President Barack Obama first appeared as a guest on The View, which garnered a total of 6.6 million viewers. In 2013, the show was reported to be averaging 3.1 million daily viewers, which outpaced rival talk show The Talk.

Does the concept of a contact force apply to both a macroscopic scale and an atomic scale? (N)
In the Standard Model of modern physics, the four fundamental forces of nature are known to be non-contact forces. The strong and weak interaction primarily deal with forces within atoms, while gravitational effects are only obvious on an ultra-macroscopic scale. Molecular and quantum physics show that the electromagnetic force is the fundamental interaction responsible for contact forces. The interaction between macroscopic objects can be roughly described as resulting from the electromagnetic interactions between protons and electrons of the atomic constituents of these objects. Everyday objects do not actually touch; rather, contact forces are the result of the interactions of the electrons at or near the surfaces of the objects.

Legal to break out of prison in Germany? (Y)
In Mexico, Belgium, Germany and Austria, the philosophy of the law holds that it is human nature to want to escape. In those countries, escapees who do not break any other laws are not charged for anything and no extra time is added to their sentence. However, in Mexico, officers are allowed to shoot prisoners attempting to escape, and an escape is illegal if violence is used against prison personnel or property, or if prison inmates or officials aid the escape.

Is the movie sand pebbles based on a true story? (N)
The Sand Pebbles is a 1966 American war film directed by Robert Wise in Panavision. It tells the story of an independent, rebellious U.S. Navy machinist's mate, first class aboard the fictional gunboat USS San Pablo in 1920s China.

Is Burberrys of London the same as Burberry? (Y)
Burberry was founded in 1856 when 21-year-old Thomas Burberry, a former draper's apprentice, opened his own store in Basingstoke, Hampshire, England. By 1870, the business had established itself by focusing on the development of outdoors attire. In 1879, Burberry introduced in his brand the gabardine, a hardwearing, water-resistant yet breathable fabric, in which the yarn is waterproofed before weaving. "Burberry" was the original name until it became "Burberrys", due to many customers from around the world began calling it "Burberrys of London". In 1999, the name was reverted to the original, "Burberry". However, the name "Burberrys of London" is still visible on many older Burberry products. In 1891, Burberry opened a shop in the Haymarket, London. Before being termed as trench, it was known as the Tielocken worn by the British officers and featured a belt with no buttons, was double breasted, and protected the body from neck to knees.

Is the Saturn Vue the same as the Chevy Equinox? (N)
Riding on the GM Theta platform, the unibody is mechanically similar to the Saturn Vue and the Suzuki XL7. However, the Equinox and the Torrent are larger than the Vue, riding on a 112.5 in (2,858mm) wheelbase, 5.9 in (150mm) longer than the Vue. Front-wheel drive is standard, with optional all-wheel drive. They are not designed for serious off-roading like the truck-based Chevrolet Tahoe and Chevrolet TrailBlazer.

Is Destin FL on the Gulf of Mexico? (Y)
The city is located on a peninsula separating the Gulf of Mexico from Choctawhatchee Bay. The peninsula was originally an island; hurricanes and sea level changes gradually connected the island to the mainland.

Figure 3: Randomly sampled examples from the BoolQ train set.