Natural Questions: a Benchmark for Question Answering Research

Tom Kwiatkowski[♣◇♠] **Jennimaria Palomaki**[♠] **Olivia Redfield**[◇♠] **Michael Collins**[♣◇♠♥]
Ankur Parikh[♡] **Chris Alberti**[♡] **Danielle Epstein**[♤♦] **Illia Polosukhin**[♤♦] **Jacob Devlin**[♤]
Kenton Lee[♡] **Kristina Toutanova**[♡] **Llion Jones**[♤] **Matthew Kelcey**[♤♦] **Ming-Wei Chang**[♡]
Andrew M. Dai[♣◇] **Jakob Uszkoreit**[♣] **Quoc Le**[♣◇] **Slav Petrov**[♣]

Google Research

natural-questions@google.com

Abstract

We present the Natural Questions corpus, a question answering dataset. Questions consist of real anonymized, aggregated queries issued to the Google search engine. An annotator is presented with a question along with a Wikipedia page from the top 5 search results, and annotates a long answer (typically a paragraph) and a short answer (one or more entities) if present on the page, or marks null if no long/short answer is present. The public release consists of 307,373 training examples with single annotations; 7,830 examples with 5-way annotations for development data; and a further 7,842 examples 5-way annotated sequestered as test data. We present experiments validating quality of the data. We also describe analysis of 25-way annotations on 302 examples, giving insights into human variability on the annotation task. We introduce robust metrics for the purposes of evaluating question answering systems; demonstrate high human upper bounds on these metrics; and establish baseline results using competitive methods drawn from related literature.

1 Introduction

In recent years there has been dramatic progress in machine learning approaches to problems such as machine translation, speech recognition, and image recognition. One major factor in these successes has been the development of neural methods that far exceed the performance of previous approaches. A second major factor has been the existence of large quantities of training data for these systems.

Open-domain question answering (QA) is a benchmark task in natural language understanding (NLU), which has significant utility to users, and in addition is potentially a challenge task that can drive the development of methods for NLU. Several pieces of recent work have introduced QA datasets (e.g. Rajpurkar et al. (2016), Reddy et al. (2018)). However, in contrast to tasks where it is relatively easy to gather naturally occurring examples,[1] the definition of a suitable QA task, and the development of a methodology for annotation and evaluation, is challenging. Key issues include the methods and sources used to obtain questions; the methods used to annotate and collect answers; the methods used to measure and ensure annotation quality; and the metrics used for evaluation. For more discussion of the limitations of previous work with respect to these issues, see section 2 of this paper.

This paper introduces Natural Questions[2] (NQ), a new dataset for QA research, along with methods for QA system evaluation. Our goals are threefold: 1) To provide large-scale end-to-end training data for the QA problem. 2) To provide a dataset that drives research in natural language understanding. 3) To study human performance in providing QA annotations for naturally occurring questions.

In brief, our annotation process is as follows. An annotator is presented with a (question, Wikipedia page) pair. The annotator returns a (long answer, short answer) pair. The long an-

To appear in Transactions of the Association of Computational Linguistics (https://www.transacl.org). Final version.

♣Project initiation; ◇Project design; ♠Data creation; ♡Model development; ♤Project support; ♥Also affiliated with Columbia University, work done at Google; ♦No longer at Google, work done at Google.

[1] For example for machine translation/speech recognition humans provide translations/transcriptions relatively easily.

[2] Available at: https://ai.google.com/research/NaturalQuestions.

swer (l) can be an HTML bounding box on the Wikipedia page—typically a paragraph or table—that contains the information required to answer the question. Alternatively, the annotator can return $l = \texttt{NULL}$ if there is no answer on the page, or if the information required to answer the question is spread across many paragraphs. The short answer (s) can be a span or set of spans (typically entities) within l that answer the question, a boolean 'yes' or 'no' answer, or NULL. If $l = \texttt{NULL}$ then $s = \texttt{NULL}$, necessarily. Figure 1 shows examples.

Natural Questions has the following properties:

Source of questions The questions consist of real anonymized, aggregated queries issued to the Google search engine. Simple heuristics are used to filter questions from the query stream. Thus the questions are "natural", in that they represent real queries from people seeking information.

Number of items The public release contains 307,373 training examples with single annotations, 7,830 examples with 5-way annotations for development data, and 7,842 5-way annotated items sequestered as test data. We justify the use of 5-way annotation for evaluation in Section 5.

Task definition The input to a model is a question together with an entire Wikipedia page. The target output from the model is: 1) a long-answer (e.g., a paragraph) from the page that answers the question, or alternatively an indication that there is no answer on the page; 2) a short answer where applicable. The task was designed to be close to an end-to-end question answering application.

Ensuring high quality annotations at scale Comprehensive guidelines were developed for the task. These are summarized in Section 3. Annotation quality was constantly monitored.

Evaluation of quality Section 4 describes post-hoc evaluation of annotation quality. Long/short answers have 90%/84% precision respectively.

Study of variability One clear finding in NQ is that for naturally occurring questions there is often genuine ambiguity in whether or not an answer is acceptable. There are also often a number of acceptable answers. Section 4 examines this variability using 25-way annotations.

Robust evaluation metrics Section 5 introduces methods of measuring answer quality that accounts for variability in acceptable answers. We demonstrate a high human upper bound on these measures for both long answers (90% precision, 85% recall), and short answers (79% precision, 72% recall).

Example 1
Question: what color was john wilkes booth's hair
Wikipedia Page: John_Wilkes_Booth
Long answer: Some critics called Booth "the handsomest man in America" and a "natural genius", and noted his having an "astonishing memory"; others were mixed in their estimation of his acting. He stood 5 feet 8 inches (1.73 m) tall, had jet-black hair , and was lean and athletic. Noted Civil War reporter George Alfred Townsend described him as a "muscular, perfect man" with "curling hair, like a Corinthian capital".
Short answer: jet-black

Example 2
Question: can you make and receive calls in airplane mode
Wikipedia Page: Airplane_mode
Long answer: Airplane mode, aeroplane mode, flight mode, offline mode, or standalone mode is a setting available on many smartphones, portable computers, and other electronic devices that, when activated, suspends radio-frequency signal transmission by the device, thereby disabling Bluetooth, telephony, and Wi-Fi. GPS may or may not be disabled, because it does not involve transmitting radio waves.
Short answer: BOOLEAN:NO

Example 3
Question: why does queen elizabeth sign her name elizabeth r
Wikipedia Page: Royal_sign-manual
Long answer: The royal sign-manual usually consists of the sovereign's regnal name (without number, if otherwise used), followed by the letter R for Rex (King) or Regina (Queen). Thus, the signs-manual of both Elizabeth I and Elizabeth II read Elizabeth R. When the British monarch was also Emperor or Empress of India, the sign manual ended with R I, for Rex Imperator or Regina Imperatrix (King-Emperor/Queen-Empress).
Short answer: NULL

Figure 1: Example annotations from the corpus.

We propose NQ as a new benchmark for research in question answering. In Section 6.4 we present baseline results from recent models developed on comparable datasets (Clark and Gardner, 2018), as well as a simple pipelined model designed for the NQ task. We demonstrate a large gap between the performance of these baselines and a human upper bound. We argue that closing this gap will require significant advances in NLU.

2 Related Work

The SQuAD (Rajpurkar et al., 2016), SQuAD 2.0 (Rajpurkar et al., 2018), NarrativeQA (Kocisky et al., 2018), and HotpotQA (Yang et al., 2018) datasets contain questions and answers written by annotators who have first read a short text containing the answer. The SQuAD datasets contain questions/paragraph/answer triples from Wikipedia. In the original SQuAD dataset, annotators often borrow part of the evidence paragraph to create a question. Jia and Liang (2017) showed that systems trained on SQuAD could be easily

fooled by the insertion of distractor sentences that should not change the answer, and SQuAD 2.0 introduces questions that are designed to be unanswerable. However, we argue that questions written to be unanswerable can be identified as such with little reasoning, in contrast to NQ's task of deciding whether a paragraph contains all of the evidence required to answer a real question. Both SQuAD tasks have driven significant advances in reading comprehension, but systems now outperform humans and harder challenges are needed. NarrativeQA aims to elicit questions that are not close paraphrases of the evidence by separate summary texts. No human performance upper bound is provided for the full task and, while an extractive system could theoretically perfectly recover all answers, current approaches only just outperform a random baseline. NarrativeQA may just be too hard for the current state of NLU. HotpotQA is designed to contain questions that require reasoning over text from separate Wikipedia pages. As well as answering questions, systems must also identify passages that contain supporting facts. This is similar in motivation to NQ's long answer task, where the selected passage must contain all of the information required to infer the answer. Mirroring our identification of acceptable variability in the NQ task definition, HotpotQA's authors observe that the choice of supporting facts is somewhat subjective. They set high human upper bounds by selecting, for each example, the score maximizing partition of four annotations into one prediction and three references. The reference labels chosen by this maximization are not representative of the reference labels in HotpotQA's evaluation set, and it is not clear that the upper bounds are achievable. A more robust approach is to keep the evaluation distribution fixed, and calculate an acheivable upper bound by approximating the expectation over annotations—as we have done for NQ in Section 5.

The QuAC (Choi et al., 2018) and CoQA (Reddy et al., 2018) datasets contain dialogues between a questioner, who is trying to learn about a text, and an answerer. QuAC also prevents the questioner from seeing the evidence text. Conversational question answering is an exciting new area, but it is significantly different from the single turn question answering task in NQ. In both QuAC and CoQA, conversations tend to explore evidence texts incrementally, progressing from the start to the end of the text. This contrasts with NQ, where individual questions often require reasoning over large bodies of text.

The WikiQA (Yang et al., 2015) and MS Marco (Nguyen et al., 2016) datasets contain queries sampled from the Bing search engine. WikiQA contains only 3,047 questions. MS Marco contains 100,000 questions with free-form answers. For each question, the annotator is presented with 10 passages returned by the search engine, and is asked to generate an answer to the query, or to say that the answer is not contained within the passages. Free-form text answers allow more flexibility in providing abstractive answers, but lead to difficulties in evaluation (BLEU score (Papineni et al., 2002) is used). MS Marco's authors do not discuss issues of variability or report quality metrics for their annotations. From our experience these issues are critical. DuReader (He et al., 2018) is a Chinese language dataset containing queries from Baidu search logs. Like NQ, DuReader contains real user queries; it requires systems to read entire documents to find answers; and it identifies acceptable variability in answers. However, as with MS Marco, DuReader is reliant on BLEU for answer scoring, and systems already out-perform a humans according to this metric.

There are a number of reading comprehension benchmarks based on multiple choice tests (Mihaylov et al., 2018; Richardson et al., 2013; Lai et al., 2017). The TriviaQA dataset (Joshi et al., 2017) contains questions and answers taken from trivia quizzes found online. A number of Cloze-style tasks have also been proposed (Hermann et al., 2015; Hill et al., 2015; Paperno et al., 2016; Onishi et al., 2016). We believe that all of these tasks are related to, but distinct from, answering information seeking questions. We also believe that, since a solution to NQ will have genuine utility, it is better equipped as a benchmark for NLU.

3 Task Definition and Data Collection

Natural Questions contains (*question*, *wikipedia page*, *long answer*, *short answer*) quadruples where: the question seeks factual information; the Wikipedia page may or may not contain the information required to answer the question; the long answer is a bounding box on this page containing all information required to infer the answer; and the short answer is one or more entities that give a short answer to the question, or a boolean

1.a	where does the nature conservancy get its funding
1.b	who is the song killing me softly written about
2	who owned most of the railroads in the 1800s
4	how far is chardon ohio from cleveland ohio
5	american comedian on have i got news for you

Table 1: Matches for heuristics in Section 3.1.

'yes' or 'no'. Both the long and short answer can be NULL if no viable candidates exist on the Wikipedia page.

3.1 Questions and Evidence Documents

All the questions in NQ are queries of 8 words or more that have been issued to the Google search engine by multiple users in a short period of time. From these queries, we sample a subset that either:

1. start with 'who', 'when', or 'where' directly followed by: a) a finite form of 'do' or a modal verb; or b) a finite form of 'be' or 'have' with a verb in some later position;
2. start with 'who' directly followed by a verb that is not a finite form of 'be';
3. contain multiple entities as well as an adjective, adverb, verb, or determiner;
4. contain a categorical noun phrase immediately preceded by a preposition or relative clause;
5. end with a categorical noun phrase, and do not contain a preposition or relative clause.[3]

Table 1 gives examples. We run questions through the Google search engine and keep those where there is a Wikipedia page in the top 5 search results. The (question, Wikipedia page) pairs are the input to the human annotation task described next.

The goal of these heuristics is to discard a large proportion of queries that are non-questions, while retaining the majority of queries of 8 words or more in length that are questions. A manual inspection showed that the majority of questions in the data, with the exclusion of question beginning with "how to", are accepted by the filters. We focus on longer queries as they are more complex, and are thus a more challenging test for deep NLU. We focus on Wikipedia as it is a very important source of factual information, and we believe that stylistically it is similar to other sources of factual information on the web; however like any dataset there may be biases in this choice. Future data-collection efforts may introduce shorter queries, "how to" questions, or domains other than Wikipedia.

[3]We pre-define the set of categorical noun phrases used in 4 and 5 by running Hearst patterns (Hearst, 1992) to find a broad set of hypernyms. Part of speech tags and entities are identified using Google's Cloud NLP API: https://cloud.google.com/natural-language

3.2 Human Identification of Answers

Annotation is performed using a custom annotation interface, by a pool of around 50 annotators, with an average annotation time of 80 seconds.

The guidelines and tooling divide the annotation task into three conceptual stages, where all three stages are completed by a single annotator in succession. The decision flow through these is illustrated in Figure 2 and the instructions given to annotators are summarized below.

Question Identification: contributors determine whether the given question is *good* or *bad*. A *good question* is a fact-seeking question that can be answered with an entity or explanation. A *bad question* is ambigous, incomprehensible, dependent on clear false presuppositions, opinion-seeking, or not clearly a request for factual information. Annotators must make this judgment solely by the content of the question; they are not yet shown the Wikipedia page.

Long Answer Identification: for good questions only, annotators select the earliest HTML bounding box containing enough information for a reader to completely infer the answer to the question. Bounding boxes can be paragraphs, tables, list items, or whole lists. Alternatively, annotators mark 'no answer' if the page does not answer the question, or if the information is present but not contained in a single one of the allowed elements.

Short Answer Identification: for examples with long answers, annotators select the entity or set of entities within the long answer that answer the question. Alternatively, annotators can flag that the short answer is 'yes', 'no', or they can flag that no short answer is possible.

3.3 Data Statistics

In total, annotators identify a long answer for 49% of the examples, and short answer spans or a yes/no answer for 36% of the examples. We consider the choice of whether or not to answer a question a core part of the question answering task, and do not discard the remaining 51% that have no answer labeled.

Annotators identify long answers by selecting the smallest HTML bounding box that contains all

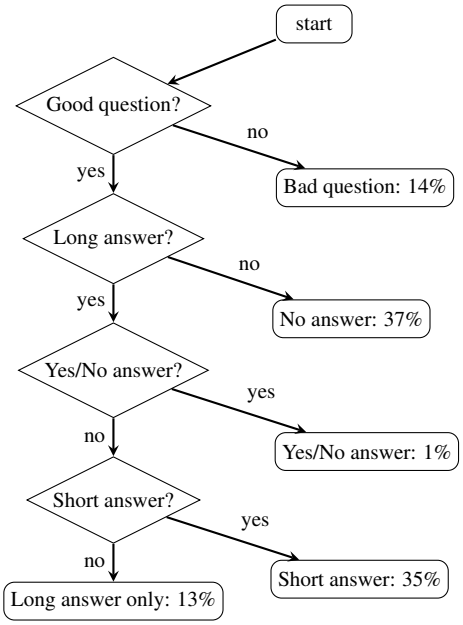

Figure 2: Annotation decision process with path proportions from NQ training data. Percentages are proportions of entire dataset. 49% of all examples have a long answer.

of the information required to answer the question. These are mostly paragraphs (73%). The remainder are made up of tables (19%), table rows (1%), lists (3%), or list items (3%).[4] We leave further subcategorization of long answers to future work, and provide a breakdown of baseline performance on each of these three types of answers in Section 6.4.

4 Evaluation of Annotation Quality

This section describes evaluation of the quality of the human annotations in our data. We use a combination of two methods: first, post-hoc evaluation of correctness of non-null answers, under consensus judgments from 4 "experts"; second, k-way annotations (with $k = 25$) on a subset of the data.

Post-hoc evaluation of non-null answers leads directly to a measure of annotation precision. As is common in information-retrieval style problems such as long-answer identification, measuring recall is more challenging. However we describe how 25-way annotated data gives useful insights into recall, particularly when combined with expert judgments.

[4]We note that both tables and lists may be used purely for the purposes of formatting text, or they may have their own complex semantics—as in the case of Wikipedia infoboxes.

4.1 Preliminaries: the Sampling Distribution

Each item in our data consists of a four-tuple (q, d, l, s) where q is a question, d is a document, l is a long answer, and s is a short answer. Thus we introduce random variables Q, D, L and S corresponding to these items. Note that L can be a span within the document, or NULL. Similarly S can be one or more spans within L, a boolean, or NULL.

For now we consider the three-tuple (q, d, l). The treatment for short answers is the same throughout, with (q, d, s) replacing (q, d, l).

Each data item (q, d, l) is IID sampled from

$$p(l, q, d) = p(q, d) \times p(l|q, d)$$

Here $p(q, d)$ is the sampling distribution (probability mass function (PMF)) over question/document pairs. It is defined as the PMF corresponding to the following sampling process:[5] first, sample a question at random from some distribution; second, perform a search on a major search engine using the question as the underlying query; finally, either: (1) return (q, d) where d is the top Wikipedia result for q, if d is in the top 5 search results for q; (2) if there is no Wikipedia page in the top 5 results, discard q and repeat the sampling process.

Here $p(l|q, d)$ is the conditional distribution (PMF) over long answer l conditioned on the pair (q, d). The value for l is obtained by: (1) sampling an annotator uniformly at random from the pool of annotators; (2) presenting the pair (q, d) to the annotator, who then provides a value for l.

Note that l is non-deterministic due to two sources of randomness: (1) the random choice of annotator; (2) the potentially random behaviour of a particular annotator (the annotator may give a different answer depending on the time of day etc.).

We will also consider the distribution

$$\begin{aligned} p(l, q, d|L \neq \texttt{NULL}) &= \frac{p(l, q, d)}{P(L \neq \texttt{NULL})} && \text{if } l \neq \texttt{NULL} \\ &= 0 && \text{otherwise} \end{aligned}$$

where $P(L \neq \texttt{NULL}) = \sum_{l,q,d:l \neq \texttt{NULL}} p(l, q, d)$. Thus $p(l, q, d|L \neq \texttt{NULL})$ is the probability of see-

[5]More formally, there is some base distribution $p_b(q)$ from which queries q are drawn, and a deterministic function $s(q)$ which returns the top-ranked Wikipedia page in the top 5 search results, or NULL if there is no Wikipedia page in the top 5 results. Define $\mathcal{Q}$ to be the set of queries such that $s(q) \neq \texttt{NULL}$, and $b = \sum_{q \in \mathcal{Q}} p_b(q)$. Then $p(q, d) = p_b(q)/b$ if $q \in \mathcal{Q}$ and $d \neq \texttt{NULL}$ and $d = s(q)$, otherwise $p(q, d) = 0$.

ing the triple (l, q, d), conditioned on L not being NULL.

We now define precision of annotations. Consider a function $\pi(l, q, d)$ that is equal to 1 if l is a "correct" answer for the pair (q, d), 0 if the answer is incorrect. The next section gives a concrete definition of π. The annotation precision is defined as

$$\Psi = \sum_{l,q,d} p(l, q, d | L \neq \texttt{NULL}) \times \pi(l, q, d)$$

Given a set of annotations $\mathcal{S} = \{(l^{(i)}, q^{(i)}, d^{(i)})\}_{i=1}^{|\mathcal{S}|}$ drawn IID from $p(l, q, d | L \neq \texttt{NULL})$, we can derive an estimate of Ψ as $\hat{\Psi} = \frac{1}{|\mathcal{S}|} \sum_{(l,q,d) \in \mathcal{S}} \pi(l, q, d)$.

4.2 Expert Evaluations of Correctness

We now describe the process for deriving "expert" judgments of answer correctness. We used four experts for these judgments. These experts had prepared the guidelines for the annotation process.[6] In a first phase each of the four experts independently annotated examples for correctness. In a second phase the four experts met to discuss disagreements in judgments, and to reach a single consensus judgment for each example.

A key step is to define the criteria used to determine correctness of an example. Given a triple (l, q, d), we extracted the passage l' corresponding to l on the page d. The pair (q, l') was then presented to the expert. Experts categorized (q, l') pairs into the following three categories:

- Correct ($\mathcal{C}$): It is clear beyond a reasonable doubt that the answer is correct.
- Correct (but debatable) ($\mathcal{C}_d$): A reasonable person could be satisfied by the answer; however a reasonable person could raise a reasonable doubt about the answer.
- Wrong ($\mathcal{W}$): There is not convincing evidence that the answer is correct.

Figure 3 shows some example judgments. We introduced the intermediate $\mathcal{C}_d$ category after observing that many (q, l') pairs are high quality answers, but raise some small doubt or quibble about whether they fully answer the question. The use of the word "debatable" is intended to be literal: (q, l') pairs falling into the $\mathcal{C}_d$ category could literally lead to some debate between reasonable people as to whether they fully answer the question or not.

[6] The first four authors of this paper.

Example 1
Question: who played will on as the world turns **Long answer:** William "Will" Harold Ryan Munson is a fictional character on the CBS soap opera As the World Turns. He was portrayed by Jesse Soffer on recurring basis from September 2004 to March 2005, after which he got a contract as a regular. Soffer left the show on April 4, 2008 and made a brief return in July 2010. **Judgment:** Correct. **Justification:** It is clear beyond a reasonable doubt that the answer is correct.

Example 2
Question: which type of rock forms on the earth's crust **Long answer:** Igneous and metamorphic rocks make up 90-95% of the top 16 km of the Earth's crust by volume. Igneous rocks form about 15% of the Earth's current land surface. Most of the Earth's oceanic crust is made of igneous rock. **Judgment:** Correct (but debatable). **Justification:** The answer goes a long way to answering the question, but a reasonable person could raise objections to the answer.

Example 3
Question: who was the first person to see earth from space **Long answer:** Yuri Alekseyevich Gagarin was a Soviet pilot and cosmonaut. He was the first human to journey into outer space when his Vostok spacecraft completed an orbit of the Earth on 12 April 1961. **Judgment:** Correct (but debatable). **Justification:** It is likely that Gagarin was the first person to see earth from space, but not guaranteed. For example it is not certain that "space" and "outer space" are the same, or that there was a window in Vostok.

Figure 3: Examples with consensus expert judgments, and justification for these judgments. See figure 6 for more examples.

Given this background, we will make the following assumption:

Answers in the $\mathcal{C}_d$ category should be very useful to a user interacting with a question-answering system, and should be considered to be high-quality answers; however an annotator would be justified in either annotating or not annotating the example.

For these cases there is often disagreement between annotators as to whether the page contains an answer or not: we will see evidence of this when we consider the 25-way annotations.

4.3 Results for Precision Measurements

We followed the following procedure to derive measurements of precision: (1) We sampled examples IID from the distribution $p(l, q, d | L \neq \texttt{NULL})$. We call this set $\mathcal{S}$. We had $|\mathcal{S}| = 139$. (2) Four experts independently classified each of the items in $\mathcal{S}$ into the categories $\mathcal{C}$, $\mathcal{C}_d$, $\mathcal{W}$. (3) The four experts met to come up with a consensus judgment for each item. For each example $(l^{(i)}, q^{(i)}, d^{(i)}) \in \mathcal{S}$, we define $c^{(i)}$ to be the consensus judgment. The above process was repeated to derive judgments for short answers.

We can then calculate the percentage of exam-

Quantity	Long answer	Short answer
$\hat{\Psi}$	90%	84%
$\hat{E}(\mathcal{C})$	59%	51%
$\hat{E}(\mathcal{C}_d)$	31%	33%
$\hat{E}(\mathcal{W})$	10%	16%

Table 2: Precision results ($\hat{\Psi}$) and empirical estimates of the proportions of $\mathcal{C}$, $\mathcal{C}_d$, and $\mathcal{W}$ items.

ples falling into the three expert categories; we denote these values as $\hat{E}(\mathcal{C})$, $\hat{E}(\mathcal{C}_d)$ and $\hat{E}(\mathcal{W})$.[7] We define $\hat{\Psi} = \hat{E}(\mathcal{C}) + \hat{E}(\mathcal{C}_d)$. We have explicitly included samples $\mathcal{C}$ and $\mathcal{C}_d$ in the overall precision as we believe that $\mathcal{C}_d$ answers are essentially correct. Table 2 shows the values for these quantities.

4.4 Variability of Annotations

We have shown that an annotation drawn from $p(l, q, d|L \neq \texttt{NULL})$ has high expected precision. Now we address the distribution over annotations for a given (q, d) pair. Annotators can disagree about whether or not d contains an answer to q—that is whether or not $L = \texttt{NULL}$. In the case that annotators agree that $L \neq \texttt{NULL}$, they can also disagree about the correct assignment to L.

In order to study variability, we collected 24 additional annotations from separate annotators for each of the (q, d, l) triples in $\mathcal{S}$. For each (q, d, l) triple, we now have a 5-tuple $(q^{(i)}, d^{(i)}, l^{(i)}, c^{(i)}, a^{(i)})$ where $a^{(i)} = a_1^{(i)} \dots a_{25}^{(i)}$ is a vector of 25 annotations (including $l^{(i)}$), and $c^{(i)}$ is the consensus judgment for $l^{(i)}$. For each i also define

$$\mu^{(i)} = \frac{1}{25} \sum_{j=1}^{25} [[a_j^{(i)} \neq \texttt{NULL}]]$$

to be the proportion of the 25-way annotations that are non-null.

We now show that $\mu^{(i)}$ is highly correlated with annotation precision. We define

$$\hat{E}[(0.8, 1.0]] = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} [[0.8 < \mu^{(i)} \leq 1]]$$

to be the proportion of examples with greater than 80% of the 25 annotators marking a non-null long answer, and

$$\hat{E}[(0.8, 1.0], \mathcal{C}] = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} [[0.8 < \mu^{(i)} \leq 1 \text{ and } c^{(i)} = \mathcal{C}]]$$

to be the proportion of examples with greater than 80% of the 25 annotators marking a non-null long answer and with $c^{(i)} = \mathcal{C}$. Similar definitions apply for the intervals $(0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$ and $(0.6, 0.8]$, and for judgments $\mathcal{C}_d$ and $\mathcal{W}$.

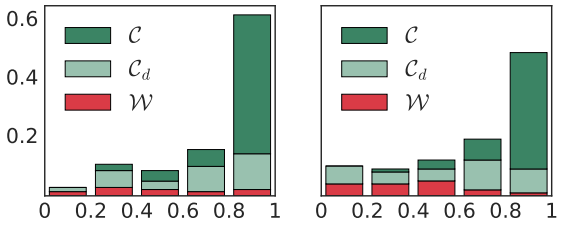

Figure 4: Values of $\hat{E}[(\theta^1, \theta^2]]$ and $\hat{E}[(\theta^1, \theta^2], \mathcal{C}/\mathcal{C}_d/\mathcal{W}]$ for different intervals $(\theta^1, \theta^2]$. The height of each bar is equal to $\hat{E}[(\theta^1, \theta^2]]$, the divisions within each bar show $\hat{E}[(\theta^1, \theta^2], \mathcal{C}]$, $\hat{E}[(\theta^1, \theta^2], \mathcal{C}_d]$, and $\hat{E}[(\theta^1, \theta^2], \mathcal{W}]$.

Figure 4 illustrates the proportion of annotations falling into the $\mathcal{C}/\mathcal{C}_d/\mathcal{W}$ categories in different regions of $\mu^{(i)}$. For those (q, d) pairs where more than 80% of annotators gave some non-null answer, our expert judgements agree that these annotations are overwhelmingly correct. Similarly, when fewer than 20% of annotators gave a non-null answer, these answers tend to be incorrect. In between these two extremes, the disagreement between annotators is largely accounted for by the $\mathcal{C}_d$ category—where a reasonable person could either be satisfied with the answer, or want more information. Later, in Section 5, we make use of the correlation between $\mu^{(i)}$ and accuracy to define a metric for the evaluation of answer quality. In that section, we also show that a model trained on (l, q, d) triples can outperform a single annotator on this metric by accounting for the uncertainty of whether or not an answer is present.

As well as disagreeing about whether (q, d) contains a valid answer, annotators can disagree about the location of the best answer. In many cases there are multiple valid long answers in multiple distinct locations on the page.[8] The most extreme example of this that we see in our 25-way annotated data is for the question *'name the substance used to make the filament of bulb'* paired with the Wikipedia page about incandescent light bulbs. Annotators identify 7 passages that discuss

[7] More formally, let $[[e]]$ for any statement e be 1 if e is true, 0 if e is false. We define $\hat{E}(\mathcal{C}) = \frac{1}{|\mathcal{S}|} \sum_{i=1}^{|\mathcal{S}|} [[c^{(i)} = \mathcal{C}]]$. The values for $\hat{E}(\mathcal{C}_d)$ and $\hat{E}(\mathcal{W})$ are calculated in a similar way.

[8] As stated earlier in this paper, we did instruct annotators to select the earliest instance of an answer when there are multiple answer instances on the page. However there are still cases where different annotators disagree on whether an answer earlier in the page is sufficient in comparison to a later answer, leading to differences between annotators.

tungsten wire filaments.

Short answers can be arbitrarily delimited and this can lead to extreme variation. The most extreme example of this that we see in the 25-way annotated data is the 11 distinct, but correct, answers for the question *'where is blood pumped after it leaves the right ventricle'*. Here, 14 annotators identify a substring of *'to the lungs'* as the best possible short answer. Of these, 6 label the entire string, 4 reduce it to *'the lungs'*, and 4 reduce it to *'lungs'*. A further 6 annotators do not consider this short answer to be sufficient and choose more precise phrases such as *'through the semilunar pulmonary valve into the left and right main pulmonary arteries (one for each lung)'*. The remaining 5 annotators decide that there is no adequate short answer.

For each question, we ranked each of the unique answers given by our 25 annotators according to the number of annotators that chose it. We found that by just taking the most popular long answer, we could account for 83% of the long answer annotations. The two most popular long answers account for 96% of the long answer annotations. It is extremely uncommon for a question to have more than three distinct long answers annotated. Short answers have greater variability, but the most popular short answer still accounts for 64% of all short answer annotations. The three most popular short answers account for 90% of all short answer annotations.

5 Evaluation Measures

NQ includes 5-way annotations on 7,830 items for development data, and we will sequester a further 7,842 items, 5-way annotated, for test data. This section describes evaluation metrics using this data, and gives justification for these metrics.

We choose 5-way annotations for the following reasons: first, we have evidence that aggregating annotations from 5 annotators is likely to be much more robust than relying on a single annotator (see Section 4). Second, 5 annotators is a small enough number that the cost of annotating thousands of development and test items is not prohibitive.

5.1 Definition of an Evaluation Measure Based on 5-Way Annotations

Assume that we have a model f_θ with parameters θ which maps an input (q, d) to a long answer $l = f_\theta(q, d)$. We would like to evaluate the accuracy of this model. Assume we have evaluation examples $\{q^{(i)}, d^{(i)}, a^{(i)}\}$ for $i = 1 \ldots n$, where $q^{(i)}$ is a question, $d^{(i)}$ is the associated Wikipedia document, and $a^{(i)}$ is a vector with components $a_j^{(i)}$ for $j = 1 \ldots 5$. Each $a_j^{(i)}$ is the output from the j'th annotator, and can be a paragraph in $d^{(i)}$, or can be NULL. The 5 annotators are chosen uniformly at random from a pool of annotators.

We define an evaluation measure based on the 5 way annotations as follows. If at least 2 out of 5 annotators have given a non-null long answer on the example, then the system is required to output a non-null answer that is seen at least once in the 5 annotations; conversely if fewer than 2 annotators give a non-null long answer, the system is required to return NULL as its output.

To make this more formal, define the function $g(a^{(i)})$ to be the number of annotations in $a^{(i)}$ that are non-null. Define a function $h_\beta(a, l)$ that judges the correctness of label l given annotations $a = a_1 \ldots a_5$. This function is parameterized by an integer β. The function returns 1 if the label l is judged to be correct, and 0 otherwise:

Definition 1 *(Definition of $h_\beta(a, l)$)* **If** $g(a) \geq \beta$ *and* $l \neq$ NULL *and* $l = a_j$ *for some* $j \in \{1 \ldots 5\}$ **Then** $h_\beta(a, l) = 1$*;* **Else If** $g(a) < \beta$ *and* $l =$ NULL **Then** $h_\beta(a, l) = 1$*;* **Else** $h_\beta(a, l) = 0$.

We used $\beta = 2$ in our experiments.[9]

The accuracy of a model is then

$$A_\beta(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} h_\beta(a^{(i)}, f_\theta(q^{(i)}, d^{(i)}))$$

The value for A_β is an estimate of accuracy with respect to the underlying distribution, which we define as $\bar{A}_\beta(f_\theta) = \mathbf{E}[h_\beta(a, f_\theta(q, d))]$. Here the expectation is taken with respect to $p(a, q, d) = p(q, d) \prod_{j=1}^{5} p(a_j | q, d)$ where $p(a_j | q, d) = P(L = a_j | Q = q, D = d)$; hence the annotations $a_1 \ldots a_5$ are assumed to be drawn IID from $p(l | q, d)$.[10]

We discuss this measure at length in this section. First, however, we make the following critical point:

[9] This is partly motivated through the results on 25-way annotations (see section 4.4), where for $\mu^{(i)} \geq 0.4$ over 93% (114/122 annotations) are in the $\mathcal{C}$ or $\mathcal{C}_d$ categories, whereas for $\mu^{(i)} < 0.4$ over 35% (11/17 annotations) are in the $\mathcal{W}$ category.

[10] This isn't quite accurate as the annotators are sampled without replacement; however it simplifies the analysis.

It is possible for a model trained on $(l^{(i)}, q^{(i)}, d^{(i)})$ triples drawn IID from $p(l, q, d)$ to exceed the performance of a single annotator on this measure.

In particular, if we have a model $p(l|q, d; \theta)$, trained on (l, q, d) triples, which is a good approximation to $p(l|q, d)$, it is then possible to use $p(l|q, d; \theta)$ to make predictions that outperform a single random draw from $p(l|q, d)$. The Bayes optimal hypothesis (see (Devroye et al., 1997)) for h_β, defined as $\arg\max_f \mathbf{E}_{q,d,a}[[h_\beta(a, f(q, d))]]$ is a function of the posterior distribution $p(\cdot|q, d)$,[11] and will generally exceed the performance of a single random annotation, $\mathbf{E}_{q,d,a}[[\sum_l p(l|q, d) \times h_\beta(a, l)]]$.

We also show this empirically, by constructing an approximation to $p(l|q, d)$ from 20-way annotations, then using this approximation to make predictions that significantly outperform a single annotator.

Precision and Recall During evaluation, it is often beneficial to separately measure false positives (incorrectly predicting an answer), and false negatives (failing to predict a answer). We define the precision (P) and recall (R) of f_θ:

$$t(q, d, a, f_\theta) = h_\beta(a, f_\theta(q, d))[[f_\theta(q, d) \neq \texttt{NULL}]]$$

$$R(f_\theta) = \frac{\sum_{i=1}^n t(q^{(i)}, d^{(i)}, a^{(i)}, f_\theta)}{\sum_{i=1}^n [[g(a^{(i)}) \geq \beta]]}$$

$$P(f_\theta) = \frac{\sum_{i=1}^n t(q^{(i)}, d^{(i)}, a^{(i)}, f_\theta)}{\sum_{i=1}^n [[f_\theta(q^{(i)}, d^{(i)}) \neq \texttt{NULL}]]}$$

5.2 Super-Annotator Upper Bound

To place an upper bound on the metrics introduced above we create a 'super-annotator' from the 25-way annotated data introduced in Section 4. From this data, we create four tuples $(q^{(i)}, d^{(i)}, a^{(i)}, b^{(i)})$. The first three terms in this tuple are the question, document, and vector of 5 reference annotations. $b^{(i)}$ is a vector of annotations $b_j^{(i)}$ for $j = 1 \ldots 20$ drawn from the same distribution as $a^{(i)}$. The super-annotator predicts NULL if $g(b^{(i)}) < \alpha$, and $l^* = \arg\max_{l \in d} \sum_{j=1}^{20} [[l = b_j]]$ otherwise.

Table 3 shows super-annotator performance for $\alpha = 8$, with 90.0% precision, 84.6% recall, and 87.2% f-measure. This significantly exceeds the performance (80.4% precision/67.6% recall/73.4% f-measure) for a single annotator. We subsequently view the super-annotator numbers as an effective upper bound on performance of a learned model.

6 Baseline Performance

The Natural Questions corpus is designed to provide a benchmark with which we can evaluate the performance of question answering systems. Every question in NQ is unique under exact string match, and we split questions randomly in NQ into separate train/development/test sets. To facilitate comparison, we introduce baselines that either make use of high level dataset regularities, or are trained on the 307k examples in the training set. Here, we present well-established baselines that were state of the art at the time of submission. We also refer readers to Alberti et al. (2019) for more recent advances in modeling. All of our baselines focus on the long and short answer extraction tasks. We leave boolean answers to future work.

6.1 Untrained Baselines

NQ's long answer selection task admits several untrained baselines. The first paragraph of a Wikipedia page commonly acts as a summary of the most important information regarding the page's subject. We therefore implement a long answer baseline that simply selects the first paragraph for all pages.

Furthermore, since 79% of the Wikipedia pages in the development set also appear in the training set, we implement two 'copying' baselines. The first of these simply selects the most frequent annotation applied to a given page in the training set. The second selects the annotation given to the train set question closest to the eval set question according to TFIDF weighted word overlap. These three baselines are reported as 'First paragraph',

[11] Specifically, for an input (q, d), if we define $l^* = \arg\max_{l \neq \texttt{NULL}} p(l|q, d)$, $\gamma = p(l^*|q, d)$, and $\bar{\gamma} = p(\texttt{NULL}|q, d)$, then the Bayes optimal hypothesis is to output l^* if $P(h_\beta(a, l^*) = 1|\gamma, \bar{\gamma}) \geq P(h_\beta(a, \texttt{NULL}) = 1|\gamma, \bar{\gamma})$, and to output NULL otherwise. Implementation of this strategy is straightforward if γ and $\bar{\gamma}$ are known; this strategy will in general give a higher accuracy value than taking a single sample l from $p(l|q, d)$ and using this sample as the prediction. In principle a model $p(l|q, d; \theta)$ trained on (l, q, d) triples can converge to a good estimate of γ and $\bar{\gamma}$. Note that for the special case $\gamma + \bar{\gamma} = 1$ we have $P(h_\beta(a, \texttt{NULL}) = 1|\gamma, \bar{\gamma}) = \bar{\gamma}^5 + 5\bar{\gamma}^4(1 - \bar{\gamma})$ and $P(h_\beta(a, l^*) = 1|\gamma, \bar{\gamma}) = 1 - P(h_\beta(a, \texttt{NULL}) = 1|\gamma, \bar{\gamma})$. It follows that the Bayes optimal hypothesis is to predict l^* if $\gamma \geq \alpha$ where $\alpha \approx 0.31381$, and to predict NULL otherwise. α is $1 - \bar{\alpha}$ where $\bar{\alpha}$ is the solution to $\bar{\alpha}^5 + 5\bar{\alpha}^4(1 - \bar{\alpha}) = 0.5$.

	Long answer Dev			Long answer Test			Short answer Dev			Short answer Test		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
First paragraph	22.2	37.8	27.8	22.3	38.5	28.3	–	–	–	–	–	–
Most frequent	43.1	20.0	27.3	40.2	18.4	25.2	–	–	–	–	–	–
Closest question	37.7	28.5	32.4	36.2	27.8	31.4	–	–	–	–	–	–
DocumentQA	47.5	44.7	46.1	48.9	43.3	45.7	38.6	33.2	35.7	40.6	31.0	35.1
DecAtt + DocReader	52.7	57.0	54.8	54.3	55.7	55.0	34.3	28.9	31.4	31.9	31.1	31.5
Single annotator†	80.4	67.6	73.4	–	–	–	63.4	52.6	57.5	–	–	–
Super-annotator†	90.0	84.6	87.2	–	–	–	79.1	72.6	75.7	–	–	–

Table 3: Precision (P), recall (R), and the harmonic mean of these (F1) of all baselines, a single annotator, and the super-annotator upper bound. The human performances marked with † are evaluated on a sample of 5 annotations from the 25-way annotated data introduced in Section 5.

'Most frequent', and 'Closest question' in Table 3 respectively.

6.2 Document-QA

We adapt the reference implementation[12] of Document-QA (Clark and Gardner, 2018) for the NQ task. This system performs well on the SQuAD and TriviaQA short answer extraction tasks, but it is not designed to represent: (*i*) the long answers that do not contain short answers, and (*ii*) the NULL answers that occur in NQ.

To address (*i*) we choose the shortest available answer span at training, differentiating long and short answers only through the inclusion of special start and end of passage tokens that identify long answer candidates. At prediction time, the model can either predict a long answer (and no short answer), or a short answer (which implies a long answer).

To address (*ii*), we tried adding special NULL passages to represent the lack of answer. However, we achieved better performance by training on the subset of questions with answers and then only predicting those answers whose scores exceed a threshold.

With these two modifications, we are able to apply Document-QA to NQ. We follow Clark and Gardner (2018) in pruning documents down to the set of passages that have highest TFIDF similarity with the question. Under this approach, we consider the top 16 passages as long answers. We consider short answers containing up to 17 words. We train Document-QA for 30 epochs with batches containing 15 examples. The post-hoc score threshold is set to 3.0. All of these values were chosen on the basis of development set performance.

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

6.3 Custom Pipeline (DecAtt + DocReader)

One view of the long answer selection task is that it is more closely related to natural language inference (NLI) (Bowman et al., 2015; Williams et al., 2018) than short answer extraction. A valid long answer must contain all of the information required to infer the answer. Short answers do not need to contain this information—they need to be surrounded by it.

Motivated by this intuition, we implement a pipelined approach that uses a model drawn from the NLI literature to select long answers. Then short answers are selected from these using a model drawn from the short answer extraction literature.

Long answer selection Let $t(d, l)$ denote the sequence of tokens in d for the long answer candidate l. We then use the Decomposable Attention model (Parikh et al., 2016) model to produce a score for each question, candidate pair $x_l = \text{DecAtt}(q, t(d, l))$. To this we add a 10 dimensional trainable embedding r_l of the long answer candidate's position in the sequence of candidates[13]; an integer u_l containing the number of the words shared by q and $t(d, l)$; and a scalar v_l containing the number of words shared by q and $t(d, l)$ weighted by inverse document frequency. The long answer score z_l is then given as a linear function of the above features $z_l = \mathbf{w}^\top [x_l, r_l, u_l, v_l] + b$ where $\mathbf{w}^\top$ and b are the trainable weight vector and bias respectively,

Short answer selection Given a long answer, the Document Reader model (Chen et al., 2017),

[13] Specifically, we have a unique learned 10 dimensional embedding for each position $1 \ldots 19$ in the sequence, and a 20th embedding used for all positions ≥ 20.

abbreviated DocReader, is used to extract short answers.

Training The long answer selection model is trained by minimizing the negative log-likelihood of the correct answer $l^{(i)}$ with a hyperparameter η that down-weights examples with the NULL label:

$$-\sum_{i=1}^{n}\left(\log\frac{\exp(z_{l^{(i)}})}{\sum_l \exp(z_l)}\right)\times(1-\eta[[l^{(i)}=\texttt{NULL}]])$$

We found that the inclusion of η is useful in accounting for the asymmetry in labels—since a NULL label is less informative than an answer location. Varying η also seems to provide a more stable method of setting a model's precision point than post-hoc thresholding of prediction scores. An analogous strategy is used for the short answer model where examples with no entity answers are given a different weight.

6.4 Results

Table 3 shows results for all baselines as well as a single annotator, and the super-annotator introduced in Section 5. It is clear that there is a great deal of headroom in both tasks. We find that Document-QA performs significantly worse than DecAtt+DocReader in long answer identification. This is likely due to the fact that Document-QA was designed for the short answer task only.

To ground these results in the context of comparable tasks, we measure performance on the subset of NQ that has non NULL labels for both long and short answers. Freed from the decision of whether or not to answer, DecAtt+DocReader gets 68.0% F1 on the long answer task, and 40.4% F1 on the short answer task. We also examine performance of the short answer extraction systems in the setting where the long answer is given, and a short answer is known to exist. With this simplification, short answer F1 increases 57.7% for DocReader. Under this restriction NQ roughly approximates the SQuAD 1.1 task. From the gap to the super-annotator upper bound we know that this task is far from being solved in NQ.

Finally, we break the long answer identification results down according to long answer type. From Table 3 we know that DecAtt+DocReader predicts long answers with 54.8% F1. If we only measure performance on examples that should have a paragraph long answer, this increases to 65.1%. For tables and table rows it is 66.4%. And for lists and list items it is 32.0%. All other examples have a NULL label. Clearly, the model is struggling to learn some aspect of list formatted data from the 6% of the non NULL examples that have this type.

7 Conclusion

We argue that progress on question answering has been hindered by a lack of appropriate training and test data. To address this, we present the Natural Questions corpus. This is the first large publicly available dataset to pair real user queries with high quality annotations of answers in documents. We also present metrics to be used with NQ, for the purposes of evaluating the performance of question answering systems. We demonstrate a high upper bound on these metrics and show that existing methods do not approach this upper bound. We argue that for them to do so will require significant advances in NLU. Figure 5 shows example questions from the dataset. Figure 6 shows example question/answer pairs from the dataset, together with expert judgments and statistics from the 25-way annotations.

when are hops added to the brewing process	what does the word china mean in chinese
when will the white house christmas tree be lit	what is the meaning of sator in latin
who lives in the imperial palace in tokyo	how old was demi lovato when she did camp rock
when does season 15 of ncis come out	what did dorothy ask the wizard of oz
where does the last name hogan come from	how many episodes in season 2 breaking bad
how many parts of 50 shades of grey are there	systemic lupus erythematosus is a condition that sometimes
when does the second season of shooter start	who is the author of the book arabian nights
where is blood pumped after it leaves the right ventricle	where is the bowling hall of fame located
who owns the rights to through the keyhole	what happens when you eat a banana and drink soda
when did the us military start hiring civilian employees	who played will on as the world turns
who won the election for mayor of cleveland	when did the soviet union entered world war ii
where is the world s largest ice sheet located today	who wrote the song then you can tell me goodbye
meaning of the cats in the cradle song	who was married to steve mcdonald in coronation street
where do dust storms occur in the us	who is the voice of tony the tiger
when did the watts riot start and end	what was a key government influence on the constitution of japan
when did kendrick lamars first album come out	who sings now you re just somebody i used to know
where does the energy in a nuclear explosion come from	where did union pacific and central pacific meet

Figure 5: Examples from the questions with 25-way annotations.

Example A1 **Question:** when are hops added to the brewing process **Wikipedia Page:** Brewing
Long answer: ($\mathcal{C}_d$) After mashing, the beer wort is boiled with hops (and other flavourings if used) in a large tank known as a "copper" or brew kettle, though historically the mash vessel was used and is still in some small breweries. The boiling process is where chemical reactions take place, including sterilization of the wort to remove unwanted bacteria, releasing of hop flavours, bitterness and aroma compounds through isomerization, stopping of enzymatic processes, precipitation of proteins, and concentration of the wort. Finally, the vapours produced during the boil volatilise off-flavours, including dimethyl sulfide precursors. The boil is conducted so that it is even and intensea continuous "rolling boil". The boil on average lasts between 45 and 90 minutes, depending on its intensity, the hop addition schedule, and volume of water the brewer expects to evaporate. At the end of the boil, solid particles in the hopped wort are separated out, usually in a vessel called a "whirlpool". **Short answer:** ($\mathcal{C}_d$) The boiling process **Long answer stats:** 13/25, 4/25; **Short answer stats:** 5/25, 1/25

Example A2 **Question:** what does the word china mean in chinese **Wikipedia Page:** Names_of_China
Long answer: ($\mathcal{C}_d$) The names of China include the many contemporary and historical appellations given in various languages for the East Asian country known as Zhongguo (/) in its official language. China, the name in English for the country, was derived from Portuguese in the 16th century, and became popular in the mid 19th century. It is believed to be a borrowing from Middle Persian, and some have traced it further back to Sanskrit. It is also generally thought that the state of Qin that later formed the Qin dynasty is the ultimate source of the name, although there are other suggestions. **Short answer:** NULL **Long answer stats:** 6/25, 3/25; **Short answer stats:** 3/25, 22/25

Example A3 **Question:** who lives in the imperial palace in tokyo **Wikipedia Page:** Tokyo_Imperial_Palace
Long answer: ($\mathcal{C}$) The Tokyo Imperial Palace (, Kkyo, literally "Imperial Residence") is the primary residence of the Emperor of Japan. It is a large park-like area located in the Chiyoda ward of Tokyo and contains buildings including the main palace (, Kyden), the private residences of the Imperial Family, an archive, museums and administrative offices. **Short answer:** ($\mathcal{C}$) The Imperial Family **Long answer stats:** 23/25, 21/25; **Short answer stats:** 22/25, 3/25

Example A4 **Question:** what did dorothy ask the wizard of oz **Wikipedia Page:** The_Wonderful_Wizard_of_Oz
Long answer: ($\mathcal{W}$) Dorothy is a young girl who lives with her Aunt Em and Uncle Henry and her little dog Toto on a farm in the Kansas prairie. One day, Dorothy and Toto are caught up in a cyclone that deposits her farmhouse into Munchkin Country in the magical Land of Oz. The falling house has killed the Wicked Witch of the East, the evil ruler of the Munchkins. The Good Witch of the North arrives with three grateful Munchkins and gives Dorothy the magical Silver Shoes that once belonged to the Wicked Witch. The Good Witch tells Dorothy that the only way she can return home is to go to the Emerald City and ask the great and powerful Wizard of Oz to help her. As Dorothy embarks on her journey, the Good Witch of the North kisses her on the forehead, giving her magical protection from harm. **Short answer:** ($\mathcal{W}$) only way she can return home is to go to the Emerald City and ask the great and powerful Wizard of Oz to help her **Long answer stats:** 9/25, 6/25; **Short answer stats:** 4/25, 1/25

Figure 6: Answer annotations for 4 examples from figure 5 that have long answers that are paragraphs (i.e., not tables or lists). We show the expert judgment ($\mathcal{C}/\mathcal{C}_d/\mathcal{W}$) for each non-null answer. **"Long answer stats"** $a/25$, $b/25$ have a = number of non-null long answers for this question, b = number of long answers the same as that shown in the figure. For example for question A1, 13 out of 25 annotators give some non-null answer, and 4 out of 25 annotators give the same long answer *After mashing* .... **"Short answer stats"** has similar statistics for short answers.

References

Chris Alberti, Kenton Lee, and Michael Collins. 2019. A BERT Baseline for the Natural Questions. *arXiv preprint*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. Association for Computational Linguistics.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879. Association for Computational Linguistics.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184, Brussels, Belgium. Association for Computational Linguistics.

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

Luc Devroye, László Györfi, and Gábor Lugosi. 1997. *A Probabilistic Theory of Pattern Recognition*, corrected 2nd edition, volume 31 of *Applications of Mathematics*. Springer.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. Dureader: a chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 15th International Conference on Computational Linguistics*.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, NIPS'15, Cambridge, MA, USA.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the International Conference on Learning Representations*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611. Association for Computational Linguistics.

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

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

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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches*.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2016. Who did what: A large-scale person-centered cloze dataset. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2230–2235, Austin, Texas. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas. Association for Computational Linguistics.

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

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

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

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 2013. MCTest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 193–203, Seattle, Washington, USA. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.