TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension

Mandar Joshi[†] **Eunsol Choi**[†] **Daniel S. Weld**[†] **Luke Zettlemoyer**[†‡]

[†] Paul G. Allen School of Computer Science & Engineering, Univ. of Washington, Seattle, WA
{mandar90, eunsol, weld, lsz}@cs.washington.edu

[‡] Allen Institute for Artificial Intelligence, Seattle, WA
lukez@allenai.org

TriviaQA는 Reading Comprehension 벤치마크

1) 복잡하고 다양한 요소로 조합됨
2) Question / Answer / Evidence 사이에 어휘적, 문법적 다양성
3) 답을 찾기 위해 상당한 Reasoning이 필요함

65만개의 Question-Answer-Evidence Triple 데이터셋을 사용
9.5만개의 Question-Answer Pair를 포함
각 Question마다 평균 6개의 Evidence가 포함됨

Abstract

We present TriviaQA, a challenging reading comprehension dataset containing over 650K question-answer-evidence triples. TriviaQA includes 95K question-answer pairs authored by trivia enthusiasts and independently gathered evidence documents, six per question on average, that provide high quality distant supervision for answering the questions. We show that, in comparison to other recently introduced large-scale datasets, TriviaQA (1) has relatively complex, compositional questions, (2) has considerable syntactic and lexical variability between questions and corresponding answer-evidence sentences, and (3) requires more cross sentence reasoning to find answers. We also present two baseline algorithms: a feature-based classifier and a state-of-the-art neural network, that performs well on SQuAD reading comprehension. Neither approach comes close to human performance (23% and 40% vs. 80%), suggesting that TriviaQA is a challenging testbed that is worth significant future study.[1]

~와 동떨어진

다양한 요소로 구성된

Question: The Dodecanese Campaign of WWII that was an attempt by the Allied forces to capture islands in the Aegean Sea was the inspiration for which acclaimed 1961 commando film?
Answer: The Guns of Navarone
Excerpt: The Dodecanese Campaign of World War II was an attempt by Allied forces to capture the Italian-held Dodecanese islands in the Aegean Sea following the surrender of Italy in September 1943, and use them as bases against the German-controlled Balkans. The failed campaign, and in particular the Battle of Leros, inspired the 1957 novel **The Guns of Navarone** and the successful 1961 movie of the same name.

발췌

Question: American Callan Pinckney's eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
Answer: Fitness
Excerpt: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney's first video release "Callanetics: 10 Years Younger In 10 Hours" outsold every other **fitness** video in the US.

Figure 1: Question-answer pairs with sample excerpts from evidence documents from TriviaQA exhibiting lexical and syntactic variability, and requiring reasoning from multiple sentences.

1 Introduction

Reading comprehension (RC) systems aim to answer any question that could be posed against the facts in some reference text. This goal is challenging for a number of reasons: (1) the questions can be complex (e.g. have highly compositional semantics), (2) finding the correct answer can require complex reasoning (e.g. combining facts from multiple sentences or background knowledge) and (3) individual facts can be difficult to recover from text (e.g. due to lexical and syntactic variation). Figure 1 shows examples of all these phenomena. This paper presents TriviaQA, a new reading comprehension dataset designed to simultaneously test all of these challenges.

Recently, significant progress has been made by introducing large new reading comprehension datasets that primarily focus on one of the challenges listed above, for example by crowdsourcing the gathering of question answer pairs (Rajpurkar et al., 2016) or using cloze-style sentences instead of questions (Hermann et al., 2015; Onishi et al., 2016) (see Table 1 for more examples). In general, system performance has improved rapidly as each resource is released. The best models of-

[1]Data and code available at http://nlp.cs.washington.edu/triviaqa/

Dataset	Large scale	Freeform Answer	Well formed	Independent of Evidence	Varied Evidence
TriviaQA	✓	✓	✓	✓	✓
SQuAD (Rajpurkar et al., 2016)	✓	✓	✓	✗	✗
MS Marco (Nguyen et al., 2016)	✓	✓	✗	✓	✓
NewsQA(Trischler et al., 2016)	✓	✓	✓	✗*	✗
WikiQA (Yang et al., 2016)	✗	✗	✗	✓	✗
TREC (Voorhees and Tice, 2000)	✗	✓	✓	✓	✓

Table 1: Comparison of TriviaQA with existing QA datasets. Our dataset is unique in that it is naturally occurring, well-formed questions collected independent of the evidences. *NewsQA uses evidence articles indirectly by using only article summaries.

ten achieve near-human performance levels within months or a year, fueling a continual need to build ever more difficult datasets. We argue that TriviaQA is such a dataset, by demonstrating that a high percentage of its questions require solving these challenges and showing that there is a large gap between state-of-the-art methods and human performance levels.

TriviaQA contains over 650K question-answer-evidence triples, that are derived by combining 95K Trivia enthusiast authored question-answer pairs with on average six supporting evidence documents per question. To our knowledge, TriviaQA is the first dataset where full-sentence questions are authored organically (i.e. independently of an NLP task) and evidence documents are collected *retrospectively* from Wikipedia and the Web. This decoupling of question generation from evidence collection allows us to control for potential bias in question style or content, while offering organically generated questions from various topics. Designed to engage humans, TriviaQA presents a new challenge for RC models. They should be able to deal with large amount of text from various sources such as news articles, encyclopedic entries and blog articles, and should handle inference over multiple sentences. For example, our dataset contains three times as many questions that require inference over multiple sentences than the recently released SQuAD (Rajpurkar et al., 2016) dataset. Section 4 present a more detailed discussion of these challenges.

Finally, we present baseline experiments on the TriviaQA dataset, including a linear classifier inspired by work on CNN Dailymail and MCTest (Chen et al., 2016; Richardson et al., 2013) and a state-of-the-art neural network baseline (Seo et al., 2017). The neural model performs best, but only achieves 40% for TriviaQA in comparison to 68% on SQuAD, perhaps due to the challenges listed above. The baseline results also fall far short of human performance levels, 79.7%, suggesting significant room for the future work. In summary, we make the following contributions.

- We collect over 650K question-answer-evidence triples, with questions originating from trivia enthusiasts independent of the evidence documents. A high percentage of the questions are challenging, with substantial syntactic and lexical variability and often requiring multi-sentence reasoning. The dataset and code are available at http://nlp.cs.washington.edu/triviaqa/, offering resources for training new reading-comprehension models.
- We present a manual analysis quantifying the quality of the dataset and the challenges involved in solving the task.
- We present experiments with two baseline methods, demonstrating that the TriviaQA tasks are not easily solved and are worthy of future study.
- In addition to the automatically gathered large-scale (but noisy) dataset, we present a clean, human-annotated subset of 1975 question-document-answer triples whose documents are certified to contain all facts required to answer the questions.

2 Overview

Problem Formulation We frame reading comprehension as the problem of answering a question q given the textual evidence provided by document set D. We assume access to a dataset of tuples $\{(q_i, a_i, D_i)|i = 1 \ldots n\}$ where a_i is a text string that defines the correct answer

to question q_i. Following recent formulations (Rajpurkar et al., 2016), we further assume that a_i appears as a substring for some document in the set D_i.[2] However, we differ by setting D_i as a *set* of documents, where previous work assumed a single document (Hermann et al., 2015) or even just a short paragraph (Rajpurkar et al., 2016).

Data and Distant Supervision Our evidence documents are automatically gathered from either Wikipedia or more general Web search results (details in Section 3). Because we gather evidence using an automated process, the documents are not *guaranteed* to contain all facts needed to answer the question. Therefore, they are best seen as a source of *distant supervision*, based on the assumption that the presence of the answer string in an evidence document implies that the document *does* answer the question.[3] Section 4 shows that this assumption is valid over 75% of the time, making evidence documents a strong source of distant supervision for training machine reading systems.

In particular, we consider two types of distant supervision, depending on the source of our documents. For web search results, we expect the documents that contain the correct answer a to be highly redundant, and therefore let each question-answer-document tuple be an independent data point. ($|D_i| = 1$ for all i and $q_i = q_j$ for many i, j pairs). However, in Wikipedia we generally expect most facts to be stated only once, so we instead pool all of the evidence documents and never repeat the same question in the dataset ($|D_i| = 1.8$ on average and $q_i \neq q_j$ for all i, j). In other words, each question (paired with the union of all of its evidence documents) is a single data point.

These are far from the only assumptions that could be made in this distant supervision setup. For example, our data would also support multi-instance learning, which makes the *at least once assumption*, from relation extraction (Riedel et al., 2010; Hoffmann et al., 2011) or many other possibilities. However, the experiments in Section 6 show that these assumptions do present a strong signal for learning; we believe the data will fuel significant future study.

[2]The data we will present in Section 3 would further support a task formulation where some documents D do not have the correct answer and the model must learn when to abstain. We leave this to future work.

[3]An example context for the first question in Figure 1 where such an assumption fails would be the following evidence string: *The Guns of Navarone is a 1961 British-American epic adventure war film directed by J. Lee Thompson.*

Total number of QA pairs	95,956
Number of unique answers	40,478
Number of evidence documents	662,659
Avg. question length (word)	14
Avg. document length (word)	2,895

Table 2: TriviaQA: Dataset statistics.

3 Dataset Collection

먼저 14개의 trivia / quiz-league 웹사이트에서 Question-Answer를 긁어옴

We collected a large dataset to support the reading comprehension task described above. First we gathered question-answer pairs from 14 trivia and quiz-league websites. We removed questions with less than four tokens, since these were generally either too simple or too vague.

We then collected textual evidence to answer questions using two sources: documents from Web search results and Wikipedia articles for entities in the question. To collect the former, we posed each question[4] as a search query to the Bing Web search API, and collected the top 50 search result URLs. To exclude the trivia websites, we removed from the results all pages from the trivia websites we scraped and any page whose url included the keywords *trivia*, *question*, or *answer*. We then crawled the top 10 search result Web pages and pruned PDF and other ill formatted documents. The search output includes a diverse set of documents such as blog articles, news articles, and encyclopedic entries.

Wikipedia pages for entities mentioned in the question often provide useful information. We therefore collected an additional set of evidence documents by applying TAGME, an off-the-shelf entity linker (Ferragina and Scaiella, 2010), to find Wikipedia entities mentioned in the question, and added the corresponding pages as evidence documents.

Finally, to support learning from distant supervision, we further filtered the evidence documents to exclude those missing the correct answer string and formed evidence document sets as described in Section 2. This left us with 95K question-answer pairs organized into (1) 650K training examples for the Web search results, each contain-

[4]Note that we did *not* use the answer as a part of the search query to avoid biasing the results.

Property	Example annotation	Statistics
Avg. entities / question	Which politician won the **Nobel Peace Prize** in 2009?	1.77 per question
Fine grained answer type	What **fragrant essential oil** is obtained from Damask Rose?	73.5% of questions
Coarse grained answer type	**Who** won the Nobel Peace Prize in 2009?	15.5% of questions
Time frame	What was photographed for the first time in **October 1959**	34% of questions
Comparisons	What is the appropriate name of the **largest** type of frog?	9% of questions

Table 3: Properties of questions on 200 annotated examples show that a majority of TriviaQA questions contain multiple entities. The boldfaced words hint at the presence of corresponding property.

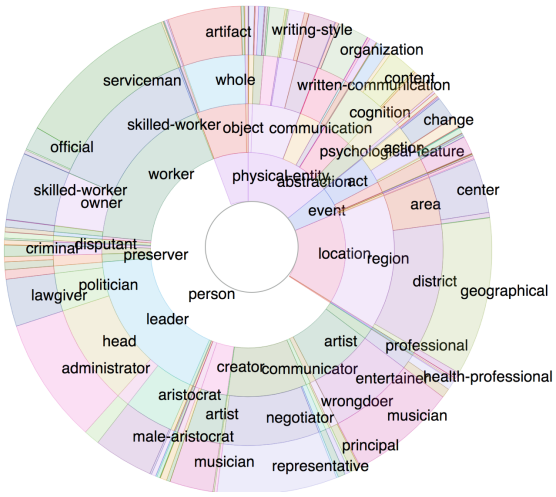

Figure 2: Distribution of hierarchical WordNet synsets for entities appearing in the answer. The arc length is proportional to the number of questions containing that category.

ing a single (combined) evidence document, and (2) 78K examples for the Wikipedia reading comprehension domain, containing on average 1.8 evidence documents per example. Table 2 contains the dataset statistics. While not the focus of this paper, we have also released the full unfiltered dataset which contains 110,495 QA pairs and 740K evidence documents to support research in allied problems such as open domain and IR-style question answering.

4 Dataset Analysis

A quantitative and qualitative analysis of TriviaQA shows it contains complex questions about a diverse set of entities, which are answerable using the evidence documents.

Question and answer analysis TriviaQA questions, authored by trivia enthusiasts, cover various topics of people's interest. The average question length is 14 tokens indicating that many questions are highly compositional. For qualitative analysis, we sampled 200 question answer pairs and manually analysed their properties. About 73.5% of these questions contain phrases that describe a fine grained category to which the answer belongs, while 15.5% hint at a coarse grained category (one of *person*, *organization*, *location*, and *miscellaneous*). Questions often involve reasoning over time frames, as well as making comparisons. A summary of the analysis is presented in Table 3.

Type	Percentage
Numerical	4.17
Free text	2.98
Wikipedia title	92.85
Person	32
Location	23
Organization	5
Misc.	40

Table 4: Distribution of answer types on 200 annotated examples.

Answers in TriviaQA belong to a diverse set of types. 92.85% of the answers are titles in Wikipedia,[5] 4.17% are numerical expressions (e.g., 9 kilometres) while the rest are open ended noun and verb phrases. A coarse grained type analysis of answers that are Wikipedia entities presented in Table 4. It should be noted that not all Wikipedia titles are named entities; many are common phrases such as *barber* or *soup*. Figure 2 shows diverse topics indicated by WordNet synsets of answer entities.

Evidence analysis A qualitative analysis of TriviaQA shows that the evidence contains answers for 79.7% and 75.4% of questions from the Wikipedia and Web domains respectively. To analyse the quality of evidence and evaluate baselines, we asked a human annotator to answer 986 and 1345 (dev and test set) questions from the Wikipedia and Web domains respectively. Trivia

[5]This is a very large set since Wikipedia has more than 11 million titles.

Reasoning	**Lexical variation (synonym)**
	Major correspondences between the question and the answer sentence are synonyms.
Frequency	41% in Wiki documents, 39% in web documents.
Examples	Q What is solid CO2 commonly called?
	S The frozen solid form of CO2, known as **dry ice** ...
	Q Who wrote the novel The Eagle Has landed?
	S The Eagle Has Landed is a book by British writer **Jack Higgins**
Reasoning	**Lexical variation and world knowledge**
	Major correspondences between the question and the document require common sense or external knowledge.
Frequency	17% in Wiki documents, 17% in web documents.
Examples	Q What is the first name of Madame Bovary in Flaubert's 1856 novel?
	S Madame Bovary (1856) is the French writer Gustave Flaubert's debut novel. The story focuses on a doctor's wife, **Emma** Bovary
	Q Who was the female member of the 1980's pop music duo, Eurythmics?
	S Eurythmics were a British music duo consisting of members **Annie Lennox** and David A. Stewart.
Reasoning	**Syntactic Variation**
	After the question is paraphrased into declarative form, its syntactic dependency structure does not match that of the answer sentence
Frequency	69% in Wiki documents, 65% in web documents.
Examples	Q In which country did the Battle of El Alamein take place?
	S The 1942 Battle of El Alamein in **Egypt** was actually two pivotal battles of World War II
	Q Whom was Ronald Reagan referring to when he uttered the famous phrase evil empire in a 1983 speech?
	S The phrase evil empire was first applied to the **Soviet Union** in 1983 by U.S. President Ronald Reagan.
Reasoning	**Multiple sentences**
	Requires reasoning over multiple sentences.
Frequency	40% in Wiki documents, 35% in web documents.
Examples	Q Name the Greek Mythological hero who killed the gorgon Medusa.
	S **Perseus** asks god to aid him. So the goddess Athena and Hermes helps him out to kill Medusa.
	Q Who starred in and directed the 1993 film A Bronx Tale?
	S **Robert De Niro** To Make His Broadway Directorial Debut With A Bronx Tale: The Musical. The actor starred and directed the 1993 film.
Reasoning	**Lists, Table**
	Answer found in tables or lists
Frequency	7% in web documents.
Examples	Q In Moh's Scale of hardness, Talc is at number 1, but what is number 2?
	Q What is the collective name for a group of hawks or falcons?

Table 5: Analysis of reasoning used to answer TriviaQA questions shows that a high proportion of evidence sentence(s) exhibit syntactic and lexical variation with respect to questions. Answers are indicated by boldfaced text.

questions contain multiple clues about the answer(s) not all of which are referenced in the documents. The annotator was asked to answer a question if the minimal set of facts (ignoring temporal references like *this year*) required to answer the question are present in the document, and abstain otherwise. For example, it is possible to answer the question, *Who became president of the Mormons in 1844, organised settlement of the Mormons in Utah 1847 and founded Salt Lake City?* using only the fact that Salt Lake City was founded by Brigham Young. We found that the accuracy (evaluated using the original answers) for the Wikipedia and Web domains was 79.6 and 75.3 respectively. We use the correctly answered questions (and documents) as verified sets for evaluation (section 6).

Challenging problem A comparison of evidence with respect to the questions shows a high proportion of questions require reasoning over multiple sentences. To compare our dataset against previous datasets, we classified 100 question-evidence pairs each from Wikipedia and the Web according to the form of reasoning required to answer them. We focus the analysis on Wikipedia since the analysis on Web documents are similar. Categories are not mutually exclusive: single example can fall into multiple categories. A summary of the analysis is presented in Table 5.

On comparing evidence sentences with their corresponding questions, we found that 69% of the questions had a different syntactic structure while 41% were lexically different. For 40% of the questions, we found that the information re-

quired to answer them was scattered over multiple sentences. Compared to SQuAD, over three times as many questions in TriviaQA require reasoning over multiple sentences. Moreover, 17% of the examples required some form of world knowledge. Question-evidence pairs in TriviaQA display more lexical and syntactic variance than SQuAD. This supports our earlier assertion that decoupling question generation from evidence collection results in a more challenging problem.

5 Baseline methods

To quantify the difficulty level of the dataset for current methods, we present results on neural and other models. We used a random entity baseline and a simple classifier inspired from previous work (Wang et al., 2015; Chen et al., 2016), and compare these to BiDAF (Seo et al., 2017), one of the best performing models for the SQuAD dataset.

5.1 Random entity baseline

We developed the random entity baseline for the Wikipedia domain since the provided documents can be directly mapped to candidate answers. In this heuristic approach, we first construct a candidate answer set using the entities associated with the provided Wikipedia pages for a given question (on average 1.8 per question). We then randomly pick a candidate that does not occur in the question. If no such candidate exists, we pick any random candidate from the candidate set.

5.2 Entity classifier

We also frame the task as a ranking problem over candidate answers in the documents. More formally, given a question q_i, an answer a_i^+, and a evidence document D_i, we want to learn a scoring function $score$, such that

$$score(a_i^+|q_i, D_i) > score(a_i^-|q_i, D_i)$$

where a_i^- is any candidate other than the answer. The function $score$ is learnt using LambdaMART (Wu et al., 2010),[6] a boosted tree based ranking algorithm.

This is similar to previous entity-centric classifiers for QA (Chen et al., 2016; Wang et al., 2015), and uses context and Wikipedia catalog based features. To construct the candidate answer set, we consider sentences that contain at least one word in common with the question. We then add every n-gram ($n \in [1, 5]$) that occurs in these sentences and is a title of some Wikipedia article.[7]

5.3 Neural model

Recurrent neural network models (RNNs) (Hermann et al., 2015; Chen et al., 2016) have been very effective for reading comprehension. For our task, we modified the BiDAF model (Seo et al., 2017), which takes a sequence of context words as input and outputs the start and end positions of the predicted answer in the context. The model utilizes an RNN at the character level, token level, and phrase level to encode context and question and uses attention mechanism between question and context.

Authored independently from the evidence document, TriviaQA does not contain the exact spans of the answers. We approximate the answer span by finding the first match of answer string in the evidence document. Developed for a dataset where the evidence document is a single paragraph (average 122 words), the BiDAF model does not scale to long documents. To overcome this, we truncate the evidence document to the first 800 words.[8]

When the data contains more than one evidence document, as in our Wikipedia domain, we predict for each document separately and aggregate the predictions by taking a sum of confidence scores. More specifically, when the model outputs a candidate answer A_i from n documents $D_{i,1}, ...D_{i,n}$ with confidences $c_{i,1}, ...c_{i,n}$, the score of A_i is given by

$$score(A_i) = \sum_k c_{i,k}$$

We select candidate answer with the highest score.

6 Experiments

An evaluation of our baselines shows that both of our tasks are challenging, and that the TriviaQA dataset supports significant future work.

[6] We use the RankLib implementation https://sourceforge.net/p/lemur/wiki/RankLib/

[7] Using a named entity recognition system to generate candidate entities is not feasible as answers can be common nouns or phrases.

[8] We found that splitting documents into smaller sub documents degrades performance since a majority of sub documents do not contain the answer.

		Train	Dev	Test
Wikipedia	Questions	61,888	7,993	7,701
	Documents	110,648	14,229	13,661
Web	Questions	76,496	9,951	9,509
	Documents	528,979	68,621	65,059
Wikipedia verified	Questions	-	297	584
	Documents	-	305	592
Web verified	Questions	-	322	733
	Documents	-	325	769

Table 6: Data statistics for each task setup. The Wikipedia domain is evaluated over questions while the web domain is evaluated over documents.

6.1 Evaluation Metrics

We use the same evaluation metrics as SQuAD – exact match (EM) and F1 over words in the answer(s). For questions that have *Numerical* and *FreeForm* answers, we use a single given answer as ground truth. For questions that have Wikipedia entities as answers, we use Wikipedia aliases as valid answer along with the given answer.

Since Wikipedia and the web are vastly different in terms of style and content, we report performance on each source separately. While using Wikipedia, we evaluate at the question level since facts needed to answer a question are generally stated only once. On the other hand, due to high information redundancy in web documents (around 6 documents per question), we report document level accuracy and F1 when evaluating on web documents. Lastly, in addition to distant supervision, we also report evaluation on the clean dev and test questions collection using a human annotator (section 4)

6.2 Experimental Setup

We randomly partition QA pairs in the dataset into train (80%), development (10%), and test set (10%). In addition to distant supervision evaluation, we also evaluate baselines on verified subsets (see section 4) of the dev and test partitions. Table 6 contains the number of questions and documents for each task. We trained the entity classifier on a random sample of 50,000 questions from the training set. For training BiDAF on the web domain, we first randomly sampled 80,000 documents. For both domains, we used only those (training) documents where the answer appears in the first 400 tokens to keep training time manageable. Designing scalable techniques that can use the entirety of the data is an interesting direction for future work.

6.3 Results

The performance of the proposed models is summarized in Table 7. The poor performance of the random entity baseline shows that the task is not already solved by information retrieval. For both Wikipedia and web documents, BiDAF (40%) outperforms the classifier (23%). The oracle score is the upper bound on the exact match accuracy.[9] All models lag significantly behind the human baseline of 79.7% on the Wikipedia domain, and 75.4% on the web domain.

We analyse the performance of BiDAF on the development set using Wikipedia as the evidence source by question length and answer type. The accuracy of the system steadily decreased as the length of the questions increased – with 50% for questions with 5 or fewer words to 32% for 20 or more words. This suggests that longer compositional questions are harder for current methods.

6.4 Error analysis

Our qualitative error analysis reveals that compositionality in questions and lexical variation and low signal-to-noise ratio in (full) documents is still a challenge for current methods. We randomly sampled 100 incorrect BiDAF predictions from the development set and used Wikipedia evidence documents for manual analysis. We found that 19 examples lacked evidence in any of the provided documents, 3 had incorrect ground truth, and 3 were valid answers that were not included in the answer key. Furthermore, 12 predictions were partially correct (*Napoleonic* vs *Napoleonic Wars*). This seems to be consistent with human performance of 79.7%.

For the rest, we classified each example into one or more categories listed in Table 8. Distractor entities refers to the presence of entities similar to ground truth. E.g., for the question, *Rebecca Front plays Detective Chief Superintendent Innocent in which TV series?*, the evidence describes all roles played by Rebecca Front.

The first two rows suggest that long and noisy documents make the question answering task more difficult, as compared for example to the short passages in SQuAD. Furthermore, a high proportion of errors are caused by paraphrasing, and the answer is sometimes stated indirectly. For

[9] A question q is considered answerable for the oracle score if the correct answer is found in the evidence D or, in case of the classifier, is a part of the candidate set. Since we truncate documents, the upper bound is not 100%.

Method	Domain	Distant Supervision						Verified					
		Dev			Test			Dev			Test		
		EM	F1	Oracle	EM	F1	Oracle	EM	F1	Oracle	EM	F1	Oracle
Random	Wiki	12.72	22.91	16.30	12.74	22.35	16.28	14.81	23.31	19.53	15.41	25.44	19.19
Classifier		23.42	27.68	71.41	22.45	26.52	71.67	24.91	29.43	80.13	27.23	31.37	77.74
BiDAF		**40.26**	**45.74**	**82.55**	**40.32**	**45.91**	**82.82**	**47.47**	**53.70**	**90.23**	**44.86**	**50.71**	**86.81**
Classifier	web	24.64	29.08	66.78	24.00	28.38	66.35	27.38	31.91	77.23	30.17	34.67	76.72
BiDAF		**41.08**	**47.40**	**82.93**	**40.74**	**47.05**	**82.95**	**51.38**	**55.47**	**90.46**	**49.54**	**55.80**	**89.99**

Table 7: Performance of all systems on TriviaQA using distantly supervised evaluation. The best performing system is indicated in bold.

Category	Proportion
Insufficient evidence	19
Prediction from incorrect document(s)	7
Answer not in clipped document	15
Paraphrasing	29
Distractor entities	11
Reasoning over multiple sentences	18

Table 8: Qualitative error analysis of BiDAF on Wikipedia evidence documents.

example, the evidence for the question *What was Truman Capote's last name before he was adopted by his stepfather?* consists of the following text *Truman Garcia Capote born Truman Streckfus Persons, was an American ... In 1933, he moved to New York City to live with his mother and her second husband, Joseph Capote, who adopted him as his stepson and renamed him Truman Garca Capote.*

7 Related work

Recent interest in question answering has resulted in the creation of several datasets. However, they are either limited in scale or suffer from biases stemming from their construction process. We group existing datasets according to their associated tasks, and compare them against TriviaQA. The analysis is summarized in Table 1.

7.1 Reading comprehension

Reading comprehension tasks aims to test the ability of a system to understand a document using questions based upon its contents. Researchers have constructed cloze-style datasets (Hill et al., 2015; Hermann et al., 2015; Paperno et al., 2016; Onishi et al., 2016), where the task is to predict missing words, often entities, in a document. Cloze-style datasets, while easier to construct large-scale automatically, do not contain natural language questions.

Datasets with natural language questions include MCTest (Richardson et al., 2013), SQuAD (Rajpurkar et al., 2016), and NewsQA (Trischler et al., 2016). MCTest is limited in scale with only 2640 multiple choice questions. SQuAD contains 100K crowdsourced questions and answers paired with short Wikipedia passages. NewsQA uses crowdsourcing to create questions solely from news article summaries in order to control potential bias. The crucial difference between SQuAD/NewsQA and TriviaQA is that TriviaQA questions have not been crowdsourced from preselected passages. Additionally, our evidence set consists of web documents, while SQuAD and NewsQA are limited to Wikipedia and news articles respectively. Other recently released datasets include (Lai et al., 2017).

7.2 Open domain question answering

The recently released MS Marco dataset (Nguyen et al., 2016) also contains independently authored questions and documents drawn from the search results. However, the questions in the dataset are derived from search logs and the answers are crowdsourced. On the other hand, trivia enthusiasts provided both questions and answers for our dataset.

Knowledge base question answering involves converting natural language questions to logical forms that can be executed over a KB. Proposed datasets (Cai and Yates, 2013; Berant et al., 2013; Bordes et al., 2015) are either limited in scale or in the complexity of questions, and can only retrieve facts covered by the KB.

A standard task for open domain IR-style QA is the annual TREC competitions (Voorhees and Tice, 2000), which contains questions from various domains but is limited in size. Many advances from the TREC competitions were used in the IBM Watson system for *Jeopardy!* (Ferrucci et al., 2010). Other datasets includes SearchQA

(Dunn et al., 2017) where *Jeopardy!* questions are paired with search engine snippets, the WikiQA dataset (Yang et al., 2015) for answer sentence selection, and the Chinese language WebQA (Li et al., 2016) dataset, which focuses on the task of answer phrase extraction. TriviaQA contains examples that could be used for both stages of the pipeline, although our focus on this paper is instead on using the data for reading comprehension where the answer is always present.

Other recent approaches attempt to combine structured high precision KBs with semi-structured information sources like OpenIE triples (Fader et al., 2014), HTML tables (Pasupat and Liang, 2015), and large (and noisy) corpora (Sawant and Chakrabarti, 2013; Joshi et al., 2014; Xu et al., 2015). TriviaQA, which has Wikipedia entities as answers, makes it possible to leverage structured KBs like Freebase, which we leave to future work. Furthermore, about 7% of the TriviaQA questions have answers in HTML tables and lists, which could be used to augment these existing resources.

Trivia questions from quiz bowl have been previously used in other question answering tasks (Boyd-Graber et al., 2012). Quiz bowl questions are paragraph length and pyramidal.[10] A number of different aspects of this problem have been carefully studied, typically using classifiers over a pre-defined set of answers (Iyyer et al., 2014) and studying incremental answering to answer as quickly as possible (Boyd-Graber et al., 2012) or using reinforcement learning to model opponent behavior (He et al., 2016). These competitive challenges are not present in our single-sentence question setting. Developing joint models for multi-sentence reasoning for questions and answer documents is an important area for future work.

8 Conclusion and Future Work

We present TriviaQA, a new dataset of 650K question-document-evidence triples. To our knowledge, TriviaQA is the first dataset where questions are authored by trivia enthusiasts, independently of the evidence documents. The evidence documents come from two domains – Web search results and Wikipedia pages – with highly differing levels of information redundancy. Results from current state-of-the-art baselines indicate that TriviaQA is a challenging testbed that deserves significant future study.

While not the focus of this paper, TriviaQA also provides a provides a benchmark for a variety of other tasks such as IR-style question answering, QA over structured KBs and joint modeling of KBs and text, with much more data than previously available.

[10] Pyramidal questions consist of a series of clues about the answer arranged in order from most to least difficult.

Acknowledgments

This work was supported by DARPA contract FA8750-13-2-0019, the WRF/Cable Professorship, gifts from Google and Tencent, and an Allen Distinguished Investigator Award. The authors would like to thank Minjoon Seo for the BiDAF code, and Noah Smith, Srinivasan Iyer, Mark Yatskar, Nicholas FitzGerald, Antoine Bosselut, Dallas Card, and anonymous reviewers for helpful comments.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*. pages 1533–1544. http://aclweb.org/anthology/D/D13/D13-1160.pdf.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR* abs/1506.02075. https://arxiv.org/abs/1506.02075.

Jordan Boyd-Graber, Brianna Satinoff, He He, and Hal Daumé III. 2012. Besting the quiz master: Crowdsourcing incremental classification games. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. Association for Computational Linguistics, Jeju Island, Korea, pages 1290–1301. http://www.aclweb.org/anthology/D12-1118.

Qingqing Cai and Alexander Yates. 2013. Large-scale semantic parsing via schema matching and lexicon extension. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Sofia, Bulgaria, pages 423–433. http://www.aclweb.org/anthology/P13-1042.

Danqi Chen, Jason Bolton, and Christopher D. Manning. 2016. A thorough examination of the

cnn/daily mail reading comprehension task. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 2358–2367. http://www.aclweb.org/anthology/P16-1223.

Matthew Dunn, Levent Sagun, Mike Higgins, Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR* https://arxiv.org/abs/1704.05179.

Anthony Fader, Luke Zettlemoyer, and Oren Etzioni. 2014. Open question answering over curated and extracted knowledge bases. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York, NY, USA, KDD '14, pages 1156–1165. https://doi.org/10.1145/2623330.2623677.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, New York, NY, USA, CIKM '10, pages 1625–1628. https://doi.org/10.1145/1871437.1871689.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building watson: An overview of the deepqa project. *AI MAGAZINE* 31(3):59–79.

He He, Jordan Boyd-Graber, Kevin Kwok, and Hal Daumé III. 2016. Opponent modeling in deep reinforcement learning. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*. PMLR, New York, New York, USA, volume 48 of *Proceedings of Machine Learning Research*, pages 1804–1813. http://proceedings.mlr.press/v48/he16.html.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems*. http://arxiv.org/abs/1506.03340.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *CoRR* https://arxiv.org/abs/1511.02301.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S. Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, pages 541–550. http://www.aclweb.org/anthology/P11-1055.

Mohit Iyyer, Jordan Boyd-Graber, Leonardo Claudino, Richard Socher, and Hal Daumé III. 2014. A neural network for factoid question answering over paragraphs. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 633–644. http://www.aclweb.org/anthology/D14-1070.

Mandar Joshi, Uma Sawant, and Soumen Chakrabarti. 2014. Knowledge graph and corpus driven segmentation and answer inference for telegraphic entity-seeking queries. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1104–1114. http://www.aclweb.org/anthology/D14-1117.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *CoRR* https://arxiv.org/abs/1704.04683.

Peng Li, Wei Li, Zhengyan He, Xuguang Wang, Ying Cao, Jie Zhou, and Wei Xu. 2016. Dataset and neural recurrent sequence labeling model for open-domain factoid question answering. *CoRR* https://arxiv.org/abs/1607.06275.

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 *Workshop in Advances in Neural Information Processing Systems*. https://arxiv.org/pdf/1611.09268.pdf.

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*. Association for Computational Linguistics, Austin, Texas, pages 2230–2235. https://aclweb.org/anthology/D16-1241.

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)*. Association for Computational Linguistics, Berlin, Germany, pages 1525–1534. http://www.aclweb.org/anthology/P16-1144.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the*

Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers. pages 1470–1480. http://aclweb.org/anthology/P/P15/P15-1142.pdf.

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*. Association for Computational Linguistics, Austin, Texas, pages 2383–2392. https://aclweb.org/anthology/D16-1264.

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*. Association for Computational Linguistics, Seattle, Washington, USA, pages 193–203. http://www.aclweb.org/anthology/D13-1020.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part III*. Springer-Verlag, Berlin, Heidelberg, ECML PKDD'10, pages 148–163. http://dl.acm.org/citation.cfm?id=1889788.1889799.

Uma Sawant and Soumen Chakrabarti. 2013. Learning joint query interpretation and response ranking. In *Proceedings of the 22Nd International Conference on World Wide Web*. ACM, New York, NY, USA, WWW '13, pages 1099–1110. https://doi.org/10.1145/2488388.2488484.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1611.01603.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *CoRR* https://arxiv.org/abs/1611.09830.

Ellen M. Voorhees and Dawn M. Tice. 2000. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, New York, NY, USA, SIGIR '00, pages 200–207. https://doi.org/10.1145/345508.345577.

Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 2015. Machine comprehension with syntax, frames, and semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 700–706. http://www.aclweb.org/anthology/P15-2115.

Qiang Wu, Christopher J. Burges, Krysta M. Svore, and Jianfeng Gao. 2010. Adapting boosting for information retrieval measures. *Inf. Retr.* 13(3):254–270. https://doi.org/10.1007/s10791-009-9112-1.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. https://arxiv.org/abs/1502.03044.

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*. Association for Computational Linguistics, Lisbon, Portugal, pages 2013–2018. http://aclweb.org/anthology/D15-1237.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1480–1489. http://www.aclweb.org/anthology/N16-1174.