DATASETS FOR MACHINE READING COMPREHENSION: A LITERATURE REVIEW

Vaishali Ingale
Department of Information Technology
Army Institute of Technology, Pune
vingale@aitpune.edu.in

Pushpender Singh
Department of Information Technology
Army Institute of Technology, Pune
Pushpender_16372@aitpune.edu.in

Aditi
Department of Information Technology
Army Institute of Technology, Pune
aditi_16437@aitpune.edu.in

ABSTRACT

Machine reading comprehension aims to teach machines to understand a text like a human, and is a new challenging direction in Artificial Intelligence. Datasets play an important role while describing or building an algorithm for machine reading comprehension. The type of answers we required from developed algorithm depends on datasets.The datasets are classified into two types, namely datasets with extractive answers and datasets with descriptive answers. This article summarize both datasets with an example of each type to get better insight of datasets in machine reading comprehension and which datasets to use depending the requirements.

***K*eywords** Reading Comprehension · Descriptive Answer · Extractive answer · Multiple-choice answers · Datasets

1 Introduction

Over past decades, there has been a growing interest in making the machine understand human languages. And recently, great progress has been made in machine reading comprehension (MRC). In one view, the recent tasks titled MRC can also be seen as the extended tasks of question answering (QA).The past decade has witnessed a huge development in the MRC field, including the soar of numbers of corpus and great progress in techniques. The fast development of the MRC field is driven by various large and realistic datasets released in recent years.The appearance of large-scale datasets above makes training an end-to-end neural MRC model possible. When competing on the leader board, many models and techniques were developed in an attempt to conquer a certain dataset. From word representations, attention mechanisms to high-level architectures, neural models evolve rapidly and even surpass human performance in some tasks. Each dataset is usually composed of documents and questions for testing the document understanding ability. The answers for the raised questions can be obtained through seeking from the documents or selecting the presented options. Here, according to the formats of answers, we classify the datasets into two types, namely datasets with extractive answers, and datasets with descriptive answers, and introduce them respectively in the following subsections.In parallel to this survey, there are also new datasets steadily coming out with more diverse task formulations, and testing more complicated understanding and reasoning abilities.

2 Datasets With Extractive Answers

To test a system's ability of reading comprehension, this kind of datasets, which originates from Cloze style[1] questions, firstly provide the system with a large amount of documents or passages, and then feed it with questions whose answers are segments of corresponding passages. A good system should select a correct text span from a given context. Such comprehension tests are appealing because they are objectively gradable and may measure a range of important abilities, from basic understanding to complex inference [2]. Either sourced by crowd workers or generated automatically from different corpus, these datasets all use a text span in the document as the answer to the proposed question. Many of them released in recent years are large enough for training strong neural models. These datasets include SQuAD, CNN/Daily Mail, CBT, NewsQA, TriviaQA, WIKIHOP. Here detailed review of SQuAD dataset is given.

2.1 SQuAD Dataset

SQuAD One of the most famous dataset of this kind is Stanford Question Answering Dataset (SQuAD) [3]. The Stanford Question Answering Dataset v1.0 (SQuAD v1.0) consists of questions posed by crowd workers on a set of Wikipedia articles, where the answer to each question is a segment of text (or span) from the corresponding reading passage. SQuAD v1.0 contains 107,785 question-answer pairs from 536 articles, which is much larger than previous manually labeled RC datasets. We quote some example question-answer pairs as in Figure 1, where each answer is a span of the document.

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

within a cloud

Figure 1: Question-answer pairs for a sample passage in the SQuAD [3]

In contrast to prior datasets, SQuAD does not provide a list of answer choices for each question.Rather, systems must select the answer from all possible spans in the passage, thus needing to cope with a fairly large number of candidates. While questions with span-based answers are more constrained than the more interpretative questions found in more advanced standardized tests, we still find a rich diversity of questions and answer types in SQuAD. SQuAD develop automatic techniques based on distances in dependency trees to quantify this diversity and stratify the questions by difficulty. The span constraint also comes with the important benefit that span-based answers are easier to evaluate than free-form answers.

2.1.1 Dataset collection

SQuAD dataset was collected in three stages: curating passages, crowd sourcing question-answers on those passages, and obtaining additional answers.

Passage curation: To retrieve high-quality articles, SQuAD used Project Nayuki's Wikipedia's internal PageRanks to obtain the top 10000 articles of English Wikipedia, from which 536 articles were sampled uniformly at random. From each of these articles, individual paragraphs was extracted, stripping away images, figures, tables, and discarding para-graphs shorter than 500 characters. The result was 23,215 paragraphs for the 536 articles covering a wide range of topics, from musical celebrities to abstract concepts. The articles were partitioned randomly into a training set (80%), a development set (10%),and a test set (10%).

Question-answer collection In next stage of data collection, crowdworkers were to create questions. SQuAD developers used the Daemo platform [4], with Amazon Mechanical Turk as its backend. Crowd workers were required to have a 97% HIT acceptance rate, a minimum of 1000 HITs, and be located in the United States or Canada. Workers were asked to spend 4 minutes on every paragraph, and paid $9 per hour for the number of hours required to complete the article.The task was reviewed favorably by crowd workers,receiving positive comments on Turkopticon.

On each paragraph, crowd workers were tasked with asking and answering up to 5 questions on the content of that paragraph. The questions had to be entered in a text field, and the answers had to be highlighted in the paragraph. To guide the workers, tasks contained a sample paragraph, and examples of good and bad questions and answers on that paragraph along with the reasons they were categorized as such. Additionally, crowd workers were encouraged to ask questions in their own words,without copying word phrases from the paragraph.On the interface, this was reinforced by a reminder prompt at the beginning of every paragraph, and by disabling copy-paste functionality on the paragraph text.

Additional answers collection To get an indication of human performance on SQuAD and to make evaluation more robust, at least 2 additional answers were obtained for each question in the development and test sets. In the secondary answer generation task, each crowd worker was shown only the questions along with the paragraphs of an article,and asked to select the shortest span in the para-graph that answered the question. If a question was not answerable by a span in the paragraph, workers were asked to submit the question without marking an answer. Workers were recommended a speed of 5 questions for 2 minutes, and paid at the same rate of $9 per hour for the number of hours required for the entire article. Over the development and test sets, 2.6% of questions were marked unanswerable by atleast one of the additional crowd workers.

Answer Type	Percentage	Example
Date	8.9%	19 October 1512
Other Numeric	10.9%	12
Person	12.9%	Thomas Coke
Location	4.4%	Germany
Other Entity	15.3%	ABC Sports
Common Noun Phrase	31.8%	property damage
Adjective Phrase	3.9%	second-largest
Verb Phrase	5.5%	returned to Earth
Clause	3.7%	to avoid trivialization
Other	2.7%	quietly

Table 1: Answer type distribution in SQuAD[3]

2.1.2 Dataset Analysis

To understand the properties of SQuAD, The questions and answers were analysed in the development set.Specifically, the (i) diversity of answer types (ii) the difficulty of questions in terms of type of reasoning required to answer them, and (iii) the degree of syntactic divergence between the question and answer sentences were explored.

Diversity in answers. The answers were categorised as follows: First the numerical and non-numerical answers were separted.The non-numerical answers are categorized using constituency parses and POS tags generated by Stanford Core NLP [5]. The proper noun phrases are further split into person, location and other entities using NER tags. In Table: 1, see dates and other numbers make up 19.8% of the data; 32.6% of the answers are proper nouns of three different types; 31.8% are common noun phrases answers; and the remaining 15.8% are made up of adjective phrases,verb phrases, clauses and other types.

Reasoning required to answer questions. To get a better understanding of the reasoning required to answer

the questions, 4 questions were sampled from each of the 48 articles in the development set, and then manually labeled the examples with the categories Table 2 . The results show that all examples have some sort of lexical or syntactic divergence between the question and the answer in the passage.

Human Performance Human performance was assessed on SQuAD's development and test sets. Recall that each of the questions in these sets has at least three answers. To evaluate human performance, the second answer to each question treated as the human prediction, and keep the other answers as ground truth answers. The resulting human performance score on the test set is 77.0% for the exact match metric, and 86.8% for F1.

Reasoning	Description	Percentage
Lexical variation(synonymy)	Major correspondences betweenthe question and the answer sentence are synonyms.	33.33%
Lexical variation(world knowledge)	Major correspondences betweenthe question and the answer sen-tence require world knowledge toresolve.	9.12%
Syntactic variation%	After the question is paraphrasedinto declarative form, its syntac-tic depen-dency structure does notmatch that of the answer sentenceeven after local modifications.	64.1%
Multiple sentencereasoning	There is anaphora, or higher-levelfusion of multiple sentences is required.	13.6%
Ambiguous	We don't agree with the crowd-workers' answer, or the questiondoes not have a unique answer	6.1 %

Table 2: SQuAD labeled 192 examples into one or more of the above categories [3]

2.1.3 Human Performance

Human performance was assessed on SQuAD's development and test sets. Recall that each of the questions in these sets has at least three answers. To evaluate human performance, the second answer to each question was treated as the human prediction, and keep the other answers as ground truth answers. The resulting human performance score on the test set is 77.0% for the exact match metric, and 86.8% for F1.

2.1.4 Some other popualar Datasets With Extractive Answers

CNN/Daily Mail CNN and Daily Mail Dataset [6], which was released by Google DeepMind and University of Oxford in 2015, is the first large-scale reading comprehension dataset constructed from natural language materials. Unlike most relevant work which uses templates or syntactic/semantic rules to extract document-query-answer triples, this work collects 93k articles from the CNN and 220k articles from the Daily Mail as the source text.

CBT The Children's Book Test [7] is a part of bAbI project of Facebook AI Research which aims at researching automatic text understanding and reasoning. Children books are chosen because they ensure a clear narrative structure which aids this task. The children stories used in CBT come from books freely available from Project Guntenberg . Questions are formed by enumerating 21 consecutive sentences from chapters in books, of which the first 20 sentences serve as context, and the last one as query after removing one word. 10 candidates are selected from words appearing in either context or query.

NewsQA [8] is a MRC dataset with over 100,000 question and span-answer pairs based off roughly 10,000 CNN news articles. The goal of the NewsQA task is to test MRC models on reasoning skills—beyond word matching

and paraphrasing. Crowd-sourced editors created the questions from the title of the articles and the summary points (provided by CNN) without access to the article itself. A 4-stage collection methodology was employed to generate a more challenging MRC task. More than 44% of the NewsQA questions require inference and synthesis,compared to SQuAD's 20%.

TriviaQA [9] Instead of relying on crowd workers to create question-answer pairs from selected passages like NewsQA and SQuAD, over 650K TriviaQA question-answer-evidence triples are generated through automatic procedures. Firstly, a huge amount of question-answer pairs from 14 trivia and quiz-league websites are gathered and filtered. Then the evidence documents for each question-answer pair are collected from either web search results or Wikipedia articles. Finally, a clean, noise free and human-annotated subset of 1975 triples from TriviaQA is given.

3 Descriptive/Narrative Answer Datasets

Instead of text spans or entities obtained from candidate documents, descriptive answers are whole, stand-alone sentences, which exhibit more fluency and integrity. In addition, in real world, many questions may not be answered simply by a text span or an entity. What's more, presenting answers with their supporting evidence and examples is preferred by human. So in light of these reasons, some descriptive answer datasets are released in recent years.
Here we mainly introduce MS MACRO in detail.

3.1 MS MACRO dataset

The MS MACRO dataset comprises of 1,010,916 anatomized questions—sampled from Bing's search query logs—each with a human generated answer and 182,669 completely human rewritten generated answers. In addition, the dataset contains 8,841,823 passages—extracted from 3,563,535 web documents retrieved by Bing—that provide the information necessary for curating the natural language answers. A question in the MS MARCO dataset may have multiple answers or no answers at all. Using this dataset, three different tasks were proposed with varying levels of difficulty: (i) predict if a question is answerable given a set of context passages, and extract and synthesize the answer as a human would (ii) generate a well-formed answer (if possible) based on the context passages that can be understood with the question and passage context, and finally (iii) rank a set of retrieved passages given a question. The size of the dataset and the fact that the questions are derived from real user search queries distinguishes MS MARCO from other well-known publicly available datasets for machine reading comprehension and question-answering. It is believed that the scale and the real-world nature of this dataset makes it attractive for bench marking machine reading comprehension and question-answering models. Several MRC and QA dataset shave also recently emerged. However, many of these existing datasets are not sufficiently large to train deep neural models with large number of parameters. Large scale existing MRC datasets,when available, are often synthetic. Furthermore, a common characteristic, shared by many of these datasets, is that the questions are usually generated by crowd workers based on provided text spans or documents. In MS MARCO, in contrast, the questions correspond to actual search queries that users submitted to Bing, and therefore may be more representative of a "natural" distribution of information need that users may want to satisfy using, say, an intelligent assistant.

3.1.1 Components of MS MACRO datasets

The MS MARCO dataset consists of six major components:

1. Questions: These are a set of anonymized question queries from Bing's search logs, where the user is looking for a specific answer. Queries with navigational and other intents are excluded from our dataset. This filtering of question queries is performed automatically by a machine learning based classifier trained previously on human annotated data. Selected questions are further annotated by editors based on whether they are answerable using the passages provided.

2. Passages: For each question, on average a set of 10 passages were included which may contain the answer to the question. These passages are extracted from relevant web documents.They are selected by a state-of-the-art passage retrieval system at Bing. The editors are in structed to annotate the passages they use to compose the final answer as is_selected. For questions, where no answer was present in any of the passages, they should all be annotated by setting is_selected to 0.

Questions contains	Percentage
YesNo	7.46%
What	34.96%
How	16.8%
Where	3.46%
When	2.71%
Why	1.67%
Who	3.33%
Which	1.79%
Other	27.83%

Table 3: Distribution of questions based on words questions contain[10]

3. Answers: For each question, the dataset contains zero, or more answers composed manually by the human editors. The editors are instructed to read and understand the questions,inspect the retrieved passages, and then synthesize a natural language answer with the correct information extracted strictly from the passages provided.

4. Well-formed Answers: For some question-answer pairs, the data also contains one or more answers that are generated by a post hoc review-and-rewrite process. This process involves a separate editor reviewing the provided answer and rewriting it if: (i) it does not have proper grammar, (ii) there is a high overlap in the answer and one of the provided passages(indicating that the original editor may have copied the passage directly), or (iii) the answer can not be understood without the question and the passage context.

5. Document: For each of the documents from which the passages were originally extracted from, (i) the URL, (ii) the body text, and (iii) the title we included. These documents were extracted from Bing's index as a separate post-processing step. Roughly 300,000 documents could not be retrieved because they were no longer in the index and for the remaining it is possible—even likely—that the content may have changed since the passages were originally extracted.

6. Question type: Each question is further automatically annotated using a machine learned classifier with one of the following segment labels: (i) NUMERIC, (ii) ENTITY, (iii) LOCATION, (iv) PERSON, or (v) DESCRIPTION (phrase). Table 4 lists the relative size of the different question segments and compares it with the proportion of questions that explicitly contain words like "what" and ""where". Note that because the questions in our dataset are based on web search queries, we are may observe a question like "what is the age of Barack Obama" be expressed simply as "Barack Obama age" in our dataset.

Questions Classification	Percentage
Description	53.12%
Numeric	26.12%
Entity	8.81%
Location	6.17%
Person	5.78%

Table 4: Classification of Questions[10]

3.1.2 The challenges

Using the MS MARCO dataset, three machine learning tasks of diverse difficulty levels were proposed:
The novice task requires the system to first predict whether a question can be answered based only on the information contained in the provided passages. If the question cannot be answered,then the system should return "No Answer Present" as response. If the question can be answered, then the system should generate the correct answer.

The intermediate task is similar to the novice task, except that the generated answer should be well-formed—such that, if the answer is read-aloud then it should make sense even with out the context of the question and retrieved

passages.

The passage re-ranking task is an information retrieval (IR) challenge. Given a question and a set of 1000 retrieved passages using BM25 [11], the system must produce a ranking of the said passages based on how likely they are to contain information relevant to answer the question. This task is targeted to provide a large scale dataset for bench marking emerging neural IR methods[12].

3.1.3 Some other popualar Datasets With Descriptive/Narrative Answers

NarrativeQA [13] dataset contains questions created by editors based on sum-maries of movie scripts and books. The dataset contains about 45,000 question-answer pairs over1,567 stories, evenly split between books and movie scripts. Compared to the news corpus usedin NewsQA, the collection of movie scripts and books are more complex and diverse—allowingthe editors to create questions that may require more complex reasoning. The movie scripts andbooks are also longer documents than the news or wikipedia article, as is the case with NewsQAand SQuAD, respectively.

4 Conclusion

In this paper we categorised datasets required for machine reading comprehension (MRC) task into two classes.We introduced recent datasets in two categories, i.e. SQuAD, CNN/Daily mail, CBT, NewsQA, TriviaQA in Extractive format, MS MARCO and NarrativeQA in descriptive in format. We explained two most widely used datasets for machine reading task.

References

[1] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism Bulletin*, 30(4):415–433, 1953.

[2] Matthew Richardson, Christopher JC 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, 2013.

[3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.

[4] Snehal Neil Gaikwad, Durim Morina, Rohit Nistala, Megha Agarwal, Alison Cossette, Radhika Bhanu, Saiph Savage, Vishwajeet Narwal, Karan Rajpal, Jeff Regino, et al. Daemo: A self-governed crowdsourcing marketplace. In *Adjunct proceedings of the 28th annual ACM symposium on user interface software & technology*, pages 101–102. ACM, 2015.

[5] Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60, 2014.

[6] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701, 2015.

[7] Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

[8] Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*, 2016.

[9] Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

[10] Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. Ms marco: A human generated machine reading comprehension dataset. *arXiv preprint arXiv:1611.09268*, 2016.

[11] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.

[12] Bhaskar Mitra, Nick Craswell, et al. An introduction to neural information retrieval. *Foundations and Trends® in Information Retrieval*, 13(1):1–126, 2018.

[13] Tomáš Kočiskỳ, 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.