A Review of Public Datasets in Question Answering Research

B. Barla Cambazoglu
RMIT University
barla.cambazogl@rmit.edu.au

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

Falk Scholer
RMIT University
falk.scholer@rmit.edu.au

Bruce Croft
University of Massachusetts Amherst
croft@cs.umass.edu

Abstract

Recent years have seen an increase in the number of publicly available datasets that are released to foster research in question answering systems. In this work, we survey the available datasets and also provide a simple, multi-faceted classification of those datasets. We further survey the most recent evaluation results that form the current state of the art in question answering research by exploring related research challenges and associated online leaderboards. Finally, we provide a discussion around the existing online challenges and provide a wishlist of datasets whose release could benefit question answering research in the future.

1 Introduction

Question answering (QA) systems have received a lot of research attention in recent years. This attention is mainly motivated by the long-sought transformation in information retrieval (IR) systems. The traditional IR systems are designed to retrieve links to documents that are relevant to short keyword queries. The users of these systems are left with the task of exploring retrieved documents to obtain the final answer. The new generation IR systems, on the other hand, focus on retrieving direct answers to well-formed, natural language questions, thus aiming to decrease the user effort. Although it is inherently a more complex task, direct QA systems are shown to be a feasible alternative to traditional document-based retrieval systems or at least couple well with them.

Together with the recent advances in deep learning techniques and increasing accessibility of very large language models, we have witnessed significant advances in QA research as well. This trend is further supported by an abundance of benchmark datasets that are specifically designed to encourage research in the field. Beside providing a common ground for large-scale experimentation, the availability of such datasets also facilitated the reproducibility of the developed techniques across different research groups.

The released datasets are often coupled with research challenges, which provide online evaluation platforms as well as leaderboards showing the results attained by different research groups on common test samples. These challenges offered two main benefits to the research community. First, the use of common evaluation scripts and metrics improved the comparability of results produced by different techniques, parameter settings or assumptions. Second, the use of leaderboards made the state-of-the-art performance results accessible to large audiences much faster than they are disseminated through scientific publications.

In this paper, we provide a survey of public datasets that are commonly used in QA research. We group these datasets based on a multi-faceted taxonomy and provide some background information for each dataset as well as a discussion of pros and cons. As another contribution, we review various research tasks associated with some of the surveyed datasets and provide pointers to online leaderboards, hoping to provide a sense of where the state-of-the-art stands in QA research. Finally, we provide a short discussion on the existing online challenges in QA research.

2 Datasets, Tasks, and Leaderboards

In this section, we provide a review of public datasets that are used in QA research. Our review excludes community QA datasets (e.g., nfL6 [Cohen and Croft, 2016],[1] ANTIQUE [Hashemi et al., 2020]), where the answers are obtained from community QA websites; knowledge-base QA datasets (e.g., CFQ [Keysers et al., 2020], RuBQ [Korablinov and Braslavski, 2020]), where the answers are extracted from factual knowledge, commonly represented as (subject, predicate, object) triplets; and quiz-style QA datasets with cloze-style tasks (e.g., BookTest [Bajgar et al., 2016], PD&CFT [Cui et al., 2016], CNN-DM [Hermann et al., 2015], CBT [Hill et al., 2016], WDW [Onishi et al., 2016]), where the goal is to predict the missing word in a given piece of text, or multi-choice tasks (e.g., RACE [Lai et al., 2017], MCTest [Richardson et al., 2013]), where the answers are selected from a small set of candidates. We exclude datasets which were released before 2015 as those datasets are small and/or outdated [Voorhees and Tice, 2000]. We omit a dataset if it is not accompanied with a publication that describes the details of the dataset (e.g., `TQuad`),[2] or the resource is no longer accessible at the link provided in the corresponding publication (e.g., SberQuAD [Efimov et al., 2020]). We also exclude QA datasets that are obtained through automatic translation from one language to another (e.g., `K-QuAD` [Lee et al., 2018] and `SQuAD-es` [Carrino et al., 2020], which are obtained by translating the original `SQuAD` dataset [Rajpurkar et al., 2018, 2016] from English to Korean and Spanish, respectively), as well as cross-lingual benchmark datasets (e.g., XQuAD [Artetxe et al., 2020], MLQA [Lewis et al., 2020]). Finally, we omit datasets which involve questions and answers, but designed for tasks other than QA (e.g., `DARC-IT` [Brunato et al., 2018], where the goal is to identify question-worthy sentences). Instead, herein, we focus on large, original, and actively used datasets that are designed for relatively complex, open-domain QA tasks.

In Table 1, we provide a multi-faceted classification of the reviewed datasets. The two "Source" columns in the table show where the questions and documents in the dataset are obtained from, while the "Context" column shows the type of the documents used when

[1] https://ciir.cs.umass.edu/downloads/nfL6

[2] Turkish NLP Q&A dataset, https://github.com/TQuad/turkish-nlp-qa-dataset.

generating the questions and answers. The "Conv." column indicates whether the dataset provides the questions and answers within a conversational context.

Table 2 shows the QA tasks that are associated with the reviewed datasets. The task names are obtained by appending a tag to the dataset name when multiple tasks are associated with a dataset (e.g., `TriviaQA-Wiki` is a task associated with the `TriviaQA` dataset). Otherwise, they are the same as the dataset name (e.g., `TyDiQA`).

The tasks are grouped and presented under three headings as abstraction, extraction, and retrieval. In abstraction, which is the most challenging form of QA, the answer is generated in natural language and in free form without necessarily relying on the vocabulary of the question or document content. Extraction is a slightly less advanced task in that the answer needs to be assembled by identifying and selecting parts of a document that potentially contains an answer to the question. The simplest form of QA is based on retrieval, where the goal is to select an answer to a given question by ranking a number of short text segments, usually passages.

The leaderboards covered in our review are displayed in Table 3. In some cases, community-based leaderboards are available on third-party websites (e.g., `NarrativeQA`[3], `WikiQA`[4], `SearchQA`[5] and `NewsQA`[6]). The metrics reported in these leaderboards are extracted from published papers. We prefer not to include such leaderboards in our review as they are not official leaderboards and tend to have limited coverage. Moreover, these leaderboards do not guarantee that the reported metrics are obtained in a common test environment.

The current leaders as well as the attained performance values are displayed in Table 4. We also display the human performance numbers in the same table when they are available for a task.

2.1 Abstractive Question Answering Datasets

2.1.1 `NarrativeQA`

Dataset. The `NarrativeQA` dataset [Kočiský et al., 2018] was released in 2017 by researchers from DeepMind and University of Oxford.[7] The document collection contains books sampled from the Project Gutenberg library and movie scripts crawled from various movie-related websites. The sampled books and movie scripts are referred to as stories. Moreover, plot summaries are obtained from Wikipedia for each story in the collection using their titles. This resulted in 1,567 story, summary pairs which are verified by human editors.

Questions were generated by crowdworkers based on presented plot summaries. Crowdworkers were instructed to create the questions in such a way that they can be answered by people who have read the full stories but not the summaries. The answers were also generated by crowdworkers based on the content of summaries. The answers were allowed to be short (e.g., phrases or short sentences) and had to contain grammatically correct, brief, and complete sentences. The final dataset contains 46,765 question, answer pairs.

[3] https://paperswithcode.com/sota/question-answering-on-narrativeqa
[4] https://paperswithcode.com/sota/question-answering-on-wikiqa
[5] https://paperswithcode.com/sota/open-domain-question-answering-on-searchqa
[6] https://paperswithcode.com/sota/question-answering-on-newsqa
[7] https://github.com/deepmind/narrativeqa

Table 1: Public datasets used in question answering research

Dataset	Year	Language	Source		Context	Conv.
			Question	Answer		
NarrativeQA	2017	English	Wikipedia	Crowdworkers	Wikipedia	No
QnA	2018	English	Bing query logs	Human editors	Web	No
NLGEN	2018	English	Bing query logs	Human editors	Web	No
DuReader	2018	Chinese	Baidu query logs	Crowdworkers	Web, Baidu Zhidao	No
TWEETQA	2019	English	Crowdworkers	Crowdworkers	Tweets	No
ELI5	2019	English	ELI5 Forum	ELI5 Forum	Forum posts	No
SQuAD	2016	English	Crowdworkers	Crowdworkers	Wikipedia	No
TriviaQA	2017	English	Trivia sites	Trivia sites	None	No
SearchQA	2017	English	Jeopardy! writers	Google	Search snippets	No
NewsQA	2017	English	Crowdworkers	Crowdworkers	News	No
KorQuAD	2018	Korean	Crowdworkers	Crowdworkers	Wikipedia	No
QuAC	2018	English	Crowdworkers	Crowdworkers	Wikipedia	Yes
CoQA	2018	English	Crowdworkers	Crowdworkers	Mixed	Yes
HotpotQA	2018	English	Crowdworkers	Crowdworkers	Wikipedia	No
DRCD	2018	Tr. Chinese	Human editors	Human editors	Wikipedia	No
CMRC	2018	Chinese	Human editors	Human editors	Wikipedia	No
NQ	2019	English	Google query logs	Human editors	Wikipedia	No
TyDiQA	2020	Mixed	Human editors	Human editors	Wikipedia	No
FQuAD	2020	French	Crowdworkers	Crowdworkers	Wikipedia	No
WikiQA	2015	English	Bing query logs	Crowdworkers	Wikipedia	No
PR	2018	English	Bing query logs	Bing	Web	No
CAsT	2020	English	Human editors	Human editors	Wikipedia, Web	Yes

Task. There are two tasks. The first task, `NarrativeQA-Summary`, is to generate free-form answers to questions using the content of given summaries. The second task, `NarrativeQA-Story`, is to generate the answers using the content of stories, instead of summaries. The latter task is much more difficult in that the stories are somewhat longer than the summaries, thus also presenting efficiency challenges for the trained QA models. The ROUGE-L, BLEU-1, BLEU-4, and METEOR measures are used for performance evaluation in both tasks.

2.1.2 QnA

Dataset. The `QnA` dataset [Nguyen et al., 2016] was released by Microsoft.[8] The first version (V1.0) was released in 2016 and contained 100,000 unique query answer pairs. The most recent version (V2.1), which was released in 2018, contained 1,010,916 queries, 8,841,823 passages extracted from 3,563,535 web pages, and 1,026,758 unique query answer pairs, including queries with no answers.

The data contains a sample of queries that were issued through Bing or Cortana. Non-question queries (e.g., those with navigational intent) are filtered out from the sample in a post-processing step using a machine-learned query intent classifier, treating the remaining queries as questions. On average, each question is associated with ten passages extracted from web pages retrieved for the question using Bing. Each question is associated with zero, one, or more answers, which are generated by crowdworkers after inspecting the content of passages retrieved for the question. The answers are expressed in natural language and are strictly limited to the information available in retrieved passages.

Task. The question answering task associated with the `QnA` dataset requires first predicting whether an answer can be provided based on an analysis of the candidate passages associated

[8] https://microsoft.github.io/msmarco

Table 2: Tasks designed for question answering research

Task	Type	Training context	Evaluation metrics
NarrativeQA-Summary	Abstraction	Wikipedia	ROUGE-L, BLEU-1, BLEU-4, METEOR
NarrativeQA-Story	Abstraction	Books, movie scripts	ROUGE-L, BLEU-1, BLEU-4, METEOR
QnA	Abstraction	Web	ROUGE-L, BLEU-1
NLGEN	Abstraction	Web	ROUGE-L, BLEU-1
DuReader	Abstraction	Web, Baidu Zhidao	ROUGE-L, BLEU-4
TWEETQA	Abstraction	Tweets	ROUGE-L, BLEU-1, METEOR
ELI5	Abstraction	Web	ROUGE-L, ROUGE-1, ROUGE-2
SQuAD	Extraction	Wikipedia	EM-L, F1
TriviaQA-Wiki	Extraction	Wikipedia	EM, F1
TriviaQA-Web	Extraction	Web	EM, F1
SearchQA	Extraction	Search snippets	Accuracy, F1
NewsQA	Extraction	News	EM, F1
KorQuAD	Extraction	Wikipedia	EM, F1
QuAC	Extraction	Wikipedia	HEQQ, HEQD, F1
CoQA	Extraction	Mixed	F1
HotpotQA-Dist	Extraction	Wikipedia	EM, F1
HotpotQA-Full	Extraction	Wikipedia	EM, F1
DRCD	Extraction	Wikipedia	EM, F1
CMRC-Test	Extraction	Wikipedia	EM, F1
CMRC-Challenge	Extraction	Wikipedia	EM, F1
NQ	Extraction	Wikipedia	F1
TyDiQA-Passage	Extraction	Wikipedia	F1
TyDiQA-Minimal	Extraction	Wikipedia	F1
FQuAD	Extraction	Wikipedia	EM, F1
WikiQA	Retrieval	Wikipedia	MAP, MRR
PR	Retrieval	Web	MRR@10
CAsT	Retrieval	Wikipedia	MAP, MRR, NDCG@3

with the query. If an answer can be provided, the system is then expected to generate a correct answer for the query in natural language. The performance is evaluated using the ROUGE-L and BLEU-1 measures.

2.1.3 NLGEN

Dataset. The NLGEN dataset [Nguyen et al., 2016] is a follow-up dataset, which is very similar to QnA, as they share the same query sample and passage collection. The two datasets mainly differ in the way the answers are generated. The dataset was released in 2018 by Microsoft.

The answers in the NLGEN dataset are well-formed human answers as in the case of QnA, but they are generated by human editors after performing a post-hoc review of answers that are previously generated by other editors. A well-formed answer is generated for a question if i) the current answer has grammar issues, ii) it appears like the answer is generated by copy-pasting a piece of text from one of the retrieved passages, or iii) understanding the current answer requires having access to the question and passage context. The ground-truth contains 182,669 well-formed query, answer pairs.

Task. The task associated with the NLGEN dataset is also very similar to that of QnA since the systems are expected to generate an answer after analyzing candidate passages of each query. The generated answers, however, are expected to be well-formed, i.e., the answers are expected to make sense even when the context of the question and their passages is not available. In this respect, the task is considerably more challenging than that in QnA. The ROUGE-L and BLEU-1 measures are used in evaluation.

Table 3: Question answering tasks and associated leaderboards

Task	Years	Entries	Leaderboard
QnA	2016–now	77	https://microsoft.github.io/msmarco
NLGEN	2018–now	86	https://microsoft.github.io/msmarco
DuReader	2018–now	243	https://ai.baidu.com/broad/subordinate?dataset=dureader
TWEETQA	2019	9	https://tweetqa.github.io
ELI5	2019	4	https://facebookresearch.github.io/ELI5
SQuAD	2016–now	167	https://rajpurkar.github.io/SQuAD-explorer
TriviaQA-Wiki	2017–now	22	https://competitions.codalab.org/competitions/17208#results
TriviaQA-Web	2017–now	14	https://competitions.codalab.org/competitions/17208#results
KorQuAD	2018–now	100	https://korquad.github.io/KorQuad%201.0
QuAC	2018–now	24	https://quac.ai
CoQA	2018–now	40	https://stanfordnlp.github.io/coqa
HotpotQA-Dist	2018–now	40	https://hotpotqa.github.io
HotpotQA-Full	2018–now	37	https://hotpotqa.github.io
CMRC-Test	2018–now	39	https://ymcui.github.io/cmrc2018
CMRC-Challenge	2018–now	39	https://ymcui.github.io/cmrc2018
NQ-Long	N/A	43	https://ai.google.com/research/NaturalQuestions
NQ-Short	N/A	43	https://ai.google.com/research/NaturalQuestions
TydiQA-Passage	2020–now	3	https://ai.google.com/research/tydiqa
TydiQA-Minimal	2020–now	3	https://ai.google.com/research/tydiqa
FQuAD	2020–now	5	https://illuin-tech.github.io/FQuAD-explorer
PR-Rerank	2018–now	59	https://microsoft.github.io/msmarco
PR-Full	2018–now	34	https://microsoft.github.io/msmarco

2.1.4 DuReader

Dataset. DuReader [He et al., 2018] is a Chinese QA dataset released by Baidu in 2018.[9] Questions are obtained from Baidu search query logs through a multi-step sampling process. First, most frequent one million queries were sampled from the query logs. A classifier was then used to automatically select 280,000 questions from that sample. Human annotators further filtered the selected questions, leaving 210,000 questions. The final question set contained 200,000 questions uniformly sampled from the previous set.

Before creating answers, some relevant documents were obtained for each question by submitting the questions to two different sources: Baidu Search (a Chinese web search engine) and Baidu Zhidao (a Chinese community QA site). The 200,000 unique questions were randomly split into two sets. Each subset was used to retrieve the top five documents from one of the sources. The answers are created through crowdsourcing. The crowdworkers were provided with a question and the set of documents relevant to the question. They were then requested to create an answer to the question in their own words by summarizing the associated documents. The answers created by the crowdworkers were further validated by 52 expert editors, who could correct or improve poor answers.

Task. The task is to analyze the set of relevant documents associated with each question and try to generate an answer to the question that resembles the human-generated answer, as much as possible. The ROUGE-L and BLEU-4 measures are used for evaluation.

2.1.5 TWEETQA

Dataset. The TWEETQA dataset [Xiong et al., 2019] was released in 2019 by UCSB and IBM Research.[10] This is the first large-scale dataset that focuses on social media text. The

[9] https://github.com/baidu/DuReader
[10] https://tweetqa.github.io

Table 4: Current leaders in online leaderboards

Task	Metric	Performance		Current leader
		System	Human	
QnA	Rouge-L	0.540	0.539	Alibaba
	Bleu-1	0.565	0.485	Alibaba
NLGEN	Rouge-L	0.498	0.632	NTT
	Bleu-1	0.501	0.530	NTT
DuReader	Rouge-L	64.38	57.4	Meituan
	Bleu-4	61.54	56.1	Alibaba
TWEETQA	BLEU-1	76.5	70.0	SCU
	METEOR	73.0	66.7	SCU
	ROUGE-L	78.3	73.5	SCU
ELI5	ROUGE-1	30.6	N/A	Facebook
	ROUGE-2	6.2	N/A	Facebook
	ROUGE-L	24.3	N/A	Facebook
SQuAD	EM-L	90.724	86.831	QI-ANXIN
	F1	93.011	89.452	QI-ANXIN
TriviaQA-Wiki	EM	80.86	N/A	Anonymous
	F1	84.50	N/A	Anonymous
TriviaQA-Web	EM	82.99	N/A	Anonymous
	F1	87.18	N/A	Anonymous
KorQuAD	EM	88.10	80.17	Samsung
	F1	95.57	91.20	Samsung
QuAC	F1	74.4	81.1	Tencent
	HEQQ	71.5	100	PAII Inc.
	HEQD	13.9	100	PAII Inc.
CoQA	F1	90.7	88.8	Zhuiyi Technology
HotpotQA-Dist	EM	70.06	N/A	Anonymous
	F1	82.20	N/A	Anonymous
HotpotQA-Full	EM	65.71	N/A	Anonymous
	F1	78.19	N/A	Anonymous
CMRC-Test	EM	74.786	92.400	Tencent
	F1	90.693	97.914	Tencent
CMRC-Challenge	EM	31.923	90.382	HIT and iFLYTEK
	F1	60.177	95.248	HIT and iFLYTEK
NQ-Long	F1	0.7778	N/A	Anonymous
NQ-Short	F1	0.6411	N/A	Anonymous
TydiQA-Passage	F1	77.65	79.9	Anonymous
TydiQA-Minimal	F1	63.40	70.1	Anonymous
FQuAD	EM	82.0	75.9	Illuin Technology
	F1	91.5	91.2	Illuin Technology
PR-Rerank	MRR@10	0.391	N/A	Google
PR-Full	MRR@10	0.419	N/A	Meituan-Dianping

dataset involves tweets and related questions generated by crowdworkers.

The collection of tweets are sampled from archived snapshots of two major news websites (CNN, NBC) as an attempt to limit the scope to more meaningful informative social media content. Uninformative tweets are further filtered out using a supervised machine learning model. The final collection contains 17,794 tweets extracted from 10,898 news articles.

The ground-truth contains tweet, question, answer triples obtained through crowdsourcing. In each human intelligence task in the study, a crowdworker read three tweets and generates two question, answer pairs for each tweet. The crowdworkers generated their answers after inspection of the content of tweets and were free to choose words that do not appear in the tweet when constructing their answers. After all pairs are obtained, further filtering was performed by removing pairs from crowdworkers who did not follow instructions, pairs with yes/no answers, and questions with less than five words. The final ground-truth contained 13,757 triples. To evaluate human performance, the questions in the development and test sets were answered further by additional crowdworkers in a separate study. In this follow-up study, the crowdworkers were presented with tweet, question pairs collected in the previous study. Questions could be labeled as unanswerable at this step.

Task. Given a short tweet and an associated question as input, the task for the `TWEETQA` dataset is to generate a textual answer. The answers could be generated in natural language, rendering the task an abstractive QA task. The evaluation is done using the ROUGE-L, BLEU-1, and METEOR measures.

2.1.6 `ELI5`

Dataset. The `ELI5` dataset [Fan et al., 2019] was released in 2019 by Facebook.[11] This dataset focuses on complex or open-ended questions that seek in-depth answers having multiple sentences.

The questions and answers in the dataset are both sampled from the ELI5 (Explain Like I'm Five) web forum. A question is included in the sample only if its score (the difference between the question's up-votes and down-votes) is at least two and the question has at least one answer with a score of at least two. The sample contained about 272,000 questions. For each question, the answer with the highest vote is assumed to be the correct answer, potentially having multiple correct answers for a question.

Each question in the dataset is associated with a supporting document, created by concatenating sentences extracted from the top 100 web pages matching the question. The sentences are selected based on their tf-idf similarity to the question. The underlying web page collection contains pages from the July 2018 archive of the Common Crawl web repository.

Task. Given a question, the task is to generate a multi-sentence answer for a given question after an inspection of the question's supporting document. Different variants of the ROUGE measure (ROUGE-L, ROUGE-1, ROUGE-2) are used in the evaluation step.

[11] https://facebookresearch.github.io/ELI5

2.2 Extractive Question Answering Datasets

2.2.1 SQuAD

Dataset. SQuAD stands for the Stanford Question Answering Dataset [Rajpurkar et al., 2018, 2016].[12] The SQuAD dataset has two versions which were released in 2016 and 2018.

The document collection is based on passages extracted from Wikipedia articles. In order to focus on high-quality articles, 10,000 English Wikipedia articles with the highest PageRank values were first identified. Then, 536 articles were sampled from this smaller set uniformly at random. Passages were selected from the sampled articles by filtering out images, figures, and tables. Passages that are shorter than 500 characters were excluded, yielding 23,215 passages on a wide range of topics.

The first version of the SQuAD dataset had 107,702 questions. The questions and answer were generated by crowdworkers. To this end, each crowdworker was presented with a Wikipedia passage and tasked with asking and answering up to five questions based on the content of the passage. The crowdworkers generated their answers by highlighting segments of text in the passage. Every question in this version of the dataset had an associated answer. For each question in the development and test sets, at least two additional answers were obtained from crowdworkers in order to get an indication of human performance.

The second version of the dataset added 53,775 new questions. These questions were obtained in a separate crowdsourcing study, where crowdworkers were asked to generate a question that cannot be answered by the presented Wikipedia passage. For each passage, crowdworkers generated up to five questions.

Task. The task for the SQuAD dataset requires making two separate decisions. First, given a question and a passage, a system has to decide whether the given passage contains enough information to answer the question. Second, if there is sufficient information, it should select parts of the passage that answers the question. Therefore, the task aims at measuring a system's ability to answer questions as well as abstaining when an answer is not available. The EM-L and F1 measures are used in the second task for evaluation.

2.2.2 TriviaQA

Dataset. The TriviaQA dataset [Joshi et al., 2017] was released in 2017 by researchers from University of Washington.[13] The questions and answers in the dataset are crawled from 14 different trivia and quiz-league websites, removing questions having less than four tokens. The questions in the dataset are mostly complex and compositional with large syntactic and lexical variability between questions. It is claimed that answering questions requires more cross-sentence reasoning compared to earlier datasets.

Question, answer pairs were associated with evidence documents obtained from two sources, the Web and Wikipedia, leading to two different types of ground-truth. The web documents were obtained by submitting questions to the Bing search API as a query and retrieving the top 50 search results URLs. After removing the trivia websites and non-HTML documents, the content of the top 10 web pages was crawled. Each question, answer, page triplet formed

[12] https://rajpurkar.github.io/SQuAD-explorer
[13] https://nlp.cs.washington.edu/triviaqa

an instance in the ground-truth data. The Wikipedia pages were obtained by mapping entities mentioned in the question to pages using an entity linker. All Wikipedia pages identified for a question, answer pair were then combined into a single evidence document for the pair.

Tasks. The are two tasks: `TriviaQA-Web` and `TriviaQA-Wiki`. The aim of the former task is to answer questions using the web page associated with the question, while using the pool of Wikipedia pages in the latter task. Both tasks relied on the EM and F1 measures.

2.2.3 `SearchQA`

Dataset. The `SearchQA` dataset [Dunn et al., 2017] was released by researchers from New York University in 2017.[14] The question, answer pairs in the dataset were obtained by crawling J! Archive, which hosts questions and answers from the television show Jeopardy! Each pair was associated with a set of search snippets retrieved from Google by submitting the question as a query. The obtained snippets went through some filtering to make sure that answers cannot be found simply by matching question words. A question, answer pair was removed if the answer has more than three words or its search snippets do not contain the answer. After all these processing, the final dataset had 140,461 question, answer pairs, and each pair has 49.6 snippets, on average.

Task. The task is to answer questions based on the context provided by the set of search snippets. The accuracy and F1 measures are the preferred measures for evaluation.

2.2.4 `NewsQA`

Dataset. The `NewsQA` dataset [Trischler et al., 2017] was released by Microsoft in 2017 with the goal of creating a more natural and challenging dataset than the previous ones.[15] This dataset specifically focuses on QA in the news domain and questions that require reasoning ability to come up with an answer.

The underlying document collection contains a sample of news articles obtained from the DeepMind QA Dataset. The original dataset contains 90,266 news articles retrieved from CNN. A total of 12,744 news articles were sampled from this collection uniformly at random.

The questions were generated by crowdworkers in natural language, using partial information from the news articles in the collection. In particular, a crowdworker was shown a news article's headline and summary, and was asked to formulate a question about the article without accessing its full content. This aimed to prevent simple reformulations of sentences in the article's text and to increase the likelihood of obtaining questions that cannot be fully answered by the article. Each crowdworker generated up to three questions for an article this way.

The answers were generated by a separate set of crowdworkers in a succeeding study. In each task of the study, a crowdworker was presented with a full news article together with its associated questions obtained in the previous study. The crowdworkers generated their answers by highlighting a single, continuous span of text within the article if an answer was

[14] https://github.com/nyu-dl/dl4ir-searchQA
[15] https://www.microsoft.com/en-us/research/project/newsqa-dataset

available. Each question received answers from multiple crowdworkers (2.73 answers per question, on average). Question, article pairs where crowdworkers provided different answers went through a validation step, in which a third set of crowdworkers, having access to the question, full article, and the set of unique answers to the question, selected the best answer from the answer set or rejected all answers. After the validation step, 86.0% of questions had answers agreed by at least two crowdworkers. The final ground-truth contained over over 100,000 question, article, answer triplets.

Task. Given a news article and a question about the article, the task is to come up with an answer that resembles the highlighted text segment in the article (the ground-truth answer) as much as possible. A sentence-finding task was also proposed as an additional task, where the goal is to identify the sentence which contains the highlighted text segment. The evaluation is done using the EM and F1 measures.

2.2.5 KorQuAD

Dataset. To the best of our knowledge, KorQuAD [Lim et al., 2019] is the only original QA datasets available in the Korean language at the moment.[16] It was released in 2018 by researchers from the AI Bigdata Research Center at LG CNS. The dataset was created following the methodology introduced in SQuAD [Rajpurkar et al., 2018, 2016].

The documents underlying the KorQuAD dataset are paragraphs that are extracted from 1,637 articles sampled from Korean Wikipedia articles. Great majority of the articles are sampled randomly while a small fraction is formed by high-quality articles. Paragraphs that are shorter than 300 characters or containing mathematical functions are removed, resulting in 11,268 paragraphs. For each paragraph in the dataset, six questions are generated by crowdworkers, who were instructed to generate their questions in natural language based on the content of the paragraphs assigned to them. The crowdworkers were encouraged to ask questions that require reasoning and to refrain from asking questions that involve simple translations. The answers were also generated by the crowdworkers by selecting the minimal text span that answers the question. A total of 70,079 question, answer pairs were generated this way.

Task. Since all questions in the dataset have an answer, the task is simply to attain a high overlap between the answers generated by the QA system and the highlighted text snippets in the ground truth. The EM and F1 measures are used for evaluation.

2.2.6 QuAC

Dataset. QuAC, which stands for Question Answering in Context, is a conversational QA dataset [Choi et al., 2018] released in 2018 by researchers from Allen Institute for Artificial Intelligence, University of Washington, Stanford University, and UMass Amherst.[17] The dataset involves human conversations in the form of question, answer pairs.

The question, answer pairs are created based on sections of Wikipedia articles. The article sections in the collection are biased to those that are related to person entities as these entities

[16] https://github.com/korquad/korquad.github.io
[17] https://quac.ai

are claimed to require less background knowledge for asking good questions. Moreover, articles with less than 100 inbound hyperlinks are removed from the collection.

The conversations involve a sequence of question, answer pairs that are generated by two crowdworkers who discuss an entity related to a section of a Wikipedia article. One crowdworker asks questions based on the title and the first paragraph of the Wikipedia article, while another crowdworker answers the question based on the entire section. During the process, the answerer also provides some hints to guide the asker towards more important or interesting questions. After all conversations are generated, four additional answers are obtained from different crowdworkers for every question in the conversations. All answers are provided simply by highlighting text segments in the Wikipedia article. In total, `QuAC` contains 98,407 questions and 13,594 conversations. Compared to other conversational QA datasets, the questions in `QuAC` are more open-ended.

Task. Given a question about an entity, the task is to extract an answer from a Wikipedia section related to the entity using the conversational context. Here, the context includes the entity, all question, answer pairs prior to the input question, the title and first paragraph of the Wikipedia article, and the section related to the entity. The evaluation is performed using the relatively uncommon HEQQ and HEQD measures, in addition to the commonly used F1 measure.

2.2.7 `CoQA`

Dataset. The `CoQA` dataset [Reddy et al., 2019] was released by the Stanford University in 2018.[18] It is the first public dataset that targets the conversational QA problem.

The dataset is composed of passages and textual human conversations that involve sequences of question, answer pairs about a passage. The passages are extracted from documents obtained in seven different domains: child stories, literature, middle and high school English exams, news articles, Wikipedia articles, Reddit articles, and science articles. In order to generate more interesting conversations, the extraction of passages is biased to those that contain multiple entities and pronominal references.

Conversations are generated by pairing two crowdworkers and letting them chat about a given passage. The two crowdworkers acted as asker and answerer, generating a question, answer pair in each turn of the conversation. The askers were instructed to refrain from using the words in the passage at hand when formulating their questions in order to increase lexical diversity. On the other hand, the answerers were instructed to stick to the vocabulary in the passage as much as possible when formulating their free-form answers in order to limit the number of possible answers. When generating an answer, the answerers highlighted a contiguous text span in the passage as evidence for the answer although this information was not used in evaluation. In total, the `CoQA` dataset contains around 127,000 question, answer pairs spread over about 8000 conversations.

Task. Given a passage and the initial part of a conversation about the passage, the task is to generate an answer to the next question in the conversation. Most of the questions in the `CoQA` dataset require the use of conversation history to generate a meaningful answer. The task is more challenging than the standard QA tasks as it requires coreference resolution and pragmatic reasoning. The performance is evaluated using only the F1 measure.

[18] https://stanfordnlp.github.io/coqa

2.2.8 HotpotQA

Dataset. HotpotQA [Yang et al., 2018] is released in 2018 by researchers from Carnegie Mellon University, Stanford University, and Université de Montréal.[19] Different from most other datasets, the HotpotQA dataset focuses on multi-hop questions, which require accessing multiple documents to generate an answer, and supporting facts for answers in order to improve the explainability of QA systems.

The document collection contains pairs of supporting paragraphs extracted from manually curated Wikipedia articles, ensuring that a multi-hop question can be asked based on the pair. Also, sentences are extracted from the supporting paragraphs to form the collection of possible supporting facts.

The questions are generated by crowdworkers. In each task, a crowdworker was presented with a pair of paragraphs and instructed to ask a question that can be answered only by using both paragraphs. The answers are also generated by crowdworkers, who accompanied their answers with supporting facts (sentences) selected from the given paragraphs. The dataset contains 112,779 question, answer pairs. The training set was split into three subsets as single hop, multi-hop, and hard multi-hop depending on the nature and difficulty of questions. The development and test sets contained only hard multi-hop questions.

Task. There are two QA tasks with different settings: distractor (HotpotQA-Dist) and full wiki (HotpotQA-Full). In the distractor setting, questions are accompanied with ten paragraphs, two of which are the gold paragraphs in the ground-truth data and the rest are the best-matching paragraphs of the question retrieved from Wikipedia using a bigram tf-idf model. In the full wiki setting, the questions are accompanied with the first paragraphs of all Wikipedia articles without specifying the gold paragraphs. There is also an additional task where the goal is to find supporting facts for a provided answer, but we do not cover that task here. The performance is evaluated using the EM and F1 measures in both tasks.

2.2.9 DRCD

Dataset. DRCD [Shao et al., 2018] is an extractive QA dataset specific to traditional Chinese. It was released in 2018 by Delta Electronics.[20] The dataset was created using the methodology proposed for the SQuAD dataset [Rajpurkar et al., 2018, 2016].

The paragraphs extracted from the top 10,000 Chinese Wikipedia pages with highest PageRank values form the context for the created questions and answers. In total, 10,014 paragraphs are extracted from 2,108 pages, removing too short or too long paragraphs. Human editors then created three to five questions per page after reading the entire page, making sure that the answer to the question is contained entirely in one of the paragraphs extracted from the page. After creating a question, the editors selected the text span that answers the question. An additional answer was obtained from a different editor for each question for the development and test sets. The final dataset contained 33,941 questions.

Task. The task is to generate answers to questions in the test set, optimizing the EM and F1 measures.

[19] https://hotpotqa.github.io
[20] https://github.com/DRCKnowledgeTeam/DRCD

2.2.10 CMRC

Dataset. CMRC [Cui et al., 2019] is a Chinese dataset released in 2018 for extractive QA.[21] The underlying document collection is composed of paragraphs extracted from Chinese Wikipedia pages. The extracted paragraphs are reviewed by human editors, and those that are hard to understand, too long, or contain many non-Chinese or special characters were removed.

Questions and answers were removed from the remaining paragraphs by human editors. Each editor created no more than five questions for each passage they received. The editors were instructed to create diverse questions while avoiding textual overlap between the paragraph and questions. While creating a question, the editors also marked the text span that forms an answer to the question. Unlike the training set, which contains a single answer per question, the development, test, and challenge sets included two additional answers obtained from different editors. The challenge set differs from the test set in that it contains questions that require comprehensive reasoning over various clues in the context. To this end, the challenge set was limited to questions whose answer can be extracted from multiple sentences in the paragraph. Also, if the expected answer to the questions is an entity, the answer had to contain at least two entities of the expected entity type. The final dataset contained 19,071 questions.

Task. The objective is to generate answers to questions in the test and challenge sets, which vary in their difficulty, thus leading to two different tasks: CMR-Test and CMR-Challenge. The usual EM and F1 measures are used for evaluation, while the latter measure is modified sightly to handle Chinese text.

2.2.11 NQ

Dataset. The NQ (natural questions) dataset [Kwiatkowski et al., 2019] was released by Google in 2019.[22] This dataset differs from SQuAD in that the answers in the ground-truth are extracted from the entire Wikipedia articles instead of sentences or passages. Therefore, the answers may be relatively longer as they can span multiple passages.

The questions are sampled from anonymized queries issued to the Google search engine. Certain heuristics were applied to select naturally occurring questions that represent real information seeking behavior. The selected questions contain at least eight words and are submitted by multiple users in a short period of time. Additional linguistic rules were used to further improve the quality of the selected questions. Finally, questions whose top five search results in Google do not include a Wikipedia page were removed from the sample.

To create the ground-truth data, annotations are obtained on question, Wikipedia article pairs, using a pool of about 50 human editors. The annotation of each pair involved three consecutive steps. First, the editors decided whether the question at hand is good (seeking facts and having an entity or explanation as a potential answer) or bad (ambiguous, incomprehensible, having false presuppositions, or not seeking facts). Second, if a question is labeled as good, the editors select the smallest HTML bounding box with sufficient information to infer an answer to the question. Bounding boxes are allowed to be paragraphs,

[21] https://github.com/ymcui/cmrc2018
[22] https://ai.google.com/research/NaturalQuestions

tables, items in lists, or entire lists. If none of the selectable HTML bounding boxes can alone provide an answer to the question, the pair was labeled as no answer. Third, if a long answer was available for a question, a short answer (a yes/no answer or one or more entities) was extracted from the long answer. Again, if no short answer can be extracted, the pair is labeled as no answer. The released data contains 323,045 (question, Wikipedia article, long answer, short answer) quadruples. The instances in the training set have a single annotation, while those in the development and test sets have five annotations each.

Task. The data has two different QA tasks, both taking a question together with an entire Wikipedia page as input. The goal of the long answer extraction task (`Long`) is to decide which part of the Wikipedia page provides a complete answer to the question if any. The goal of the short answer extraction task (`Short`) is to extract one or more entity names as answer or to provide a simple yes/no answer. In both tasks, the systems are also expected to decide whether the question is answerable. The performance is evaluated using only the F1 measure.

2.2.12 `TyDiQA`

Dataset. The `TyDiQA` dataset [Clark et al., 2020] was released by Google in 2020 in order to foster QA research for a typologically diverse set of languages.[23] It provides around questions and answers for 11 languages: English, Arabic, Bengali, Finnish, Indonesian, Japanese, Kiswahili, Korean, Russian, Telugu, and Thai.

The questions are generated by human editors. They were shown text segments containing the first 100 characters of Wikipedia articles and asked to generate questions that are loosely related to the displayed text segment. Each question is matched with the top Wikipedia article retrieved by Google search restricted to the Wikipedia domain of question's language. The editors were then given a question, article pair and asked to select the passage that best answers the question or indicate that no answer is available in any of the passages in the article. If there is a passage that provides an answer, however, the editors were asked to select the shortest character span that still forms a satisfactory answer to the question. In total, the dataset contains around 200,000 question, answer pairs.

Task. There are two separate tasks. The first task (passage answer task) is to predict the passage the answer is extracted from. The second task (minimal answer task) is to predict the shortest character span selected by the editors. In both tasks, the performance is evaluated by the F1 measure, while the precision and recall measures are also reported. The final measure is computed by macro-averaging the F1 measures computed for different languages, excluding English.

2.2.13 `FQuAD`

Dataset. `FQuAD` is a QA dataset [d'Hoffschmidt et al., 2020] released in 2020 by researchers from Illuin Technology and ETH Zurich.[24] The questions and answers in the dataset are in French and created by following the methodology proposed for `SQuAD`. The dataset has two versions (1.0 and 1.1).

[23] https://ai.google.com/research/tydiqa
[24] https://github.com/illuin-tech/FQuAD-explorer

The document collection is sampled from 1,769 high-quality French Wikipedia articles. The 1.0 version of the dataset includes 145 articles randomly sampled from this initial set. The 1.1 version has 181 additional, randomly sampled articles, resulting in a total of 326 articles. Paragraphs are extracted from each article, removing articles which have less than 500 characters, resulting in 6,242 and 14,908 paragraphs for the 1.0 and 1.1 versions, respectively. The crowdworkers created their questions based on the content of the paragraphs assigned to them, and they also selected the smallest text span which contains the answer. For each paragraph, three to five questions, answer pairs are created. For each question in the development and test sets, two additional answers are collected to reduce a potential annotation bias. The process resulted in 26,108 and 62,003 question, answer pairs for the 1.0 and 1.1 versions, respectively. The questions in the 1.1 version were relative more difficult than those in the 1.0 version since the crowdworkers were specifically instructed to come up with complex questions when creating the 1.1 version of the dataset.

Task. Like all other `SQuAD` variants, the task was to generate an answer as similar as possible to the one(s) in the ground-truth dataset. The evaluation is performed using the EM and F1 measures.

2.3 Retrieval-Based Question Answering Datasets

2.3.1 `WikiQA`

Dataset. The `WikiQA` dataset [Yang et al., 2015] was released by Microsoft in 2015.[25] It is one of the earliest datasets in QA research. The document collection is based on Wikipedia as each question is associated with sentences extracted from Wikipedia articles.

The questions in the `WikiQA` dataset are sampled from 15 months of Bing query logs after a several steps. Question-like queries were first selected using simple heuristics (e.g., based on the presence of question words and question marks in the query). Entity queries that form false positives were then filtered out. To restrict the focus to factoid questions and improve question quality, queries that were submitted by less than five unique users or led to no clicks on at least one Wikipedia page were removed. Finally, the most frequent 3,050 questions were sampled from the remaining questions.

Each question in the sample is matched with a number of candidate sentences (potential answers) extracted from the summary section of the Wikipedia pages corresponding to the question. A question, sentence pair was labeled as correct by crowdworkers if the sentence answers the question in isolation (assuming coreferences are already resolved). Each question was labeled by three crowdworkers. Sentences with inconsistent labels were verified by additional crowdworkers, and the final labels were assigned based on majority voting. The final dataset included 3,047 questions and 29,258 sentences, of which 1,473 were labeled as correct answers for their respective questions. The dataset includes questions for which there are no correct sentences among the candidate set.

Task. The main task associated with the `WikiQA` dataset is to rank candidate sentences (answers) to a question in decreasing likelihood of correctness. The evaluation is done using the MAP and MRR measures. The answer triggering problem is also considered as a separate task.

[25] https://gist.github.com/shagunsodhani/7cf3677ff2b0028a33e6702fbd260bc5

2.3.2 PR

Dataset. The PR dataset [Nguyen et al., 2016] was released by Microsoft in 2018. This dataset is based on the most recent version (V2.1) of the QnA dataset, described before in Section 2.1.2. It contains the same 1,010,916 user queries issued through Bing and Cortana platforms, and 8,841,823 passages (potential answers) extracted from 3,563,535 web pages.

The ground-truth involves relevance labels associated with query, passage pairs. In order to create the ground-truth, the top 1000 passages matching a query were retrieved by the BM25 heuristic. A subset of retrieved passages were then assessed by human editors. Every passage that can provide an answer to the query was labeled as relevant. Since the passages were not assessed exhaustively, unlabeled passages could be relevant or irrelevant.

Task. The ranking objective associated with the PR dataset comes in two different flavors. In the first task (Rerank), the goal is to rerank 1000 candidate passages retrieved by BM25 in decreasing order of their relevance to the query. In the second task (Full), the goal is to identify the most relevant passages by performing retrieval on the entire passage collection. The MRR@10 measures is used for evaluation.

2.3.3 CAsT

Dataset. CAsT is a dataset for conversational QA, but specifically targeting a retrieval scenario.[26] The first and second versions of the dataset was released by TREC in 2019 and 2020, respectively. The first version of the dataset uses passages from the QnA dataset, Wikipedia articles from TREC CAR, and news articles from Washington Post as corpus. Only the first two collections are used in the second version of the dataset.

The conversations in the dataset are information centric in nature and involve pairs of questions and passages (answers) retrieved from the above-mentioned collections. The questions in the conversation are generated by human editors.

Task. The task involves understanding the conversation context and performing retrieval over the collection of passages to select and rank passages in decreasing order of the likelihood that they form an answer to a given question. Three different evaluation measures are computed: MAP, MRR, and NDCG@3.

3 Discussion

Measures. Interestingly, we observe that the measures used in the evaluation are clearly separated depending on the type of the QA task (Table 2). All abstractive QA tasks use versions of the ROUGE, BLEU, and METEOR measures: ROUGE-L (7 tasks), BLEU-1 (5 tasks), BLEU-4 (3 tasks), METEOR (3 tasks), ROUGE-1 (1 task), and ROUGE-2 (1 task). The extractive QA tasks rely on completely different measures. The most common are F1 and EM, which are used in 17 and 10 tasks, respectively. Finally, retrieval-based QA tasks employ some well-known IR measures, such as MAP and MRR. It is unclear why there is such a clear cut difference in the choice of measures for abstractive and extractive QA tasks, as technically it is possible to interchange the measures.

[26] http://www.treccast.ai

Current leaders. 15 of the 41 task leaders shown in Table 4 are anonymous or could not be resolved to a specific institution. Interestingly, all remaining 26 leaders are research labs of companies, not including even a single university. This may point at the significant progress done in the industry in the field of artificial intelligence, compared to the academia. At the country level, China and the USA have the largest numbers of leaders with 14 and 6 leaders, respectively. Korea, Japan, and France each contribute with two leaders. These results validate the leading role of China and the USA.

System performance. The system performance values reported in the leaderboards vary considerably (Table 4). As an example, the reported ROGUE-L measure values vary between 24.3 and 78.3 for the abstractive QA tasks, while the F1 measure varies between 60.18 and 95.57 for the extractive QA tasks. The high variation is potentially due to differences in the underlying document collections and question generation process, as well as factors such as answer length and language. Similarly, a comparison between abstractive and extractive QA tasks is not possible due the disjoint set of measures used.

Human performance. A comparison of reported human performance values across different tasks also appears infeasible. This is not only because of the intrinsic features of the datasets, but is also due to the many different ways the human performance values are computed. For example, in certain cases, the human performance is evaluated using non-professional assessors over professionally labeled ground-truth data while, in other cases, a subset of the assessors who created the ground-truth data are used to compute the human performance. The number of assessors and averaging is also done in different ways. Not having a standard way to compute the human performance renders comparisons between system and human performance difficult as well. Currently, in about half of the cases where a human performance measure is reported (14 of 26 cases), the system performance is reported to exceed the human performance. It is unclear whether the human performance values reported in all those cases are actually strong baselines. We clearly see an urgent need to standardize the process to compute the human performance in QA tasks.

Wishlist. For many years, the industry had been quite reluctant to release datasets (in particular, large query/question samples) that would facilitate QA research. This has started to change in recent years as major search engine companies, such as Bing, Google, and Baidu, started releasing large, real-life datasets for QA research. We hope this trend to continue in the future. We also hope to see richer datasets including query-level and page-level user engagement data, such as clicks and dwell time, which may facilitate further research in QA. Releasing large-scale mouse tracking data obtained from a real-life QA system is another interesting possibility. Finally, large-scale conversational QA datasets, which can be released by social media companies, will be a great contribution to the research community.

4 Conclusion

To the best of our knowledge, we have presented the most recent and largest review of publicly available datasets used in QA research, providing pointers to related online information. We have provided background information about the way questions and answers in each dataset are created. Moreover, we surveyed the available online QA challenges and the current leaders in each challenge as well as reported human performance values. We hope our review to be useful to the NLP and IR communities.

Acknowledgment

We thank Yiqun Liu and Zhijing Wu from Tsinghua University for their feedback on Chinese question answering datasets.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the cross-lingual transferability of monolingual representations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637. Association for Computational Linguistics, 2020.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. Embracing data abundance: Booktest dataset for reading comprehension. *CoRR*, abs/1610.00956, 2016.

Dominique Brunato, Martina Valeriani, and Felice Dell'Orletta. DARC-IT: a DAtaset for reading comprehension in ITalian. In Elena Cabrio, Alessandro Mazzei, and Fabio Tamburini, editors, *Proceedings of the Fifth Italian Conference on Computational Linguistics*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2018.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. Automatic spanish translation of SQuAD dataset for multi-lingual question answering. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5515–5523. European Language Resources Association, 2020.

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

Jonathan H. Clark, Jennimaria Palomaki, Vitaly Nikolaev, Eunsol Choi, Dan Garrette, Michael Collins, and Tom Kwiatkowski. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470, 2020.

Daniel Cohen and W. Bruce Croft. End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, page 143?146, New York, NY, USA, 2016. Association for Computing Machinery.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. Consensus attention-based neural networks for chinese reading comprehension. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1777–1786, Osaka, Japan, 2016.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5883–5889, Hong Kong, China, November 2019. Association for Computational Linguistics.

Martin d'Hoffschmidt, Wacim Belblidia, Quentin Heinrich, Tom Brendlé, and Maxime Vidal. FQuAD: French question answering dataset. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1193–1208, Online, November 2020. Association for Computational Linguistics.

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

Pavel Efimov, Andrey Chertok, Leonid Boytsov, and Pavel Braslavski. SberQuAD - russian reading comprehension dataset: Description and analysis. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, Linda Cappellato, and Nicola Ferro, editors, *Proceedings of 11th International Conference of the CLEF Association - Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 12260 of *Lecture Notes in Computer Science*, pages 3–15. Springer, 2020.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy, July 2019. Association for Computational Linguistics.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. ANTIQUE: A non-factoid question answering benchmark. In Joemon M. Jose, Emine Yilmaz, João Magalhães, Pablo Castells, Nicola Ferro, Mário J. Silva, and Flávio Martins, editors, *Proceedings of the 42nd European Conference on IR Research - Advances in Information Retrieval*, volume 12036 of *Lecture Notes in Computer Science*, pages 166–173. Springer, 2020.

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. DuReader: a chinese machine reading comprehension dataset from real-world applications. In Eunsol Choi, Minjoon Seo, Danqi Chen, Robin Jia, and Jonathan Berant, editors, *Proceedings of the Workshop on Machine Reading for Question Answering@ACL 2018, Melbourne, Australia, July 19, 2018*, pages 37–46. Association for Computational Linguistics, 2018.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, pages 1693–1701, Cambridge, MA, USA, 2015. MIT Press.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. In Yoshua Bengio and

Yann LeCun, editors, *Proceedings of the 4th International Conference on Learning Representations*, 2016.

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

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations*. OpenReview.net, 2020.

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

Vladislav Korablinov and Pavel Braslavski. RuBQ: A russian dataset for question answering over wikidata. *CoRR*, abs/2005.10659, 2020.

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, and Slav Petrov. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March 2019.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 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, September 2017. Association for Computational Linguistics.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), May 2018.

Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. MLQA: Evaluating cross-lingual extractive question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online, July 2020. Association for Computational Linguistics.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. KorQuAD1.0: Korean QA dataset for machine reading comprehension, 2019.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated MAchine reading COmprehension dataset.

In Tarek Richard Besold, Antoine Bordes, Artur S. d'Avila Garcez, and Greg Wayne, editors, *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org, December 2016.

Takeshi Onishi, Hai Wang, Mohit Bansal, Kevin Gimpel, and David McAllester. 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, November 2016. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 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, November 2016. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 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, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7: 249–266, March 2019.

Matthew Richardson, Christopher J.C. Burges, and Erin Renshaw. 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, October 2013. Association for Computational Linguistics.

Chih-Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. DRCD: a chinese machine reading comprehension dataset. *CoRR*, abs/1806.00920, 2018.

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

Ellen M. Voorhees and Dawn M. Tice. Building a question answering test collection. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 200–207, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132263.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. TWEETQA: A social media focused question answering dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5020–5031, Florence, Italy, July 2019. Association for Computational Linguistics.

Yi Yang, Wen-tau Yih, and Christopher Meek. 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, September 2015. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 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, October-November 2018. Association for Computational Linguistics.