English Machine Reading Comprehension Datasets: A Survey

Daria Dzendzik
ADAPT Centre
Dublin City University
Dublin, Ireland
daria.dzendzik
@adaptcentre.ie

Carl Vogel
School of Computer
Science and Statistics
Trinity College Dublin
the University of Dublin
Dublin, Ireland
vogel@tcd.ie

Jennifer Foster
School of Computing
Dublin City University
Dublin, Ireland
jennifer.foster
@dcu.ie

Abstract

This paper surveys 54 English Machine Reading Comprehension datasets, with a view to providing a convenient resource for other researchers interested in this problem. We categorize the datasets according to their question and answer form and compare them across various dimensions including size, vocabulary, data source, method of creation, human performance level, and first question word. Our analysis reveals that Wikipedia is by far the most common data source and that there is a relative lack of *why*, *when*, and *where* questions across datasets.

1 Introduction

Reading comprehension is often tested by measuring a person or system's ability to answer questions on a given text. Machine reading comprehension (MRC) datasets have proliferated in recent years, particularly for the English language – see Fig. 1. The aim of this paper is to make sense of this landscape by providing as extensive as possible a survey of English MRC datasets.

Similar surveys have been carried out previously (Liu et al., 2019; Zhang et al., 2019; Wang, 2020; Ingale et al., 2019; Qiu et al., 2019; Lakshmi and Arivuchelvan, 2019; Baradaran et al., 2020; Zeng et al., 2020) but ours differs in its breadth – 54 datasets compared to the two next largest, 47 (Zeng et al., 2020) and 29 (Baradaran et al., 2020) – and its focus on MRC *datasets* rather than on MRC *systems*. Our survey takes a mostly structured form, with the following information presented for each dataset: size, data source, creation method, human performance level, whether the dataset has been "solved", availability of a leaderboard, the most frequent first question token, and whether the dataset is publicly available. We also categorise each dataset by its question/answer type.

The study contributes to the field as follows:

1. it describes and teases apart the ways in which MRC datasets can vary according to their question and answer types;
2. it provides analysis in a structured and visual form (tables and figures) to facilitate easy comparison between datasets;
3. by providing a systematic comparison, and by reporting the "solvedness" status of a dataset, it brings the attention of the community to less popular and relatively understudied datasets;
4. it contains per-dataset statistics such as number of instances, average question/passage/answer length, vocabulary size and text domain can be used to estimate the computational requirements for training an MRC system.

This paper has been written with the following readers in mind: (1) those who are new to the field and would like to get a quick yet informative overview of English MRC datasets; (2) those who are planning to create a new MRC dataset; (3) MRC system developers, interested in designing the appropriate architecture for a particular dataset, choosing appropriate datasets for a particular architecture, or finding compatible datasets for use in transfer or joint learning.

2 Question, Answer, and Passage Types

All MRC datasets in this survey have three components: *passage*, *question*, and *answer*.[1] We begin with a categorisation based on the types of answers and the way the question is formulated. We divide questions into three main categories: *Statement*, *Query*, and *Question*. Answers are divided into the following categories: *Cloze*, *Multiple Choice*, *Boolean*, *Extractive*, *Generative*. The relationships

[1] We mention the datasets which do not meet this criteria in the supplementary materials Section B and explain why we exclude them.

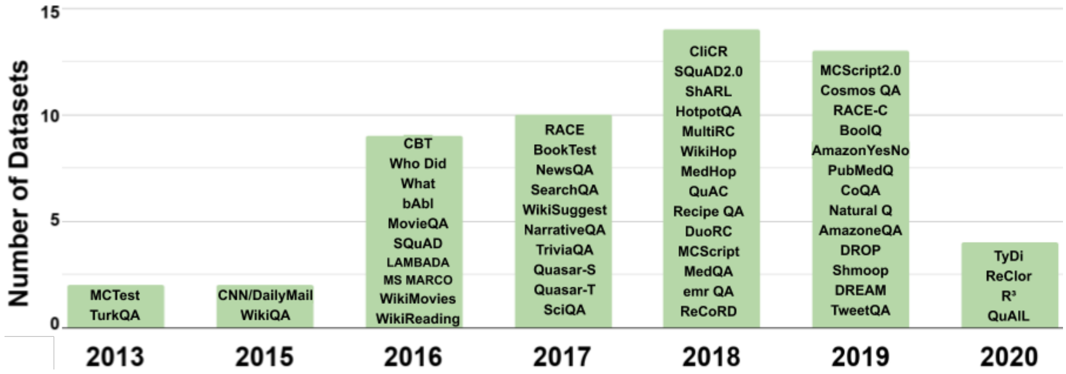

Figure 1: English MRC datasets released per year

between question and answer types are illustrated in Fig. 2. In what follows we briefly describe each question and answer category, followed by a discussion of passage types, and dialog-based datasets.

2.1 Answer Types

Cloze The question is formulated as a sentence with a missing word(s) and the correct entity should be inserted according to the context. We consider the Cloze task in a broader sens as it is not only *insert word* task but also sentence completion. The answer candidates may be included as in (1) from ReciteQA (Yagcioglu et al., 2018), and may not, as in (2) from CliCR (Šuster and Daelemans, 2018).

(1) **Passage (P):** *You will need 3/4 cup of blackberries ... Pour the mixture into cups and insert a popsicle stick in it or pour it in a posicle maker. Place the cup ... in the freezer. ...* **Question (Q):** *Choose the best title for the missing blank to correctly complete the recipe. Ingredients, _, Freeze, Enjoying* **Candidates (AC):** *(A) Cereal Milk Ice Cream (B) Ingredients (C) Pouring (D) Oven* **Answer (A):** C

(2) **P:** ... *intestinal perforation in dengue is very rare and has been reported only in eight patients until today. ...* **Q:** *Perforation peritonitis is a _* **Possible A:** *very rare complication of dengue*

Selective or Multiple Choice (MC) A number of options is given for each question, and the correct one (or a number of correct answers) should be selected, e.g. (3) from MCTest (Richardson et al., 2013).

(3) **P:** *It was Jessie Bear's birthday. She ...* **Q:** *Who was having a birthday?* **AC:** *(A) Jessie Bear (B) no one (C) Lion (D) Tiger* **A:** A

We distinguish cloze multiple choice datasets from other multiple choice datasets. The difference is the form of question: in the cloze datasets, the answer is a missing part of the question context and, combined together, they form a grammatically correct sentence, whereas for other multiple choice datasets, the question has no missing words.

Boolean A "Yes/No" answer is expected, e.g. (4) from the BoolQ dataset (Clark et al., 2019). Some datasets which we put in this category have the third option *"Cannot be answered"* or *"Maybe"*, e.g. (5) from PubMedQuestions (Jin et al., 2019).

(4) **P:** *The series is filmed partially in Prince Edward Island as well as locations in ...* **Q:** *Is anne with an e filmed on pei?* **A:** *Yes*

(5) **P:** ... *Young adults whose families were abstainers in 2000 drank substantially less across quintiles in 2010 than offspring of nonabstaining families. The difference, however, was not statistically significant between quintiles of the conditional distribution. Actual drinking levels in drinking families were not at all or weakly associated with drinking in offspring. ...* **Q:** *Does the familial transmission of drinking patterns persist into young adulthood?* **A:** *Maybe*

Extractive or Span Extractive The answer is a substring of the passage. In other words the task is to determine the start and end index of the characters in the original passage. The string between those two indexes is the answer, as shown in (6) from SQuAD (Rajpurkar et al., 2016).

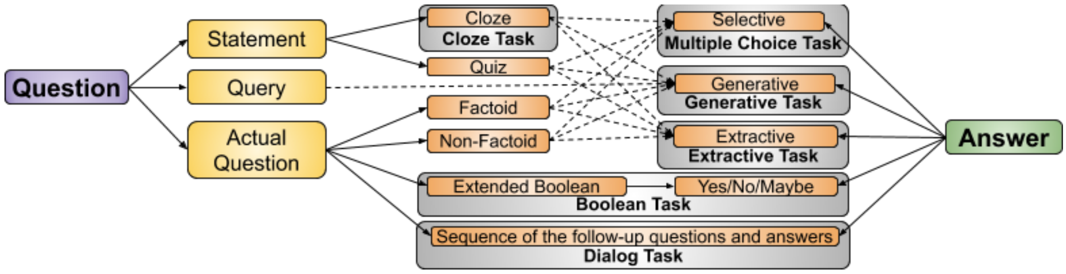

Figure 2: Hierarchy of types of question and answer and the relationships between them. → indicates a sub type whereas ⇢ indicates inclusion.

(6) **P:** *With Rivera having been a linebacker with the Chicago Bears in Super Bowl XX, ...* **Q:** *What team did Rivera play for in Super Bowl XX?* **A:** *46-59: Chicago Bears*

Generative or Free Form Answer The answer must be generated based on information presented in the passage. Although the answer might be in the text, as illustrated in (7) from NarrativeQA (Kočiský et al., 2018), no passage index connections are provided.

(7) **P:** *...Mark decides to broadcast his final message as himself. They finally drive up to the crowd of protesting students, ... The police step in and arrest Mark and Nora....* **Q:** *What are the students doing when Mark and Nora drive up?* **A:** *Protesting.*

2.2 Question Type

Statement The question is an affirmative sentence and used in cloze, e.g. (1-2), and quiz questions, e.g. (8) from SearchQA (Dunn et al., 2017).

(8) **P:** *Jumbuck (noun) is an Australian English term for sheep, ...* **Q:** *Australians call this animal a jumbuck or a monkey* **A:** *Sheep*

Question is an actual question in the standard sense of the word, e.g. (3)-(7). Usualy questions are divided into Factoid (*Who? Where? What? When?*), Non-Factoid (*How? Why?*), and Yes/No.

Query The question is formulated to obtain a particular property of a particular object. It is similar to a knowledge graph query, and, in order to be answered, a part of the passage might involve additional sources as a knowledge graph, or the dataset may have been created using a knowledge graph, e.g. (9) from WikiReading (Hewlett et al., 2016).

(9) **P:** *Cecily Bulstrode (1584-4 August 1609), was a courtier and ... She was the daughter ...* **Q:** *sex or gender* **A:** *female*

We put datasets with more than one type of question into a separate **Mixed** category.

2.3 Passage Type

Passages can take the form of a *one-document* or *multi-document* passage. They can be categorised based on the type of reasoning required to answer a question: *Simple Evidence* where the answer to a question is clearly presented in the passage, e.g. (3) and (6), *Multihop Reasoning* with questions requiring that several facts from different parts of the passage or different documents are combined to obtain the answer, e.g. (10) from the HotpotQA (Yang et al., 2018), and *Extended Reasoning* where general knowledge or common sense reasoning is required, e.g. (11) from the Cosmos dataset (Huang et al., 2019):

(10) **P:** *...2014 S\/S is the debut album of South Korean group WINNER. ... WINNER, is a South Korean boy group formed in 2013 by YG Entertainment and debuted in 2014. ...*
Q: *2014 S\/S is the debut album of a South Korean boy group that was formed by who?*
A: *YG Entertainment*

(11) **P:** *I was a little nervous about this today, but they felt fantastic. I think they'll be a very good pair of shoes. This time I'm going to keep track of the miles on them.*
Q: *Why did the writer feel nervous?*
AC: *(A) None of the above choices. (B) Because the shoes felt fantastic.*
(C) Because they were unsure if the shoes would be good quality. (D) Because the writer thinks the shoes will be very good. **A:** C

Dataset	Size (questions)	Data Source	Q/A Source	LB	Human Performance	Sol-ved	TM FW	PAD
Cloze Datasets								
CNN/Daily Mail(Hermann et al., 2015)	387k/997k	CNN/DailyMail	AG	✻	-	✗	-	✓
Children BookTest (Hill et al., 2016)	687k	Project Gutenberg	AG	✻	82	✓	-	✓
Who Did What (Onishi et al., 2016)	186k	Gigaword	AG	✓	84	✗	-	✉
BookTest (Bajgar et al., 2017)	14M	Project Gutenberg	AG	✗	-	✗	-	✗
Quasar-S (Dhingra et al., 2017)	37k	Stack Overflow	AG	✗	46.8/50.0	✗	-	✓
RecipeQA (Yagcioglu et al., 2018)	9.8k	instructibles.com	AG	✓	73.6	✗	-	✓
CliCR (Šuster and Daelemans, 2018)	105k	Clinical Reports	AG	✻	53.7/45.1 (F1)	✗	-	✉
ReCoRD (Zhang et al., 2018a)	121k	CNN	AG	✓✻	91.3/91.7 (F1)	✓	-	✓
Shmoop (Chaudhury et al., 2019)	7.2k	Project Gutenberg	ER, AG	✗	-	✗	-	✉
Multiple Choice Datasets								
MCTest (Richardson et al., 2013)	2k/640	Stories	CRW	✓✻	95.3	✗	*what*	✓
WikiQA (Yang et al., 2015)	3k	Wikipedia	UG, CRW	✻	-	✗	*what*	✓
bAbI (Weston et al., 2016)	40k	AG	AG	✻	100	✓	*what*	✓
MovieQA (Tapaswi et al., 2016)	15k	Wikipedia	annotators	✓	-	✗	*what*	✉
RACE (Lai et al., 2017)	98k	ER	experts	✓✻	73.3/94.5	✗	*what*	✓
SciQ (Welbl et al., 2017)	12k	Science Books	CRW	✗	87.8	✗	*what*	✓
MultiRC (Khashabi et al., 2018)	10k	reports, News, Wikipedia, ...	CRW	✓✻	81.8(F1)	✓	*what*	✓
MedQA (Zhang et al., 2018b)	235k	Medical Books	expert	✗	-	✓	-	✗
MCScript (Ostermann et al., 2018)	14k	Scripts, CRW	CRW	✗	98.0	✗	*how*	✓
MCScript2.0 (Ostermann et al., 2019)	20k	Scripts, CRW	CRW	✗	97.0	✗	*what*	✓
RACE-C (Liang et al., 2019)	14k	ER	experts	✗	-	✗	*the*	✓
DREAM(Sun et al., 2019)	10k	ER	experts	✓	98.6	✗	*what*	✓
Cosmos QA (Huang et al., 2019)	36k	Blogs	CRW	✓	94	✗	*what*	✓
ReClor (Yu et al., 2020)	6k	ER	experts	✓	63.0	✗	*which*	✓
QuAIL (Rogers et al., 2020)	15k	News, Stories, Fiction, Blogs, UG	CRW, experts	✓	60.0	✗	-	✓
Boolean Questions								
BoolQ (Clark et al., 2019)	16k	Wikipedia	UG, CRW	✓✻	89	✓	*is*	✓
AmazonYesNo (Dzendzik et al., 2019)	80k	Reviews	UG	✗	-	✗	*does*	✉
PubMedQA (Jin et al., 2019)	211k	PubMed	CRW	✓	78	✗	*does*	✓
Extractive Datasets								
SQuAD (Rajpurkar et al., 2016)	108k	Wikipedia	CRW	✓✻	86.8(F1)	✓	*what*	✓
SQuAD2.0 (Rajpurkar et al., 2018)	151k	Wikipedia	CRW	✓✻	89.5(F1)	✓	*what*	✓
NewsQA (Trischler et al., 2017)	120k	CNN	CRW	✻	69.4(F1)	✓	*what*	✓
SearchQA (Dunn et al., 2017)	140k	CRW, AG	J!Archive	✻	57.6(F1)	✓	*this*	✓
Generative Datasets								
MS MARCO (Nguyen et al., 2016)	100k	Web documents	UG, HG	✓✻	-	✗	*what*	✓
LAMBADA (Paperno et al., 2016)	10k	BookCorpus	CRW, AG	✗	-	✗	-	✓
WikiMovies (Miller et al., 2016; Watanabe et al., 2017)	116k	Wikipedia, KG	CRW, AG, KG	✗	93.9 (hit@1)	✗	*what*	✓
WikiSuggest (Choi et al., 2017)	3.47M	Wikipedia	CRW, AG	✗	-	✗	-	✗

Continued on next page

Table 1 – *Continued from previous page*

Dataset	Size	Data Source	Q/A Source	LB	HP	Sol-ved	TM FW	PAD
TriviaQA (Joshi et al., 2017)	96k	Wikipedia, Web docs	Trivia, CRW	✓✻	79.7/75.4 wiki/web	✗	*which*	✓
NarrativeQA (Kočiský et al., 2018)	47k	Wikipedia, Project Gutenberg, movie, HG	HG	✻	19.7 BLEU4	✓	*what*	✓
TweetQA (Xiong et al., 2019)	14k	News, Twitter, HG	CRW	✓	70.0 BLEU1	✓	*what*	✓
Conversational Datasets								
ShARC (Saeidi et al., 2018)	32k	Legal web sites	CRW	✓	93.9	✗	*can*	✓
CoQA (Reddy et al., 2019)	127k	Books, News, Wikipedia, ER	CRW	✓✻	88.8	✓	*what*	✓
Mixed Datasets								
TurkQA (Malon and Bai, 2013)	54k	Wikipedia	CRW	✗	-	✗	*what*	✓
WikiReading (Hewlett et al., 2016)	18.9M	Wikipedia	AG, KG	✗	-	✗	-	✓
Quasar-T (Dhingra et al., 2017)	43k	Trivia ClueWeb09	AG	✻	60.4/60.6 (F1)	✗	*what*	✓
HotpotQA (Yang et al., 2018)	113k	Wikipedia	CRW	✓✻	96.37(F1)	✗	*what*	✓
QAngaroo WikiHop (Welbl et al., 2018)	51k	Wikipedia	CRW, KG	✓✻	85.0	✗	-	✓
QAngaroo MedHop (Welbl et al., 2018)	2.5k	Medline abstracts	CRW, KG	✓	-	✗	-	✓
QuAC (Choi et al., 2018)	98k	Wikipedia	CRW	✓✻	81.1(F1)	✗	*what*	✓
DuoRC (Saha et al., 2018)	86k	Wikipedia + IMDB	CRW	✓	-	✗	*who*	✓
emr QA (Pampari et al., 2018)	456k	Clinic Records	expert, AG	✗	-	✗	*does*	✉
DROP (Dua et al., 2019)	97k	Wikipedia	CRW	✓	96.4(F1)	✗	*how*	✓
NaturalQuestions (Kwiatkowski et al., 2019)	323k	Wikipedia	UG, CRW	✓✻	87/76 L/S(F1)	✗	*who*	✓
AmazonQA (Gupta et al., 2019b)	570k	UG Review	UG	✗	53.5	✗	*does*	✓
TyDi (Clark et al., 2020)	11k	Wikipedia	CRW	✓	54.4(F1)	✗	*what*	✓
R^3 (Wang et al., 2020b)	60k	Wikipedia	CRW	✗	-	✗	-	✗

Table 1: Reading comprehension datasets comparison. Where: **LB** – leader board available; **Human Performance** – (expert/non-expert if other not specified): accuracy if other is not specified; **TMFW** – the most frequent first word; **PAD** – publicly available data; **k/M** – thousands/millions; **CRW** – crowdsourcing; **AG** – automatically generated; **KG** – knowledge graph; **ER** – educational resources; **UG** – user generated; **HG** – human generated (UG + annotators, crw, experts); **L/S** – long/short answer; ✓ – available/“solved”; ✗ – unavailable/not “solved”; ✻ – the leader board is presented at https://paperswithcode.com/; ✉ – the dataset is available by request. The information is verified in June 2020.

2.4 Conversational MRC

We put **Conversational** or **Dialog** datasets in a separate category as there is a unique combination of passage, question, and answer. The passage has a particular context and is then completed by a number of follow-up questions and answers. The full passage is presented as a conversation and the question should be answered based on previous utterances as illustrated in (12) from ShARC (Saeidi et al., 2018), where the scenario is an additional part of the passage unique for each dialog. The question asked before and its answer becomes a part of the passage for the following question. [2]

(12) **P:** *Eligibility. You'll be able to claim the new State Pension if you're: a man born on or after 6 April 1951, a woman born on or after 6 April 1953*
Scenario: *I'm female and I was born in 1966*
Q: *Am I able to claim the new State Pension?*
Follow ups: (1)*Are you a man born on or after 6 April 1951? – No* (2)*Are you a woman born on or after 6 April 1953? – Yes* **A:** *Yes*

3 Datasets

All datasets and their properties of interest are listed in Table 1.[3] We indicate the number of questions per dataset (size), the text sources, the method of creation, whether there are a leaderboard and data publicly available, and whether the dataset is *solved*, i.e. the performance of a MRC system exceeds the reported human performance (also shown). We will discuss each of these aspects.

3.1 Data Sources

A significant proportion of datasets (21 out of 54) use Wikipedia as a passage source. Six of those use Wikipedia along with additional sources. Other popular sources of text data are news (CNN/DailyMail, WhoDidWhat, NewsQA, CoQA, MultiRC, ReCoRD, QuAIL), books, including Project Gutenberg and BookCorpus[4] (Zhu et al., 2015), (ChildrenBookTest, BookTest, LAMBADA, partly CoQA, Shmoop, SciQ), movie scripts (MovieQA, WikiMovies, DuoRC), and a combination of these (MultiRC and NarrativeQA). Five datasets (CliCR, PubMedQuestions, MedQA, emrQA, QAngaroo MedHop) were created in the medical domain based on clinical reports, medical books, MEDLINE abstracts, and PubMed. ShARL is based on a legal resource websites.[5] Some datasets make use of exam questions. RACE, RACE-C, and DREAM use data from English as a Foreign Language examinations, ReClor from the Graduate Management Admission Test (GMAT) and The Law School Admission Test (LSAT),[6] and MedQA from medical exams. Other sources of data include personal narratives from the Spinn3r Blog Dataset (Burton et al., 2009) (MCScript, MCScript2.0, CosmosQA), StackOverflow.com (Quasar-S), Quora.com (QuAIL) Twitter.com (TweetQA),[7] Amazon.com user reviews and questions (AmazonQA, AmazonYesNo), and a cooking website[8] (RecipeQA).

Fig. 3 shows the domains used by datasets as well as any overlaps between datasets. Some datasets share not only text sources but the actual samples. SQuAD2.0 extends SQuAD with unanswerable questions. AmazonQA and AmazonYesNo overlap in questions and passages with slightly different processing. BoolQ shares 3k questions and passages with the NaturalQuestions dataset. The R^3 dataset is fully based on DROP with a focus on reasoning.

3.2 Dataset Creation

Rule-based approaches have been used to automatically obtain questions and passages for the MRC task by generating the sentences (e.g. bAbI) or, in the case of cloze type questions, excluding a word from the context. We call those methods *automatically generated* (AG). Most dataset creation, however, involves a human in the loop. We distinguish three types of people: *experts* are professionals in a specific domain; *crowdworkers* (CRW) are casual workers who normally meet certain criteria (for example a particular level of proficiency in the dataset language) but are not experts in the subject area; *users* who voluntarily create content based on their personal needs and interests.

[2]We do not include DREAM (Sun et al., 2019) in this category as, even though the passages are in dialog form, the questions are about the dialog but not a part of it. That is why DREAM is in the Multiple-Choice category.

[3]Extra features are in the supplementary materials Table 3.

[4]Gutenberg: `www.gutenberg.org` BookCorpus: `yknzhu.wixsite.com/mbweb` – all links last verified (l.v.) 02/2020

[5]For example: `www.benefits.gov/` `www.gov.uk/`, `www.usa.gov/` – all links l.v. 06/2020

[6]GMAT: `www.mba.com/exams/gmat/`, LSAT: `www.lsac.org/lsat` – all links l.v. 03/2020

[7]Xiong et al. (2019) selected tweets featured in the news.

[8]`www.instructables.com/cooking` – l.v. 03/2020

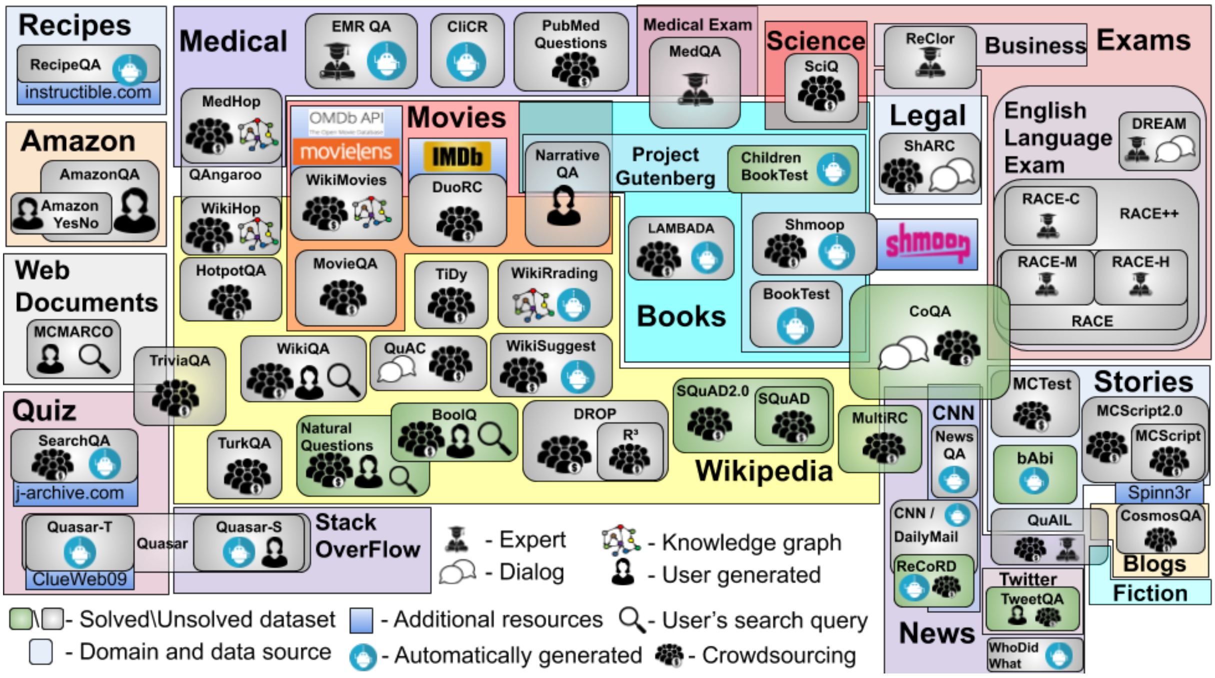

Figure 3: Question Answering Reading Comprehension datasets overview.

More than half of the datasets (33 out of 54) were created using crowdworkers. In one scenario, crowdworkers have access to the passage and must formulate questions based on it. For example, MovieQA, ShaRC, SQuAD, and SQuAD2.0 were created in this way. In contrast, another scenario involves finding a passage containing the answer for a given question. That works well for datasets where questions are taken from already existing resources such as trivia and quiz questions: TriviaQA, Quasar-T, and SearchQA, or using web search queries and results from Google and Bing as a source of questions and passages: BoolQ, NaturalQuestions, MS MARCO.

In an attempt to avoid word repetition between passages and questions, some datasets used different texts about the same topic as a passage and a source of questions. For example, DuoRC takes descriptions of the same movie from Wikipedia and IMDB. One description is used as a passage while another is used for creating the questions. NewsQA uses only a title and a short news article summary as a source for questions while the whole text becomes the passage. Similarly, in NarrativeQA, only the abstracts of the story were used for the question creation. For MCScript and MCScript 2.0, questions and passages were created by different sets of crowdworkers given the same script.

3.3 Quantitative Analysis

Each dataset's size is shown in Table 1. The majority of datasets contain 100k+ questions which makes them suitable for training and/or fine tuning a deep learning model. A few datasets contain fewer than 10k samples: MultiRC (9.9k), Shmoop (7.2k), ReClor (6.1k), QAngaroo MedHop (2.5k), WikiQA (2k). Every dataset has its own structure and data format but we processed all datasets the same way extracting lists of questions, passages, and answers, including answer candidates, using the *spaCy*[9] tokenizer.

Question/Passage/Answer Length The graphs in Fig. 4 provide more insight into the differences between the datasets in terms of answer, question, and passage length, as well as vocabulary size.[10] The outliers are highlighted.[11] The majority of datasets have a passage length under 1500 tokens with the median being 329 tokens but due to seven outliers the average number tokens is 1250 (Fig. 4 (a)). Some datasets (MS MARCO, SearchQA, AmazonYesNo, AmazonQA, MedQA) have a collection of documents as a passage but others contain just a few sentences. The number of tokens in a question lies mostly between 5 and

[9] `spacy.io/api/tokenizer` – l.v. 03/2020

[10] See supplementary materials Table 4 for details.

[11] We use `matplotlib` for calculation and visualisation: `https://matplotlib.org/` – l.v. 10/2020

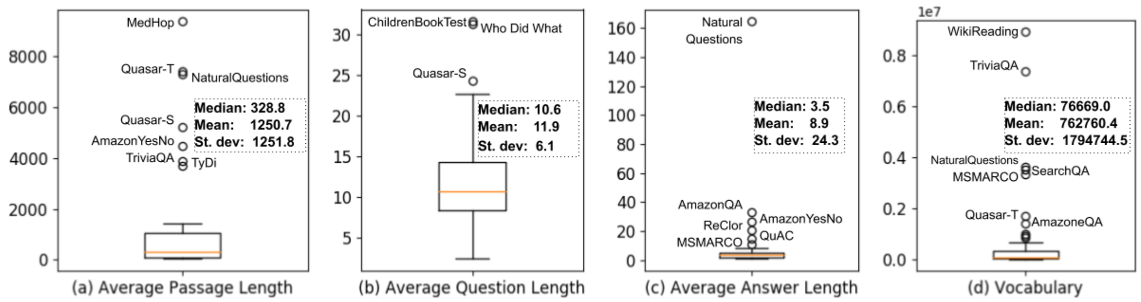

Figure 4: The average length in tokens of (a) passages, (b) questions, (c) answers, and (d) vocabulary size in unique lower-cased lemmas of datasets with the **median**, **mean** value, and standard deviation (**St. dev**). Outliers are highlighted.

20. Two datasets, ChildrenBookTest and WhoDidWhat, have on average more than 30 tokens per question while WikiReading, QAngaroo MedHop, and WikiHope have only 2 – 3.5 average tokens per question (Fig. 4 (b)). The majority of datasets contain fewer than 8 tokens per answer with the average being 3.5 tokens per answer. The NaturalQuestions is an outlier with average 164 tokens per answer[12] (Fig. 4 (c)).

Vocabulary Size To obtain a vocabulary size we calculate the number of unique lower cased lemmas of tokens. A vocabulary size distribution is presented in Fig. 4 (d). There is a moderate correlation[13] between the number of questions in a dataset and its vocabulary size (see Fig. 5). WikiReading has the largest number of questions as well as the richest vocabulary. bAbI is a synthetic dataset with 40k questions but only 152 lemmas in its vocabulary.

Language Detection We ran a language detector over all datasets using the `pyenchant` for American and British English, and `langid` libraries.[14] In 36 of the 54 datasets, more than 10% of the words are reported to be non-English.[15] We inspected 200 randomly chosen samples from a subset of these. For Wikipedia datasets (HotPotQA, QAngoroo WikiHop), around 70-75% of those words are named entities; 10-12% are specific terms borrowed from other languages such as names of plants, animals, etc.; another 8-10% are foreign words, e.g. the word *"dialetto"* from HotPotQA *"Bari dialect (dialetto barese) is a dialect of Neapolitan ..."*; about 1.5-3% are misspelled words and tokenization errors. In contrast, for the user-generated dataset, AmazonQA, 67% are tokenization and spelling errors. This aspect of a dataset's vocabulary is useful to bear in mind when, for example, fine-tuning a pre-trained language model which has been trained on less noisy text.

Question	All Questions		Unique Questions	
	Count	%	Count	%
what	1497009	22.39	1069275	24.23
when	137865	2.06	116158	2.63
where	154990	2.32	119250	2.70
which	275454	4.12	123731	2.80
why	95493	1.43	68217	1.55
how	456559	6.83	230948	5.23
who/whose	392166	5.87	293130	6.64
boolean	2236356	33.45	1259287	28.53
other	1439241	21.53	975681	22.11

Table 2: Frequency of the first token of the questions across datasets.

First Question Word A number of datasets come with a breakdown of question types based on the first token (Nguyen et al., 2016; Ostermann et al., 2018, 2019; Kočiský et al., 2018; Clark et al., 2019; Xiong et al., 2019). We inspect the most frequent first word in a dataset's questions excluding cloze-style questions. Table 1 shows the most frequent first word per dataset and Table 2 shows the same information over all datasets.[16] The most popular first word is *what* – 22% of all questions analysed and over half of questions in WikiQA, WikiMovies, MCTest, CosmosQA, and DREAM start with *what*. The majority of questions in ReClor (56.5%) start with the word *which*, and RACE has 23.1%. DROP mostly focused on *how much/many, how old* questions (60.4%). DuoRC has a significant proportion of *who/whose* ques-

[12] We focus on short answers, considering a long ones only if the short answer is not available.

[13] As data has non-normal distribution we calculated the Spearman correlation coefficient=0.58 and p-value is 1.3e-05.

[14] `pypi.org/project/pyenchant/` and `github.com/saffsd/langid.py` – all links l.v. 10/2020.

[15] See supplementary materials Section A.2 and Table 5.

[16] See the supplementary materials Section A.3.

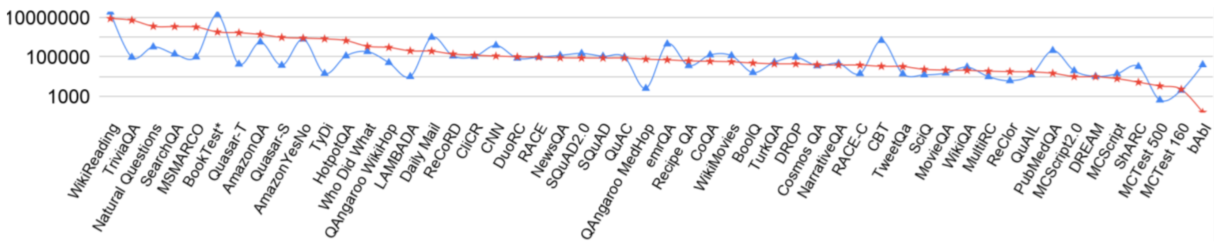

Figure 5: The number of questions and vocabulary size (unique lower-cased lemmas). The values for the BookTest (Bajgar et al., 2017) and WhoDidWhat (Onishi et al., 2016) are borrowed.

tions (39.5%). *Why*, *When*, and *Where* questions are under-represented – only 1.4%, 2%, and 2.3% of all questions respectively. Only CosmosQA has a significant proportion (34.2%) of *Why* questions, MCScript2 (27.9%) and TyDi (20.5%) of *When* questions, and bAbI (36.9%) of *Where* questions.

3.4 Human Performance

Human performance figures have been reported for some datasets – see Table 1. This is useful in two ways. Firstly, it gives some indication of the difficulty of the questions in the dataset. Contrast, for example, the low human performance score reported for the Quasar and CliCR datasets with the very high scores for DREAM, DROP, and MCScript. Secondly, it provides a comparison point for automatic systems, which may serve to direct researchers to under-studied datasets where the gap between state-of-the-art machine performance and human performance is large, e.g. CliCR (33.9 vs. 53.7), RecipeQA (29.07 vs 73.63), ShaRC (78.3 vs 93.9) and HotpotQA (82.20 vs 96.37).

Although useful, the notion of human performance is problematic and has to be interpreted with caution. It is usually an average over individual humans, whose reading comprehension abilities will vary depending on age, ability to concentrate, interest in, and knowledge of the subject area. Some datasets (CliCR, QUASAR) take the latter into account by distinguishing between expert and non-expert human performance, while RACE distinguishes between crowd-worker and author annotations. The authors of MedQA, which is based on medical examinations, use a passing mark (of 60%) as a proxy for human performance. It is important to know this when looking at its "solved" status since state-of-the-art accuracy on this dataset is only 75.3% (Zhang et al., 2018b).

Finally, Dunietz et al. (2020) call into question the importance of comparing human and machine performance on the MRC task and argue that the questions that MRC systems need to be able to answer are not necessarily the questions that people find difficult to answer.

4 Concluding Remarks and Recommendations

This paper represents an up-to-date, one-stop-shop picture of 54 English MRC datasets. We compare the datasets by question and answer type, size, data source, creation method, vocabulary, question type, "solvedness", and human performance level. Seeing the history of dataset creation we can observe the tendency of moving from smaller datasets towards large collections of questions, and from synthetically generated data through crowdsourcing towards spontaneously created. We also observe a scarcity of *why*, *when*, and *where* questions.

Gathering and processing the data for this survey was a painstaking task, from which we emerge with some very practical recommendations for future MRC dataset creators. In order to 1) compare to existing datasets, 2) highlight possible limitation for applicable methods, and 3) indicate the computational resources required to process the data, some basic statistics such as average passage/question/answer length, vocabulary size and frequency of question words should be reported; the data itself should be stored in consistent, easy-to-process fashion, ideally with an API provided; any data overlap with existing datasets should be reported; human performance on the dataset should be measured and what it means clearly explained; and finally, if the dataset is for the English language and its design does not differ radically from those surveyed here, e.g. the recent Template of Understanding approach (Dunietz et al., 2020), it is crucial to explain why this new dataset is needed.

For any future datasets, we suggest a move away from Wikipedia given the volume of existing datasets that are based on it and its use in pre-trained language models such as BERT (Devlin et al., 2019). As shown by Petroni et al. (2019), its use in MRC dataset and pre-training data brings with it the problem that we cannot always determine whether a system's ability to answer a question comes from its comprehension of the relevant passage or from the underlying language model.

The medical domain is well represented in the collection of English MRC datasets, indicating a demand for understanding of this type of text. Datasets may be required for other domains, such as retail, law and government.

Some datasets are designed to test the ability of systems to tell if a question cannot be answered, by including a "*no answer*" label. Building upon this, we suggest that datasets be created for the more complex task of answering questions differently based on different possible interpretations of the question, or determining whether contradictory information is given, i.e. similar to dialog datasets such as ShARC but in a non-dialog scenario.

Acknowledgements

This research is supported by Science Foundation Ireland in the ADAPT Centre for Digital Content Technology, funded under the SFI Research Centres Programme (Grant 13/RC/2106) and the European Regional Development Fund.

We also thank Andrew Dunne, Koel Dutta Chowdhury, Valeriia Filimonova, Victoria Serga, Marina Lisuk, Ke Hu, Joachim Wagner, and Alberto Poncelas.

References

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv:1809.03275*.

Ondrej Bajgar, Rudolf Kadlec, and Jan Kleindienst. 2017. Embracing data abundance. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Razieh Baradaran, Razieh Ghiasi, and Hossein Amirkhani. 2020. A survey on machine reading comprehension systems. *CoRR*, abs/2001.01582.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv:1506.02075*.

Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r Dataset. In *Third Annual Conference on Weblogs and Social Media (ICWSM 2009)*, San Jose, CA. AAAI.

Casimiro Pio Carrino, Marta R. Costa-jussà, and José A. R. Fonollosa. 2020. Automatic Spanish translation of SQuAD dataset for multi-lingual question answering. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5515–5523, Marseille, France. European Language Resources Association.

Atef Chaudhury, Makarand Tapaswi, Seung Wook Kim, and Sanja Fidler. 2019. The Shmoop Corpus: A Dataset of Stories with Loosely Aligned Summaries. *arXiv:1912.13082*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. CODAH: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

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

Eunsol Choi, Daniel Hewlett, Jakob Uszkoreit, Illia Polosukhin, Alexandre Lacoste, and Jonathan Berant. 2017. Coarse-to-fine question answering for long documents. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–220, Vancouver, Canada. Association for Computational Linguistics.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2018. Neural learning for question answering in italian. In *AI*IA 2018 – Advances in Artificial Intelligence*, pages 389–402, Cham. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv:1707.03904*.

Martin d'Hoffschmidt, Maxime Vidal, Wacim Belblidia, and Tom Brendlé. 2020. FQuAD: French Question Answering Dataset. *arXiv:2002.06071*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dunietz, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci. 2020. To test machine comprehension, start by defining comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7839–7859, Online. Association for Computational Linguistics.

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

Daria Dzendzik, Carl Vogel, and Jennifer Foster. 2019. Is it dish washer safe? Automatically answering "Yes/No" questions using customer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 1–6, Minneapolis, Minnesota. Association for Computational Linguistics.

Pavel Efimov, Leonid Boytsov, and Pavel Braslavski. 2019. SberQuAD–Russian reading comprehension dataset: Description and analysis. *arXiv:1912.09723*.

Ahmed Elgohary, Chen Zhao, and Jordan Boyd-Graber. 2018. A dataset and baselines for sequential open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1077–1083, Brussels, Belgium. Association for Computational Linguistics.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2019a. A deep neural network framework for english hindi question answering. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(2).

Mansi Gupta, Nitish Kulkarni, Raghuveer Chanda, Anirudha Rayasam, and Zachary C. Lipton. 2019b. AmazonQA: A review-based question answering task. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 4996–5002. International Joint Conferences on Artificial Intelligence Organization.

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

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 1693–1701, Cambridge, MA, USA. MIT Press.

Daniel Hewlett, Alexandre Lacoste, Llion Jones, Illia Polosukhin, Andrew Fandrianto, Jay Han, Matthew Kelcey, and David Berthelot. 2016. WikiReading: A novel large-scale language understanding task over Wikipedia. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1545, Berlin, Germany. Association for Computational Linguistics.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2016. The goldilocks principle: Reading children's books with explicit memory representations. In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, San Juan, Puerto Rico.

Lifu Huang, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Cosmos QA: Machine reading comprehension with contextual commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2391–2401, Hong Kong, China. Association for Computational Linguistics.

Vaishali Ingale, Pushpender Singh, and Aditi Bhardwaj. 2019. Datasets for machine reading comprehension: A literature review. *SSRN Electronic Journal*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

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

A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5376–5384.

Tom Kenter, Llion Jones, and Daniel Hewlett. 2018. Byte-level machine reading across morphologically varied languages. In *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5820–5827.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262, New Orleans, Louisiana. Association for Computational Linguistics.

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

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

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

K Lakshmi and K.M. Arivuchelvan. 2019. A survey on datasets for machine reading comprehension. In *Proceedings of International Conference on Recent Trends in Computing, Communication & Networking Technologies (ICRTCCNT) 2019*. SSRN Electronic Journal (2019).

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

Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Proceedings of The Eleventh Asian Conference on Machine Learning*, volume 101 of *Proceedings of Machine Learning Research*, pages 742–757, Nagoya, Japan. PMLR.

Seungyoung Lim, Myungji Kim, and Jooyoul Lee. 2019. KorQuAD1.0: Korean qa dataset for machine reading comprehension. *arXiv:1909.07005*.

Shanshan Liu, Xin Zhang, Sheng Zhang, Hui Wang, and Weiming Zhang. 2019. Neural machine reading comprehension: Methods and trends. *Applied Sciences*, 9(18):3698.

Kateřina Macková and Milan Straka. 2020. Reading comprehension in czech via machine translation and cross-lingual transfer. In *Text, Speech, and Dialogue*, pages 171–179, Cham. Springer International Publishing.

Christopher Malon and Bing Bai. 2013. Answer extraction by recursive parse tree descent. In *Proceedings of the Workshop on Continuous Vector Space Models and their Compositionality*, pages 110–118, Sofia, Bulgaria. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1400–1409, Austin, Texas. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. Lsdsem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics, LSDSem@EACL 2017, Valencia, Spain, April 3, 2017*, pages 46–51.

Kiet Van Nguyen, Khiem Vinh Tran, Son T. Luu, Anh Gia-Tuan Nguyen, and Ngan Luu-Thuy Nguyen. 2020. A pilot study on multiple choice machine reading comprehension for vietnamese texts. *arXiv:2001.05687*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MAchine Reading COmprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

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

Simon Ostermann, Ashutosh Modi, Michael Roth, Stefan Thater, and Manfred Pinkal. 2018. MCScript: A novel dataset for assessing machine comprehension using script knowledge. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Simon Ostermann, Michael Roth, and Manfred Pinkal. 2019. MCScript2.0: A machine comprehension corpus focused on script events and participants. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 103–117, Minneapolis, Minnesota. Association for Computational Linguistics.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

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

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Boyu Qiu, Xu Chen, Jungang Xu, and Yingfei Sun. 2019. A survey on neural machine reading comprehension. *arXiv:1906.03824*.

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

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

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

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

Anna Rogers, Olga Kovaleva, Matthew Downey, and Anna Rumshisky. 2020. Getting closer to AI complete question answering: A set of prerequisite real tasks. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8722–8731. AAAI Press.

Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of natural language rules in conversational machine reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2087–2097, Brussels, Belgium. Association for Computational Linguistics.

Amrita Saha, Rahul Aralikatte, Mitesh M. Khapra, and Karthik Sankaranarayanan. 2018. DuoRC: Towards complex language understanding with paraphrased reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1683–1693, Melbourne, Australia. Association for Computational Linguistics.

Guo Shangmin, Liu Kang, He Shizhu, Liu Cao, Zhao Jun, and Wei Zhuoyu. 2017. Ijcnlp-2017 task 5: Multi-choice question answering in examinations. In *Proceedings of the IJCNLP 2017, Shared Tasks*, page 34–40. Asian Federation of Natural Language Processing.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. DREAM: A challenge dataset and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7(0):217–231.

Simon Šuster and Walter Daelemans. 2018. CliCR: a dataset of clinical case reports for machine reading comprehension. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1551–1563, New Orleans, Louisiana. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4631–4640. IEEE Computer Society.

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

Bingning Wang, Ting Yao, Qi Zhang, Jingfang Xu, and Xiaochuan Wang. 2020a. Reco: A large scale chinese reading comprehension dataset on opinion. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9146–9153.

Chao Wang. 2020. A study of the tasks and models in machine reading comprehension. *arXiv:2001.08635*.

Ran Wang, Kun Tao, Dingjie Song, Zhilong Zhang, Xiao Ma, Xi'ao Su, and Xinyu Dai. 2020b. R3: A reading comprehension benchmark requiring reasoning processes. *arXiv:2004.01251*.

Yusuke Watanabe, Bhuwan Dhingra, and Ruslan Salakhutdinov. 2017. Question answering from unstructured text by retrieval and comprehension. *arXiv:1703.08885*.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. 2017. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, Copenhagen, Denmark. Association for Computational Linguistics.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Qizhe Xie, Guokun Lai, Zihang Dai, and Eduard Hovy. 2018. Large-scale cloze test dataset created by teachers. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2344–2356, Brussels, Belgium. Association for Computational Linguistics.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. 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. Association for Computational Linguistics.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1358–1368, Brussels, Belgium. Association for Computational Linguistics.

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

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

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. ReClor: A reading comprehension dataset requiring logical reasoning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*.

Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. 2018. SWAG: A large-scale adversarial dataset for grounded commonsense inference. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 93–104, Brussels, Belgium. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Chengchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension: Tasks, evaluation metrics, and benchmark datasets. *arXiv:2006.11880*.

Sheng Zhang, Xiaodong Liu, Jingjing Liu, Jianfeng Gao, Kevin Duh, and Benjamin Van Durme. 2018a. ReCoRD: Bridging the gap between human and machine commonsense reading comprehension. *arXiv:1810.12885*.

Xiao Zhang, Ji Wu, Zhiyang He, Xien Liu, and Ying Su. 2018b. Medical exam question answering with large-scale reading comprehension. In *Proceedings of the The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5706–5713.

Xin Zhang, An Yang, Sujian Li, and Yizhong Wang. 2019. Machine reading comprehension: a literature review. *CoRR*, abs/1907.01686.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, page 19–27, USA. IEEE Computer Society.

A Supplementary Materials

A.1 Extra Features and Statistics

For some datasets it is not straightforward to collect those numerical characteristics. Here we explain how we calculated those numbers and some features of counting we have done for certain datasets.

A.1.1 Instances

The concept of an instance might not be straightforward so we will explain it here. Some datasets based on books, movies or Wikipedia have divided the original source of data into different passages. For example, for SQuAD, SQuAD2.0, BoolQ, and NaturalQuestions the instance is a Wikipedia Article; for AmazonQA and AmazonYesNo the instance would be a product; in case of RecipeQA, instance is a type of recipe; etc. In other words *instance* is a level up item which might contain multiple passages. For some datasets the number of instances is equal to the number of passages. For example MovieQA, the instance is a movie and the passage is a movie plot. In case of bAbI dataset we consider the task as an instance. Not all datasets have the concept of instances, that is why for some of those there is "-" in the table.

A.1.2 Statistics Calculation

Some data we took from original and related papers. Those datasets are: BookTest (Bajgar et al., 2017), MedQA (Zhang et al., 2018b), and R^3 (Wang et al., 2020b).

We calculated all characteristics based on publicly available data. That means some datasets do not have the test set available so we based our calculations on training and development sets only. Those datasets are: AmazonQA, CoQA, DROP, MovieQA, QAngaroo (WikiHop and MedHop), QuAC, ShARC, SQuAD, SQuAD2, TyDi.

As mention in Section 3.3 we processed all datasets in the same way with *spiCy*. If the data is tokenized or split by sentences we simply join it back using Python `" ".join(tokens/sentences_list)`. Based on *spiCy* implementation[17] we would not expect significant differences between the originally provided tokenization and the results of *spiCy* the tokenization of the joined tokens. This ensures the consistency of the processing of all datasets.

There are a few particular features about certain datasets we would like to mention:

- **ShARC**: in this case instance is a *tree_id*. There are several scenarios for the same snippet. We consider concatenation of the snippet, scenario and follow up questions with answers as a passage;
- **HotpotQA**: we consider passage as a concatenation of all supporting facts, and instance is a title of supporting fact. In this case there a multiple instances for the same question.
- **WikiQA**: we consider passage as a concatenation of all sentences, we did calculations based on publicly available data and code from the github page;[18]
- **TriviaQA**: to obtain the data we modified the script provided by the authors on the github page;[19]
- **MSMARCO**: we consider every passage separately, so in this case, there are multiple passages for one question.
- **TyDi**: we calculated the statistic for joined data from English subset for both the Minimal answer span task and the Gold passage task for training and development set.
- **Who Did What** dataset we looked into relaxed setting. We do not have a licence to get gigaword data so we calculated only the average length of the questions and answers. The vocabulary size is provided by the original paper (Onishi et al., 2016).

A.2 Vocabulary

See Table 5 for detailed vocabulary analysis per dataset.

[17] `spacy.io` – l.v. 06/2020

[18] `github.com/RaRe-Technologies/gensim-data/issues/31` – l.v. 06/2020

[19] `github.com/mandarjoshi90/triviaqa` – l.v. 06/2020

Dataset	Dataset contains							Extra Data
	YesNo	Non-Factoid	Query	Multi Hop	Multi Doc	Dialogs	No Answer	
AmazoneQA	✓	✓	✗	✗	◆	✗	✗	✗
AmazonYesNo	✓	✗	✗	◆	✓	✗	◆	✗
bAbI	✓	✗	✗	✓	✗	✗	✗	✗
BookTest	✗	✗	✗	◆	✗	◆	✗	✗
BoolQ	✓	✗	✗	◆	✗	✗	✗	✗
CBT	✗	✗	✗	✗	✗	✗	✗	✗
CliCR	✗	✗	✗	✗	✗	✗	✗	✗
CNN/Daily Mail	✗	✗	✗	✗	✗	✗	✗	✗
Cosmos QA	✗	✓	✗	✓	✗	✗	✗	✗
CoQA	✓	✗	✗	◆	✗	✓	✓	✗
DREAM	◆	✓	✗	✓	✗	✓	✗	✗
DROP	◆	✓	✗	✓	✓	✗	✗	✗
DuoRC	◆	✓	✗	✓	✗	✗	✓	✗
emrQA	✓	✓	◆	✓	✗	✗	✓	✓
HotpotQA	✓	✓	✗	✓	✓	✗	✓	✗
LAMBADA	✗	✗	✗	✗	✗	✗	✗	✗
MCScript	✓	✓	✗	✗	✗	✗	✗	✗
MCScript2.0	✗	✓	✗	✗	✗	✗	◆	✗
MCTest	✓	◆	✗	◆	✗	✗	✗	✗
MedQA	✗	✓	✗	✓	✓	✗	✗	✗
MovieQA	◆	✓	✗	◆	✗	✗	✗	✓
MSMARCO	✓	✓	✗	✓	✓	✗	✓	✓
MultiRC	◆	✓	✗	✓	✓	✗	✗	✗
NarrativeQA	✗	✓	✗	✓	◆	✓	✗	✗
NaturalQuestions	✓	✓	✗	✗	✗	✗	✓	✗
NewsQA	◆	✓	✗	✓	✗	✗	✓	✗
PubMedQA	✓	✗	✗	✓	✗	✗	◆	✗
RACE	✗	✓	✗	✓	✗	✗	✗	✗
RACE-C	✗	✗	✗	✓	✗	✗	✗	✗
Recipe QA	✗	✗	✗	✗	✗	✗	✗	✓
ReClor	✗	✓	✗	✓	✗	✗	✗	✗
ReCoRD	✗	✗	✗	✓	✗	✗	✗	✗
SciQ	✗	✗	✗	✗	✗	✗	✗	✓
SQuAD	✗	✓	✗	✗	✗	✗	✗	✗
SQuAD2.0	✗	✓	✗	✗	✗	✗	✓	✗
SearchQA	✗	✓	✗	✓	✓	✗	✗	✗
ShARC	✓	✗	✗	◆	✗	✓	✗	✗
TriviaQA	✗	◆	✗	✓	✓	✗	✓	✗
TurkQA	✓	✓	✗	✗	✗	✗	✗	✗
TweetQA	✗	✓	✗	✗	✗	✗	✗	✗
TyDi	✓	✗	✗	✗	✗	✗	✓	✓
QAngarooWikiHop	✗	✗	✓	✓	✓	✗	✗	✗
QAngarooMedHop	✗	✗	✓	✓	✓	✗	✗	✗
QuAC	✓	✓	✗	✓	✗	✓	✓	✗
QuAIL	✗	✓	✗	✓	✗	✗	✓	✗
Quasar-S	◆	✗	✗	✗	✗	✗	✗	✗
Quasar-T	✗	✗	✗	✗	✗	✗	✗	✗
Who Did What	✗	✗	✗	✗	✗	✗	✗	✗
WikiMovies	✗	✓	✓	◆	✓	✗	✗	✓
WikiReading	✗	✗	✓	✓	✗	✗	✓	✗
WikiQA	✗	✗	✗	✗	✗	✗	✓	✗

Table 3: Datasets in alphabetical order and additional properties. Where extra data means the English RC task is only one part of bigger dataset with additional resources such as images or video, or there is an availability of resources in other languages. ✓ – presented; ◆ – presented in a limited form; ✗ – not presented.

A.3 Questions

Some questions could be formulated with a question word inside, for example: *"About how much does each box of folders weigh?"* or *"According to the narrator, what may be true about their employer?"*. We analyse 6.7M questions excluding all cloze datasets (ChildrenBookTest, CNN/DailyMail, WhoDidWhat, CliCR, LAMBADA, RecipeQA, Quasar-S, some cloze style questions from MS MARCO, DREAM, Quasar-T, RACE, RACE-C, SearchQA, TriviaQA, emrQA) (there are all together approximately 2.5M cloze questions) and WikiReading, WikiHop, and MedHop (almost 19 million questions-queries), as the queries are not formulated in question form. As mentioned in section 3.1 some datasets shared the questions and some datasets have the same questions asked more than once within a different context (for example, question *"Where is Daniel?"* asked 2007 times in bAbI), or same questions asked with different answer options (for example in CosmosQA dataset). We calculated the frequency of question words for both scenarios: *all questions* and *unique questions* (see Table 6).

To separate boolean questions we used the same list of words as Clark et al. (2019): *"did", "do", "does", "is", "are", "was", "were", "have", "has", "can", "could", "will", "would"*. Apart from datasets which contain only yes/no/maybe questions a significant portion of boolean questions are in ShaRC (85.4%), emrQA (74.0 %) AmazonQA (55.3%), QuAC (36.6%), MCSCript (28.6%), TurcQA (25.7%), bAbI (25.0 %) and CoQA (20.7%).

Almost a third of all questions and more than a quarter of unique questions are boolean. Another quarter of unique questions (26.57%) contain the word *"What"*, 6.64% of questions asks *"Who"* and *"Whose"*, and 4.49% *"Which"*, about 3% of questions are *"When"* and *"Where"*. Only 5.95% ask the question *"how"* excluding (*"how many/much"* and *"how old"*). Other questions which do not contain any of these question words constitute 16,73% of unique questions. There are datasets where more than 20 % of questions are formulated in such a way that the first token is not one of the considered words: Quasar-S (98.8 %), SearchQA (98.3 %), RACE-C (64.1 %), TriviaQA (49.6 %), HotPotQA (42.0 %), Quasar-T (40.7 %), MSMARCO (26.6 %), NaturalQuestions(23.4 %), AmazonQA (22.8 %), and SQuAD (21.1 %).

See Table 7 for more detailed information.

B Other Datasets

There are a number of datasets we did not include into our analysis as we would like to stay focused on the Question Answering Machine Reading Comprehension task. In this section we mention those works and explain why they are excluded.

B.1 Question Answering Datasets

CLOTH ((Xie et al., 2018)) and **Story Cloze Test**, ((Mostafazadeh et al., 2016, 2017)) are cloze-style datasets and it is a missing word from the context task without a specific query. As well as this, we did not include a number of RC datasets where the story should be completed such as **ROCStories** ((Mostafazadeh et al., 2016)), **CODAH** ((Chen et al., 2019)), **SWAG** ((Zellers et al., 2018)), and **HellaSWAG** ((Zellers et al., 2019)) because there are no question so they are not question answering datasets. In contrast, cloze question answering datasets considered in this work have separate coherent text (passage) and separate sentence which can be treated as *question* with a missed word.

QBLink ((Elgohary et al., 2018)) is technically a RC QA dataset but for every question there is only the name of a wiki page available. The "lead in" information is not enough to answer the question without additional resources. In the other words QBLink is a more general QA dataset, like **CommonSenceQA** ((Talmor et al., 2019)), rather than RC.

Textbook Question Answering (TQA) ((Kembhavi et al., 2017)) is a multi-modal dataset requiring not only text understanding but also picture processing.

MCQA is a Multiple Choice Question Answering dataset in English and Chinese based on examination questions introduced as Shared Task on IJCNLP 2017 by (Shangmin et al., 2017). The authors do not provide any supportive documents which can be considered as a passage so it is not a reading comprehension task.

A number of datasets, such as **SimpleQuestions** ((Bordes et al., 2015)) and **WebQuestion** ((Berant et al., 2013)), were created with the idea of extracting answers from a knowledge graph. Even though the additional resources are involved it is presented in a structured form rather than a free natural text so we do not consider those dataset in the current chapter.

B.2 Non-English Datasets

In the paper we focus on English dataset but there are a number of RC datasets in other languages, and in this section we will briefly mention some of them.

B.2.1 Chinese datasets

DuReader ((He et al., 2018)) is a Chinese RC dataset. It contains mixed types of questions based on Baidu Search and Baidu Zhidao.[20] **ReCO** ((Wang et al., 2020a)) **Re**ading **C**omprehension dataset on **O**pinion) is the largest human-curated Chinese reading comprehension dataset containing 300k questions with *"Yes/No/Unclear"* answers.

B.2.2 Other Languages

The extended version of **WikiReading** ((Kenter et al., 2018)) apart of 18M English questions also contains 5M Russian and about 600K Turkish examples.

TyDi ((Clark et al., 2020)) is a question answering corpus of 11 Typologically Diverse languages (Arabic, Bengali, Korean, Russian, Telugu, Thai, Finnish, Indonesian, Kiswahili, Japanese, and English). It contains 200k+ question answers pairs based on the Wikipedia articles in those languages.

ViMMRC ((Nguyen et al., 2020)) is a multiple-choice questions RC dataset in Vietnamese language. It contains 2,783 questions based on a set of 417 texts.

Following the approach of SQuAD datased constraction there were a few more datasets created: **FQuAD** ((d'Hoffschmidt et al., 2020)) is a 25,000+ question French Native Reading Comprehension dataset; **KorQuAD** ((Lim et al., 2019)) has 70,000 original questions in Korean. Both datasets are based on Wikipedia.

B.2.3 Datasets Translation

SQuAD has been semi-automatically translated into several other languages as: Korean **K-QuAD** ((Lee et al., 2018)); Italian **SQuAD-it** ((Croce et al., 2018)); Japanese and French ((Asai et al., 2018)); Spanish **SQuAD-es** ((Carrino et al., 2020)); Hindi ((Gupta et al., 2019a)); Russian **SberQuAD** ((Efimov et al., 2019)); and Czech ((Macková and Straka, 2020)).

[20] `zhidao.baidu.com` – last verified February2020

Dataset	# of instances	# of passages	# A/Q	AVG Q len	AVG P len	AVG A len	Vocabulary Size
AmazonQA	139,905	830,959	-	16.6	558.2	32.8	1,395,460
AmazonYesNo	40,806	40,806	-	13.2	4398.2	-	864,929
bAbI	20	1,2534	-	6.3	67.2	1.1	152
BookTest	14,062	14,140,825	10		522	1	1,860,394
BoolQ	8208	12,697	2	8.8	109.4	-	49,117
CBT	108	687,343	10	30	440	1	53,628
CliCR	11,846	11,846	-	22.6	1411.7	3.4	122,568
CosmosQA	35,210	35,210	4	10.6	70.4	8.1	40,067
CoQA	-	7,699	-	6.5	328.0	2.9	59,840
CNN	-	107,122	-	12.8	708.4	1.4	111,198
DailyMail	-	218,017	-	14.8	854.4	1.5	197,388
DREAM	6,138	6,444	3	8.8	86.4	5.3	9,850
DROP	-	6147	-	12.2	246.2	4	44,430
DuoRC	7,477	7,477	-	8.6	1,260.9	3.1	119,547
emrQA	2427	2,427	-	7.9	1328.4	2.0	70,837
HotpotQA	534,433	105,257	-	20.0	1100.7	2.4	741,974
LAMBADA	5,325	10,022	-	15.4	58.5	1	203,918
MedQA	5	243,712	5	27.4	4.2	43.2	-
MCScript	110	2,119	2	6.7	196.0	3.6	7,867
MCScript2.0	200	3,487	2	8.2	164.4	3.4	11.890
MCTest 160	160	160	4	9.2	241.8	3.7	2,246
MCTest 500	500	500	4	8.9	251.6	3.8	3,334
MovieQA	408	408	3-5	9.34	727.91	5.6	21,322
MSMARCO	-	10,087,677	-	6.5	65.9	11.1	3,324,030
MultiRC	871	871	5.4	4.8	92.4	5.5	23,331
NarrativeQA	1,572	1,572	2	9.9	673.9	4.8	38,870
NaturalQuestions	109,715	315,203	-	9.36	7312.13	164.56	3,635,821
NewsQA	12,744	12,744	-	7.8	749.2	5.0	90,854
PubMedQA	-	3,358	3	15.1	73.8	-	14,751
QAngarooWikiHop	-	48,867	-	3.5	1381	1.8	304,322
QAngarooMedHop	-	1962	-	3	9366.7	1	76,954
QuAC	8853	13,594	-	5.6	401	14.1	99,912
QuAIL	680	680	4	9.70	388.29	4.36	17271
Quasar-S	-	37,362	-	24.3	(S)1995.9 (L)5210.1	1.5	(S)660,425 (L)987,380
Quasar-T	-	43,012	-	11.1	(S)2256.2 (L)7372.6	1.9	(S)1,021,823 (L)2,019,336
RACE	-	27,933	4	12.0	329.5	6.3	98,482
RACE-C	-	2,708	4	13.8	423.8	7.4	38,399
Recipe QA	-	9,761	4	10.8	580.0	3.3	62,938
ReClor	-	6,138	4	17.0	73.6	20.6	17,865
ReCoRD	-	73190	-	24.72	193.64	1.5	139724
SciQ	-	12,252	-	14.6	87.1	1.5	23,320
SearchQA	27,995	13,796,295	-	16.7	58.7	2.1	3,506,501
ShARC	697	24,160	-	8.6	87.2	4.0	5,231
SQuAD	490	20,963	-	11.4	137.1	3.5	87,765
SQuAD2.0	477	20,239	-	11.2	137.0	3.5	88,081
TyDi	-	14,378	-	8.3	3,694.2	4.6	848,524
TriviaQA	-	801,194	-	16.4	3867.6	2.3	7,366,586
TurkQA	-	13,425	-	10.3	41.6	2.9	44,677
TweetQA	-	13757	-	8.02	31.93	2.70	32542
WhoDidWhat	-	205,978	3.5	31.2	N/A	2.1	347,406
WikiMovies	-	186,444	-	8.7	77.9	6.8	56,893
WikiQA	-	1,242	-	6.5	252.6	-	20,686
WikiReading	4,313,786	18,807,888	-	2.35	569.0	2.2	8,928,645

Table 4: Basic statistics of RC datasets. **Q** - question; **P** - passage; **A** - answer; **S** - short passages; **L** - long passages; Average length of Q/P/A are measured in tokens. Vocabulary size is measured in lower-cased unique lemmas. Note: some numbers are aggregated and might be slightly different from other sources. The aim of this paper is to show an estimation rather than an exact value. For PubMedQuestions we count statistic for labeled data (1000 questions).

Dataset	English Words	Numbers	Not English Words	Not Ascii	Web Links
AmazonQA	1065795 (76.4%)	38323 (2.7%)	235019 (16.8%)	6240 (0.4%)	49765 (3.6%)
AmazonYesNo	736037 (81.3%)	17931 (2.0%)	144761 (16.0%)	45 (0.0%)	6345 (0.7%)
bAbI	145 (95.4%)	0(0%)	7 (4.6%)	0(0%)	0(0%)
BoolQ	36940 (75.2%)	3050 (6.2%)	7081 (14.4%)	2007 (4.1%)	41 (0.1%)
CBTest	29630 (88.4%)	167 (0.5%)	3651 (10.9%)	58 (0.2%)	0(0%)
CNN	75523 (67.9%)	6290 (5.7%)	27250 (24.5%)	726 (0.7%)	1408 (1.3%)
CliCR	82981 (67.7%)	7798 (6.4%)	30809 (25.1%)	890 (0.7%)	85 (0.1%)
CoQA	45112 (75.4%)	2605 (4.4%)	10270 (17.2%)	1748 (2.9%)	93 (0.2%)
CosmosQA	34466 (86.0%)	934 (2.3%)	4617 (11.5%)	6 (0.0%)	42 (0.1%)
DREAM	8653 (87.8%)	711 (7.2%)	469 (4.8%)	11 (0.1%)	2 (0.0%)
DROP	27458 (61.8%)	7564 (17.0%)	7545 (17.0%)	1840 (4.1%)	13 (0.0%)
DailyMail	130062 (65.9%)	13919 (7.1%)	49752 (25.2%)	1457 (0.7%)	2197 (1.1%)
DuoRC	73800 (72.5%)	1235 (1.2%)	22937 (22.5%)	3715 (3.6%)	33 (0.0%)
emrQA	48174 (68.0%)	12287 (17.3%)	10060 (14.2%)	2 (0.0%)	0(0%)
HotPotQA	341142 (50.2%)	29140 (4.3%)	199911 (29.4%)	107605(15.8%)	1901 (0.3%)
LAMBADA	144310 (70.8%)	4828 (2.4%)	49745 (24.4%)	2846 (1.4%)	2186 (1.1%)
MCScript	7544 (95.9%)	101 (1.3%)	198 (2.5%)	15 (0.2%)	6 (0.1%)
MCScript2	9467 (94.4%)	138 (1.4%)	395 (3.9%)	17 (0.2%)	12 (0.1%)
MCTest 160	2135 (95.1%)	31 (1.4%)	74 (3.3%)	1 (0.0%)	0(0%)
MCTest 500	3145 (94.3%)	35 (1.0%)	147 (4.4%)	1 (0.0%)	0(0%)
MSMARCO	2046615 (61.6%)	261290 (7.9%)	703298 (21.2%)	246936 (7.4%)	65825 (2.0%)
MovieQA	18166 (85.2%)	385 (1.8%)	2768 (13.0%)	1 (0.0%)	0(0%)
MultiRC	16034 (84.9%)	896 (4.7%)	1821 (9.6%)	106 (0.6%)	14 (0.1%)
NarrativeQA	31058 (79.9%)	631 (1.6%)	6213 (16.0%)	927 (2.4%)	1 (0.0%)
NaturalQuestions	1177894 (32.4%)	891487 (24.5%)	757428 (20.8%)	364341(10.0%)	444670(12.2%)
NewsQA	65487 (72.1%)	4316 (4.7%)	19370 (21.3%)	716 (0.8%)	950 (1.0%)
PubMedQA	11139 (75.4%)	2531 (17.1%)	941 (6.4%)	148 (1.0%)	1 (0.0%)
QAngoroo MedHop	59186 (77.2%)	4858 (6.3%)	10877 (14.2%)	1722 (2.2%)	26 (0.0%)
QAngoroo WikiHop	173858 (57.1%)	22415 (7.4%)	93948 (30.9%)	13860 (4.6%)	345 (0.1%)
QuAC	63683 (72.6%)	3499 (4.0%)	20315 (23.2%)	101 (0.1%)	107 (0.1%)
Quasar-S	622534 (63.0%)	109403 (11.1%)	210401 (21.3%)	1 (0.0%)	45475 (4.6%)
Quasar-T	941480 (55.5%)	167738(9.9%)	479864 (28.3%)	1 (0.0%)	107374(6.3%)
RACE-C	30697 (79.9%)	1248 (3.3%)	3988 (10.4%)	2420 (6.3%)	30 (0.1%)
RACE	75342 (76.5%)	6277 (6.4%)	15863 (16.1%)	1 (0.0%)	889 (0.9%)
ReClor	16364 (91.6%)	326 (1.8%)	1174 (6.6%)	1 (0.0%)	0(0%)
RecipeQA	48929 (77.0%)	1031 (1.6%)	10560 (16.6%)	1181 (1.9%)	835 (1.3%)
SQuAD	58444 (66.6%)	5708 (6.5%)	16827 (19.2%)	6706 (7.6%)	55 (0.1%)
SQuAD2	58793 (66.8%)	5724 (6.5%)	16935 (19.2%)	6548 (7.4%)	54 (0.1%)
SearchQA	2129356 (60.7%)	313517 (8.9%)	957977 (27.3%)	392 (0.0%)	105207 (3.0%)
ShaRC	4703 (90.6%)	303 (5.8%)	161 (3.1%)	15 (0.3%)	1 (0.0%)
TriviaQA	3269469 (44.3%)	421543 (5.7%)	1566735 (21.2%)	1806003(24.5%)	293086 (4.0%)
TurkQA	32225 (72.1%)	1660 (3.7%)	10778 (24.1%)	1 (0.0%)	25 (0.1%)
TyDi	532336 (61.8%)	31785 (3.7%)	184915 (21.5%)	83113(9.6%)	828 (0.1%)
WhoDidWhat	79056 (63.5%)	2658 (2.1%)	42670 (34.3%)	40 (0.0%)	52 (0.0%)
WikiMovies	39249 (69.0%)	447 (0.8%)	15310 (26.9%)	1880 (3.3%)	3 (0.0%)
WikiQA	17074 (82.5%)	1081 (5.2%)	2041 (9.9%)	477 (2.3%)	12 (0.1%)
WikiReading	3431134(38.4%)	823603(9.2%)	2777734(31.1%)	1801580(20.2%)	94594(1.1%)

Table 5: Types of lemmas in datasets vocabulary in percentage listed in decreasing order according to the vocabulary size.

Question	All Questions		Unique Questions	
	First token	**Contains**	**First token**	**Contains**
# of questions	6,685,252		4,413,218	
what	1497009 (22.39%)	1687898 (25.25%)	1069275 (24.23%)	1172454 (26.57%)
when	137865 (2.06%)	156628 (2.34%)	116158 (2.63%)	131509 (2.98%)
where	154990 (2.32%)	166866 (2.50%)	119250 (2.70%)	128067 (2.90%)
which	275454 (4.12%)	541835 (8.10%)	123731 (2.80%)	198335 (4.49%)
why	95493 (1.43%)	98649 (1.48%)	68217 (1.55%)	71185 (1.61%)
how	258961 (3.87%)	298526 (4.47%)	230948 (5.23%)	262646 (5.95%)
who/whose	392166 (5.87%)	392166 (5.87%)	293130 (6.64%)	293130 (6.64%)
how much/many/old	197598 (2.96%)	197598 (2.96%)	157427 (3.57%)	157427 (3.57%)
boolean	2236356 (33.45%)	–	1259287 (28.53%)	–
other	1439241 (21.53%)	907748 (13.58%)	975681 (22.11%)	738245 (16.73%)

Table 6: Frequency of first token of the questions and question words inside the question across datasets.

Dataset	**# of Q**	**what**	**when**	**where**	**which**	**why**	**how**	**who/ whose**	**how old/ much/ many**	**boolean**	**other**
AmazonQA	830954	10.2%	0.4%	1.2%	0.4%	0.6%	6.8%	0.1%	2.3%	55.3%	22.8%
AmazonYesNo	80391	0.2%	0.2%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	88.3%	11.1%
bAbI	40000	21.7%	0	36.9%	0	3.1%	5.0%	3.3%	5.0%	25.0%	0
BoolQ	15942	0.1%	0.1%	0.0%	0	0	0.0%	0.0%	0	97.5%	2.3%
CoQA	116630	29.7%	4.1%	6.6%	1.6%	2.7%	4.6%	14.7%	5.4%	20.7%	9.9%
CosmosQA	35210	54.6%	0.2%	1.6%	0.7%	34.2%	5.2%	1.1%	0.3%	1.2%	0.9%
DREAM	9934	56.3%	4.7%	10.0%	2.9%	8.5%	6.5%	3.6%	4.1%	1.1%	2.2%
DROP	86945	6.5%	0.6%	0.5%	18.2%	0.1%	0.7%	8.1%	60.4%	1.7%	3.2%
DuoRC	100966	33.1%	1.0%	8.6%	1.2%	2.5%	3.4%	39.5%	2.6%	1.8%	6.3%
emrQA	1980621	16.1%	0.5%	0.0%	0	1.4%	1.0%	0.0%	0.2%	74.0%	6.8%
HotPotQA	105253	22.6%	2.6%	1.9%	13.5%	0.0%	0.6%	8.6%	1.1%	6.9%	42.0%
MCScript	13939	13.9%	5.8%	9.4%	0.5%	11.6%	13.4%	12.2%	3.8%	28.6%	0.7%
MCScript2	19821	42.0%	27.9%	11.0%	0.2%	0.7%	3.8%	8.4%	0.7%	0.0%	5.2%
MCTest 160	639	51.3%	1.4%	6.9%	2.2%	12.1%	3.9%	13.3%	3.9%	1.6%	3.4%
MCTest 500	2000	52.1%	1.7%	8.5%	3.2%	12.0%	3.5%	12.6%	3.9%	0.8%	1.7%
MSMARCO	1009035	35.6%	2.7%	3.5%	1.8%	1.7%	11.1%	3.4%	5.8%	7.9%	26.6%
MovieQA	29888	46.3%	1.2%	6.8%	0.9%	11.0%	9.4%	19.3%	1.4%	1.9%	1.8%
MultiRC	7903	36.6%	2.3%	3.9%	4.0%	7.0%	6.7%	14.3%	4.6%	7.2%	13.4%
NarrativeQA	46764	38.3%	1.6%	7.5%	2.2%	9.8%	8.3%	24.4%	2.2%	0.1%	5.6%
NaturalQuestions	315104	15.5%	13.1%	10.1%	2.9%	1.2%	2.3%	25.3%	3.8%	2.6%	23.4%
NewsQA	119632	44.3%	4.1%	7.1%	2.2%	0.1%	0.9%	19.8%	5.9%	3.9%	11.7%
PubMedQA	1000	0	0	0	0	0	0	0	0	64.1%	35.9%
QuAC	90922	35.0%	5.2%	3.5%	0.7%	2.8%	6.6%	5.3%	1.4%	36.6%	2.9%
Quasar-S	37362	0.0%	0.0%	0.0%	0	0	0.0%	0.0%	0	1.1%	98.8%
Quasar-T	41102	32.0%	0.5%	2.1%	10.6%	0.2%	0.6%	11.3%	1.4%	0.5%	40.7%
RACE-C	11909	17.7%	1.1%	0.3%	10.0%	4.4%	1.4%	0.6%	0.3%	0.1%	64.1%
RACE	51526	35.6%	1.8%	2.3%	23.1%	8.5%	4.3%	2.6%	3.0%	0.4%	18.4%
ReClor	6138	0.2%	0	0	56.5%	0.0%	0.0%	0	0	0.0%	43.2%
SQuAD	98160	43.4%	6.3%	3.8%	4.7%	1.4%	3.3%	9.7%	6.1%	1.2%	20.1%
SQuAD2	142183	46.0%	6.1%	3.6%	4.3%	1.4%	3.2%	9.9%	5.8%	1.0%	18.7%
SearchQA	163981	0.1%	0.7%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.7%	98.3%
ShaRC	24160	0	0	0	0	0	0	0	0	85.4%	14.6%
TriviaQA	800827	18.6%	0.3%	0.8%	19.0%	0.0%	0.3%	9.9%	1.2%	0.3%	49.6%
TurkQA	53700	34.8%	5.7%	6.9%	1.2%%	0.2%	0.9%	6.9%	1.7%	25.7%	16.0%
TyDi	14378	29.0%	20.5%	4.8%	1.4%	0.8%	9.0%	11.8%	13.5%	8.6%	0.6%
WikiMovies	216453	50.6%	1.7%	0.0%	10.4%	0.0%	7.7%	17.3%	0	2.2%	10.1%
WikiQA	1242	54.5%	9.1%	8.9%	0	0	6.5%	13.5%	7.5%	0	0

Table 7: The percentage of question words per dataset.