# A Survey on Machine Reading Comprehension—Tasks, Evaluation Metrics and Benchmark Datasets

**Changchang Zeng [1,2,3]**, **Shaobo Li [1,4,\***], **Qin Li [5]**, **Jie Hu [5]** and **Jianjun Hu [6,\***]

[1]   Chengdu Institute of Computer Application, Chinese Academy of Sciences, Chengdu 610041, China; zengchangchang16@mails.ucas.ac.cn
[2]   University of Chinese Academy of Sciences, Beijing 100049, China
[3]   Department of Computer Science and Engineering, Chengdu Neusoft University, Chengdu 611844, China
[4]   School of Mechanical Engineering, Guizhou University, Guiyang 550025, China
[5]   College of Big Data Statistics, GuiZhou University of Finance and Economics, Guiyang 550025, China; qinlee85@126.com (Q.L.); jiehu@mail.gufe.edu.cn (J.H.)
[6]   Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208, USA
\*   Correspondence: lishaobo@gzu.edu.cn (S.L.); jianjunh@cse.sc.edu (J.H.); Tel.: +1-803-777-7304 (J.H.)

check for updates

**Abstract:** Machine Reading Comprehension (MRC) is a challenging Natural Language Processing (NLP) research field with wide real-world applications. The great progress of this field in recent years is mainly due to the emergence of large-scale datasets and deep learning. At present, a lot of MRC models have already surpassed human performance on various benchmark datasets despite the obvious giant gap between existing MRC models and genuine human-level reading comprehension. This shows the need for improving existing datasets, evaluation metrics, and models to move current MRC models toward "real" understanding. To address the current lack of comprehensive survey of existing MRC tasks, evaluation metrics, and datasets, herein, (1) we analyze 57 MRC tasks and datasets and propose a more precise classification method of MRC tasks with 4 different attributes; (2) we summarized 9 evaluation metrics of MRC tasks, 7 attributes and 10 characteristics of MRC datasets; (3) We also discuss key open issues in MRC research and highlighted future research directions. In addition, we have collected, organized, and published our data on the companion website where MRC researchers could directly access each MRC dataset, papers, baseline projects, and the leaderboard.

**Keywords:** machine reading comprehension; survey; dataset

## 1. Introduction

### 1.1. Overview

In the long history of Natural Language Processing (NLP), teaching computers to read the text and understand the meaning of the text is a major research goal that has not been fully realized. In order to accomplish this task, researchers have conducted machine reading comprehension (MRC) research in many aspects recently with the emergence of the large-scale datasets, higher computing power, and the deep learning techniques, which have boosted the whole NLP research [1–3]. The concept of MRC comes from the human understanding of text. The most common way to test whether a person can fully understand a piece of text is to require she/he answer questions about the text. Just like the human language test, reading comprehension is a natural way to evaluate a computer's language understanding ability.

In the NLP community, machine reading comprehension has received extensive attention in recent years [4–8]. The goal of a typical MRC task is to require a machine to read a (set of) text passage(s) and then answers questions about the passage(s), which is very challenging [9].

Machine reading comprehension could be widely applied in many NLP systems such as search engines and dialogue systems. For example, as shown in Figure 1, nowadays, when we enter a question into the search engine Bing, sometimes the Bing can directly return the correct answer by highlight it in the context (if the question is simple enough). Moreover, if we open the "Chat with Bing" in the website of Bing, as shown in the right part of the browser in Figure 1, we can also ask it questions such as "How large is the pacific?", the Bing chatbot will directly give the answer "63.78 million square miles". And on Bing's App, we can also open this "Chat with Bing", as shown in the right part of Figure 1. It is clear that MRC can help improve the performances of search engines and dialogue systems, which can allow users to quickly get the right answer to their questions, or to reduce the workload of customer service staff.
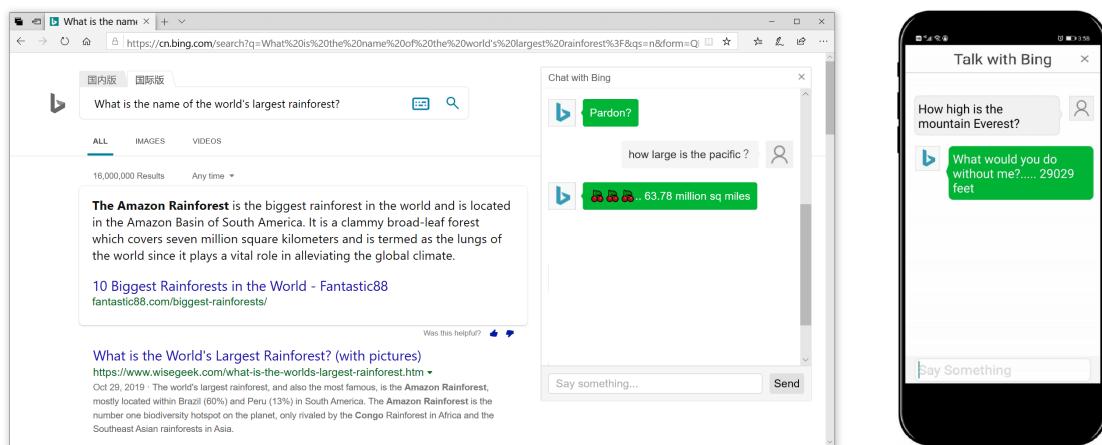


**Figure 1.** Examples of machine reading comprehension applied to search engine and dialogue system.

## 1.2. History

Machine reading comprehension is not newly proposed. As early as 1977, Lehnert et al. [10] had already built a question answering program called the QUALM which was used by two story understanding systems. In 1999, Hirschman et al. [11] constructed a reading comprehension system with a corpus of 60 development and 60 test stories of 3rd to 6th-grade material. The accuracy of the baseline system is between 30% and 40% on 11 sub-tasks. Most of MRC systems in the same period were rule-based or statistical models [12,13]. However, due to the lack of high quality MRC datasets, this research field has been neglected for a long time [14]. In 2013, Richardson et al. [15] created the MCTest [15] dataset which contained 500 stories and 2000 questions. Later, many researchers began to apply machine learning models on MCTest [15–18] despite that the original baseline of MCTest [15] is a rule-based model and the number of training samples in the MCTest [15] dataset is not large. A turning point for this field came in 2015 [14]. In order to resolve these bottlenecks, Hermann et al. [19] defined a new dataset generation method that provides large-scale supervised reading comprehension datasets in 2015. They also developed a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure. Since 2015, with the emergence of various large-scale supervised datasets and neural network models, the field of machine reading comprehension has entered a period of rapid development. Figure 2 shows the numbers of research papers on MRC since 2013. As is seen, the number of papers on MRC has been growing at an impressive rate.
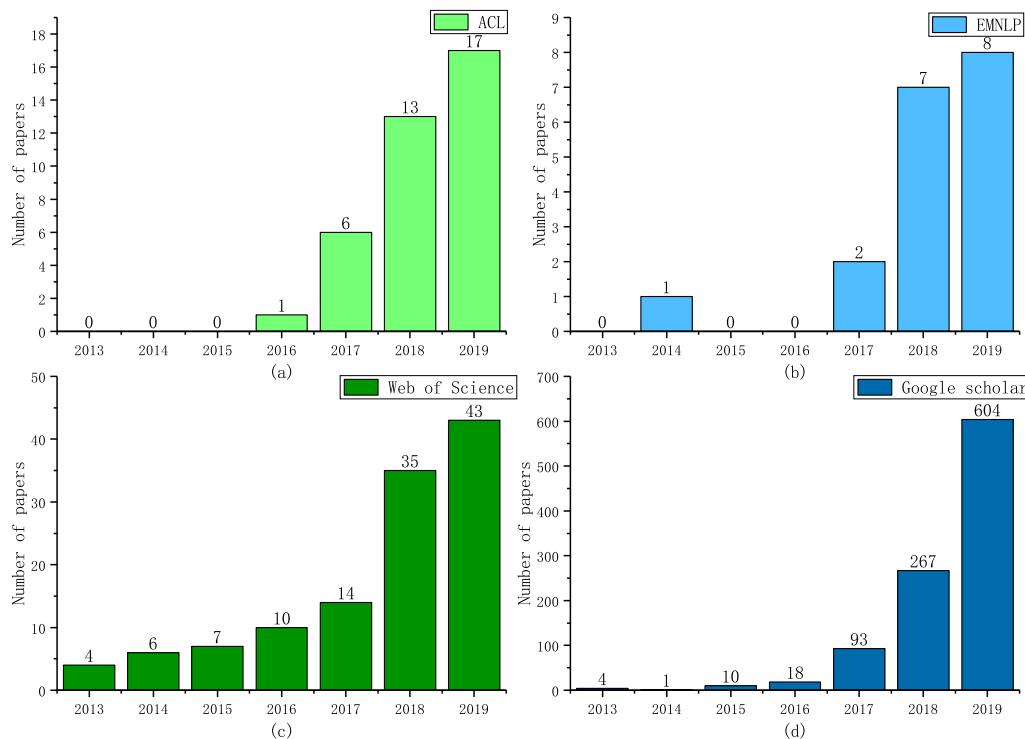
**Figure 2.** The number of research papers for machine reading comprehension each year: (**a**) The number of research papers on machine reading comprehension (MRC) in ACL from 2013 to 2019. (**b**) The number of research papers on MRC in ENMLP from 2013 to 2019. (**c**) The number of research papers on MRC in Web of Science from 2013 to 2019. (**d**) The number of research papers on MRC in Google scholar from 2013 to 2019.

*1.3. Motivation*

The benchmark datasets play a crucial role in speeding up the development of better neural models. In the past few years, we have witnessed an explosion of work that brings various MRC benchmark datasets [4–8]. Figure 3a shows the cumulative number of MRC datasets from the beginning of 2014 to the beginning of 2020. It shows that the number of MRC datasets has increased exponentially in recent years. And these novel datasets inspired a large number of new neural MRC models, such as those shown in Figure 3b, just take SQuAD 1.1 [19] for example, we can see that many neural network models were created in recent years, such as BiDAF [20], ELMo [21], BERT [22], RoBERTa [23] and XLNet [24]. The performance of the state-of-the-art neural network models have already exceeded human performance over the related MRC benchmark datasets.

Despite the critical importance of MRC datasets, most of the existing MRC reviews have focused on MRC algorithms for improving system performance [25,26], performance comparisons [7], or general review that has limited coverage of datasets [6]. In addition, there is also a need for systematic categorization/classification of task types. For example, MRC tasks are usually divided into four categories—cloze style, multiple-choice, span prediction and free form [14,26,27]. But this classification method is not precise because the same MRC task could belong to both cloze style and multiple-choice style at the same time, such as the CBT [28] task in the Facebook bAbi project [29]. Moreover, most researchers focus on few popular MRC datasets while most other MRC datasets are not widely known and studied by the community. To address these gaps, a comprehensive survey of existing MRC benchmark datasets, evaluation metrics and tasks is strongly needed.
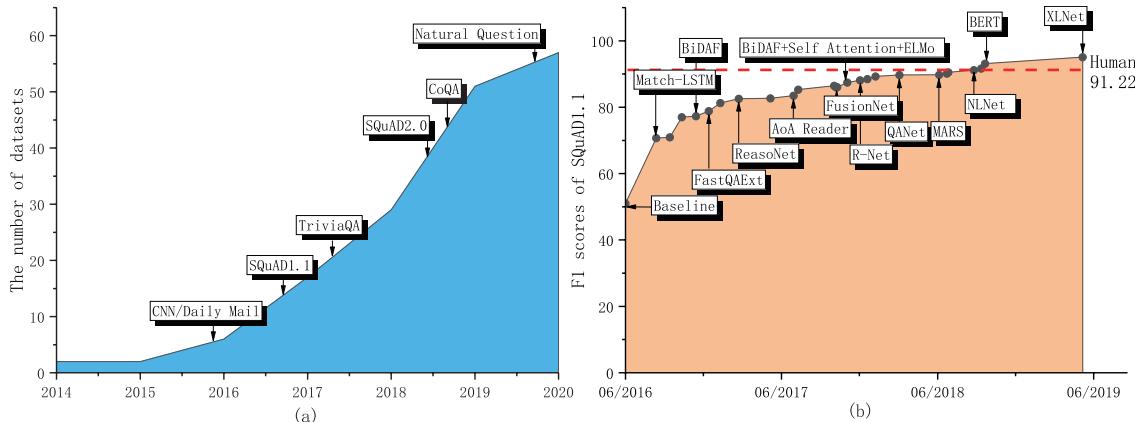
**Figure 3.** The number of MRC datasets created in recent years and the F1 scores of state-of-the-art models on SQuAD 1.1 [19]: (**a**) The cumulative number of MRC datasets from the beginning of 2014 to the end of 2019. (**b**) The progress of state-of-the-art models on SQuAD 1.1 since this dataset was released. The data points are taken from the leaderboard at https://rajpurkar.github.io/SQuAD-explorer/.

At present, a lot of neural MRC models have already surpassed human performance on many MRC datasets, but there is still a giant gap between existing MRC and real human comprehension [30]. This shows the need of improving existing MRC datasets in terms of both question and answer challenges and related evaluation criteria. In order to build more challenging MRC datasets, we need to understand existing MRC tasks, evaluation metrics and datasets better.

*1.4. Outline*

In Section 2, we focus on the MRC tasks. We first give a definition of typical MRC task. Then we compare multi-modal MRCs with textual MRCs, and discuss the differences between question answering tasks and machine reading comprehension tasks. Next, we analyze the existing classification method of MRC tasks which is widely used by the community. We argue that the existing classification method is inadequate and has potential problems. In order to solve the above problems, we propose a more adequate classification method of MRC tasks. We summarize 4 different attributes of MRC tasks. Each of these attributes can be divided into several categories. We give a detailed definition of each category with examples and explain why the new classification method is more adequate. After that, we collect totally 57 different MRC tasks and categorize them according to the new classification method. Finally, we analyze these MRC tasks and make statistical tables and charts of them.

In Section 3, we discuss the MRC evaluation metrics. Nine evaluation metrics of MRC tasks have been analyzed. We begin by presenting an overview of MRC evaluation metrics. Then we discuss the computing methods of each evaluation metric, including several sub-metrics such as token-level F1 and question-level F1. Next, we analyze the usage of each evaluation metric in different MRC tasks. After that, we make statistics on the usages of different evaluation metrics in the 57 MRC tasks. Finally, we analyze the relationship between the MRC task types and the evaluation metrics they used.

In Section 4, we present the family of MRC datasets. We begin by analyzing the size of each MRC datasets. Here, we have counted the total number of questions in each MRC dataset along with the sizes of its training set, development set and testing set, as well as the proportion of training set. Then we discuss the generation method of datasets which can be roughly described as several categories: Crowdsourcing, Expert, and Automated. Next, we conduct an in-depth analysis of the source of corpus and the type of context of MRC datasets. After that, we try to find all the download links, leaderboards and baseline projects of MRC datasets, all of which have been published on our website. Then, we present a statistical analysis of prerequisite skills and citations of the papers in which each dataset was proposed. Next, we summarize 10 characteristics of MRC datasets. Finally, we give a detailed description of each MRC dataset.

In Section 5, we discuss several open issues that remain unsolved in this field. Firstly, We believe that many important aspects have been overlooked which merit additional research, such as multi-modal MRC, commonsense and world knowledge, complex reasoning, robustness, interpretability, evaluation of the quality of MRC datasets. Secondly, we talk about understanding from the perspective of cognitive neuroscience. Finally, we share some of the latest research results of cognitive neuroscience and the inspiration of these results for NLP research.

In Section 6, we present a comprehensive conclusion of this survey.

Finally, we have published all the data on the website, the researchers could directly access each MRC datasets, papers, and baseline projects, or browse the leaderboards by clicking the hyperlinks. It is hoped that the research community could quickly access the comprehensive information of MRC datasets and their tasks. The address of the website is https://mrc-datasets.github.io/.

## 2. Tasks

### 2.1. Definition of Typical MRC Tasks

In our survey, machine reading comprehension is considered as a special research field, which includes some specific tasks, such as multi-modal machine reading comprehension, textual machine reading comprehension, and so forth. Since most of the existing machine reading comprehension tasks are in the form of question answering, the textual QA-based machine reading comprehension task is considered to be the typical machine reading comprehension task. According to previous review papers on MRC [14,27], the definition of a typical MRC task is:

**Definition 1.** *Typical machine reading comprehension task could be formulated as a supervised learning problem. Given a collection of textual training examples* $\left\{ \left( p_i, q_i, a_i \right) \right\}_{i=1}^{n}$, *where p is a passage of text, and q is a question regarding the text p. The goal of typical machine reading comprehension task is to learn a predictor f which takes a passage of text p and a corresponding question q as inputs and gives the answer a as output, which could be formulated as the following formula [14]:*

$$a = f(p, q) \tag{1}$$

*and it is necessary that a majority of native speakers would agree that the question q does regarding that text p, and the answer a is a correct one which does not contain information irrelevant to that question.*

### 2.2. Discussion on MRC Tasks

In this section, we first compare multi-modal MRCs with textual MRCs, and then discuss the relationship between question answering tasks and machine reading comprehension tasks.
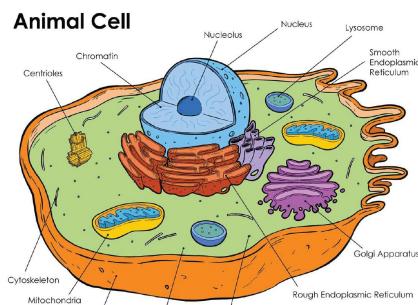
#### 2.2.1. Multi-Modal MRC vs. Textual MRC

Multi-modal MRC is a new challenging task that has received increasing attention from both the NLP and the CV communities. Compared with existing MRC tasks which are mostly textual, multi-modal MRC requires a deeper understanding of the text and visual information such as images and videos. When human reads, illustrations can help to understand the text. Experiments showed that children with higher mental imagery skills outperformed children with lower mental imagery skills on story comprehension after reading the experimental narrative [4]. These results emphasize the importance of mental imagery skills for explaining individual variability in reading development [4]. Therefore, if we want the machine to acquire human-level reading comprehension ability, multi-modal machine reading comprehension is a promising research direction.

In fact, there are already many tasks and datasets in this field, such as the TQA [31], MovieQA [32], COMICS [33] and RecipeQA [34]. As seen in Figure 4, TQA is a multi-modal MRC dataset that aims at answering multi-modal questions given a context of text, diagrams and images.

**Passage with illustration:**
This diagram shows the anatomy of an Animal cell. Animal Cells have an outer boundary known as the plasma membrane. The nucleus and the organelles of the cell are bound by this membrane. The cell organelles have a vast range of functions to perform like hormone and enzyme production to providing energy for the cells. They are of various sizes and have irregular shapes. Most of the cells size range between 1 and 100 micrometers and are visible only with help of microscope.

**Question with illustration:**
What is the outer surrounding part of the Nucleus?

**Choices:**
(1) Nuclear Membrane √
(2) Golgi Body
(3) Cell Membrane
(4) Nucleolus
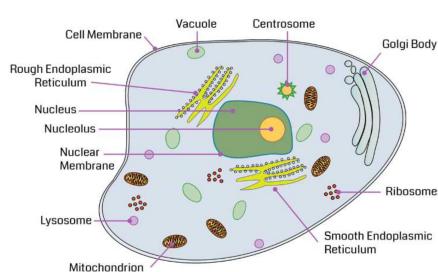
**Figure 4.** An example of multi-modal MRC task. The illustrations and questions are taken from the TQA [31] dataset.

## 2.2.2. Machine Reading Comprehension vs. Question Answering

The relationship between question answering and machine reading comprehension is very close. Some researchers consider MRC as a kind of specific QA task [14,27]. Compared with other QA tasks such as open-domain QA, MRC is characterized by that the computer is required to answer questions according to the specified text. However, other researchers regard the machine reading comprehension as a kind of method to solve QA tasks. For example, in order to answer open-domain questions, Chen et al. [35] first adopted document retrieval to find the relevant articles from Wikipedia, then used MRC to identify the answer spans from those articles. Similarly, Hu [36] regarded machine reading as one of the four methods to solve QA tasks. The other three methods are rule-based method, information retrieval method and knowledge-based method.

However, although the typical machine reading comprehension task is usually in the form of textual question answering, the forms of MRC tasks are usually diverse. Lucy Vanderwende [37] argued that machine reading could be defined as an automatic understanding of text. "One way in which human understanding of text has been gauged is to measure the ability to answer questions pertaining to the text. An alternative way of testing human understanding is to assess one's ability to ask sensible questions for a given text".

In fact, there are many such benchmark datasets for evaluating such techniques. For example, ShARC [38] is a conversational MRC dataset. Unlike other conversational MRC datasets, when answering questions in the ShARC, the machine needs to use background knowledge that is not in the context to get the correct answer. The first question in a ShARC conversation is usually not fully explained and does not provide enough information to answer directly. Therefore, the machine needs to take the initiative to ask the second question, and after the machine has obtained enough information, it then answers the first question.

Another example is RecipeQA [34] which is a dataset for multi-modal comprehension of illustrated recipes. There are four sub-tasks in RecipeQA, one of which is ordering task. Ordering task tests the ability of a model in finding a correctly ordered sequence given a jumbled set of representative images of a recipe [34]. As in previous visual tasks, the context of this task consists of the titles and descriptions of a recipe. To successfully complete this task, the model needs to understand the

temporal occurrence of a sequence of recipe steps and infer temporal relations between candidates, that is, boiling the water first, putting the spaghetti next, so that the ordered sequence of images aligns with the given recipe. In addition, in the MS MARCO [39], ordering tasks are also included.

In summary, although most machine reading comprehension tasks are in the form of question answering, it does not mean that machine reading comprehension tasks belong to the question answering. In fact, as mentioned above, the forms of MRC tasks are diverse. Question answering also includes a lot of tasks that do not emphasize that the system must read a specific context to get an answer, such as rule-based question answering systems and knowledge-based question answering systems (KBQA). Figure 5 illustrates the relation between machine reading comprehension (MRC) tasks and question answering (QA) tasks. As shown in Figure 5, we regard the general machine reading comprehension and the question answering as two subfields in the research field of natural language processing, both of which contain various specific tasks, such as Visual Question Answering (VQA) tasks, multi-modal machine reading comprehension tasks, and so forth. Among them, some of these tasks belong to both natural language processing and computer vision research fields, such as the VQA task and the multi-mode reading comprehension task. Lastly, most of the existing MRC tasks are textual question answering tasks, so we regard this kind of machine reading comprehension task as a typical machine reading comprehension task, and its definition is shown in Definition 1 above.
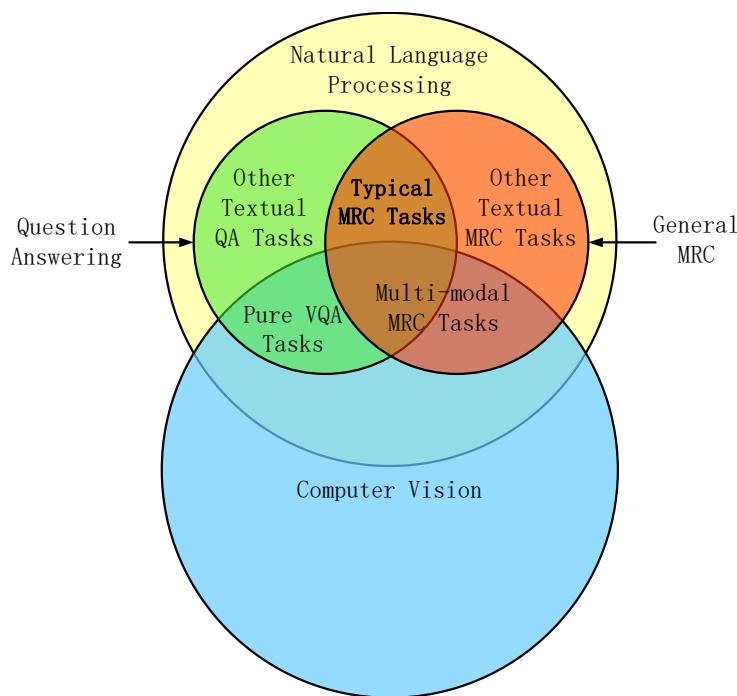


**Figure 5.** The relations between machine reading comprehension (MRC), question answering (QA), natural language processing (NLP) and computer version (CV).

### 2.2.3. Machine Reading Comprehension vs. Other NLP Tasks

There is a close and extensive relationship between machine reading comprehension and other NLP tasks. First of all, many useful methods in the field of machine reading comprehension can be introduced into other NLP tasks. For example, the stochastic answer network (SAN) [40,41] is first applied to MRC tasks and achieved results competitive to the state of the art on many MRC tasks such as the SQuAD and the MS MARCO. At the same time, the SAN can also be used in natural language processing (NLP) benchmarks [42], such as Stanford Natural Language Inference (SNLI), MultiGenre Natural Language Inference (MultiNLI), SciTail, and Quora Question Pairs datasets. For another example, Yin et al. (2017) [43] regards the document-level multi-aspect sentiment

classification task as a machine understanding task, and proposed a hierarchical iterative attention model. The experimental result of this model outperforms the classical baseline in TripAdvisor and BeerAdvocate datasets.

Secondly, some other NLP research results can also be introduced into the MRC area. Asai et al. (2018) [44] solved the task of non-English reading comprehension through a neural network translation (NMT) model based on attention mechanism. In detail, the paragraph question pair of non-English language is translated into English using the neural machine translation model, so that the English extraction reading comprehension model can output its answer, and then use the attention weights of the neural machine translation model to align the answers in the target text. Extra knowledge can also be introduced into MRC tasks. The authors of SG-Net [45] used syntax information to constrain attention in the MRC task. They used the syntactic dependency of interest (SDOI) to form an SDOI-SAN and have achieved state-of-the-art results on SQuAD 2.0 challenge. Minaee et al. (2020) [46] summarized more than 150 deep learning text classification methods and their performance on more than 40 popular datasets. Many of the methods mentioned in this article have been applied to MRC tasks.

Thirdly, MRC can be used as a step or component in the pipeline of some complex NLP tasks. For example, machine reading comprehension can be used as a step in open domain QA [35]. And in many dialogue tasks, machine reading comprehension can also be regarded as a part of pipeline [8,38,47].

### 2.3. Classification of MRC Tasks

In order to have a better understanding of MRC tasks, in this section, we analyze existing classification methods of tasks and identify potential limitations of these methods. After analyzing 57 MRC tasks and datasets, we propose a more precise classification method of MRC tasks which has 4 different attributes and each of them could be divided into several types. The statistics of the 57 MRC tasks are shown in the table in this section.

#### 2.3.1. Existing Classification Methods of MRC Tasks

In many research papers [14,26,27], MRC tasks are divided into four categories: cloze style, multiple-choice, span prediction, and free-form answer. Their relationship is shown in Figure 6:
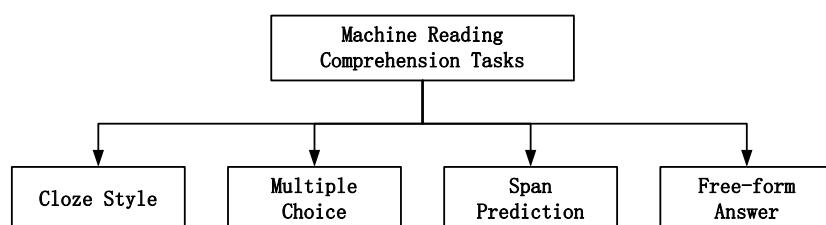


**Figure 6.** Existing classification method of machine reading comprehension tasks.

- Cloze style

In a cloze style task, there are some placeholders in the question. The MRC system needs to find the most suitable words or phrases which can be filled in these placeholders according to the context content.

- Multiple-choice

In a multiple-choice task, the MRC system needs to select a correct answer from a set of candidate answers according to the provided context.

- Span prediction

In a span prediction task, the answer is a span of text in the context. That is, the MRC system needs to select the correct beginning and end of the answer text from the context.

- Free-form answer

This kind of tasks allows the answer to be any free-text forms, that is, the answer is not restricted to a single word or a span in the passage [14].

### 2.3.2. Limitations of Existing Classification Method

However, the above task classification method does have certain limitations. Here are the reasons:

First, an adequate classification method should be precise at least or can classify each MRC task distinctly. But the existing classification method is a bit ambiguous or indistinct, that is, according to this classification method, a MRC task may belong to multiple task types. For instance, as seen in Figure 7, a sample in the "Who did What" task [48] are both in the form of "Cloze style" and "Multiple-choice", and we can see that the answer is a span of a text in the context so that it can also be classified to "Span prediction".

| |
|---|
| **Passage:** Tottenham won 2-0 at Hapoel Tel Aviv in UEFA Cup action on Thursday night in a defensive display which impressed Spurs skipper Robbie Keane... Keane scored the first goal at the Bloomfield Stadium with **Dimitar Berbatov**, who insisted earlier on Thursday he was happy at the London club, heading a second. The 26-year-old Berbatov admitted the reports linking him with a move had affected his performances ... Spurs manager Juande Ramos has won the UEFA Cup in the last two seasons ... |
| **Question:** Tottenham manager Juande Ramos has hinted he will allow _____ to leave if the Bulgaria striker makes it clear he is unhappy. |
| **Choices:** (A) Robbie Keane (B) **Dimitar Berbatov** $\sqrt{}$ |

**Figure 7.** An example of MRC task. The question-answer pair and passage are taken from the "Who did What" [48].

Secondly, with the rapid development of MRC, a large number of novel MRC tasks have emerged in recent years. One example is multi-modal MRC, such as MovieQA [32], COMICS [33], TQA [31] and RecipeQA [34]. Compared with the traditional MRC task which only requires understanding a text, the multi-modal MRC task requires the model to understand the semantics behind the text and visual images at the same time. A fundamental characteristic of human language understanding is multimodality. Our observation and experience of the world bring us a lot of common sense and world knowledge, and the multi-modal information is extremely important for us. In essence, real world information is multi-modal and widely exists in texts, voices, and images. But these multi-modal tasks are ignored by the existing classification method.

In addition, as seen in Figure 8, we list several tasks that belong to the fuzzy classification mentioned above, such as ReviewQA, Qangaroo, Who-did-What, MultiRC, LAMBADA, ReCoRD. Due to the limited space, we only list a few of them in the figure. According to our statistics, among the 57 MRC tasks we collected, 29 tasks fall into this situation.
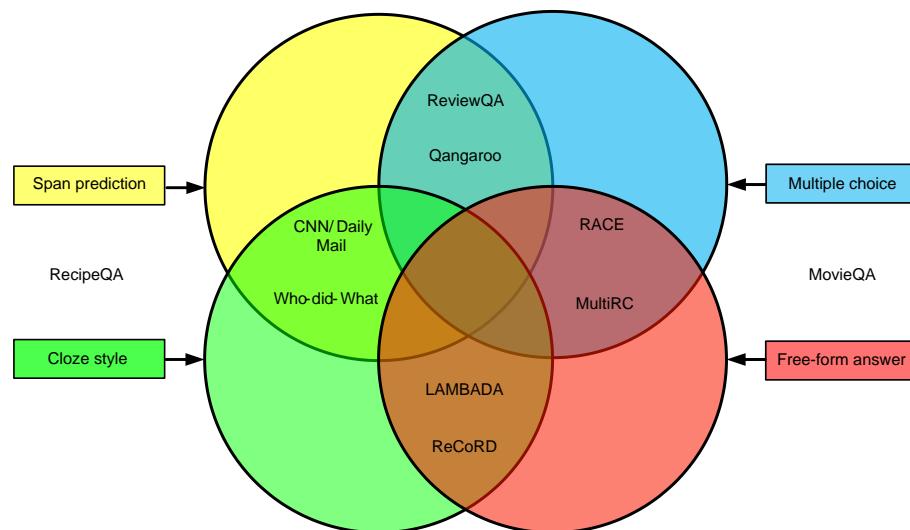
**Figure 8.** The indistinct classification caused by existing classification method.

### 2.3.3. A New Classification Method

In this section, we propose a new classification method of MRC tasks. As shown in Figure 9, we summarize four different attributes of MRC tasks, including the type of corpus, the type of questions, the type of answers, and the source of answers. Each of these attributes can be divided into several different categories. These categories are: (1) Type of corpus: textual, multi-modal. (2) Type of questions: natural form, cloze style, synthetic form. (3) Type of answers: natural form, multiple-choices. (4) Source of answers: spans, free-form.
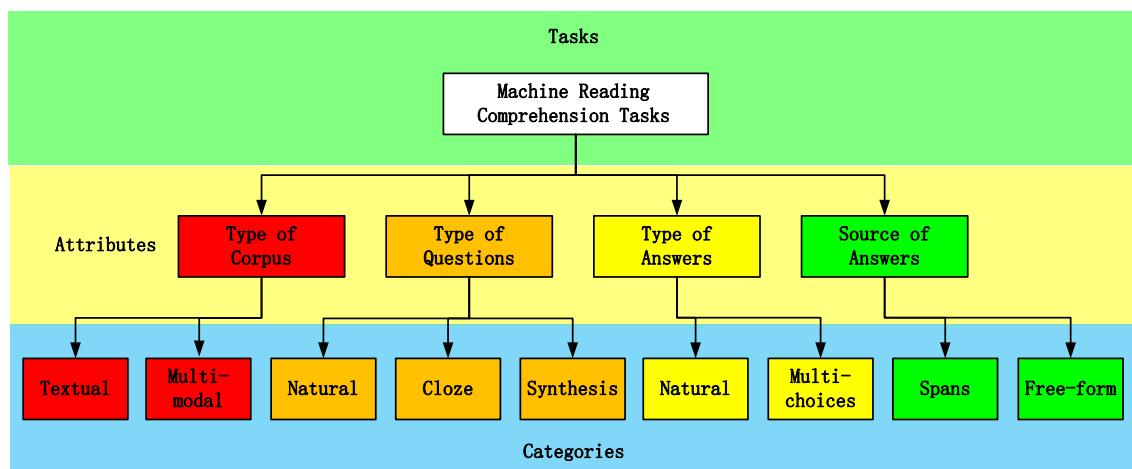


**Figure 9.** A new classification method of machine reading comprehension tasks.

In order to explain the new classification method more clearly, we make a sunrise statistical chart for the MRC task classification, as seen in Figure 10. We collect 57 different specific MRC tasks. Finally, according to the new classification method, the sunrise chart is divided into four layers of rings, representing the four attributes of tasks. The most central blue layer represents the 'Type of Corpus'. Among them, light blue indicates that the type of the task's corpus belongs to 'Textual', and dark blue means that the type of the task's corpus belongs to 'Multi-modal'. The magnitude of different color blocks is set according to the proportion of 57 MRC tasks we collected. Among them, the 'Textual' tasks still account for the vast majority of tasks (89.47%). Currently, the proportion of MRC tasks is still very small, about 10.53%. Therefore, as can be seen, the range of light blue color blocks is large, while the range of dark blue color blocks is small. The second green layer represents the 'Type of

Question', the third red and pink layer represents the 'Type of Answer', and the outermost yellow layer represents the 'Source of Answer'.
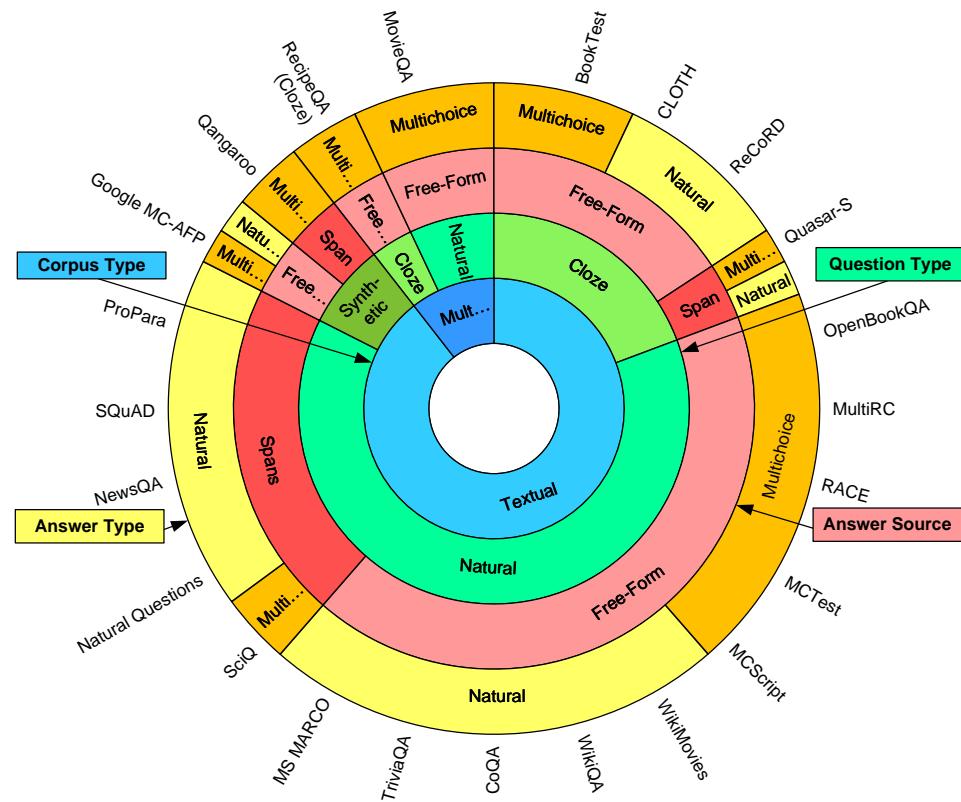


**Figure 10.** A sunburst chart on the proportion of different types of machine reading comprehension tasks.

Take the BookTest for an example, as seen in the top of Figure 10, the BookTest is a 'Textual' MRC task, its question is in the 'Cloze' form, and the answer comes from 'Spans' in the context, its answer form is 'Multiple-choice'. Another example is the cloze subtask in the RecipeQA dataset, which is also in the top of Figure 10. In this task, the answer form is 'Multiple-choice', and question types is 'Cloze' form. Moreover, the context corpus of RecipeQA contains images, so it is a 'Multi-modal' task. The answer types include textual 'Multiple-choice' and image 'Multiple-choice'. Therefore, the type of each MRC task is determined according to four different attributes, which eliminates the fuzzy situation that the same task belongs to multiple types in the traditional classification method.

However, it must be pointed out that although the new classification method fits precisely to the existing datasets, it may suffer from the lack of future generalization. We believe that with the continuous development of MRC field, new MRC tasks will certainly appear, and the classification methods of MRC tasks will also keep pace with them.

## 2.4. Definition of Each Category in the New Classification Method

As mentioned above, we propose a new classification method of MRC tasks. As shown in Figure 9 above, we summarize four different attributes of MRC tasks, including the type of corpus, the type of questions, the type of answers, and the source of answers. Each of these attributes can be divided into several different categories. In this subsection, we will give detailed definitions of each category with examples.

Here are some assumptions or notations we need before the formal definitions:

**Assumption 1.** *Suppose V is a pure textual vocabulary, and M is a multi-modal dataset which consists of images or other non-text information.*

**Assumption 2.** *Suppose in a MRC corpus, $C_i$ is the i-th context, $Q_i$ is the i-th question, and $A_i$ is the answer to question $Q_i$ according to context $C_i$. Let the context $C_i = \{c_0, c_1, \ldots, c_{lci}\}$, and the question $Q_i = \{q_0, q_1, \ldots, q_{lqi}\}$, and $A_i = \{a_0, a_1, \ldots, a_{lai}\}$, where $l_{ci}$, $l_{qi}$ and $l_{ai}$ denote the length of the i-th context $C_i$, question $Q_i$, and answer $A_i$ respectively. While $c_i$, $q_i$ and $a_i$ is usually a word or a image, that is, $c_k \in V \cup M$, $q_k \in V \cup M$ and $a_k \in V \cup M$.*

2.4.1. Type of Corpus

According to whether or not the corpus contains information other than text, such as pictures, the MRC tasks can be divided into two categories: multi-modal (the combination of graphics and text) and textual.

- Multi-modal

In multi-modal MRC Corpus, multi-modal information includes context, questions, or answers. It can be defined as:

**Definition 2.** *In a MRC task with multi-modal corpus, the corpus P can be formalized as a collection of training examples, that is, $P = \{C_i, Q_i, A_i\}_{i=1}^{n}$, where $C_i$ is the context, $Q_i$ is a question, and $A_i$ is the answer to question $Q_i$ according to context $C_i$. In the multi-modal corpus P, the entities in the corpus consists of text and images at the same time, therefore, $P \cap V \neq \varnothing$ and $P \cap M \neq \varnothing$.*

An example of the multi-modal corpus can be seen in Figure 4 above. There is a certain similarity between multi-modal MRC tasks and Visual Question Answering (VQA) tasks. But multi-modal MRC tasks focus more on natural language understanding, and their context contains more text that needs to be read, and the VQA task usually does not have much context and gives the image directly.

- Textual

Most MRC tasks belong to this category. Their context, questions and answers are all plain texta. It can be defined as:

**Definition 3.** *In a MRC task with textual corpus, the corpus P can be formalized as a collection of training examples, that is, $P = \{C_i, Q_i, A_i\}_{i=1}^{n}$, where $C_i$ is the context, $Q_i$ is a question, and $A_i$ is the answer to question $Q_i$ according to context $C_i$. In the textual corpus P, all the entities in the context, questions and answers are in pure text, therefore, $P \cap V \neq \varnothing$ and $P \cap M = \varnothing$.*

Example of textual corpus can be seen in Figure 11 below:

| | |
|---|---|
| **Passage:** | 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, in- tense periods of rain in scattered locations are called "showers". |
| **Question:** | What causes precipitation to fall? |
| **Answer:** | **gravity** |

**Figure 11.** An example of textual MRC task.

2.4.2. Type of Questions

According to the type of question, a MRC task can be classified into three categories: cloze style, natural form, and synthetic form:

- Cloze style

The cloze question is usually a sentence with a placeholder. Its sentence pattern may be a declarative sentence, an imperative sentence, and so forth, and is not necessarily an interrogative sentence. In addition, the sentence may also contain image information. The system is required to find a correct word, phrase or image that is suitable to be filled in the placeholder so that the sentence is complete. The cloze question can be defined as:

**Definition 4.** *Given the context $C = \{c_0, c_1, \ldots c_j, \ldots c_{j+n} \ldots, c_{lc}\}$ $(0 \le j \le lc, 0 \le n \le lc - 1, c_j \in V \cup M)$, where $l_c$ denotes the length of this context C. $A = \{c_j, \ldots c_{j+n}\}$ is a short span in context C. After replaced A with a placeholder X, a cloze style question Q for context C is formed, it can be formulated as $Q = \{c_0, c_1, \ldots X \ldots, c_{lc}\}$, in which the X is a placeholder. The answer to question Q is the $A = \{c_j, \ldots c_{j+n}\}$.*

According to the type of corpus, cloze questions also can be divided into textual and multi-modal. A textual cloze question is usually a sentence with a placeholder. The MRC system is required to find a correct word or phrase that is suitable to be filled in the placeholder so that the sentence is complete. An example of textual cloze question has been shown in Figure 7.

A multi-modal cloze question is a natural sentence with visual information such as images, but some parts of these images are missing, and the MRC system is required to fill in the missing images. For example, a sample of visual cloze question in the RecipeQA [34] dataset is shown in Figure 12:

**Passage**
Last-Minute Lasagna:
1. Heat oven to 375 degrees F. Spoon a thin layer of sauce over the bottom of a 9-by-13-inch baking dish.
2. Cover with a single layer of ravioli.
3. Top with half the spinach half the mozzarella and a third of the remaining sauce.
4. Repeat with another layer of ravioli and the remaining spinach mozzarella and half the remaining sauce.
5. Top with another layer of ravioli and the remaining sauce not all the ravioli may be needed. Sprinkle with the Parmesan.
6. Cover with foil and bake for 30 min. Uncover and bake until bubbly, 5 to 10 min.
7. Let cool 5 min before spooning onto individual plates.

**Question:** Choose the best image for the missing blank to correctly complete the recipe.



**Choices:**



| **(A)** ✓ | **(B)** | **(C)** | **(D)** |

**Figure 12.** An example of multi-modal cloze style question. The images and questions are taken from the RecipeQA [34] dataset.

- Natural form

A question in natural form is a natural question that conforms to the grammar of natural language. Different from the cloze question, which contains placeholder, a natural form question is a complete sentence and a question that conforms to the grammatical rules. It could be defined as:

**Definition 5.** *In a MRC task, given a 'Natural' question Q, it could be formulated as* $Q_i = \{q_0, q_1, \ldots q_i \ldots, q_{lq}\}$, *where* $q_i \in V \cup M (0 \le i \le lq)$. *Q denotes a complete sentence (may also contain images) that conforms to the natural language grammar and* $l_q$ *denotes the length of the question Q.*

In most cases, a 'Natural' question $Q$ is an interrogative sentence that asks a direct question and is punctuated at the end with a question mark. However, in some cases, $Q$ may not be an interrogative sentence but an imperative sentence, for example, "please find the correct statement from the following options."

In addition, according to the type of corpus, natural form questions can be divided into textual and multi-modal. Textual natural question is usually a natural question or imperative sentence. With some graphics or video, the multi-modal natural question is also a natural question or imperative. Example of textual natural question is shown in Figure 13 below, and example of multi-modal natural question has been shown in Figure 4.

| | |
|---|---|
| **Passage:** | 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, in- tense periods of rain in scattered locations are called "showers". |
| **Question:** | What causes precipitation to fall? |
| **Answer:** | **gravity** |

**Figure 13.** An example of textual natural question.

- Synthetic style

The synthetic form of the question is just a list of words and do not necessarily conform to normal grammatical rules. Common datasets with synthetic form questions are Qangaroo, WikiReading, and so on. Take Qangaroo as an example, in the Qangaroo dataset, the question is replaced by a collection of attribute words. The 'question' here is not a complete sentence that fully conforms to the natural language grammar, but a combination of words. The synthetic form of the question can be defined as:

**Definition 6.** *In a MRC task, given a 'Synthetic style' question Q, it could be formulated as* $Q_i = \{q_0, q_1, \ldots q_i \ldots, q_{lq}\}$, *where* $q_i \in V \cup M (0 \le i \le lq)$. *Q denotes a series of words (may also contain images) that do not conforms to the natural language grammar and* $l_q$ *denotes the length of the Q.*

The example of synthetic style question is in shown in the following:

2.4.3. Type of Answers

According to the type of answers, MRC tasks can be divided into two categories: multiple-choice forms, natural forms.

- Multiple-choice answer

In a MRC task, when the type of answers is 'Multi-choice', there is a series of candidate answers for each question. and it can be defined as:

**Definition 7.** *Given the candidate answers $A = \{A_1, \ldots A_i \ldots, A_n\}$, where n denotes the number of candidate answers for each question, and $A_i(0 \le i \le n)$ denotes an optional answer. The goal of the task is to find the right answer $A_j(0 \le j \le n)$ from A, and one or more answer options in A is correct.*

Examples of textual multiple-choices form of answers have been shown in Figures 7 and 14, and multi-modal example has been shown in Figure 12 above.

| |
|---|
| **Passage:** The hanging Gardens, in [Mumbai], also known as Pherozeshah Mehta Gardens, are terraced gardens … They provide sunset views over the [Arabian Sea] … <br> Mumbai (also known as Bombay, the official name until 1995) is the capital city of the Indian state of Maharashtra. It is the most populous city in ***India*** … <br> The Arabian Sea is a region of the northern Indian Ocean bounded on the north by Pakistan and Iran, on the west by northeastern Somalia and the Arabian Peninsula, and on the east by India … |
| **Synthetic Question:**    (Hanging gardens of Mumbai, country, ?) |
| **Choices:**    (A)Iran,    (B)**India**√,    (C)Pakistan,    (D) Somalia |

**Figure 14.** An example of synthetic style question. The passage and question are taken from the Qangaroo [49] dataset.

- Natural form of answers

The answer is a natural word, phrase, sentence or image but it does not have to be in the form the multiple options. It could be defined as follows:

**Definition 8.** *In a MRC task, when the type of answers is 'Natural', it means the answer A can be a word, a phrase or a natural sentence, or even images. The answer A could be formulated as: $A = \{a_1, \ldots a_k \ldots, a_l\}$, where l denotes the length of answer A. $a_k \in V \cup M(0 \le k \le l)$.*

The example of natural textual answers has been shown in Figure 13 above, and the example of natural multi-modal answer has not been found by us, that is, all the multi-modal MRC datasets we collected in this survey contain only multiple-choice answers.

2.4.4. Source of Answers

According to different sources of answers, we divide the MRC tasks into two categories: span and free-form.

- Span answer

In a MRC task, when the source of answer is 'Spans', it means that the answers come from context and are spans of context, and it can be defined as:

**Definition 9.** *Given the context $C = \{c_0, \ldots c_k \ldots, c_l\}$, where l denotes the length of the context. $c_k \in V \cup M(0 \le k \le l)$. The 'Span' answer A could be formulated as $A = \{c_m, \ldots, c_n\}(0 \le m \le n \le l)$.*

The example of textual span answer is shown in Figure 7 above. It should be noted that, in this paper, we do not provide example for multi-modal span answers, because such tasks already exist in the field of computer vision, such as semantic segmentation, object detection, or instance segmentation.

- Free-form answer

A free-form answer may be any phrase, word, or even image (not necessarily from the context). In a MRC task, when the source of answer is 'Free-form', it means that the answers can be any free-text or images, and there is no limit to where the answer comes from. It could be defined as follows:

**Definition 10.** *Given the context C, the 'Free-form' answer A may or may not come from context C, that is, either $A \subseteq C$ or not. The 'Free-form' answer A could be formulated as $A = \{w_0, w_1, \ldots, w_{l-1}, w_l\}$ where $l$ denotes the length of the context. $w_k \in V \cup M (0 \leq k \leq l)$.*

Example of multi-modal free-form answer are shown in Figure 12 and example of textual free-form answer are shown in Figure 15 below:

| | |
|---|---|
| **Passage:** | That year, his Untitled (1981), a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artist's signature scrawls, was sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate. |
| **Question:** | How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation? |
| **Answer:** | 4300000 |

**Figure 15.** An example of textual free-form answer. The question-answer pair and passage are taken from the DROP [50] dataset.

*2.5. Statistics of MRC Tasks*

In this section, we collected 57 different MRC tasks and made a statistical chart of MRC task classification according to four attributes, as shown in Figure 16. We can see that for the type of corpus, the textual task still accounts for a large proportion which is 89.47%. At present, the proportion of multi-modal reading comprehension tasks is still small, about 10.53%, which shows that the field of multi-modal reading comprehension still has many challenge problems for future research. In terms of question types, the most common type is the natural form of questions, followed by cloze type and synthetic type. In terms of answer types, the proportion of natural type and multiple-choice type are 52.63% and 47.37% respectively. In terms of answer source, 29.82% of the answers are of spans type, and 70.18% of the answers are of free-form.
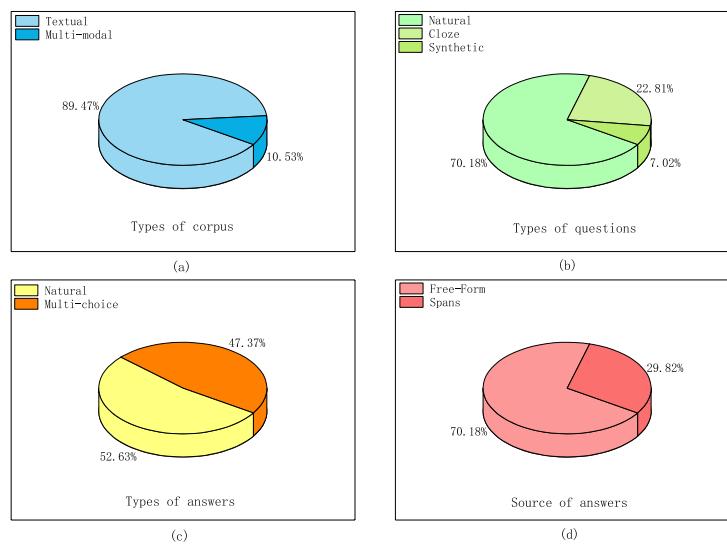


**Figure 16.** A pie chart on the proportion of different types of machine reading comprehension tasks: (**a**) Type of corpus. (**b**) Type of questions. (**c**) Type of answers. (**d**) Source of answers.

As shown in Table 1. The tasks in the table are ordered by the year the dataset was published. It should be noted that note that the names of many specific MRC tasks are often the same as the names of the datasets they may utilize. And the name of a certain category of MRC task and the name of a specific MRC task are two different concepts. For example, the RecipeQA [34] dataset contains two different tasks which are RecipeQA-Coherence and RecipeQA-Cloze.

**Table 1.** Different machine reading comprehension tasks.

| Year | MRC Tasks | Corpus Type | Question Type | Answer Source | Answer Type |
|------|-----------|-------------|---------------|---------------|-------------|
| 2013 | MCTest [15] | Textual | Natural | Free-Form | Multi-choice |
| 2015 | CNN/Daily Mail [19] | Textual | Cloze | Spans | Natural |
| 2015 | CuratedTREC [51] | Textual | Natural | Free-Form | Natural |
| 2015 | WikiQA [52] | Textual | Natural | Free-Form | Natural |
| 2016 | WikiMovies [53] | Textual | Natural | Free-Form | Natural |
| 2016 | SQuAD 1.1 [19] | Textual | Natural | Spans | Natural |
| 2016 | Who-did-What [48] | Textual | Cloze | Spans | Natural |
| 2016 | MS MARCO [39] | Textual | Natural | Free-Form | Natural |
| 2016 | NewsQA [54] | Textual | Natural | Spans | Natural |
| 2016 | LAMBADA [55] | Textual | Cloze | Free-Form | Natural |
| 2016 | WikiReading [56] | Textual | Synthetic | Free-Form | Natural |
| 2016 | Facebook CBT [28] | Textual | Cloze | Free-Form | Multi-choice |
| 2016 | BookTest [57] | Textual | Cloze | Free-Form | Multi-choice |
| 2016 | Google MC-AFP [58] | Textual | Synthetic | Free-Form | Multi-choice |
| 2016 | MovieQA [32] | Multi-modal | Natural | Free-Form | Multi-choice |
| 2017 | TriviaQA-Web [59] | Textual | Natural | Free-Form | Natural |
| 2017 | TriviaQA-Wiki [59] | Textual | Natural | Free-Form | Natural |
| 2017 | RACE [60] | Textual | Natural | Free-Form | Multi-choice |
| 2017 | Quasar-S [61] | Textual | Cloze | Spans | Multi-choice |
| 2017 | Quasar-T [61] | Textual | Natural | Spans | Natural |
| 2017 | SearchQA [62] | Textual | Natural | Free-Form | Natural |
| 2017 | NarrativeQA [63] | Textual | Natural | Free-Form | Natural |
| 2017 | SciQ [64] | Textual | Natural | Spans | Multi-choice |
| 2017 | Qangaroo-MedHop [49] | Textual | Synthetic | Spans | Multi-choice |
| 2017 | Qangaroo-WikiHop [49] | Textual | Synthetic | Spans | Multi-choice |
| 2017 | TQA [31] | Multi-modal | Natural | Free-Form | Multi-choice |
| 2017 | COMICS-Coherence [33] | Multi-modal | Natural | Free-Form | Multi-choice |
| 2017 | COMICS-Cloze [33] | Multi-modal | Cloze | Free-Form | Multi-choice |
| 2018 | QuAC [65] | Textual | Natural | Free-Form | Natural |
| 2018 | CoQA [47] | Textual | Natural | Free-Form | Natural |
| 2018 | SQuAD 2.0 [66] | Textual | Natural | Spans | Natural |
| 2018 | HotpotQA-Distractor [67] | Textual | Natural | Spans | Natural |
| 2018 | HotpotQA-Fullwiki [67] | Textual | Natural | Spans | Natural |
| 2018 | DuoRC-Self [68] | Textual | Natural | Free-Form | Natural |
| 2018 | DuoRC-Paraphrase [68] | Textual | Natural | Free-Form | Natural |
| 2018 | CLOTH [69] | Textual | Cloze | Free-Form | Natural |
| 2018 | ReCoRD [70] | Textual | Cloze | Free-Form | Natural |
| 2018 | CliCR [71] | Textual | Cloze | Free-Form | Natural |
| 2018 | ReviewQA [25] | Textual | Natural | Spans | Multi-choice |
| 2018 | ARC-Challenge Set [72] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | ARC-Easy Set [72] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | OpenBookQA [73] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | SciTail [74] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | MultiRC [75] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | RecipeQA-Cloze [34] | Multi-modal | Cloze | Free-Form | Multi-choice |
| 2018 | RecipeQA-Coherence [34] | Multi-modal | Natural | Free-Form | Multi-choice |
| 2018 | PaperQA-Title [76] | Textual | Cloze | Free-Form | Multi-choice |
| 2018 | PaperQA-Last [76] | Textual | Cloze | Free-Form | Multi-choice |
| 2018 | PaperQA(Hong et al.) [77] | Textual | Natural | Spans | Natural |
| 2018 | MCScript [78] | Textual | Natural | Free-Form | Multi-choice |
| 2018 | ProPara [79] | Textual | Natural | Spans | Natural |
| 2019 | Natural Questions-Short [80] | Textual | Natural | Spans | Natural |
| 2019 | Natural Questions-Long [80] | Textual | Natural | Spans | Natural |
| 2019 | DREAM [81] | Textual | Natural | Free-Form | Multi-choice |
| 2019 | ShARC [38] | Textual | Natural | Free-Form | Multi-choice |
| 2019 | CommonSenseQA [82] | Textual | Natural | Free-Form | Multi-choice |
| 2019 | DROP [50] | Textual | Natural | Free-Form | Natural |

Form of Task vs. Content of Task

The discussion above is mainly about the form of MRC tasks. However, it should be noted that, besides the form of the MRC task, the content of the context/passage and the question also determine the type of a task. As shown in Figure 17, in the FaceBook BAbi dataset [29], there are many different types of MRC tasks depending on the content of the passages and questions. But because classifying tasks based on the content is a very subjective matter without established standards, herein, we mainly analyze the forms of tasks rather than the content.

| Task 1: Yes/No Questions | Task 2: Counting |
|---|---|
| John moved to the playground. Daniel went to the bathroom. John went back to the hallway. | Daniel picked up the football. Daniel dropped the football. Daniel got the milk. Daniel took the apple. |
| Is John in the playground? Answer:no Is Daniel in the bathroom? Answer:yes | How many objects is Daniel holding? Answer:two |
| **Task 3: Lists/Sets** | **Task 4: Indefinite Knowledge** |
| Daniel picks up the football. Daniel drops the newspaper. Daniel picks up the milk. John took the apple. | John is either in the classroom or the playground. Sandra is in the garden. |
| What is Daniel holding? Answer:milk, football | Is John in the classroom? Answer:maybe Is John in the office? Answer:no |

**Figure 17.** Task examples in the Facebook BAbi dataset [29], the types of these tasks are determined by the the content of passages and questions.

## 3. Evaluation Metrics

### 3.1. Overview of Evaluation Metrics

The most commonly used evaluation metric for MRC models is accuracy. However, in order to more comprehensively compare the performances of MRC models, the models should be evaluated by various evaluation metrics. In this section, we introduce the calculation methods of commonly used evaluation metrics in machine reading comprehension, which include: Accuracy, Exact Match, Precision, Recall, F1, ROUGE, BLEU, HEQ and Meteor. For multiple-choice or cloze style tasks, Accuracy is usually used to evaluate MRC models. For span prediction tasks, Exact Match, Precision, Recall, and F1 are usually used as evaluation metrics. Currently, many of the evaluation metrics for MRC tasks are derived from other research areas in NLP (natural language processing) such as machine translation and text summaries. Similar to machine translation tasks, the goal of a MRC task is also to generate some text and compare it with the correct answer. So the evaluation metrics of machine translation tasks can also be used for MRC tasks. In the following sections, we will give detailed calculation methods of these evaluation metrics.

### 3.2. Accuracy

Accuracy represents the percentage of the questions that a MRC system accurately answers. For example, suppose a MRC task contains $N$ questions, each question corresponds to one correct answer, the answers can be a word, a phrases, or a sentence, and the number of questions that the system answers correctly is $M$. The equation for the accuracy is as follows:

$$Accuracy = \frac{M}{N} \qquad (2)$$

### 3.3. Exact Match

If the correct answer to the question is a sentence or a phrase, it is possible that some of the words in the system-generated answer are correct answers, and the other words are not correct answers. In this case, Exact Match represents the percentage of questions that the system-generated answer exactly matches the correct answer, which means every word is the same. Exact Match is often abbreviated as EM.

For example, if a MRC task contains $N$ questions, each question corresponds to one right answer, the answers can be a word, a phrases or a sentence, and the number of questions that the system answers correctly is $M$. Among the remaining $N - M$ answers, some of the answers may contain some ground truth answer words, but not exactly match the ground truth answer. The Exact Match can then be calculated as follows:

$$\text{Exact Match} = \frac{M}{N} \tag{3}$$

Therefore, for the span prediction task, Exact Match and Accuracy are exactly the same. But for a multi-choice task, Exact Match is usually not used because there is no situation where the answer includes a portion of the correct answer. In addition, to make the evaluation more reliable, it is also common to collect multiple correct answers for each question. Therefore, the exact match score is only required to match any of the correct answers [14].

### 3.4. Precision

#### 3.4.1. Token-Level Precision

The token-level precision represents the percentage of token overlap between the tokens in the correct answer and the tokens in the predicted answer. Following the evaluation method in SQuAD [19,66], we treat the predicted answer and correct answer as bags of tokens, while ignoring all punctuation marks and the article words such as "a" and "an" or "the". In order to get the token-level Precision, we first need to understand the token-level true positive (TP), false positive (FP), true negative (TN), and false negative (FN), as shown in Figure 18:
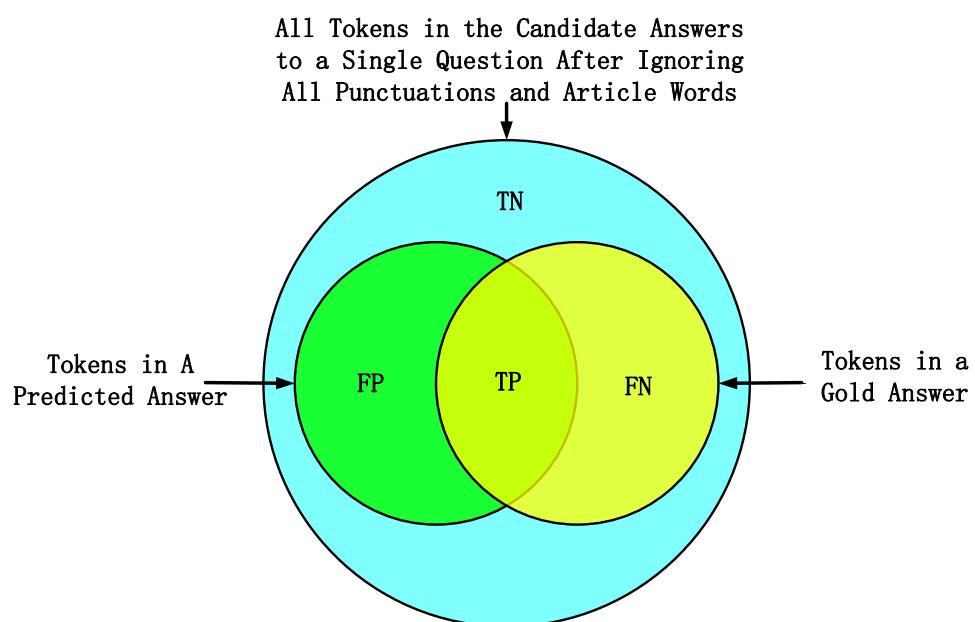


**Figure 18.** The token-level true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

As seen in Figure 18, for a single question, the token-level true positive (TP) denotes the same tokens between the predicted answer and the correct answer. The token-level false positive (FP) denotes the tokens which are not in the correct answer but the predicted answer, while the false negative (FN) denotes the tokens which are not in the predicted answer but the correct answer. A token-level Precision for a single question is computed as follows:

$$Precision_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FP_T)} \tag{4}$$

where $Precision_{TS}$ denotes the token-level Precision for a single question, and $Num(TP_T)$ denotes the number of token-level true positive (TP) tokens and $Num(FP_T)$ denotes the number of token-level false positive (FP) tokens.

For example, if a correct answer is "a cat in the garden" and the predicted answer is "a dog in the garden". We can see, after ignoring the article word "a" and "the", the number of the shared tokens between the predicted answer and the correct answer is 2, which is also the $Num(TP_T)$ , and $Num(FP_T)$ is 1, so the token-level Precision for this answer is 2/3.

### 3.4.2. Question-Level Precision

The question-level precision represents the average percentage of answer overlaps (not token overlap) between all the correct answers and all the predicted answers in a task [52]. The question-level true positive (TP), false positive (FP), true negative (TN), and false negative (FN) are shown in Figure 19:
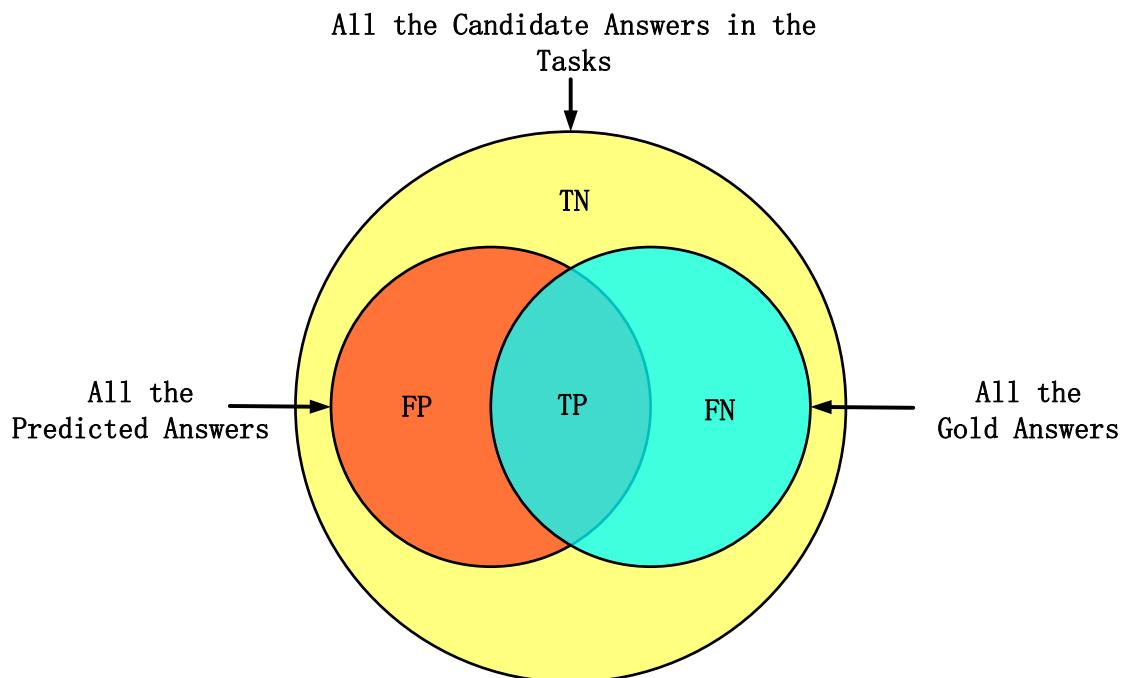


**Figure 19.** The question-level true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

As seen in Figure 19, the question-level true positive (TP) denotes the shared answers between all predicted answers and all correct answers, in which one answer is treated as one entity, no matter how many words it consists of. And the question-level false positive (FP) denotes these predicted answers which do not belong to the set of correct answers, while the question-level false negative (FN) denotes those correct answers which do not belong to the set of predicted answers. A question-level Precision for a task is computed as follows:

$$Precision_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FP_Q)} \tag{5}$$

where $Precision_Q$ denotes the question-level Precision for a task, $Num(TP_Q)$ denotes the number of question-level true positive (TP) answers and $Num(FP_Q)$ denotes the number of question-level false positive (FP) answers.

*3.5. Recall*

3.5.1. Token-Level Recall

The Recall represents the percentage of tokens in a correct answer that have been correctly predicted in a question. Following the definitions of the token-level true positive (TP), false positive (FP), and false negative (FN) above, A token-level Recall for a single answer is computed as follows:

$$Recall_{TS} = \frac{Num(TP_T)}{Num(TP_T) + Num(FN_T)} \tag{6}$$

where $Recall_{TS}$ denotes the token-level Recall for a single question, $Num(TP_T)$ denotes the number of token-level true positive (TP) tokens and $Num(FN_T)$ denotes the number of token-level false negative (FN) tokens.

3.5.2. Question-Level Recall

The question-level Recall represents the percentage of the correct answers that have been correctly predicted in a task [52]. Following the definitions of the token-level true positive (TP), false positive (FP), and false negative (FN), A token-level Recall for a single answer is computed as follows:

$$Recall_Q = \frac{Num(TP_Q)}{Num(TP_Q) + Num(FN_Q)} \tag{7}$$

where $Recall_Q$ denotes the question-level Recall for a task, $Num(TP_Q)$ denotes the number of question-level true positive (TP) answers and $Num(FN_Q)$ denotes the number of question-level false negative (FN) answers.

*3.6. F1*

3.6.1. Token-Level F1

Token-level F1 is a commonly used MRC task evaluation metrics. The equation of token-level F1 for a single question is:

$$F1_{TS} = \frac{2 \times Precision_{TS} \times Recall_{TS}}{Precision_{TS} + Recall_{TS}} \tag{8}$$

where $F1_{TS}$ denotes the token-level F1 for a single question, $Precision_{TS}$ denotes the token-level Precision for a single question and $Recall_{TS}$ denotes the token-level Recall for a single question.

To make the evaluation more reliable, it is also common to collect multiple correct answers to each question [14]. Therefore, to get the average token-level F1, we first have to compute the maximum token-level F1 of all the correct answers of a question, and then average these maximum token-level F1 over all of the questions [14]. The equation of average token-level F1 for a task is:

$$F1_T = \frac{\sum Max(Precision_{TS})}{Num(Questions)} \tag{9}$$

where $F1_T$ denotes the average token-level F1 for a task, and $Max(Precision_{TS})$ denotes the maximum token-level F1 of all the correct answers for a single question, $\sum Max(Precision_{TS})$ denotes the sum of for every question in the task. $Num(Questions)$ denotes the number of questions in the task.

### 3.6.2. Question-Level F1

The equation of question-level F1 for a task is:

$$F1_Q = \frac{2 \times Precision_Q \times Recall_Q}{Precision_Q + Recall_Q} \tag{10}$$

where $F1_Q$ denotes the question-level F1, $Precision_Q$ denotes the question-level Precision for a task and $Recall_Q$ denotes the question-level Recall for a task.

### *3.7. ROUGE*

ROUGE stands for Recall-Oriented Understudy for Gisting Evaluation, which was first proposed by Chin-Yew Lin [83]. In this paper, ROUGE was used to evaluate the performance of text summary systems. Currently, ROUGE is also used in the evaluation of MRC systems.

ROUGE-N is a *n*-gram Recall between a candidate summary and a set of reference summaries [83]. According to the value of *n*, ROUGE is specifically divided into ROUGE-1, ROUGE-2, ROUGE-3, and so on. The ROUGE-N is computed as follows:

$$\text{ROUGE-N} = \frac{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{RS\}} \sum_{gram_n \in S} Count(gram_n)} \tag{11}$$

where $n$ is the length of the $n$-gram, $Count(gram_n)$ is the maximum number of times the $n$-gram appears in the candidate text and predicted text generated by the algorithm, and $RS$ is an abbreviation of $ReferenceSummaries$.

### *3.8. BLEU*

BLEU (Bilingual Evaluation Understudy) was proposed by Papineni et al. [84]. In the original paper, BLEU was used to evaluate the performance of machine translation systems. Currently, BLEU is also used in the performance evaluation of MRC.

The computation method of BLEU is to take the geometric mean of the modified Precision and then multiply the result by an exponential brevity penalty factor. Currently, case folding is the only text normalization performed before computing the precision. First, we compute the geometric average of the modified $n$-gram precision, $P_n$, using $n$-grams up to length $N$ and positive weights $w_n$ summing to one [84].

Next, let $C$ be the length of the candidate sentence and $r$ be the length of the effective reference corpus. The brevity penalty $BP$ is computed as follows [bib BLEU]:

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r. \end{cases} \tag{12}$$

Then:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{13}$$

*3.9. Meteor*

Meteor was first proposed by Banerjee and Lavie [85] in order to evaluate the machine translation system. Unlike the BLEU using only Precision, the Meteor indicator uses a combination of Recall and Accuracy to evaluate the system. In addition, Meteors also include features such as synonym matching.

Besides Meteor, Denkowski and Lavie also proposed Meteor-next [86] and Meteor 1.3 [87], the new metric features include improved text normalization, higher-precision paraphrase matching, and discrimination between content and function words. Currently, some MRC datasets use Meteor as one of their evaluation metrics, such as the NarrativeQA [63] dataset. The Meteor score for the given alignment is computed as follows:

$$Meteor = F_{mean} \times (1 - Penalty) \tag{14}$$

where $F_{mean}$ is combined by the *Precision* and *Recall* via a harmonic-mean [88] that places most of the weight on *Recall*, and the formula of $F_{mean}$ is:

$$F_{mean} = \frac{Precision \times Recall}{\alpha \times Precision + (1 - \alpha) \times Recall} \tag{15}$$

And *Penalty* is a fragmentation penalty to account for differences and gaps in word order, which is calculated using the total number of matched words ($m$, average over hypothesis and reference) and number of chunks ($ch$):

$$Penalty = \gamma \times \left( \frac{ch}{m} \right)^{\beta} \tag{16}$$

where the parameters $\alpha$, $\beta$, and $\gamma$ are tuned to maximize correlation with human judgments [87]. It should be noted that the *Precision* and *Recall* in Meteor 1.3 is improved by text normalization, we can see the original paper of Denkowski and Lavie for the detailed calculation method of *Precision* and *Recall* in Meteor 1.3 [87].

*3.10. HEQ*

The HEQ stands for Human Equivalence Score, which is a new MRC evaluation metric that can be used in conversational reading comprehension datasets, such as QuAC [65]. For these datasets in which questions with multiple valid answers, the F1 may be misleading. Therefore, HEQ was introduced. The HEQ is an evaluation metric for judging whether the output of the system is as good as the output of an ordinary person. For example, suppose a MRC task contains $N$ questions, and the number of questions for which the token-level $F1$ performance of algorithm exceeds or reaches the token-level $F1$ of humans is $M$. The HEQ score is computed as follows [65]:

$$HEQ = \frac{M}{N} \tag{17}$$

*3.11. Statistics of Evaluation Metrics*

In this section, we collated the evaluation metrics of 57 MRC tasks. As seen in Table 2, the typical MRC dataset evaluation metrics are Accuracy, Exact Match, F1 score, ROUGE, BLEU, HEQ, and Meteor. Many datasets use more than one evaluation metric. Moreover, some datasets adopt detailed evaluation metrics according to their own characteristics. For example, the HotpotQA [67] dataset adopts evaluation metrics such as Exact Match of Supportings, F1 of Supportings, Exact Match of Answer, F1 of Answer, and so forth. And the Facebook CBT [28] dataset adopts Accuracy on Named Entities, Accuracy on Common Nouns, Accuracy on Verbs, Accuracy on Prepositions.

**Table 2.** Evaluation metrics of different machine reading comprehension tasks.

| Year | MRC Tasks | Metric 1 | Metric 2 | Metric 3 | Metric 4 |
|------|-----------|----------|----------|----------|----------|
| 2013 | MCTest [15] | Accuracy | N/A | N/A | N/A |
| 2015 | CNN/Daily Mail [19] | Accuracy | N/A | N/A | N/A |
| 2015 | CuratedTREC [51] | Exact Match | N/A | N/A | N/A |
| 2015 | WikiQA [52] | Question-level Precision | Question-level Recall | Question-level F1 | N/A |
| 2016 | BookTest [57] | Accuracy | N/A | N/A | N/A |
| 2016 | Facebook CBT [28] | Accuracy on Named Entities | Accuracy on Common Nouns | Accuracy on Verbs | Accuracy on Prepositions |
| 2016 | Google MC-AFP [58] | Accuracy | N/A | N/A | N/A |
| 2016 | LAMBADA [55] | Accuracy | N/A | N/A | N/A |
| 2016 | MovieQA [32] | Accuracy of Video Clips | Accuracy of Plots and Subtitles | N/A | N/A |
| 2016 | MS MARCO [39] | Rouge-L | BLEU-1 | N/A | N/A |
| 2016 | NewsQA [54] | Exact Match | Token-level F1 | N/A | N/A |
| 2016 | SQuAD 1.1 [19] | Token-level F1 | Exact Match | N/A | N/A |
| 2016 | Who-did-What [48] | Accuracy | N/A | N/A | N/A |
| 2016 | WikiMovies [53] | Accuracy | N/A | N/A | N/A |
| 2016 | WikiReading [56] | Question level F1 | N/A | N/A | N/A |
| 2017 | COMICS-Cl [33] | Accuracy of Text Cloze | Accuracy of Visual Cloze | N/A | N/A |
| 2017 | COMICS-Co [33] | Accuracy of Coherence | N/A | N/A | N/A |
| 2017 | NarrativeQA [63] | ROUGE-L | BLEU-1 | BLEU-4 | Meteor |
| 2017 | Qangaroo-M [49] | Accuracy | N/A | N/A | N/A |
| 2017 | Qangaroo-W [49] | Accuracy | N/A | N/A | N/A |
| 2017 | Quasar-S [61] | Accuracy | N/A | N/A | N/A |
| 2017 | Quasar-T [61] | Exact Match | Token-level F1 | N/A | N/A |
| 2017 | RACE [60] | Accuracy | N/A | N/A | N/A |
| 2017 | SciQ [64] | Accuracy | N/A | N/A | N/A |
| 2017 | SearchQA [62] | F1 score (for n-gram) | Accuracy | N/A | N/A |
| 2017 | TQA [31] | Accuracy of All | Accuracy of Diagram | N/A | N/A |
| 2017 | TriviaQA-Wiki [59] | Exact Match | Question-level F1 | Verified-EM | Verified-F1 |
| 2017 | TriviaQA-Web [59] | Exact Match | Document-level F1 | Verified-EM | Verified-F1 |
| 2018 | ARC-C [72] | Accuracy | N/A | N/A | N/A |
| 2018 | ARC-E [72] | Accuracy | N/A | N/A | N/A |
| 2018 | CliCR [71] | Exact Match | Token-level F1 | BLEU-2 | BLEU-4 |
| 2018 | CLOTH [69] | Accuracy | N/A | N/A | N/A |
| 2018 | CoQA [47] | Token-level F1 | F1 out of domain | F1 in domain | N/A |
| 2018 | DuoRC-P [68] | Accuracy | Token-level F1 | N/A | N/A |
| 2018 | DuoRC-S [68] | Accuracy | Token-level F1 | N/A | N/A |
| 2018 | HotpotQA-D [67] | EM of Answer | F1 of Answer (Token-level) | EM of Supportings | F1 of Supportings |
| 2018 | HotpotQA-F [67] | EM of Answer | F1 of Answer (Token-level) | EM of Supportings | F1 of Supportings |
| 2018 | MCScript [78] | Accuracy | N/A | N/A | N/A |
| 2018 | MultiRC [75] | F1m | Exact Match | N/A | N/A |
| 2018 | OpenBookQA [73] | Accuracy | N/A | N/A | N/A |
| 2018 | PaperQA(Hong et al.) [77] | F1 | N/A | N/A | N/A |
| 2018 | PaperQA-LS [76] | Accuracy | N/A | N/A | N/A |
| 2018 | PaperQA-T [76] | Accuracy | N/A | N/A | N/A |
| 2018 | ProPara [79] | Accuracy | N/A | N/A | N/A |
| 2018 | QuAC [65] | Token-level F1 | HEQ-Q | HEQ-D | N/A |
| 2018 | RecipeQA-Cl [34] | Accuracy of Textual Cloze | Accuracy of Visual Cloze | N/A | N/A |
| 2018 | RecipeQA-Co [34] | Accuracy-VO | Accuracy-VC | N/A | N/A |
| 2018 | ReCoRD [70] | Exact Match | Token-level F1 | N/A | N/A |
| 2018 | ReviewQA [25] | Accuracy | N/A | N/A | N/A |
| 2018 | SciTail [74] | Accuracy | N/A | N/A | N/A |
| 2018 | SQuAD 2.0 [66] | Token-level F1 | EM | N/A | N/A |
| 2019 | CommonSenseQA [82] | Accuracy | N/A | N/A | N/A |
| 2019 | DREAM [81] | Accuracy | N/A | N/A | N/A |
| 2019 | DROP [50] | EM | Token-level F1 | N/A | N/A |
| 2019 | Natural Questions-Long [80] | Precision | Recall | N/A | N/A |
| 2019 | Natural Questions-Short [80] | Precision | Recall | F1 | N/A |
| 2019 | ShARC [38] | Micro Accuracy | Macro Accuracy | BLEU-1 | BLEU-4 |

Table 3 shows the statistics on the usage of different evaluation metrics in the 57 MRC tasks collected in this paper. Among them, Accuracy is the most widely used evaluation metric, and 61.40% of MRC tasks collected in this paper used it. It is followed by F1 (36.84%) and Exact Match (22.81%). The rest of these evaluation metrics are less used, as shown in Table 3:

**Table 3.** Statistics on the usage of different evaluation metrics in 57 machine reading comprehension tasks.

| Metrics | Accuracy | F1 | EM | BLEU | Recall | Precision | ROUGE-L | HEQ-D | Meteor |
|---------|----------|-----|-----|------|--------|-----------|---------|-------|--------|
| Usage | 61.40% | 36.84% | 22.81% | 7.02% | 5.26% | 5.26% | 3.51% | 1.75% | 1.75% |

We also analyzed the relationship between the evaluation metrics and task types. Figure 20 shows the usage of evaluation metrics with different types of tasks. Taking the "Accuracy" in Figure 20b as an example, a total of 35 MRC tasks use the "Accuracy" as the evaluation metric. Among them, 25 tasks have the "Multi-choice" type of answers, and the remaining 10 tasks have the "Natural" type

of answers. It can be seen from Figure 20b that tasks with the "Multi-choice" type of answers prefer to use the "Accuracy" evaluation metric rather than other evaluation metrics. This is because it is impossible to calculate the EM, Precision, BLEU or F1 score of a typical "Multi-choice" question which has only one correct answer in the candidates. Among the "Multi-choice" tasks we collected, only the MultiRC [75] task does not use Accuracy, but F1 and Exact Match as the evaluation metric. That is because there are multiple correct answers in the candidates of the MultiRC task. As can be seen from Figure 20a, tasks with "Cloze" questions prefer to use the "Accuracy" as evaluation metrics rather than other evaluation metrics, which is because "Cloze" tasks tend to have "Multi-choice" answers. From Figure 20c, we can see that tasks with "Spans" answers and tasks with "Free-form" answers have no special preference in selecting evaluation metrics.
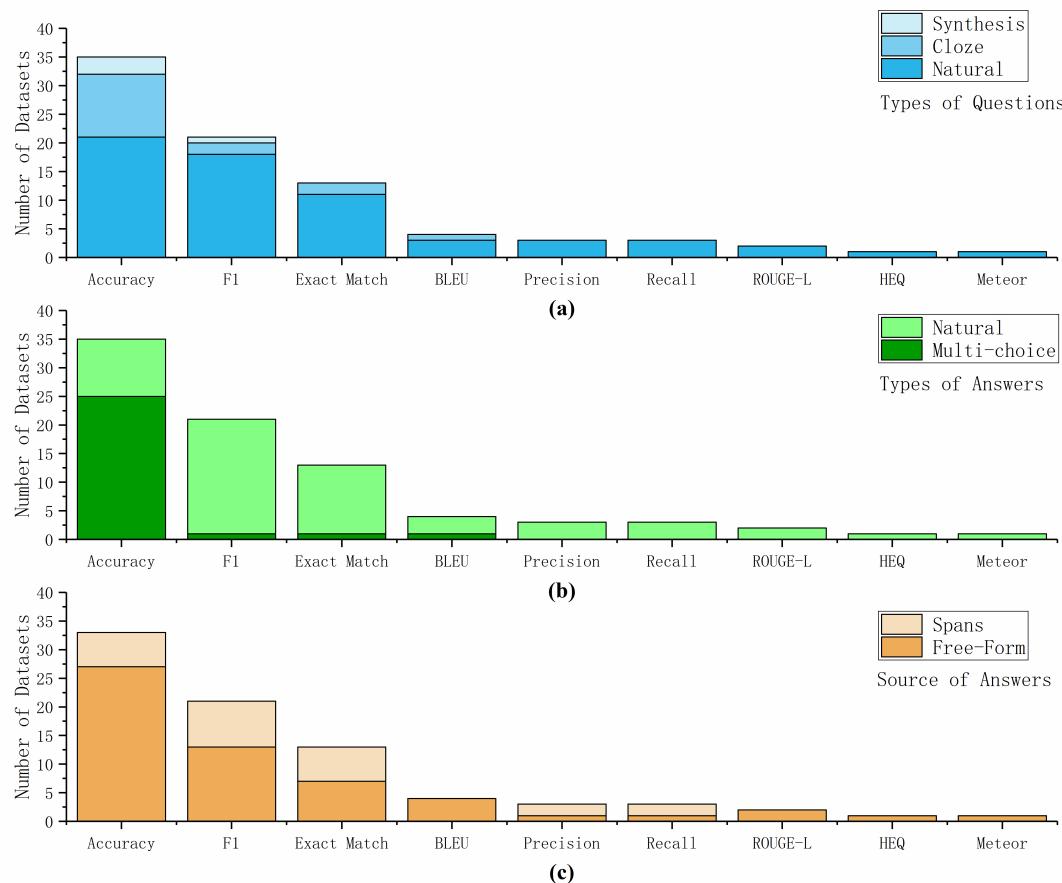


**Figure 20.** The usage of evaluation metrics with different types of MRC tasks: (**a**) The MRC tasks are classified according to the type of questions. Dark blue indicates that the question type is natural, blue represents that the question type is cloze, and light blue represents that the question type is synthetic. (**b**) The MRC tasks are classified according to the type of answers. Dark green indicates that the answer type is natural, and light green represents that the answer type is multi-choice. (**c**) The MRC tasks are classified according to the source of answers. Dark orange indicates that the answer source is free-form, and light red represents that the answer source is spans;

## 4. Benchmark Dataset

In this section, we analyze various attributes of 57 MRC benchmark datasets, including dataset size, generation method, source of corpus, context type, availability of leaderboards and baselines, prerequisite skills, and citations of related papers. We have provided the timeline figure of the MRC datasets, as seen in Figure 21.
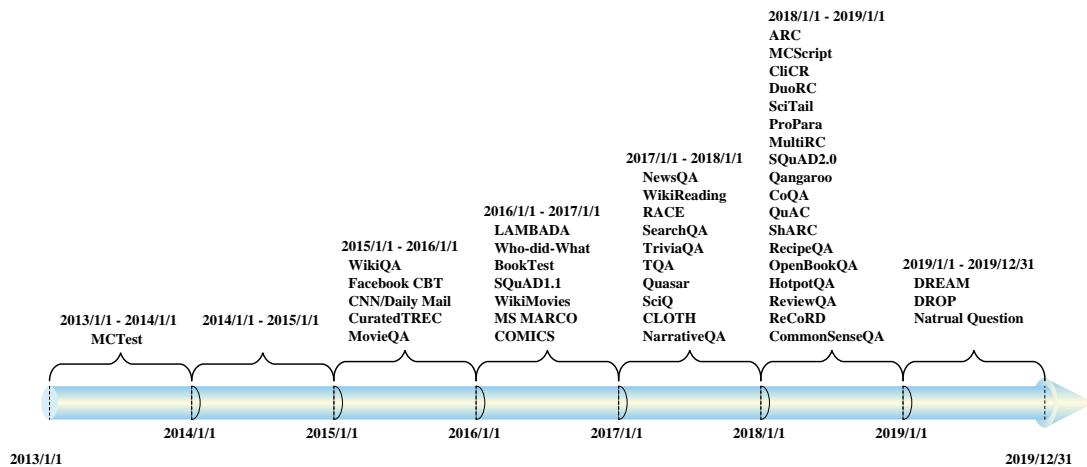


**Figure 21.** The timeline of MRC datasets discussed in this survey.
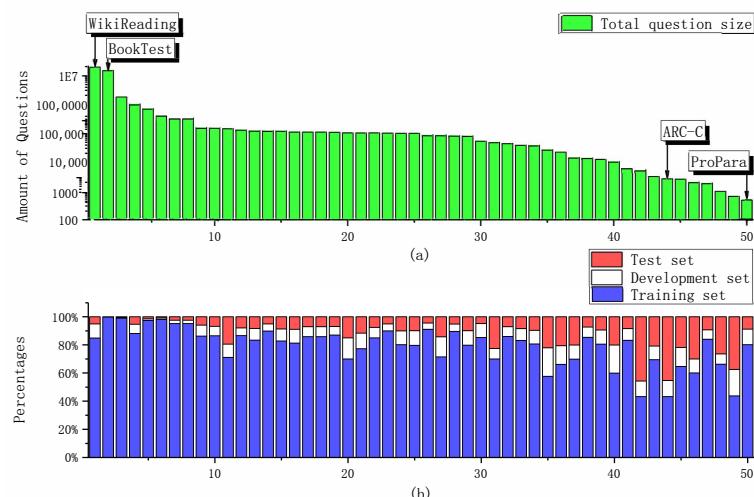
### 4.1. The Size of Datasets

The recent success of machine reading comprehension is driven largely by both large-scale datasets and neural models [14]. The size of a dataset affects the generalization ability of the MRC model and determines whether the model is useful in the real world. Early MRC datasets tend to of small sizes. With the continuous development of MRC datasets in recent years, the question set sizes of newly created MRC datasets are generally more than 10K. Here, we have counted the total number of questions in each MRC dataset along with the sizes of its training set, development set, and testing set, as well as the proportion of training set to the total number of questions. The data is shown in Table 4 which is sorted by the question set size of the datasets.

We also use the data in Table 4 to make a statistical chart where the Y coordinate is logarithmic, as shown in Figure 22, we can see that the WikiReading is the dataset with the largest question size [56] of a total of 18.87M questions; BookTest [57] is ranked second, and ProPara [79] is the smallest which has only 488 questions. When it comes to the proportion of training sets, BookTest has the highest proportion, 99.86%, while the ARC (challenge set) has the lowest proportion which is 43.20%. The development set is generally slightly smaller than the testing set.

**Table 4.** The question set size of machine reading comprehension datasets.

| Year | Datasets | Question Size | #Training Questions | #Dev Questions | #Test Questions | Percentage of Training Set |
|------|----------|---------------|---------------------|----------------|-----------------|----------------------------|
| 2016 | WikiReading [56] | 18.87M | 16.03M | 1.89M | 0.95M | 84.95% |
| 2016 | BookTest [57] | 14,160,825 | 14,140,825 | 10,000 | 10,000 | 99.86% |
| 2016 | Google MC-AFP [58] | 1,742,618 | 1,727,423 | 7602 | 7593 | 99.13% |
| 2015 | Daily Mail [19] | 997,467 | 879,450 | 64,835 | 53,182 | 88.17% |
| 2016 | Facebook CBT [28] | 687K | 669,343 | 8000 | 10,000 | 97.38% |
| 2018 | ReviewQA [25] | 587,492 | 528,665 | N/A | 58,827 | 89.99% |
| 2015 | CNN [19] | 387,420 | 380,298 | 3924 | 3198 | 98.16% |
| 2019 | Natural Questions [80] | 323,045 | 307,373 | 7830 | 7842 | 95.15% |
| 2016 | Who-did-What [48] | 147,786 | 127,786 | 10,000 | 10,000 | 86.47% |
| 2018 | SQuAD 2.0 [66] | 151,054 | 130,319 | 11,873 | 8862 | 86.27% |
| 2017 | SearchQA [62] | 140,461 | 99,820 | 13,393 | 27,248 | 71.07% |
| 2018 | CoQA [47] | 127K | 110K | 7K | 10K | 86.61% |
| 2018 | ReCoRD [70] | 120,730 | 100,730 | 10,000 | 10,000 | 83.43% |
| 2016 | NewsQA [54] | 119K | 107K | 6K | 6K | 89.92% |
| 2018 | HotpotQA [67] | 105,374 | 90,564 | 7405 | 7405 | 85.95% |
| 2018 | CliCR [71] | 104,919 | 91,344 | 6391 | 7184 | 87.06% |
| 2018 | DuoRC-P [68] | 100,316 | 70K | 15K | 15K | 70.00% |
| 2016 | SQuAD 1.1 [19] | 107,702 | 87,599 | 10,570 | 9533 | 81.33% |
| 2016 | WikiMovies [53] | 116K | 96K | 10K | 10K | 82.76% |
| 2018 | CLOTH [69] | 99,433 | 76,850 | 11,067 | 11,516 | 77.29% |
| 2018 | QuAC [65] | 98,275 | 83,568 | 7354 | 7353 | 85.03% |
| 2017 | RACE [60] | 97,687 | 87,866 | 4887 | 4934 | 89.95% |
| 2019 | DROP [50] | 96,567 | 77,409 | 9536 | 9622 | 80.16% |
| 2017 | TriviaQA-Web [59] | 95,956 | 76,496 | 9951 | 9509 | 79.72% |
| 2018 | PaperQA-T [76] | 84,803 | 77,298 | 3752 | 3753 | 91.15% |
| 2018 | DuoRC-S [68] | 84K | 60K | 12K | 12K | 70.00% |
| 2018 | PaperQA-L [76] | 80,118 | 71,804 | 4179 | 4135 | 89.62% |
| 2017 | TriviaQA-Wiki [59] | 77,582 | 61,888 | 7993 | 7701 | 79.77% |
| 2017 | Qangaroo-W [49] | 51,318 | 43,738 | 5129 | 2451 | 85.23% |
| 2017 | NarrativeQA [63] | 46,765 | 32,747 | 3461 | 10,557 | 70.02% |
| 2017 | Quasar-T [61] | 43,013 | 37,012 | 3000 | 3000 | 86.05% |
| 2017 | Quasar-S [61] | 37,362 | 31,049 | 3174 | 3139 | 83.10% |
| 2018 | RecipeQA [34] | 36K | 29,657 | 3562 | 3567 | 80.62% |
| 2017 | TQA [31] | 26,260 | 15,154 | 5309 | 5797 | 57.71% |
| 2016 | MovieQA [32] | 21,406 | 14,166 | 2844 | 4396 | 66.18% |
| 2018 | MCScript [78] | 13,939 | 9731 | 1411 | 2797 | 69.81% |
| 2017 | SciQ [64] | 13,679 | 11,679 | 1000 | 1000 | 85.38% |
| 2019 | CommonSenseQA [82] | 12,102 | 9741 | 1221 | 1140 | 80.49% |
| 2019 | DREAM [81] | 10,197 | 6116 | 2040 | 2041 | 59.98% |
| 2018 | OpenBookQA [73] | 5957 | 4957 | 500 | 500 | 83.21% |
| 2018 | ARC-Easy Set [72] | 5197 | 2251 | 570 | 2376 | 43.31% |
| 2015 | WikiQA [52] | 3047 | 2118 | 296 | 633 | 69.51% |
| 2018 | ARC-Challenge Set [72] | 2590 | 1119 | 299 | 1172 | 43.20% |
| 2017 | Qangaroo-M [49] | 2508 | 1620 | 342 | 546 | 64.59% |
| 2013 | MCTest-mc500 [15] | 2000 | 1200 | 200 | 600 | 60.00% |
| 2018 | SciTail [74] | 1834 | 1542 | 121 | 171 | 84.08% |
| 2019 | ShARC [38] | 948 | 628 | 69 | 251 | 66.24% |
| 2013 | MCTest-mc160 [15] | 640 | 280 | 120 | 240 | 43.75% |
| 2018 | ProPara [79] | 488 | 391 | 54 | 43 | 80.12% |



**Figure 22.** The size of machine Reading Comprehension datasets: (**a**) Total question size of each dataset. (**b**) Percentages of training sets, development sets and test sets.

Because different MRC datasets contain different corpora, we also give details of the corpus used in each MRC dataset, including the size of corpus and the unit of corpus, as well as the size of training set, development set, and testing set. As seen in Table 5, The units of corpus in MRC datasets are various, such as paragraphs, documents, and so forth.

**Table 5.** The corpus size of machine reading comprehension datasets.

| Year | Datasets | Corpus Size | #Train Corpus | #Dev Corpus | #Test Corpus | Unit of Corpus |
|------|----------|-------------|---------------|-------------|--------------|----------------|
| 2016 | WikiReading [56] | 4.7M | N/A | N/A | N/A | Article |
| 2016 | SQuAD 1.1 [19] | 536 | 442 | 48 | 46 | Article |
| 2018 | SQuAD 2.0 [66] | 505 | 442 | 35 | 28 | Article |
| 2016 | BookTest [57] | 14,062 | N/A | N/A | N/A | Book |
| 2017 | COMICS [33] | 3948 | N/A | N/A | N/A | Book |
| 2016 | Facebook CBT [28] | 108 | 98 | 5 | 5 | Book |
| 2019 | DREAM [81] | 6444 | 3869 | 1288 | 1287 | Dialogue |
| 2016 | NewsQA [54] | 1,010,916 | 909,824 | 50,546 | 50,546 | Document |
| 2017 | TriviaQA-Web [59] | 662,659 | 528,979 | 68,621 | 65,059 | Document |
| 2015 | Daily Mail [19] | 219,506 | 196,961 | 12,148 | 10,397 | Document |
| 2017 | TriviaQA-Wiki [59] | 138,538 | 110,648 | 14,229 | 13,661 | Document |
| 2018 | ReviewQA [25] | 100,000 | 90,000 | N/A | 10,000 | Document |
| 2015 | CNN [19] | 92,579 | 90,266 | 1220 | 1093 | Document |
| 2017 | NarrativeQA [63] | 1572 | 1102 | 115 | 355 | Document |
| 2017 | TQA [31] | 1076 | 666 | 200 | 210 | Lesson |
| 2016 | MovieQA [32] | 548 | 362 | 77 | 109 | Movie |
| 2016 | Google MC-AFP [58] | 1,742,618 | 1,727,423 | 7602 | 7593 | Passage |
| 2016 | Who-did-What [48] | 147,786 | 127,786 | 10,000 | 10,000 | Passage |
| 2017 | SearchQA [62] | 140,461 | 99,820 | 13,393 | 27,248 | Passage |
| 2018 | ReCoRD [70] | 80,121 | 65,709 | 7133 | 7279 | Passage |
| 2017 | Quasar-T [61] | 43,012 | 37,012 | 3000 | 3000 | Passage |
| 2017 | Quasar-S [61] | 37,362 | 31,049 | 3174 | 3139 | Passage |
| 2017 | RACE [60] | 27,933 | 25,137 | 1389 | 1407 | Passage |
| 2018 | SciTail [74] | 27,026 | 23,596 | 1304 | 2126 | Passage |
| 2016 | LAMBADA [55] | 12,684 | 2662 | 4869 | 5153 | Passage |
| 2018 | CoQA [47] | 8399 | 7199 | 500 | 700 | Passage |
| 2018 | CLOTH [69] | 7131 | 5513 | 805 | 813 | Passage |
| 2017 | Qangaroo-W [49] | 51,318 | 43,738 | 5129 | 2451 | Passage |
| 2019 | DROP [50] | 6735 | 5565 | 582 | 588 | Passage |
| 2017 | Qangaroo-M [49] | 2508 | 1620 | 342 | 546 | Passage |
| 2018 | RecipeQA [34] | 19,779 | 15,847 | 1963 | 1969 | Recipe |
| 2015 | WikiQA [52] | 29,258 | 20,360 | 2733 | 6165 | Sentence |
| 2013 | MCTest-mc500 [15] | 500 | 300 | 50 | 150 | Story |
| 2013 | MCTest-mc160 [15] | 160 | 70 | 30 | 60 | Story |
| 2018 | QuAC [65] | 8845 | 6843 | 1000 | 1002 | Unique section |
| 2019 | ShARC [38] | 32,436 | 21,890 | 2270 | 8276 | Utterance |
| 2019 | Natural Questions [80] | 323,045 | 307,373 | 7830 | 7842 | Wikipedia Page |

## 4.2. The Generation Method of Datasets

The generation method of datasets can be roughly described into several categories: Crowdsourcing, Expert, and Automated. "Crowdsourcing" is evolving as a distributed problem-solving and business production model in recent years [89]. An example of crowdsourcing website is Amazon Mechanical Turk. Today, many MRC datasets are posed by the distributed workforce on such crowdsourcing websites. The "Expert" generation method means that question and answer pairs in the dataset are generated by people with professional knowledge in some fields. For example, in the ARC dataset [72], there are 7787 science questions covered by US elementary and middle schools. The "Automated" generation method means that question and answer pairs are automatically generated based on corpus, such as many cloze datasets.

### 4.3. The Source of Corpus

The source of corpus affects the readability and complexity of machine reading comprehension datasets. According to the source of corpus, the MRC datasets can be described as the following types: Exam Text, Wikipedia, News articles, Abstract of Scientific Paper, Crafted story, Technical documents, Text Book, Movie plots, Recipe, Government Websites, Search engine query logs, Hotel Comments, Narrative text, and so forth.

### 4.4. The Type of Context

The type of context can affect the training method of machine reading comprehension model, which produces many special models, such as the multi-hop reading comprehension, and multi-document reading comprehension. There are many types of context in MRC datasets, including Paragraph, Multi-paragraph, Document, Multi-document, URL, Paragraphs with diagrams or images. As shown in Table 6, we give details of the generation method, corpus source, and context type of each machine's reading comprehension dataset.

**Table 6.** The generation method of datasets, source of corpus and type of context.

| Year | Datasets | Generation Method | Source of Corpus | Type of Context |
|------|----------|-------------------|------------------|-----------------|
| 2013 | MCTest-mc160 [15] | Crowd-sourcing | Factoid stories | Paragraph |
| 2013 | MCTest-mc500 [15] | Crowd-sourcing | Factoid stories | Paragraph |
| 2015 | CNN [19] | Automated | News | Document |
| 2015 | CuratedTREC [51] | Crowd-sourcing | Factoid stories | Paragraph |
| 2015 | Daily Mail [19] | Automated | News | Document |
| 2015 | WikiQA [52] | Crowd-sourcing | Wikipedia | Paragraph |
| 2016 | BookTest [57] | Automated | Factoid stories | Paragraph |
| 2016 | Facebook CBT [28] | Automated | Factoid stories | Paragraph |
| 2016 | Google MC-AFP | Automated | The Gigaword corpus | Paragraph |
| 2016 | LAMBADA [55] | Crowd-sourcing | Book Corpus | Paragraph |
| 2016 | MovieQA [32] | Crowd-sourcing | Movie | Paragraph with Images and Videos |
| 2016 | MS MARCO | Automated | The Bing | Paragraph |
| 2016 | NewsQA [54] | Crowd-sourcing | News | Document |
| 2016 | SQuAD 1.1 [19] | Crowd-sourcing | Wikipedia | Paragraph |
| 2016 | Who-did-What [48] | Automated | News | Document |
| 2016 | WikiMovies [53] | Automated | Movie | Document |
| 2016 | WikiReading [56] | Automated | Wikipedia | Document |
| 2017 | COMICS [33] | Automated | Comics | Paragraph with Images |
| 2017 | NarrativeQA [63] | Crowd-sourcing | Movie | Document |
| 2017 | Qangaroo-M [49] | Crowd-sourcing | Wikipedia | Paragraph |
| 2017 | Qangaroo-W [49] | Crowd-sourcing | Scientic paper | Paragraph |
| 2017 | Quasar-S [61] | Crowd-sourcing | Stack Overflow | Paragraph |
| 2017 | Quasar-T [61] | Crowd-sourcing | Stack Overflow | Paragraph |
| 2017 | RACE [60] | Expert | English Exam | Document |
| 2017 | SciQ [64] | Crowd-sourcing | School science curricula | Paragraph |
| 2017 | SearchQA [62] | Crowd-sourcing | J! Archive and Google | Paragraph & URL |
| 2017 | TQA [31] | Expert | School science curricula | Paragraph with Images |
| 2017 | TriviaQA-Wiki [59] | Automated | The Bing | Paragraph |
| 2017 | TriviaQA-Web [59] | Automated | The Bing | Paragraph |
| 2018 | ARC-Challenge Set [72] | Expert | School science curricula | Paragraph |
| 2018 | ARC-Easy Set [72] | Expert | School science curricula | Paragraph |
| 2018 | CliCR [71] | Automated | BMJ Case Reports | Paragraph |
| 2018 | CLOTH [69] | Expert | English Exam | Document |
| 2018 | CoQA [47] | Crowd-sourcing | Jeopardy | Paragraph |
| 2018 | DuoRC-Paraphrase [68] | Crowd-sourcing | Movie | Paragraph |
| 2018 | DuoRC-Self [68] | Crowd-sourcing | Movie | Paragraph |
| 2018 | HotpotQA-D [67] | Crowd-sourcing | Wikipedia | Multi-paragraph |
| 2018 | HotpotQA-F [67] | Crowd-sourcing | Wikipedia | Multi-paragraph |
| 2018 | MCScript [78] | Crowd-sourcing | Narrative texts | Paragraph |
| 2018 | MultiRC [75] | Crowd-sourcing | News and other web pages | Multi-sentence |
| 2018 | OpenBookQA [73] | Crowd-sourcing | School science curricula | Paragraph |
| 2018 | PaperQA(Hong et al.) [77] | Crowd-sourcing | Scientic paper | Paragraph |
| 2018 | PaperQA-L [76] | Automated | Scientic paper | Paragraph |
| 2018 | PaperQA-T [76] | Automated | Scientic paper | Paragraph |
| 2018 | ProPara [79] | Crowd-sourcing | Process Paragraph | Paragraph |
| 2018 | QuAC [65] | Crowd-sourcing | Wikipedia | Document |
| 2018 | RecipeQA [34] | Automated | Recipes | Paragraph with Images |

**Table 6.** *Cont.*

| Year | Datasets | Generation Method | Source of Corpus | Type of Context |
|---|---|---|---|---|
| 2018 | ReCoRD [70] | Crowd-sourcing | News | Paragraph |
| 2018 | ReviewQA [25] | Crowd-sourcing | Hotel Comments | Paragraph |
| 2018 | SciTail [74] | Crowd-sourcing | School science curricula | Paragraph |
| 2018 | SQuAD 2.0 [66] | Crowd-sourcing | Wikipedia | Paragraph |
| 2019 | CommonSenseQA [82] | Crowd-sourcing | Narrative texts | Paragraph |
| 2019 | DREAM [81] | Crowd-sourcing | English Exam | Dialogues |
| 2019 | DROP [50] | Crowd-sourcing | Wikipedia | Paragraph |
| 2019 | Natural Questions-L [80] | Crowd-sourcing | Wikipedia | Paragraph |
| 2019 | Natural Questions-S [80] | Crowd-sourcing | Wikipedia | Paragraph |
| 2019 | ShARC [38] | Crowd-sourcing | Government Websites | Paragraph |

## 4.5. The Availability of Datasets, Leaderboards and Baselines

The release of MRC baseline projects and leaderboards can help the researchers evaluate the performance of their models. In this section, we try to find all the MRC dataset download links, leaderboards, and baseline projects. As shown in Table 7, all the download links of MRC datasets are available except PaperQA [76]. Most of the datasets provide leaderboards and baseline projects except only 19.3% of the datasets. We have published all the download links, leaderboards, and the baseline projects on our website https://mrc-datasets.github.io/.

**Table 7.** The availability of datasets, leaderboards and baselines.

| Year | Datasets | Dataset Availability | Leaderboard Availability | Baseline Availability |
|---|---|---|---|---|
| 2019 | CommonSenseQA [82] | √ | √ | √ |
| 2018 | MCScript [78] | √ | √ | × |
| 2018 | OpenBookQA [73] | √ | √ | × |
| 2018 | ReCoRD [70] | √ | √ | × |
| 2018 | ARC-Challenge Set [72] | √ | √ | √ |
| 2018 | ARC-Easy Set [72] | √ | √ | √ |
| 2018 | CLOTH [69] | √ | √ | √ |
| 2016 | Facebook CBT [28] | √ | × | √ |
| 2016 | NewsQA [54] | √ | × | × |
| 2018 | ProPara [79] | √ | √ | × |
| 2017 | RACE [60] | √ | √ | √ |
| 2016 | SQuAD 1.1 [19] | √ | √ | √ |
| 2017 | TriviaQA-Wiki [59] | √ | √ | √ |
| 2017 | TriviaQA-Web [59] | √ | √ | √ |
| 2019 | DROP [50] | √ | √ | √ |
| 2017 | NarrativeQA [63] | √ | × | √ |
| 2019 | ShARC [38] | √ | √ | × |
| 2018 | CoQA [47] | √ | √ | √ |
| 2019 | DREAM [81] | √ | √ | √ |
| 2018 | QuAC [65] | √ | √ | √ |
| 2013 | MCTest-mc160 [15] | √ | √ | √ |
| 2013 | MCTest-mc500 [15] | √ | √ | √ |
| 2015 | WikiQA [52] | √ | × | × |
| 2018 | CliCR [71] | √ | × | √ |
| 2018 | PaperQA(Hong et al.)[77] | √ | × | × |
| 2018 | PaperQA-L [76] | × | × | × |
| 2018 | PaperQA-T [76] | × | × | × |
| 2018 | ReviewQA [25] | √ | × | × |
| 2017 | SciQ [64] | √ | × | × |
| 2016 | WikiMovies [53] | √ | × | √ |
| 2016 | BookTest [57] | √ | × | × |
| 2015 | CNN [19] | √ | × | √ |

**Table 7.** *Cont.*

| Year | Datasets | Dataset Availability | Leaderboard Availability | Baseline Availability |
|------|----------|:--------------------:|:------------------------:|:---------------------:|
| 2015 | Daily Mail [19] | ✓ | ✗ | ✓ |
| 2016 | Who-did-What [48] | ✓ | ✓ | ✓ |
| 2016 | WikiReading [56] | ✓ | ✗ | ✓ |
| 2016 | Google MC-AFP [58] | ✓ | ✗ | ✗ |
| 2016 | LAMBADA [55] | ✓ | ✗ | ✓ |
| 2018 | SciTail [74] | ✓ | ✓ | ✗ |
| 2018 | DuoRC-Paraphrase [68] | ✓ | ✓ | ✓ |
| 2018 | DuoRC-Self [68] | ✓ | ✓ | ✓ |
| 2015 | CuratedTREC [51] | ✓ | ✓ | ✓ |
| 2017 | Quasar-S [61] | ✓ | ✗ | ✓ |
| 2017 | Quasar-T [61] | ✓ | ✗ | ✓ |
| 2017 | SearchQA [62] | ✓ | ✗ | ✗ |
| 2019 | Natural Questions-L [80] | ✓ | ✓ | ✓ |
| 2019 | Natural Questions-S [80] | ✓ | ✓ | ✓ |
| 2018 | SQuAD 2.0 [66] | ✓ | ✓ | ✓ |
| 2016 | MS MARCO [39] | ✓ | ✓ | ✓ |
| 2017 | Qangaroo-MEDHOP [49] | ✓ | ✓ | ✗ |
| 2017 | Qangaroo-WIKIHOP [49] | ✓ | ✓ | ✗ |
| 2018 | MultiRC [75] | ✓ | ✓ | ✓ |
| 2018 | HotpotQA-Distractor [67] | ✓ | ✓ | ✓ |
| 2018 | HotpotQA-Fullwiki [67] | ✓ | ✓ | ✓ |
| 2017 | COMICS [33] | ✓ | ✗ | ✓ |
| 2016 | MovieQA [32] | ✓ | ✓ | ✓ |
| 2018 | RecipeQA [34] | ✓ | ✓ | ✗ |
| 2017 | TQA [31] | ✓ | ✓ | ✗ |

## 4.6. Statistical Analysis

Figure 23 demonstrates the statistical analysis of the attributes of datasets in Table 6. As seen in Figure 23a, the most common way to generate datasets is "Crowdsourcing", by which we can generate question and answer pairs that need complex reasoning abilities. The second is the "Automated" method which can help us quickly create large-scale MRC datasets. The "Expert" generate method is the least used because it is usually expensive. When it comes to context type, as seen in Figure 23b, the main context type is the "Paragraph" type, followed by "Document" type, "Paragraph with images", "Multi-Paragraph" and so on. Figure 23c shows the source of corpus which is very diverse. Among them, "Wikipedia" is the most common context source, but only accounts for 19.30%. Figure 23d illustrated the availability of leaderboard and baseline. As can be seen in Figure 23d, 45.61% of the datasets provide both leaderboards and baseline project, only 19.3% of the datasets neither provide leaderboards nor baseline projects. For the availability of dataset, all the download links of MRC datasets are available except PaperQA [76].
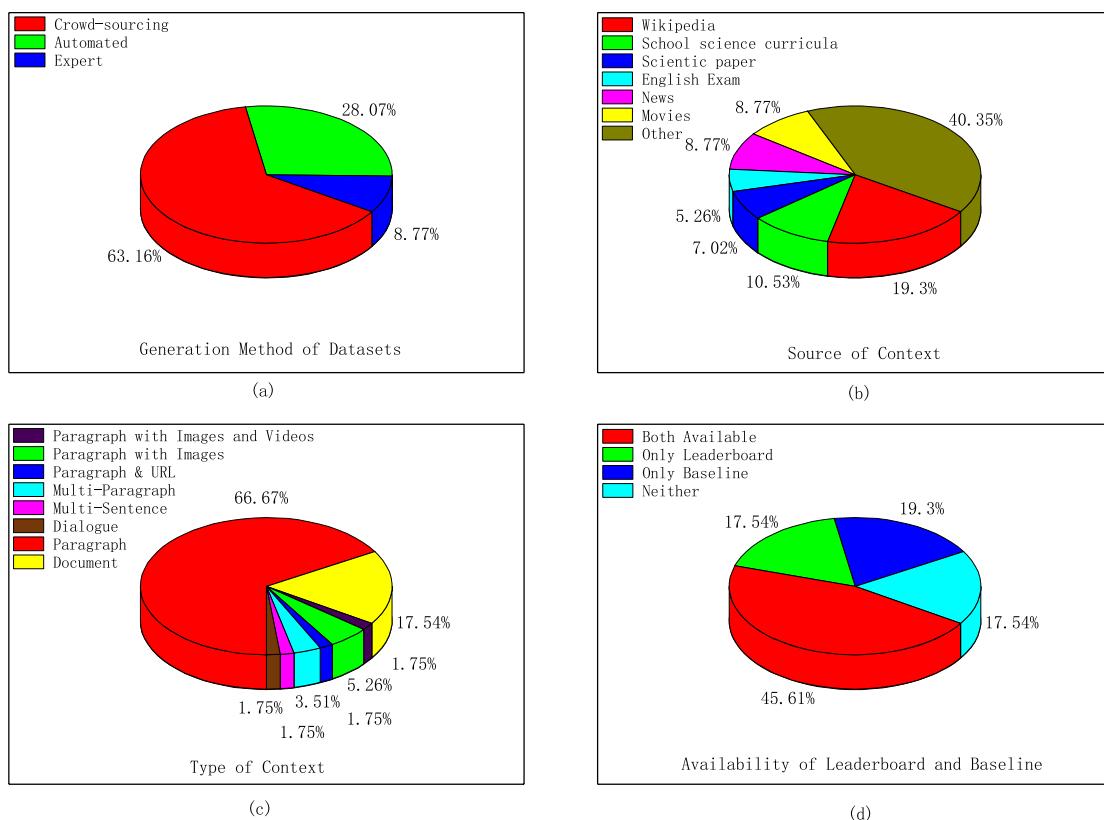
**Figure 23.** Statistical analysis of the datasets: (**a**) The generation method of datasets. (**b**) The source of corpus. (**c**) The type of context. (**d**) The availability of leaderboards and baselines.

## 4.7. Prerequisite Skills

When humans read passages and answer questions, we need to master various prerequisite skills to answer them correctly. The analysis of these prerequisite skills may help us understand the intrinsic properties of the MRC datasets. In Table 8, we quote the descriptions and examples of prerequisite skills proposed by Sugawara et al. [90]. They defined 10 kinds of prerequisite skills, including List/Enumeration, Mathematical operations, Coreference resolution, Logical reasoning, and so forth. By manually annotate questions in the MCTest [15] and SQuAD 1.1 [19], they got the frequencies of each prerequisite skill in the two MRC datasets. As seen in Table 8. However, the definition and classification of these prerequisite skills are often subjective and changeable. Many definitions have been drawn [5,90,91] , but they are still hard to give a standard mathematical definition of them, which is the same as natural language understanding.

**Table 8.** Prerequisite skills with descriptions or examples [90], and their frequencies (in percentage) in SQuAD 1.1 [19] and MCTest [15] (MC160 development set).

| Prerequisite Skills | Descriptions or Examples | Frequency SQuAD | Frequency MCTest |
|---|---|---|---|
| List/Enumeration | Tracking, retaining, and list/enumeration of entities or states | 5.00% | 11.70% |
| Mathematical operations | Four basic operations and geometric comprehension | 0.00% | 4.20% |
| Coreference resolution | Detection and resolution of coreferences | 6.20% | 57.50% |
| Logical reasoning | Induction, deduction, conditional statement, and quantifier | 1.20% | 0.00% |
| Analogy | Trope in figures of speech, e.g., metaphor | 0.00% | 0.00% |
| Spatiotemporal relations | Spatial and/or temporal relations of events | 2.50% | 28.30% |
| Causal relations | Why, because, the reason, etc. | 6.20% | 18.30% |
| Commonsense reasoning | Taxonomic/qualitative knowledge, action and event change | 86.20% | 49.20% |
| Complex sentences | Coordination or subordination of clauses | 20.00% | 15.80% |
| Special sentence structure | Scheme in figures of speech, constructions, and punctuation marks | 25.00% | 10.00% |

*4.8. Citation Analysis*

The number of citations of the paper in which a dataset was proposed reveals the dataset's impact to some extent. As shown in Table 9, we analyze how many times each paper was cited and make a statistical table. We count both the total number of citations and the monthly average citations since they were published. Except for the two PaperQA datasets [76,77], the number of citations of all other papers have been found in Google Scholar. Besides, we make a Table 9 in which the datasets are sorted by the monthly average citations. As expected, the dataset with the highest monthly average citations is SQuAD 1.1 [19], followed by CNN/Daily Mail [19] and SQuAD 2.0 [66]. It shows that these datasets are widely used as a benchmark.

**Table 9.** Citation analysis of the paper in which each dataset was proposed.

| Year | Datasets | Average Monthly Citations | Total Citations | Months after Publication | Date of Publication | Date of Statistics |
|------|----------|---------------------------|-----------------|--------------------------|---------------------|---------------------|
| 2016 | SQuAD 1.1 [19] | 33.35 | 1234 | 37 | 2016-10-10 | 2019-12-01 |
| 2015 | CNN/Daily Mail [19] | 25.21 | 1210 | 48 | 2015-11-19 | 2019-12-01 |
| 2018 | SQuAD 2.0 [66] | 14.65 | 249 | 17 | 2018-06-11 | 2019-12-01 |
| 2019 | Natural Questions [80] | 9.00 | 45 | 5 | 2019-07-01 | 2019-12-01 |
| 2017 | TriviaQA [59] | 7.97 | 239 | 30 | 2017-05-13 | 2019-12-01 |
| 2018 | CoQA [47] | 7.93 | 119 | 15 | 2018-08-21 | 2019-12-01 |
| 2016 | WikiMovies [53] | 7.73 | 286 | 37 | 2016-10-10 | 2019-12-01 |
| 2016 | CBT [28] | 6.92 | 332 | 48 | 2015-11-07 | 2019-12-01 |
| 2016 | MS MARCO [39] | 6.65 | 246 | 37 | 2016-10-31 | 2019-12-01 |
| 2015 | WikiQA [52] | 6.43 | 328 | 51 | 2015-09-01 | 2019-12-01 |
| 2018 | HotpotQA [67] | 5.71 | 80 | 14 | 2018-09-25 | 2019-12-01 |
| 2016 | NewsQA [54] | 5.21 | 172 | 33 | 2017-02-07 | 2019-12-01 |
| 2016 | MovieQA [32] | 5.00 | 235 | 47 | 2015-12-09 | 2019-12-01 |
| 2017 | RACE [60] | 4.87 | 151 | 31 | 2017-04-15 | 2019-12-01 |
| 2018 | QuAC [65] | 4.73 | 71 | 15 | 2018-08-27 | 2019-12-01 |
| 2013 | MCTest [15] | 4.69 | 347 | 74 | 2013-10-01 | 2019-12-01 |
| 2017 | Qangaroo [49] | 4.59 | 78 | 17 | 2018-06-11 | 2019-12-01 |
| 2018 | SciTail [74] | 4.16 | 79 | 19 | 2018-04-27 | 2019-12-01 |
| 2017 | NarrativeQA [63] | 3.74 | 86 | 23 | 2017-12-19 | 2019-12-01 |
| 2019 | DROP [50] | 3.00 | 27 | 9 | 2019-03-01 | 2019-12-01 |
| 2018 | ARC | 2.90 | 58 | 20 | 2018-03-14 | 2019-12-01 |
| 2017 | SearchQA [62] | 2.81 | 87 | 31 | 2017-04-18 | 2019-12-01 |
| 2018 | OpenBookQA [73] | 2.64 | 37 | 14 | 2018-09-08 | 2019-12-01 |
| 2016 | WikiReading [56] | 2.41 | 77 | 32 | 2017-03-15 | 2019-12-01 |
| 2019 | CommonSenseQA [82] | 2.33 | 28 | 12 | 2018-11-02 | 2019-12-01 |
| 2017 | Quasar [61] | 1.82 | 51 | 28 | 2017-07-12 | 2019-12-01 |
| 2016 | Who-did-What [48] | 1.69 | 66 | 39 | 2016-08-18 | 2019-12-01 |
| 2018 | MultiRC [75] | 1.67 | 30 | 18 | 2018-06-01 | 2019-12-01 |
| 2017 | TQA [31] | 1.55 | 45 | 29 | 2017-07-01 | 2019-12-01 |
| 2019 | DREAM [81] | 1.50 | 15 | 10 | 2019-01-31 | 2019-12-01 |
| 2018 | ReCoRD [70] | 1.39 | 18 | 13 | 2018-10-30 | 2019-12-01 |
| 2016 | LAMBADA [55] | 1.29 | 53 | 41 | 2016-6-20 | 2019-12-01 |
| 2019 | ShARC [38] | 1.27 | 19 | 15 | 2018-08-28 | 2019-12-01 |
| 2018 | MCScript [78] | 1.10 | 22 | 20 | 2018-03-14 | 2019-12-01 |
| 2015 | CuratedTREC [51] | 0.98 | 47 | 48 | 2015-11-20 | 2019-12-01 |
| 2018 | RecipeQA [34] | 0.93 | 13 | 14 | 2018-09-04 | 2019-12-01 |
| 2017 | COMICS [33] | 0.86 | 31 | 36 | 2016-11-16 | 2019-12-01 |
| 2018 | ProPara [79] | 0.83 | 15 | 18 | 2018-05-17 | 2019-12-01 |
| 2017 | SciQ [64] | 0.79 | 22 | 28 | 2017-07-19 | 2019-12-01 |
| 2016 | BookTest [57] | 0.73 | 27 | 37 | 2016-10-04 | 2019-12-01 |
| 2018 | DuoRC [68] | 0.63 | 12 | 19 | 2018-04-21 | 2019-12-01 |
| 2018 | CliCR [71] | 0.55 | 11 | 20 | 2018-03-26 | 2019-12-01 |
| 2018 | CLOTH [69] | 0.42 | 10 | 24 | 2017-11-09 | 2019-12-01 |
| 2018 | ReviewQA [25] | 0.08 | 1 | 13 | 2018-10-29 | 2019-12-01 |

We also analyze the monthly average citations. As seen in Figure 24, on the whole, there is a correlation between the monthly average citations and the total citations of the MRC dataset. For example, the top two citations of the total citations and the monthly average citations are the same which are SQuAD 1.1 [19] and CNN/Daily Mail [19]. However, some papers with lower total citations have higher monthly citations. This shows that these papers have been published for a short time, but they have received a lot of attention from the community, such as SQuAD 2.0 [66]. In addition, some papers with higher total citations have relatively low monthly average citations. Because these datasets have been published for a long time, but are rarely used in recent years.
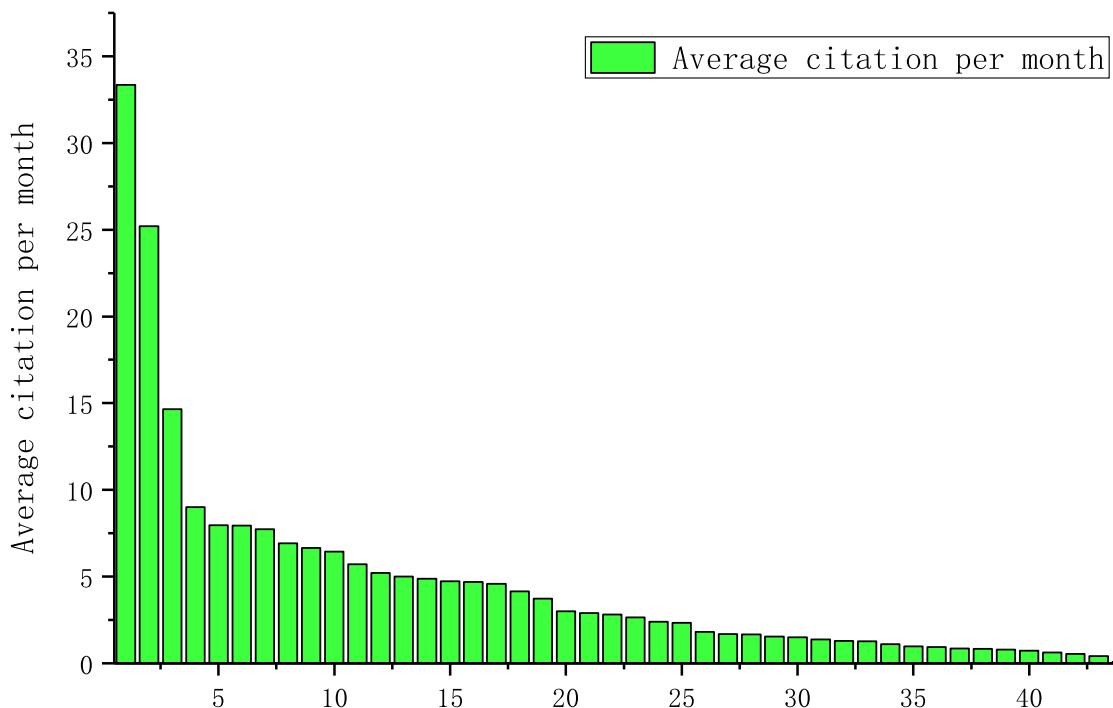
**Figure 24.** The average number of citations per month of the papers presenting the MRC datasets.

## 4.9. Characteristics of Datasets

### 4.9.1. Overview

In recent years, various large-scale MRC datasets have been created. The growth of large-scale datasets greatly promoted the research process of the machine reading comprehension.

In this section, we analyze the characteristics of existing MRC datasets, including MRC with unanswerable questions, multi-hop MRC, MRC with paraphrased paragraph, MRC which require commonsense (world knowledge), complex reasoning MRC, large-scale dataset, domain-specific dataset, multi-modal MRC, MRC dataset for open-domain QA, and conversational MRC.

It should be noted that many MRC datasets have multiple characteristics. A typical example is the DuoRC [68] dataset, which has the following four characteristics: 1. DuoRC contains two versions of context, and the meanings of different versions of context are the same while the authors try to avoid words overlap between the two versions, so the DuoRC is a MRC dataset with paraphrased paragraphs. 2. DuoRC requires the use of commonsense and world knowledge. 3. It requires complex reasoning across multiple sentences to infer the answer. 4. There are unanswerable questions in DuoRC [68].

Finally, we summarize the characteristics of each dataset in Table 10. In the following sections, we will describe each of them separately.

**Table 10.** The characteristics of each MRC dataset.

| Year | Datasets | Characteristics |
|---|---|---|
| 2015 | WikiQA [52] | With Unanswerable Questions |
| 2018 | SQuAD 2.0 [66] | With Unanswerable Questions |
| 2019 | Natural Question [80] | With Unanswerable Questions |
| 2016 | MS MARCO [39] | With Unanswerable Questions; Multi-hop MRC |
| 2018 | DuoRC [68] | With Paraphrased Paragraph; Require Commonsense (World knowledge); Complex Reasoning; With Unanswerable Questions |
| 2016 | Who-did-What [48] | With Paraphrased Paragraph; Complex Reasoning |
| 2018 | ARC [72] | Require Commonsense (World knowledge); Complex Reasoning |
| 2018 | MCScript [78] | Require Commonsense (World knowledge) |
| 2018 | OpenBookQA [73] | Require Commonsense (World knowledge) |
| 2018 | ReCoRD [70] | Require Commonsense (World knowledge) |
| 2019 | CommonSenseQA [82] | Require Commonsense (World knowledge) |
| 2016 | WikiReading [56] | Require Commonsense (External knowledge); Large Scale Dataset |
| 2016 | WikiMovies [53] | Require Commonsense (External knowledge); Domain-specific |
| 2016 | MovieQA [32] | Multi-Modal MRC |
| 2017 | COMICS [33] | Multi-Modal MRC |
| 2017 | TQA [31] | Multi-Modal MRC |
| 2018 | RecipeQA [34] | Multi-Modal MRC |
| 2018 | HotpotQA [67] | Multi-hop MRC; Complex Reasoning |
| 2017 | NarrativeQA [63] | Multi-hop MRC; Complex Reasoning |
| 2017 | Qangaroo [49] | Multi-hop MRC |
| 2018 | MultiRC [75] | Multi-hop MRC |
| 2015 | CNN/Daily Mail [19] | Large-scale Dataset |
| 2016 | BookTest [57] | Large-scale Dataset |
| 2013 | MCTest [15] | For Open-domain QA |
| 2015 | CuratedTREC [51] | For Open-domain QA |
| 2017 | Quasar [61] | For Open-domain QA |
| 2017 | SearchQA [62] | For Open-domain QA |
| 2017 | SciQ [64] | Domain-specific |
| 2018 | CliCR [71] | Domain-specific |
| 2018 | PaperQA(Hong et al.) [77] | Domain-specific |
| 2018 | PaperQA(Park et al.) [76] | Domain-specific |
| 2018 | ReviewQA [25] | Domain-specific |
| 2018 | SciTail [74] | Domain-specific |
| 2019 | DROP [50] | Complex Reasoning |
| 2016 | Facebook CBT [28] | Complex Reasoning |
| 2016 | Google MC-AFP | Complex Reasoning |
| 2016 | LAMBADA [55] | Complex Reasoning |
| 2016 | NewsQA [54] | Complex Reasoning |
| 2016 | SQuAD 1.1 [19] | Complex Reasoning |
| 2017 | RACE [60] | Complex Reasoning |
| 2017 | TriviaQA [59] | Complex Reasoning |
| 2018 | CLOTH [69] | Complex Reasoning |
| 2018 | ProPara [79] | Complex Reasoning |
| 2019 | DREAM [81] | Conversational MRC; Require Commonsense (World knowledge) |
| 2018 | CoQA [47] | Conversational MRC; With Unanswerable Questions |
| 2018 | QuAC [65] | Conversational MRC; With Unanswerable Questions |
| 2019 | ShARC [38] | Conversational MRC |

### 4.9.2. MRC with Unanswerable Questions

The existing MRC datasets often lack training sets for unanswerable questions, which weaken the robustness of the MRC systems. As a result, when the MRC models answer unanswerable questions, the models always try to give a most likely answer, rather than refuse to answer these unanswered questions. In this way, no matter how the model answers, the answers must be wrong.

To solve this problem, the researchers proposed many MRC datasets with unanswerable questions which were more challenging. Among the datasets collected by us, the datasets that contain unanswerable questions include: SQuAD 2.0, MS MARCO [39], Natural Questions [80] and NewsQA [54]. We will give a detailed description of these datasets in section in Section 4.10.

### 4.9.3. Multi-Hop Reading Comprehension

In most MRC dataset, the answer to a question usually can be found in a single paragraph or a document. However, in real human reading comprehension, when reading a novel, we are very likely to extract answers from multiple paragraphs. Compared with single passage MRC, the multi-hop machine reading comprehension is more challenging and requires multi-hop searching and reasoning over confusing passages or documents.

In different papers, multi-hop MRC is named in different ways such as multi-document machine reading comprehension [92], multi-paragraph machine reading comprehension [93], multi-sentence machine reading comprehension [75]. Compared with single paragraph MRC, multi-hop MRC is more challenging and is naturally suitable for unstructured information processing. Among the datasets collected by us, the datasets that contain unanswerable questions including SQuAD 2.0 [66], MS MARCO [39], Natural Questions [80], and NewsQA [54].

### 4.9.4. Multi-Modal Reading Comprehension

When humans read, they often do it in a multi-modal way. For example, in order to understand the information and answer the questions, sometimes, we need to read both the texts and illustrations, and we also need to use our brains to imagine, reconstruct, reason, calculate, analyze or compare. Currently, most of the existing machine reading comprehension datasets belong to plain textual machine reading comprehension, which has some limitations. some complex or precise concepts can not be described or communicated only via text. For example, if we need the computer to answer some precise questions related to aircraft engine maintenance, we may have to input the image of the aircraft engine.

Multi-modal machine reading comprehension is a dynamic interdisciplinary field that has great application potential. Considering the heterogeneity of data, multi-modal machine reading comprehension brings unique challenges to NLP researchers, because the model has to understand both texts and images. In recent years, due to the availability of large-scale internet data, many multi-modal MRC datasets have been created, such as TQA [31], RecipeQA [34], COMICS [33], and MovieQA [32].

### 4.9.5. Reading Comprehension Require Commonsense or World Knowledge

Human language is complex. When answering questions, we often need to draw upon our commonsense or world knowledge. Moreover, in the process of human language, many conventional puns and polysemous words have been formed. The use of the same words in different scenes also requires the computer to have a good command of the relevant commonsense or world knowledge.

Conventional MRC tasks usually focus on answering questions about given passages. In the existing machine reading comprehension datasets, only a small proportion of questions need to be answered with commonsense knowledge. In order to build MRC models with commonsense or world knowledge, many Commonsense Reading Comprehension (CRC) datasets have been created, such as CommonSenseQA [82], ReCoRD [70] and OpenBookQA [73].

### 4.9.6. Complex Reasoning MRC

The reasoning is an innate ability of human beings, which can be embodied in logical thinking, reading comprehension, and other activities. The reasoning is also a key component in artificial intelligence and a fundamental goal of MRC. In recent years, reasoning has been an essential topic among the MRC community. We hope that the MRC system can not only read and learn the representation of the language but also can really understand the context and answer complex questions. In order to push towards complex reasoning MRC system, many datasets have been generated, such as Facebook bAbI [29], DROP [50], RACE [60], and CLOTH [69].

### 4.9.7. Conversational Reading Comprehension

It is a natural way for human beings to exchange information through a series of conversations. In the typical MRC tasks, different question and answer pairs are usually independent of each other. However, in real human language communication, we often achieve an efficient understanding of complex information through a series of interrelated conversations. Similarly, in human communication scenarios, we often ask questions on our own initiative, to obtain key information that helps us understand the situation. In the process of conversation, we need to have a deep understanding of the previous conversations in order to answer each other's questions correctly or ask meaningful new questions. Therefore, in this process, historical conversation information also becomes a part of the context.

In recent years, conversational machine reading comprehension (CMRC) has become a new research hotspot in the NLP community, and there emerged many related datasets, such as CoQA [47], QuAC [65], DREAM [81] and ShARC [38].

### 4.9.8. Domain-Specific Datasets

In this paper, a domain-specific dataset refers to the MRC dataset whose context comes from a particular domain, such as science examinations, movies, clinical reports. Therefore, the neural network models trained by those datasets usually can be directly applied to a certain field. For example, CliCR [71] is a cloze MRC dataset in the medical domain. There are approximately 100,000 cloze questions about the clinical case reports. SciQ [64] is a multiple-choice MRC dataset containing 13.7K crowdsourced science exam questions about physics, chemistry and biology, and others. The context and questions of SciQ are derived from scientific exam questions. In addition, domain-specific datasets also include ReviewQA [25], SciTail [74], WikiMovies [53], PaperQA [76].

### 4.9.9. MRC with Paraphrased Paragraph

Paragraph paraphrasing refers to rewriting or rephrasing a paragraph using different words, while still conveying the same messages as before. The MRC dataset with paraphrased paragraph has at least two versions of context which expresses the same meanings while there is little word overlap between the different versions of context. The task of paraphrased MRC requires the computer to answer questions about contexts. To answer these questions correctly, the computer needs to understand the true meaning of different versions of context. So far, we only find that the DuoRC [68] and Who-did-What [48] are datasets of this type.

### 4.9.10. Large-Scale MRC Dataset

The size of the early MRC dataset is usually not very large, such as QA4MRE, CuratedTREC [51], MCTest [15]. With the emergence of large-scale datasets, MRC is greatly promoted due to the possibility of neural network models training.

### 4.9.11. MRC Dataset for Open-Domain QA

The open-domain question answering was originally defined as finding answers in collections of unstructured documents [35]. With the development of MRC research, many MRC datasets tend to be used to solve open-domain QA. The release of new MRC datasets such as MCTest [15], CuratedTREC [51], Quasar [61], SearchQA [62] greatly promotes open-domain QA recently.

### *4.10. Descriptions of Each Mrc Dataset*

In Section 4.9, we introduced the characteristics of various machine reading comprehension datasets. In this section, we will give a detailed description of the 47 MRC datasets collected in our survey with their download links available. Then we will describe them according to the order of datasets in Table 10.

### 4.10.1. WikiQA

The WikiQA [52] dataset contains a large number of real Bing query logs as the question-answer pair and provided links to Wikipedia passages that might have answers in the dataset. Differs from previous datasets such as QASENT, questions in WIKIQA were sampled from real queries of Bing without editorial revision. The WikiQA dataset also contains questions that can not actually be answered from the given passages, so the machine is required to detect these unanswerable questions. The WikiQA was created by crowd-workers and contains 3047 questions and 29,258 sentences, in which 1473 sentences were marked as answer sentences for the question [52]. The WikiQA dataset is available on https://www.microsoft.com/en-us/download/details.aspx?id=52419.

### 4.10.2. SQuAD 2.0

SQuAD 2.0 [66] is the latest version of the Stanford Question Answering Dataset (SQuAD). SQuAD 2.0 combines the data from the existing version of SQuAD 1.1 [19] with more than 50,000 unanswerable questions written by crowd workers. To acquire a good performance on SQuAD 2.0, the MRC model not only needs to answer questions when possible, but also needs to identify issues without correct answers in the context and not to answer them [66]. For existing models, SQuAD 2.0 is a challenging natural language understanding task. The author also compares the test data of similar model architecture in SQuAD 1.1. Compared with SQuAD 1.1, the gap between human accuracy and machine accuracy in SQuAD 2.0 is much larger, which confirms that square 2.0 is a more difficult data set for existing models. As mentioned in the authors' paper, the powerful nervous model that achieved 86% F1 on SQuAD 1.1 received only 66% of F1 on SQuAD 2.0. Data for both SQuAD 1.1 and SQuAD 2.0 are available on https://rajpurkar.github.io/SQuAD-explorer/.

### 4.10.3. Natural Questions

Natural Questions [80] is a MRC dataset with unanswerable questions. The samples in this dataset come from real anonymous questions and answers in the Google search engine. The dataset is manually generated by the crowd workers. The MRC model presents the crowd worker with a question and related Wikipedia pages and requires the crowd worker to mark a long answer (usually a paragraph) and a short answer (usually one or more entities) on the page or mark null if there is no correct answer. The Natural Questions dataset consists of 307,373 training samples with single annotations, 7830 samples with 5-way annotations for development data, and 7842 test examples with 5-way annotations [80]. The dataset can be downloaded at https://github.com/google-research-datasets/natural-questions, which also has a link to the leaderboard.

### 4.10.4. MS MARCO

MS MARCO [39] is a large-scale machine reading comprehension dataset containing unanswerable questions. The dataset consists of 1,010,916 questions and answers collected from Bing's search query logs. Besides, the dataset contains 8,841,823 paragraphs extracted from 3,563,535 Web documents retrieved by Bing, which provide the information for answering questions. MS MARCO contains three different tasks: (1) Identify unanswerable questions; (2) Answer the question if it is answerable; (3) Rank a set of retrieved passages given a question [39]. The MRC model needs to estimate whether these paragraphs contain correct answers, and then sort them depending on how close they are to the answers. The dataset and leaderboard of MS MARCO are available on http://www.msmarco.org/.

### 4.10.5. DuoRC

DuoRC [68] is a MRC dataset which contains 186,089 question-answer pairs generated from 7680 pairs of movie plots. Each pair of movie plots reflects two versions of the same movie—one from

Wikipedia and the other from IMDb. The texts of these two versions are written by two different authors. In the process of building question-answer pairs, the authors require crowd workers to create questions from one version of the story and a different set of crowd workers to extract or synthesize answers from another version. This is the unique feature of DuoRC in which there is almost no vocabulary overlap between the two versions. Additionally, the narrative style of the paragraphs generated from the movie plots (compare to the typical descriptive paragraphs in the existing dataset) indicates the need for complex reasoning of events in multiple sentences [39]. DuoRC is a challenging dataset, and the authors observed that the state-of-the-art model on the SQuAD 1.1 [19] also performed poorly on DuoRC, with F1 score of 37.42% while 86% on SQuAD 1.1. The dataset, paper and, leaderboard of DuoRC can be obtained at https://duorc.github.io/.

### 4.10.6. Who-Did-What

The Who-did-What [48] dataset contains more than 200,000 fill-in-the-gap (cloze) multiple-choice reading comprehension questions constructed from the LDC English Gigaword newswire corpus. Compared to other existing machine reading comprehension datasets, such as CNN/Daily Mail [19], the Who-did-What dataset avoided using the same article summaries to create a sample in the dataset. Instead, each sample is formed by two separate articles. One article is given as the passage to be read and the other article on the same events is used to form the question. Second, the authors avoided anonymization—each choice is a person named entity. Third, the questions that can be easily solved by simple baselines have been removed, while humans can still solve 84% of the questions [48]. The dataset and leaderboard of Who-did-What are available on https://tticnlp.github.io/who_did_what/index.html.

### 4.10.7. ARC

AI2 Reasoning Challenge (ARC) [72] is a MRC dataset and task to encourage AI research in question answering that requires deep reasoning. To finish the ARC task, the MRC model requires far more powerful knowledge and reasoning than previous challenges such as SQuAD [19,66] or SNLI [94]. The ARC dataset contains 7787 elementary-level scientific questions that are in the form of multiple-choices. The dataset is divided into a Challenge Set and an Easy Set, where the Challenge Set only contains questions that are not correctly answered by both a retrieval-based algorithm and a word co-occurrence algorithm. The ARC dataset contains only natural, primary-level science questions (written for the human exam) and is the largest collection of such datasets. The authors tested several baselines on the Challenge Set, including state-of-the-art models from the SQuAD and SNLI, and found that none of them were significantly better than the random baseline, reflecting the difficulty of the task. The author also publishes the ARC corpus, which is a corpus of 14M scientific sentences related to this task, and the implementation of three neural baseline models tested [72]. Information about the ARC dataset and leaderboards is available on http://data.allenai.org/arc/.

### 4.10.8. MCScript

MCScript [78] is a large-scale MRC dataset with narrative texts and questions that require reasoning using commonsense knowledge. The dataset focuses on narrative texts about everyday activities, and the commonsense knowledge are required to answer multiple-choice questions based on these texts. The feature of the MCScript dataset is to evaluate the contribution of script knowledge to machine understanding. A script is a series of events (also called scenarios) that describe human behavior. The MCScript dataset also forms the basis of a shared task on commonsense and script knowledge organized at SemEval 2018 [78]. The official web page and CodaLab competition page of the SemEval 2018 Shared Task 11 are available on https://competitions.codalab.org/competitions/17184.

### 4.10.9. OpenBookQA

OpenBookQA [73] consists of about 6000 elementary level science questions in the form of multi-choice (4957 training sets, 500 validation sets, and 500 test sets). Answering the questions in OpenBookQA requires broad common knowledge. OpenBookQA also requires a deeper understanding of both the topic (in the context of common knowledge) and the language it is expressed in Reference [73]. The baseline model provided by the authors has reached about 50% in this dataset, but many state-of-the-art pre-trained QA methods perform surprisingly even worse [73]. Dataset and leaderboard of OpenBookQA are available on https://leaderboard.allenai.org/open_book_qa/.

### 4.10.10. ReCoRD

ReCoRD [70] is a large-scale MRC dataset that requires deep commonsense reasoning. Experiments on the ReCoRD show that the performance of the state-of-the-art MRC model lags far behind human performance. The ReCoRD represents the challenge of future research to bridge the gap between human and machine commonsense reading comprehension. The ReCoRD dataset contains more than 120,000 queries from over 70,000 news articles. Each query has been verified by crowd workers [70]. The feature of the data set is that all queries and paragraphs in the records are automatically mined from news articles, which minimizes the artificially induced bias. So most records need deep commonsense reasoning. Since July 2019, the ReCoRD has been added to SuperGLUE as an evaluation suite. The ReCoRD dataset and leaderboard are available on https://sheng-z.github.io/ReCoRD-explorer/.

### 4.10.11. CommonSenseQA

CommonSenseQA [82] is a MRC dataset that requires different types of commonsense knowledge to predict the correct answer. It contains 12,247 questions. The CommonSenseQA dataset is split into a training set, validation set and, test set. The authors performed two types of splits: "Random split" which is the main evaluation split, and "Question token split" where each of the three sets has disjoint question concepts [82]. To capture common sense beyond association, the authors of CommonSenseQA extracted multiple target concepts from Conceptnet 5.5 [95] that have the same semantic relationship to a single source concept. Crowd workers were asked to propose multiple-choice questions, mention source concepts, and then distinguished each goal concept. This encouraged crowd workers to ask questions with complex semantics that often require prior knowledge [82]. The dataset and leaderboard of CommonSenseQA are available on https://www.tau-nlp.org/commonsenseqa.

### 4.10.12. WikiReading

WikiReading [56] is a large-scale machine reading comprehension dataset that contains 18 million instances. The dataset consists of 4.7 million unique Wikipedia articles, which means that about 80% of the English language Wikipedia is represented. The WikiReading dataset is composed of a variety of challenging classification and extraction subtasks, which makes it very suitable for neural network models. In the WikiReading dataset, multiple instances can share the same document, with an average of 5.31 instances per article (median: 4, maximum: 879). The most common document categories are humans, categories, movies, albums, and human settlements, accounting for 48.8% of documents and 9.1% of instances respectively. The average and median document lengths are 489.2 and 203 words [56]. The WikiReading dataset is available on https://github.com/google-research-datasets/wiki-reading.

### 4.10.13. WikiMovies

WikiMovies [53] is a MRC dataset with Wikipedia documents. To compare using Knowledge Bases (KBs), information extraction or Wikipedia documents directly in a single framework, the author built the WikiMovies dataset which contains raw texts and preprocessed KBs.

WikiMovies is part of FaceBook's bAbI project, and information about the BABi project is available on https://research.fb.com/downloads/babi/, and the WikiMovies dataset is available on http://www.thespermwhale.com/jaseweston/babi/movieqa.tar.gz.

### 4.10.14. MovieQA

The MovieQA [32] dataset is a multi-modal machine reading comprehension dataset designed to evaluate the automatic understanding of both pictures and texts. The dataset contains 14,944 questions from 408 movies. The types of questions in the MovieQA dataset are multiple-choice, and the questions range from simpler "Who" did "What" to "Whom", to "Why" and "How" certain events occurred. The MovieQA dataset is unique because it contains multiple sources of information-video clips, episodes, scripts, subtitles, and DVS [32]. Download links and evaluation benchmarks of the MovieQA dataset can be obtained for free from http://movieqa.cs.toronto.edu/home/.

### 4.10.15. COMICS

COMICS [33] is a multi-modal machine reading comprehension dataset, which is composed of more than 1.2 million comic panels (120 GB) and automatic text box transcriptions. In the COMICS task, the machine is required to read and understand the text and images in the comic panels at the same time. Besides the traditional textual cloze tasks, the authors also designed two novel MRC tasks (visual cloze, and character coherence) to test the model's ability to understand narratives and characters in a given context [33]. The dataset and baseline of COMICS are available on https://obj.umiacs.umd.edu/comics/index.html.

### 4.10.16. TQA

The TQA [31] (Textbook Question Answering) challenge encourages multi-modal machine reading (M3C) tasks. Compared with Visual Question Answering (VQA) [96], the TQA task provides the multi-modal context and question-answer pair which consists of text and images. TQA dataset is constructed from the science curricula of middle school. The textual and diagrammatic content in middle school science reference fairly complex phenomena that occur in the world. Many questions need not only simple search, but also complex analysis and reasoning of multi-mode context.

The TQA dataset consists of 1076 courses and 26,260 multi-modal questions [31]. The analysis shows that a high proportion of questions in the TQA dataset require complex text analysis, graphing, and reasoning, which indicates that the TQA dataset is related to previous machine understanding and VQA dataset [96] The TQA dataset and leaderboards are available on http://vuchallenge.org/tqa.html.

### 4.10.17. RecipeQA

RecipeQA [34] is a MRC dataset for multi-modal comprehension of recipes. It consists of about 20 K instructional recipes with both texts and images and more than 36 K automatically generated question-answer pairs. RecipeQA is a challenging multi-modal dataset for evaluating reasoning on real-life cooking recipes. The RecipeQAtask consists of many specific tasks. A sample in RecipeQA contains a multi-modal context, such as headings, descriptions, or images. To find an answer, the model needs (i) a joint understanding of the pictures and texts; (ii) capturing the temporal flow of events; and (iii) understanding procedural knowledge [34]. The dataset and leaderboard of RecipeQA are available on http://hucvl.github.io/recipeqa.

### 4.10.18. HotpotQA

HotpotQA [67] is a multi-hop MRC dataset with multi-paragraphs. There are 113k Wikipedia-based QA pairs in HotpotQA. Different from other MRC datasets, In the HotpotQA, the model is required to perform complex reasoning and provide explanations for answers from multi-paragraphs. HotpotQA has four key features: (1) the questions require the machine to read

and reason over multiple supporting documents to find the answer; (2) The questions are diverse and not subject to any pre-existing knowledge base; (3) The authors provided sentence-level supporting facts required for reasoning; (4) The authors offered a new type of factoid comparison questions to test QA systems' ability to extract relevant facts and perform necessary comparison [67]. Dataset and leaderboard of HotpotQA are publicly available on https://hotpotqa.github.io/.

### 4.10.19. NarrativeQA

NarrativeQA [63] is a multi-paragraph machine reading comprehension dataset and a set of tasks. To encourage progress on deeper comprehension of language, the authors designed the NarrativeQA dataset. Unlike other datasets in which the questions can be solved by selecting answers using superficial information, in the NarrativeQA, the machine is required to answer questions about the story by reading the entire book or movie script. In order to successfully answer questions, the model needs to understand the underlying narrative rather than relying on shallow pattern matching or salience [63]. NarrativeQA is available on https://github.com/deepmind/narrativeqa.

### 4.10.20. Qangaroo

Qangaroo [49] is a multi-hop machine reading comprehension dataset. Most reading comprehension methods limit themself to questions that can be answered using a single sentence, paragraph, or document [49]. Therefore, the authors of Qangaro proposed a new task and dataset to encourage the development of text understanding models across multiple documents and to study the limitations of existing methods. In the Qangaroo task, the model is required to seek and combine evidence—effectively performing multihop, alias multi-step, inference [49]. The dataset, papers, and leaderboard of Qangaroo are publicly available on http://qangaroo.cs.ucl.ac.uk/index.html.

### 4.10.21. MultiRC

MultiRC (Multi-Sentence Reading Comprehension) [75] is a MRC dataset in which questions can only be answered by considering information from multiple sentences. The purpose of creating this dataset is to encourage the research community to explore more useful methods than complex lexical matching. MultiRC consists of about 6000 questions from more than 800 paragraphs across 7 different areas (primary science, news, travel guides, event stories, etc.) [75]. MultiRC is available on http://cogcomp.org/multirc/. Since May 2019, MultiRC is part of SuperGLUE, so the authors will no longer provide the leaderboard on the above website.

### 4.10.22. CNN/Daily Mail

In order to solve the problem of lack of large-scale datasets, Hermann et al. [19] created a new dataset generation method that provided a large-scale supervised reading comprehension dataset in 2015. They also extracted text from the websites of CNN and Daily Mail and created two MRC datasets, which is the CNN/Daily Mail [19] dataset. In the CNN dataset, there are 90,266 documents and 380,298 questions. The Daily Mail dataset consist of 196,961 documents and 879,450 questions. The creation of the CNN/Daily Mail dataset allows the community to develop a class of attention based deep neural networks that learn to read real documents and answer complex questions with minimal prior knowledge of language structure [19]. The CNN/Daily Mail dataset and related materials are available on https://github.com/deepmind/rc-data.

### 4.10.23. BookTest

The BookTest [57] is a large-scale MRC dataset with 14,140,825 training examples and 7,917,523,807 tokens. The BookTest dataset is derived from books available through the project Gutenberg [97]. The training dataset contains the original CBT NE and CN data [28] and extends the new NE and CN examples. The authors of BookTest extracted 10,507 books for NE instances from the

project Gutenberg and also used 3555 copyright-free books to extract CN instances [57]. The BookTest dataset can be downloaded from https://ibm.biz/booktest-v1.

### 4.10.24. MCTest

In MCTest [15] dataset, the model is required to answer multiple-choice questions about fictional stories, directly tackling the high-level goal of open-domain machine comprehension. The stories and questions of MCTest are also carefully limited to those a young child would understand, reducing the world knowledge that is required for the task [15]. The data in MCTest was gathered using Amazon Mechanical Turk. Since the answer is a fictional story, the content of the answer is very broad and not limited to a certain field. Therefore, the MRC model trained by MCTest is helpful for the open-domain question answering research [15]. The MCTest dataset and leaderboards are available on https://mattr1.github.io/MCTest/.

### 4.10.25. CuratedTREC

The CuratedTREC [51] dataset is a curated version of the TREC corpus [98]. The Text REtrieval Conference (TREC) [98] was started in 1992 by the U.S. Department of Defense and the National Institute of Standards and Technology (NIST). Its purpose was to support the research of the information retrieval system. The large version of CuratedTREC is based on the QA tasks of TREC 1999, 2000, 2001 and 2002 which have been curated by Baudiš and JŠedivý [51] and contains a total of 2180 questions. CuratedTREC is also used to evaluate the ability of the machine reading comprehension model to answer open-domain questions [35,99,100]. The TREC corpus is available in https://github.com/brmson/dataset-factoid-curated.

### 4.10.26. Quasar

Quasar [61] is a MRC dataset for open-domain questions, it contains two sub-datasets—Quasar-T and Quasar-S. Quasar is designed to evaluate the model's ability of understanding natural language queries and extract answers from large amounts of texts. The Quasar-S dataset consists of 37,000 cloze-style questions, and the Quasar-T dataset contains 43,000 open-domain trivia issues questions. ClueWeb09 [101] serves as a background corpus for extracting these answers. The Quasar dataset is a challenge to two related sub-tasks of the factoid questions: (1) searching for relevant text segments containing the correct answers to the query, and (2) reading the retrieved passages to answer the questions [61]. The dataset and paper of Quasar are available on https://github.com/bdhingra/quasar.

### 4.10.27. SearchQA

SearchQA [62] is a MRC dataset with retrieval systems. To answer open-domain questions in SearchQA, the model needs to read the text retrieved by the search engine, so it can also be regarded as a machine reading comprehension dataset. The question-answer pairs in the SearchQA dataset are all collected from the J!Archive, and the context is retrieved from Google. SearchQA consists of more than 140k QA pairs, with an average of 49.6 clips per pair. Each QA environment tuple in SearchQA comes with additional metadata, such as the URL of the fragment, which the authors believe will be a valuable resource for future research. The authors perform a manual evaluation on SearchQA and test two baseline methods, one simple word selection, and another deep learning [62]. The paper suggests that the SearchQA can be obtained at https://github.com/nyu-dl/SearchQA.

### 4.10.28. SciQ

SciQ [64] is a domain-specific multiple-choice MRC dataset containing 13.7K crowdsourced science questions about Physics, Chemistry, and Biology, and so forth. The context and questions are derived from real 4th and 8th-grade exam questions. The questions are in the form of multiple-choices,

with an average of four choices for each question. For the majority of the questions, an additional paragraph with supporting evidence for the correct answer is provided. In addition, the authors proposed a new method for generating domain-specific multiple-choice MRC datasets from crowd workers [64]. The SciQ dataset can be downloaded at http://data.allenai.org/sciq/.

### 4.10.29. CliCR

CliCR [71] is a cloze MRC dataset in the medical domain. There are approximately 100,000 cloze questions about the clinical case reports. The authors applied several baselines and a state-of-the-art neural model and observed the performance gap (20% F1) between the human and the best neural models [71]. They also analyzed the skills required to correctly answer the question and explained how the model's performance changes based on the applicable skills, and they found that reasoning using domain knowledge and object tracking is the most frequently needed skill, and identifying missing information and spatiotemporal reasoning is the most difficult for machines [71]. The code of the baseline project can be publicly available on https://github.com/clips/clicr, where the author claims that the CliCR dataset can be obtained by contacting the author via email.

### 4.10.30. PaperQA (Hong et al.,2018)

PaperQA [77] created by Hong et al. is a MRC dataset containing more than 6000 human-generated question-answer pairs about academic knowledge. To build the PaperQA, crowd workers have provided questions based on more than 1000 abstracts of the research paper on deep learning, and their answers that consist of text spans of the related abstracts. The authors collected the PaperQA through a four-stage process to acquire QA pairs that require reasoning. And they have proposed a semantic segmentation model to solve this task [77]. PaperQA is publicly available on http://bit.ly/PaperQA.

### 4.10.31. PaperQA (Park et al.,2018)

In order to measure the machine's ability of understanding professional-level scientific papers, a domain-specific MRC dataset called PaperQA [76] was created. PaperQA consists of over 80,000 cloze questions from research papers. The authors of PaperQA performed fine-grained linguistic analysis and evaluation to compare PaperQA and other conventional question and answering (QA) tasks on general literature (e.g., books, news, and Wikipedia), and the results indicated that the PaperQA task is difficult, showing there is ample room for future research [76]. According to the authors' paper, PaperQA had been published on http://dmis.korea.ac.kr/downloads?id=PaperQA, but when we visited this website, it was not available at that moment.

### 4.10.32. ReviewQA

ReviewQA [25] is a domain-specific MRC dataset about hotel reviews. ReviewQA contains over 500,000 natural questions and 100,000 hotel reviews. The authors hope to improve the relationship understanding ability of the machine reading comprehension model by constructing the ReviewQA dataset. Each question in ReviewQA is related to a set of relationship understanding capabilities that the model is expected to master [25]. The ReviewQA dataset, summary of the tasks, and results of models are available on https://github.com/qgrail/ReviewQA/.

### 4.10.33. SciTail

The SciTail [74] is a textual entailment dataset which consists of multiple-choices QA pairs about scientific exams and web sentences. The dataset consists of 27,026 examples, of which 10,101 examples contain entails labels and 16,925 examples contain neutral labels. Different from existing datasets, SciTail was created solely from natural sentences that already exist independently "in the wild" rather than sentences authored specifically for the entailment task [74]. The authors also generated hypotheses from questions, the relevant answer options, and premises from related

web sentences from a large corpus [74]. Baseline and leaderboard of SciTail are available on https://leaderboard.allenai.org/scitail/submissions/public. The SciTail dataset is available on http://data.allenai.org/scitail/.

### 4.10.34. DROP

DROP [50] is an English MRC dataset that requires the Discrete Reasoning Over the content of Paragraphs. The DROP dataset contains 96k questions created by crowd workers. Unlike the existing MRC task, in the DROP, the MRC model is required to resolve references in a question, and perform discrete operations on them (such as adding, counting, or sorting) [50]. These operations require a deeper understanding of the content of paragraphs than what was necessary for prior datasets [50]. The dataset of DROP can be downloaded at https://s3-us-west-2.amazonaws.com/allennlp/datasets/drop/drop_dataset.zip. The Leaderboard is available on https://leaderboard.allenai.org/drop.

### 4.10.35. Facebook CBT

Children's Book Test (CBT) [28] is a MRC dataset that uses children's books as context. Each sample in the CBT dataset contains 21 consecutive sentences, the first 20 sentences become the context, and a word is deleted from the 21st sentence, so it becomes a cloze question. MRC model is required to identify the answer word among a selection of 10 candidate answers appearing in the context sentences and the question. Different from standard language-modeling tasks, CBT distinguishes the task of predicting syntactic function words from that of predicting lower-frequency words, which carry greater semantic content [28]. The CBT dataset is part of FaceBook's bAbI project which is available on https://research.fb.com/downloads/babi/.The Children's Book Test (CBT) dataset can be downloaded at http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz.

### 4.10.36. Google MC-AFP

Google MC-AFP [58] is a MRC dataset which has about 2 million examples. It is generated from the AFP portion of LDC's English Gigaword corpus [102]. The authors of MC-AFP also provided a new method for creating large-scale MRC datasets using paragraph vector models. In the MC-AFP, the upper limit of accuracy achieved by human testers is approximately 91%. Among all models tested by the authors, the authors' hybrid neural network architecture achieves the highest accuracy of 83.2%. The remaining gap to the human-performance ceiling provides enough room for future model improvements [58]. Google MC-AFP is available on https://github.com/google/mcafp.

### 4.10.37. LAMBADA

The main task of the LAMBADA [55] is to read the text and predict the missing last word. The authors of LAMBADA hoped to encourage the development of new models capable of genuine understanding of broad context in natural language text [55], therefore, it can also be understood as a MRC task. The LAMBADA dataset consists of narrative passages in which human subjects could guess the last word if they read the whole paragraph, but not if they only read the last sentence preceding the answer word. In order to get high scores in LAMBADA, the models have to track information in a wider discourse. For the above reasons, LAMBADA as a challenging dataset and exemplifies a wide range of linguistic phenomena [55]. LAMBADA can be obtained at https://zenodo.org/record/2630551.

### 4.10.38. NewsQA

NewsQA [54] is a new MRC dataset that contains more than 100,000 natural instances. Crowd workers provide questions and answers based on more than 10,000 news articles from CNN, in which the answer is a text span of the related news article. The authors collected NewsQA

through a four-stage process to seek exploratory question-answer pairs that require reasoning. The authors also stratified reasoning categories in NewsQA, including word matching, paraphrasing, inference, synthesis, ambiguous/insufficient. The NewsQA requires the ability to go beyond simple word matching and recognizing textual entailment. The authors measured human performance on NewsQA and compared it to several powerful neural models. The performance gap between humans and the MRC model (0.198 in F1) suggested that significant progress could be made on NewsQA through future research [54]. The NewsQA dataset and model leaderboards are available for free at https://www.microsoft.com/en-us/research/project/newsqa-dataset/.

### 4.10.39. SQuAD 1.1

The Stanford Question Answering Dataset (SQuAD) [19] is a well-known machine reading comprehension dataset that contains more than 100,000 questions generated by crowd-workers, in which the answer of each question is a segment of text from the related paragraph [19]. Since it was released in 2016, SQuAD 1.1 quickly became the most widely used MRC dataset. Now it has been updated to SQuAD 2.0 [66]. In the leaderboards of SQuAD 1.1 and SQuAD 2.0, we have witnessed the birth of a series of state-of-the-art neural models, such as BiDAF [20], BERT [22], RoBERTa [23] and XLNet [24], and so forth. The data and leaderboard of SQuAD 1.1 and SQuAD 2.0 are available on https://rajpurkar.github.io/SQuAD-explorer/.

### 4.10.40. RACE

RACE [60] is a MRC dataset collected from the English exams for Chinese students. There are approximately 28,000 articles and 100,000 questions provided by humans (English teachers), covering a variety of carefully designed topics to test students' understanding and reasoning ability. Different from the existing MRC dataset, the proportion of questions that need reasoning ability in RACE is much large than other MRC datasets, and there is a great gap between the performance of the state-of-the-art models (43%) and the best human performance (95%) [60]. The authors hope that this new dataset can be used as a valuable resource for machine understanding research and evaluation [60]. The dataset of RACE is available on http://www.cs.cmu.edu/glai1/data/race/. The baseline project is available on https://github.com/qizhex/RACE_AR_baselines.

### 4.10.41. TriviaQA

TriviaQA [59] is a challenging MRC dataset, which contains more than 650k question-answer pairs and their evidence. TriviaQA has many advantages over other existing MRC datasets: (1) relatively complex combinatorial questions; (2) considerable syntactic and lexical variability between the questions and the related passages; (3) more cross sentence reasoning is required to answer the question [59]. The TriviaQA dataset and baseline project are available on http://nlp.cs.washington.edu/triviaqa/ and information about the Codalab competition of TriviaQA is available on https://competitions.codalab.org/competitions/17208.

### 4.10.42. CLOTH

CLOTH [69] is a large-scale cloze MRC dataset with 7131 passages and 99,433 questions collected from English examinations. CLOTH requires a deeper language understanding of multiple aspects of natural language including reasoning, vocabulary and grammar. In addition, CLOTH can be used to evaluate language models' abilities in modeling long text [69]. CLOTH's leaderboard is available on http://www.qizhexie.com/data/CLOTH_leaderboard and dataset can be downloaded from http://www.cs.cmu.edu/glai1/data/cloth/. The code of baseline project can be downloaded at https://github.com/qizhex/Large-scale-Cloze-Test-Dataset-Created-by-Teachers.

### 4.10.43. ProPara

ProPara [79] is a MRC dataset for understanding contexts about processes (such as photosynthesis). In the ProPara task, the model is required to identify the actions described in the procedural text and tracking the state changes that have occurred to the entities involved. The ProPara dataset contains 488 paragraphs and 3300 sentences (about 81,000 notes) generated by crowd workers. The purpose of creating ProPara is to predict the presence and location of each participant based on the sentences in the context [79]. The dataset of Propara can be downloaded from http://data.allenai.org/propara, and the leaderboard of Propara is available on https://leaderboard.allenai.org/propara/submissions/public.

### 4.10.44. DREAM

DREAM [81] is a conversational, multiple-choice MRC dataset. The dataset was collected from English exam questions designed by human experts to evaluate the reading comprehension level of English learners. The DREAM dataset consists of 10,197 questions in the form of multiple-choice with a total of 6444 dialogues. Compared to the existing conversational reading comprehension (CRC) dataset, DREAM is the first to focus on in-depth multi-turn multi-party dialogue understanding [81]. In the DREAM dataset, 84% of answers are non-extractive, 85% require more than one sentence of reasoning, and 34% of questions involve common sense knowledge. DREAM's authors applied several neural models on DREAM that used surface information in the text and found that they could barely surpass rule-based methods. In addition, the authors also studied the effects of incorporating dialogue structures and different types of general world knowledge into several models on the DREAM dataset. The experimental results demonstrated the effectiveness of the dialogue structure and general world knowledge [81]. DREAM is available on: https://dataset.org/dream/.

### 4.10.45. CoQA

CoQA [47] is a conversational MRC dataset that contains 127K questions and answers from 8k dialogues in 7 different fields. Through an in-depth analysis of CoQA, the authors showed that conversational questions in CoQA have challenging phenomena that are not presented in existing MRC datasets, such as coreference and pragmatic reasoning. The authors also evaluated a set of state-of-the-art conversational MRC models on CoQA. The best F1 score achieved by those models is 65.1%, and human performance is 88.8%, indicating that there was plenty of room for future advance [47]. Dataset and leaderboard of CoQA can be found at https://stanfordnlp.github.io/coqa/.

### 4.10.46. QuAC

QuAC [65] is a conversational MRC dataset containing about 100K questions from 14K information-seeking QA dialogs. Each dialogue in QuAC involves two crowd workers: (1) One act like a student who asks a few questions to learn a hidden passage from Wikipedia, and (2) the other one act as a teacher to answer questions by providing a brief excerpt from the Wikipedia passage. The QuAC dataset introduced the challenges that not present in existing MRC datasets—its questions are often more open-ended, unanswerable, or meaningful only in a dialog environment [65]. The authors also reported the performance of many state-of-the-art models on QuAC, and the best result was 20% lower than human F1, suggesting there was ample room for future research [65]. Dataset, baseline and leaderboard of QuAC can be found at http://quac.ai.

### 4.10.47. ShARC

ShARC [38] is a conversational MRC dataset. Unlike existing conversational MRC datasets, when answering questions in the ShARC, the model needs to use background knowledge that is not in the context to get the correct answer. The first question in a ShARC conversation is usually not fully explained and does not provide enough information to answer directly. Therefore, the model needs to take the initiative to ask the second question, and after the model has got enough information,

it then answers the first question [38]. The dataset, paper, and leaderboard of ShARC are available on https://sharc-data.github.io.

## 5. Open Issues

In recent years, great progress has been made in the field of MRC due to large-scale datasets and effective deep neural network approaches. However, there are still many issues remaining in this field. In this section, we describe these issues in the following aspects:

### 5.1. What Needs to Be Improved?

Nowadays, the neural machine reading models have exceeded the human performance scores on many MRC datasets. However, state-of-the-art models are still far from human-level language understanding. What needs to be improved on existing tasks and datasets? We believe that there are many important aspects that have been overlooked which merit additional research. Here we list several areas as below:

#### 5.1.1. Multi-Modal MRC

A fundamental characteristic of human language understanding is multimodality. Psychologists examined the role of mental imagery skills on story comprehension in fifth graders (10- to 12-year-olds). Experiments showed that children with higher mental imagery skills outperformed children with lower mental imagery skills on story comprehension after reading the experimental narrative [4]. Our observation and experience of the world bring us a lot of common sense and world knowledge, and the multi-modal information is extremely important for us to acquire such common sense and world knowledge. However, it is currently not clear how our brains store, encode, represent, and process knowledge, which is an important scientific problem in cognitive neuroscience, philosophy, psychology, artificial intelligence and other fields. At present, the research in the field of natural language processing mainly focuses on the pure textual corpus, but in neuroscience, the research methods are very different. Since the 1990s, cognitive neuroscientists have found that knowledge extraction could activate the widely distributed cerebral cortex, including the sensory cortex and the motor cortex [103]. More and more cognitive neuroscientists believe that concepts are rooted in modality-specific representations [103]. This is usually called Grounded Cognition Model [104,105], or Embodied Cognition Model [103,106,107]. The key idea is that semantic knowledge does not exist in an abstract domain completely separated from perception and behavior, but overlaps these capabilities to some extent [103,108,109]. In that case, can we still make computers really understand human languages only by the neural network training of pure textual corpus? Nowadays, although there are already a few of multi-modal MRC datasets, the related research is still insufficient. The number of current multi-modal MRC datasets are still small, and these datasets simply put pictures and texts together, lacking detailed annotations and internal connections. How to make better use of multi-modal information is an important research area in the future.

#### 5.1.2. Commonsense and World Knowledge

Commonsense and world knowledge are the main bottlenecks in machine reading comprehension. Among different kinds of commonsense and world knowledge, two types of commonsense knowledge are considered fundamental for human reasoning and decision making—intuitive psychology and intuitive physics [110]. Although there are some MRC datasets about commonsense, such as CommonSenseQA [82], ReCoRD [70], DREAM [81], OpenBookQA [73], this field is still in a very early stage. In these datasets, there is no strict division of commonsense types, nor research on commonsense acquisition methods combined with psychology. Understanding how the commonsense knowledge is acquired in the process of human growth may help to reveal the computing model of commonsense. Observing the world is the first step for us to acquire commonsense and world knowledge. For example, "this book can't be put into a school bag, it's too small" and "this

book can't be put into a schoolbag, it's too big". In these two sentences, human beings can know from commonsense that the former "it" refers to a school bag, and the latter "it" refers to a book. But this is not intuitive for computers. Human beings receive a great deal of multi-modal information in our daily life, which forms commonsense. When the given information is insufficient, we can make up the gap by predicting. Correct prediction is the core function of our commonsense. In order to gain real understanding ability comparable to human beings, machine reading comprehension models must need massive data to provide commonsense and world knowledge. Algorithms are needed to get a better commonsense corpus and we need to create multi-modal MRC datasets to help machines acquire commonsense and world knowledge.

### 5.1.3. Complex Reasoning

Many of the existing MRC datasets are relatively simple. In these datasets, the answers are short, usually a word or a phrase. Many of the questions can be answered by understanding a single sentence in the context, and there are very few datasets that need multi-sentences reasoning [14]. This shows that most of the samples in existing MRC datasets are lack of complex reasoning. In addition, researchers found that after input-ablation, many of the answers in existing MRC datasets are still correct [5]. This shows that many existing benchmark datasets do not really require the machine reading comprehension model to have reasoning skills. From this perspective, high-quality MRC datasets that need complex reasoning is needed to test the reasoning skill of MRC modals.

### 5.1.4. Robustness

Robustness is one of the key desired properties of a MRC model. Jia and Liang [30] found that existing benchmark datasets are overly lenient on models that rely on superficial cues [14,27]. They tested whether MRC systems can answer questions that contain distracting sentences. In their experiment, a distracting sentence that contains words that overlap with the question was added at the end of the context. These distracting sentences will not mislead human understanding, but the average scores of the sixteen models on SQuAD will be significantly reduced. This shows that these state-of-the-art MRC models still rely too much on superficial cues, and there is still a huge gap between MRC and human-level reading comprehension [30]. How to avoid the above situation and improve the robustness of MRC model is an important challenge.

### 5.1.5. Interpretability

In the existing MRC tasks, the model is only required to give the answer to the question directly, without explaining why it gets the answer. So it is very difficult to really understand how the model makes decisions [14,27]. Regardless of whether the complete interpretability of these models is absolutely necessary, it is fair to say that a certain degree of understanding of the internal model can greatly guide the design of neural network structure in the future. In future MRC datasets, sub-tasks can be set up to let the model give the reasoning process, or the evidence used in reasoning.

### 5.1.6. Evaluation of the Quality of MRC Datasets?

There are many evaluation metrics for machine reading comprehension models, such as F1, EM, accuracy, and so forth. However, different MRC datasets also need to be evaluated. How to evaluate the quality of MRC datasets? One metric of MRC dataset is the readability. The classical measures of readability are based on crude approximations of the syntactic complexity (using the average sentence length as a proxy) and lexical complexity (average length in characters or syllables of words in a sentence). One of the most well-known measures along these lines is the Flesch-Kincaid readability index [111] which combines these two measures into a global score [112]. However, recent studies have shown that the readability of MRC dataset is not directly related to the question difficulty [91]. The experiment results suggest that while the complexity of datasets is decreasing, the performance of MRC model will not be improved to the same extent and the correlation is quite small [112]. Another

possible metric is the frequencies of different prerequisite skills needed in MRC datasets. Sugawara et al. defined 10 prerequisite skills [91], including Object tracking, Mathematical reasoning, Coreference resolution, Analog, Causal relation, and so forth. However, the definition of prerequisite skills is often arbitrary and changeable. Different definitions can be drawn from different perspectives [5,90,91]. Moreover, at present, the frequency of prerequisite skills is still manually counted, and there is no automated statistical method. In summary, how to evaluate the quality of MRC datasets is still an unsolved problem.

*5.2. Have We Understood Understanding?*

The word "understanding" has been used by human beings for thousands of years [113,114]. But, what is the exact meaning of understanding? What are the specific neural processes of understanding?

Many researchers attempted to give definitions of understanding. For example, Hough and Gluck [113] conducted an extensive survey of literature about understanding. They summarized:

> "In an attempt to summarize the preceding review, we propose the following general definition for the process and outcome of understanding: The acquisition, organization, and appropriate use of knowledge to produce a response directed towards a goal, when that action is taken with awareness of its perceived purpose."

But understanding is too natural and complex for us as it is difficult to define, especially from different perspectives such as philosophy, psychology, pedagogy, neuroscience, computer science, and so forth. In the field of NLP, we still lack a comprehensive definition of understanding of language and also lack of specific metrics to evaluate the real understanding capabilities of MRC models.

In recent years, great progress has been made in the field of cognitive neuroscience of language. Thanks to the advanced neuroimaging technologies such as PET and fMRI, contemporary cognitive neuroscientists have been able to study and describe large-scale cortical networks related to language in various ways, and they have found many interesting findings. Just taking understanding object nouns as an example. How are these object nouns represented in the brain? As David Kemmerer summarized in his book [103]:

> "From roughly the 1970s through the 1990s, the dominant theory of conceptual knowledge was the Amodal Symbolic Model. It emerged from earlier developments in logic, formal linguistics, and computer science, and its central claim was that concepts, including word meanings, consist entirely of abstract symbols that are represented and processed in an autonomous semantic system that is completely separate from the modality specific systems for perception and action [115–117].
>
> Since 1990s, the Grounded Cognition Model has been attracting increasing interest. The key idea is that semantic knowledge does not reside in an abstract realm that is totally segregated from perception and action, but instead overlaps with those capacities to some degree. To return to the banana example mentioned above, understanding this object noun is assumed to involve activating modality-specific records in long-term memory that capture generalizations about how bananas look, how they taste, how they feel in one's hands, how they are manipulated, and so forth. This theory maintains that conceptual processing amounts to recapitulating modality-specific states, albeit in a manner that draws mainly on high-level rather than low-level components of the perceptual and motor systems [103]."

In addition, a recent study [118] published in the Cell reveals that the two hypothesis theories mentioned above are both right. The authors studied the brain basis of color knowledge in sighted individuals and congenitally blind individuals whose color knowledge can only be obtained through language descriptions. Their experiments show that congenitally blind individuals can obtain knowledge representation similar to healthy people through language without any sensory experience. And more importantly, they also found that there are two different coding systems in

the brain of sighted individuals: one is directly related to the sense, in the visual color processing brain area; the other is in the left anterior temporal lobe dorsal side, the same as the memory brain area of knowledge obtained only through language in congenitally blind individuals [118]. According to their study, there are (at least) two forms of object knowledge representations in the human brain—sensory-derived and cognitively-derived knowledge, supported by different brain systems [118]. It also shows that human language is not only used to express symbols for communication, but also to encode conceptual knowledge.

So, can we get a more effective MRC model through training multi-modal corpus? Probably. But, due to the complexity of the human brain, cognitive neuroscientists are still unable to fully understand the details of natural language understanding. But these cognitive neuroscience studies have brought a lot of inspiration to the NLP community. We could make full use of the existing research results of cognitive neuroscience to design novel MRC systems.

## 6. Conclusions

We conducted a comprehensive survey of recent efforts on the tasks, evaluation metrics, and benchmark datasets of machine reading comprehension (MRC). We discussed the definition and taxonomy of MRC tasks and proposed a new classification method for MRC tasks. The computing methods of different MRC evaluation metrics have been introduced with their usage in each type of MRC tasks also analyzed. We also introduced attributes and characteristics of MRC datasets, with 47 MRC datasets described in detail. Finally, we discussed the open issues for future research of MRC and we argued that high-quality multi-modal MRC datasets and the research findings of cognitive neuroscience may help us find better ways to construct more challenging datasets and develop related MRC algorithms to achieve the ultimate goal of human-level machine reading comprehension. To facilitate the MRC community, we have published the above data on the companion website (https://mrc-datasets.github.io/), from where MRC researchers could directly access the MRC datasets, papers, baseline projects and browse the leaderboards.

## References

1. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
2. Li, S.; Hu, J.; Cui, Y.; Hu, J. DeepPatent: Patent classification with convolutional neural networks and word embedding. *Scientometrics* **2018**, *117*, 721–744. [CrossRef]
3. Li, Q.; Li, S.; Hu, J.; Zhang, S.; Hu, J. Tourism Review Sentiment Classification Using a Bidirectional Recurrent Neural Network with an Attention Mechanism and Topic-Enriched Word Vectors. *Sustainability* **2018**, *10*, 3313. [CrossRef]
4. Boerma, I.E.; Mol, S.E.; Jolles, J. Reading pictures for story comprehension requires mental imagery skills. *Front. Psychol.* **2016**, *7*, 1630. [CrossRef] [PubMed]
5. Sugawara, S.; Stenetorp, P.; Inui, K.; Aizawa, A. Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. *arXiv* **2019**, arXiv:1911.09241.
6. Zhang, X.; Yang, A.; Li, S.; Wang, Y. Machine reading comprehension: A literature review. *arXiv* **2019**, arXiv:1907.01686.

7.   Baradaran, R.; Ghiasi, R.; Amirkhani, H. A Survey on Machine Reading Comprehension Systems. *arXiv* **2020**, arXiv:2001.01582.

8.   Gupta, S.; Rawat, B.P.S. Conversational Machine Comprehension: A Literature Review. *arXiv* **2020**, arXiv:2006.00671.

9.   Gao, J.; Galley, M.; Li, L. Neural Approaches to Conversational AI. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts, Melbourne, Australia, 15–20 July 2018; pp. 2–7. [CrossRef]

10.   Lehnert, W.G. The Process of Question Answering. Ph.D. Thesis, Yale Univ New Haven Conn Dept Of Computer Science, New Haven, CT, USA, 1977.

11.   Hirschman, L.; Light, M.; Breck, E.; Burger, J.D. Deep Read: A Reading Comprehension System. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD, USA, 20–26 June 1999; pp. 325–332. [CrossRef]

12.   Riloff, E.; Thelen, M. A Rule-Based Question Answering System for Reading Comprehension Tests. In Proceedings of the ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, New Orleans, LA, USA, 2–7 February 2000; pp. 13–19. [CrossRef]

13.   Charniak, E.; Altun, Y.; de Salvo Braz, R.; Garrett, B.; Kosmala, M.; Moscovich, T.; Pang, L.; Pyo, C.; Sun, Y.; Wy, W.; et al. Reading Comprehension Programs in a Statistical-Language-Processing Class. In Proceedings of the ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems, Seattle, WA, USA, 29 April–4 May 2000.

14.   Chen, D. Neural Reading Comprehension and Beyond. Ph.D. Thesis, Stanford University, Palo Alto, CA, USA, 2018.

15.   Richardson, M.; Burges, C.J.; Renshaw, E. 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, Seattle, WA, USA, 18–21 October 2013; pp. 193–203.

16.   Wang, H.; Bansal, M.; Gimpel, K.; McAllester, D. Machine comprehension with syntax, frames, and semantics. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Beijing, China, 26–31 July 2015; pp. 700–706.

17.   Sachan, M.; Dubey, K.; Xing, E.; Richardson, M. Learning answer-entailing structures for machine comprehension. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; pp. 239–249.

18.   Narasimhan, K.; Barzilay, R. Machine comprehension with discourse relations. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Beijing, China, 26–31 July 2015; pp. 1253–1262.

19.   Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. Squad: 100,000+ questions for machine comprehension of text. *arXiv* **2016**, arXiv:1606.05250.

20.   Seo, M.; Kembhavi, A.; Farhadi, A.; Hajishirzi, H. Bidirectional attention flow for machine comprehension. *arXiv* **2016**, arXiv:1611.01603.

21.   Peters, M.E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; Zettlemoyer, L. Deep contextualized word representations. *arXiv* **2018**, arXiv:1802.05365.

22.   Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.

23.   Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv* **2019**, arXiv:1907.11692.

24.   Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J.; Salakhutdinov, R.R.; Le, Q.V. Xlnet: Generalized autoregressive pretraining for language understanding. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 5754–5764.

25.   Grail, Q.; Perez, J. ReviewQA: A relational aspect-based opinion reading dataset. *arXiv* **2018**, arXiv:1810.12196.

26.   Qiu, B.; Chen, X.; Xu, J.; Sun, Y. A Survey on Neural Machine Reading Comprehension. *arXiv* **2019**, arXiv:1906.03824.

27. Liu, S.; Zhang, X.; Zhang, S.; Wang, H.; Zhang, W. Neural machine reading comprehension: Methods and trends. *Appl. Sci.* **2019**, *9*, 3698. [CrossRef]

28. Hill, F.; Bordes, A.; Chopra, S.; Weston, J. 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, 2–4 May 2016.

29. Weston, J.; Bordes, A.; Chopra, S.; Mikolov, T. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. In Proceedings of the 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, 2–4 May 2016.

30. Jia, R.; Liang, P. Adversarial Examples for Evaluating Reading Comprehension Systems. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 2021–2031. [CrossRef]

31. Kembhavi, A.; Seo, M.; Schwenk, D.; Choi, J.; Farhadi, A.; Hajishirzi, H. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4999–5007.

32. Tapaswi, M.; Zhu, Y.; Stiefelhagen, R.; Torralba, A.; Urtasun, R.; Fidler, S. Movieqa: Understanding stories in movies through question-answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4631–4640.

33. Iyyer, M.; Manjunatha, V.; Guha, A.; Vyas, Y.; Boyd-Graber, J.; Daume, H.; Davis, L.S. The amazing mysteries of the gutter: Drawing inferences between panels in comic book narratives. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7186–7195.

34. Yagcioglu, S.; Erdem, A.; Erdem, E.; Ikizler-Cinbis, N. RecipeQA: A Challenge Dataset for Multimodal Comprehension of Cooking Recipes. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 1358–1368. [CrossRef]

35. Chen, D.; Fisch, A.; Weston, J.; Bordes, A. Reading Wikipedia to Answer Open-Domain Questions. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1870–1879. [CrossRef]

36. Hu, M. Research on Machine Reading Comprehension and Textual Question Answering. Ph.D. Thesis, National University of Defense Technology of China, Changsha, China, 2019.

37. Vanderwende, L. Answering and Questioning for Machine Reading. In Proceedings of the AAAI 2007 Spring Symposium, Palo Alto, CA, USA, 26–28 March, 2007; p. 91.

38. Saeidi, M.; Bartolo, M.; Lewis, P.; Singh, S.; Rocktäschel, T.; Sheldon, M.; Bouchard, G.; Riedel, S. Interpretation of Natural Language Rules in Conversational Machine Reading. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2087–2097. [CrossRef]

39. Nguyen, T.; Rosenberg, M.; Song, X.; Gao, J.; Tiwary, S.; Majumder, R.; Deng, L. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. *arXiv* **2016**, arXiv:1611.09268.

40. Liu, X.; Shen, Y.; Duh, K.; Gao, J. Stochastic Answer Networks for Machine Reading Comprehension. *arXiv* **2018**, arXiv:1611.09268.

41. Liu, X.; Li, W.; Fang, Y.; Kim, A.; Duh, K.; Gao, J. Stochastic Answer Networks for SQuAD 2.0. *arXiv* **2018**, arXiv:1809.09194.

42. Liu, X.; Duh, K.; Gao, J. Stochastic Answer Networks for Natural Language Inference. *arXiv* **2018**, arXiv:1804.07888.

43. Yin, Y.; Song, Y.; Zhang, M. Document-Level Multi-Aspect Sentiment Classification as Machine Comprehension. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017.

44. Asai, A.; Eriguchi, A.; Hashimoto, K.; Tsuruoka, Y. Multilingual Extractive Reading Comprehension by Runtime Machine Translation. *arXiv* **2018**, arXiv:1809.03275.

45. Zhang, Z.; Wu, Y.; Zhou, J.; Duan, S.; Wang, R. SG-Net: Syntax-Guided Machine Reading Comprehension. In Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2020), New York, NY, USA, 7–12 February 2020.

46. Minaee, S.; Kalchbrenner, N.; Cambria, E.; Nikzad, N.; Chenaghlu, M.; Gao, J. Deep Learning Based Text Classification: A Comprehensive Review. *arXiv* **2020**, arXiv:2004.03705.

47. Reddy, S.; Chen, D.; Manning, C.D. Coqa: A conversational question answering challenge. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 249–266. [CrossRef]

48. Onishi, T.; Wang, H.; Bansal, M.; Gimpel, K.; McAllester, D. Who did What: A Large-Scale Person-Centered Cloze Dataset. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 2230–2235. [CrossRef]

49. Welbl, J.; Stenetorp, P.; Riedel, S. Constructing datasets for multi-hop reading comprehension across documents. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 287–302. [CrossRef]

50. Dua, D.; Wang, Y.; Dasigi, P.; Stanovsky, G.; Singh, S.; Gardner, M. DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *arXiv* **2019**, arXiv:1903.00161.

51. Baudiš, P.; Šedivỳ, J. Modeling of the question answering task in the yodaqa system. In *International Conference of the Cross-Language Evaluation Forum for European Languages*; Springer: Berlin/Heidelberg, Germany, 2015; pp. 222–228.

52. Yang, Y.; Yih, W.T.; Meek, C. Wikiqa: A challenge dataset for open-domain question answering. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 2013–2018.

53. Miller, A.H.; Fisch, A.; Dodge, J.; Karimi, A.H.; Bordes, A.; Weston, J. Key-Value Memory Networks for Directly Reading Documents (EMNLP16). *arXiv* **2016**, arXiv:1606.03126.

54. Trischler, A.; Wang, T.; Yuan, X.; Harris, J.; Sordoni, A.; Bachman, P.; Suleman, K. NewsQA: A Machine Comprehension Dataset. In Proceedings of the 2nd Workshop on Representation Learning for NLP, Vancouver, BC, Canada, 30 July–4 August 2017; pp. 191–200. [CrossRef]

55. Boleda, G.; Paperno, D.; Kruszewski, G.; Lazaridou, A.; Pham, Q.N.; Bernardi, R.; Pezzelle, S.; Baroni, M.; Fernandez, R. 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), Berlin, Germany, 7–12 August 2016; pp. 1525–1534.

56. Hewlett, D.; Lacoste, A.; Jones, L.; Polosukhin, I.; Fandrianto, A.; Han, J.; Kelcey, M.; Berthelot, D. 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), Berlin, Germany, 7–12 August 2016; pp. 1535–1545. [CrossRef]

57. Bajgar, O.; Kadlec, R.; Kleindienst, J. Embracing data abundance: Booktest dataset for reading comprehension. *arXiv* **2016**, arXiv:1610.00956.

58. Soricut, R.; Ding, N. Building Large Machine Reading-Comprehension Datasets using Paragraph Vectors. *arXiv* **2016**, arXiv:1612.04342.

59. Joshi, M.; Choi, E.; Weld, D.; Zettlemoyer, L. 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), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 1601–1611. [CrossRef]

60. Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; Hovy, E. RACE: Large-scale ReAding Comprehension Dataset From Examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 9–11 September 2017; pp. 785–794. [CrossRef]

61. Dhingra, B.; Mazaitis, K.; Cohen, W.W. Quasar: Datasets for Question Answering by Search and Reading. *arXiv* **2017**, arXiv:1707.03904.

62. Dunn, M.; Sagun, L.; Higgins, M.; Guney, V.U.; Cirik, V.; Cho, K. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv* **2017**, arXiv:1704.05179.

63. Kočiskỳ, T.; Schwarz, J.; Blunsom, P.; Dyer, C.; Hermann, K.M.; Melis, G.; Grefenstette, E. The narrativeqa reading comprehension challenge. *Trans. Assoc. Comput. Linguist.* **2018**, *6*, 317–328. [CrossRef]

64. Welbl, J.; Liu, N.F.; Gardner, M. Crowdsourcing Multiple Choice Science Questions. In Proceedings of the 3rd Workshop on Noisy User-generated Text, Copenhagen, Denmark, 7 September 2017; pp. 94–106. [CrossRef]

65. Choi, E.; He, H.; Iyyer, M.; Yatskar, M.; Yih, W.T.; Choi, Y.; Liang, P.; Zettlemoyer, L. QuAC: Question Answering in Context. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 2174–2184. [CrossRef]

66. Rajpurkar, P.; Jia, R.; Liang, P. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, VIC, Australia, 2018; pp. 784–789. [CrossRef]

67. Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; Manning, C.D. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 2369–2380. [CrossRef]

68. Saha, A.; Aralikatte, R.; Khapra, M.M.; Sankaranarayanan, K. 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), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1683–1693. [CrossRef]

69. Xie, Q.; Lai, G.; Dai, Z.; Hovy, E. Large-scale Cloze Test Dataset Created by Teachers. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; Association for Computational Linguistics: Brussels, Belgium, 2018; pp. 2344–2356. [CrossRef]

70. Zhang, S.; Liu, X.; Liu, J.; Gao, J.; Duh, K.; Van Durme, B. Record: Bridging the gap between human and machine commonsense reading comprehension. *arXiv* **2018**, arXiv:1810.12885.

71. Šuster, S.; Daelemans, W. 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), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1551–1563. [CrossRef]

72. Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; Tafjord, O. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv* **2018**, arXiv:1803.05457.

73. Mihaylov, T.; Clark, P.; Khot, T.; Sabharwal, A. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. *arXiv* **2018**, arXiv:1809.02789.

74. Khot, T.; Sabharwal, A.; Clark, P. Scitail: A textual entailment dataset from science question answering. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018.

75. Khashabi, D.; Chaturvedi, S.; Roth, M.; Upadhyay, S.; Roth, D. 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), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 252–262. [CrossRef]

76. Park, D.; Choi, Y.; Kim, D.; Yu, M.; Kim, S.; Kang, J. Can Machines Learn to Comprehend Scientific Literature? *IEEE Access* **2019**, *7*, 16246–16256. [CrossRef]

77. Hong, Y.; Wang, J.; Zhang, X.; Wu, Z. Learning to Read Academic Papers; In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18), New Orleans, LA, USA, 2–7 February 2018.

78. Ostermann, S.; Modi, A.; Roth, M.; Thater, S.; Pinkal, M. 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, 7–12 May 2018; European Language Resources Association (ELRA): Miyazaki, Japan, 2018.

79. Dalvi, B.; Huang, L.; Tandon, N.; Yih, W.T.; Clark, P. Tracking State Changes in Procedural Text: A Challenge Dataset and Models for Process Paragraph 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), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: New Orleans, LA, USA, 2018; pp. 1595–1604. [CrossRef]

80. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural questions: A benchmark for question answering research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 453–466. [CrossRef]

81. Sun, K.; Yu, D.; Chen, J.; Yu, D.; Choi, Y.; Cardie, C. Dream: A challenge data set and models for dialogue-based reading comprehension. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 217–231. [CrossRef]

82. Talmor, A.; Herzig, J.; Lourie, N.; Berant, J. 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), New Orleans, LA, USA, 1–6 June 2018; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 4149–4158. [CrossRef]

83. Lin, C.Y. *ROUGE: A Package for Automatic Evaluation of Summaries*; Text Summarization Branches Out; Association for Computational Linguistics: Barcelona, Spain, 2004; pp. 74–81.

84. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA, USA, 7–12 July 2002; Association for Computational Linguistics: Philadelphia, PA, USA, 2002; pp. 311–318. [CrossRef]

85. Banerjee, S.; Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Ann Arbor, MI, USA, 25–30 June 2005; Association for Computational Linguistics: Ann Arbor, MI, USA, 2005; pp. 65–72.

86. Denkowski, M.; Lavie, A. METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support for Five Target Languages. In Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, 15–16 July 2010; Association for Computational Linguistics: Uppsala, Sweden, 2010; pp. 339–342.

87. Denkowski, M.; Lavie, A. Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; Association for Computational Linguistics: Edinburgh, UK, 2011; pp. 85–91.

88. Van Rijsbergen, C.J. *Information Retrieval*; Butterworth: Ann Arbor, MI , USA, 1979.

89. Yuen, M.C.; King, I.; Leung, K.S. A survey of crowdsourcing systems. In Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing, Boston, MA, USA, 9–11 October 2011; pp. 766–773.

90. Sugawara, S.; Aizawa, A. An analysis of prerequisite skills for reading comprehension. In Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods, Austin, TX, USA, 5 November 2016; pp. 1–5.

91. Sugawara, S.; Kido, Y.; Yokono, H.; Aizawa, A. Evaluation metrics for machine reading comprehension: Prerequisite skills and readability. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 806–817.

92. Yan, M.; Xia, J.; Wu, C.; Bi, B.; Zhao, Z.; Zhang, J.; Si, L.; Wang, R.; Wang, W.; Chen, H. A deep cascade model for multi-document reading comprehension. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 7354–7361.

93. Wang, Y.; Liu, K.; Liu, J.; He, W.; Lyu, Y.; Wu, H.; Li, S.; Wang, H. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; Association for Computational Linguistics: Melbourne, Australia, 2018; pp. 1918–1927. [CrossRef]

94. Bowman, S.R.; Angeli, G.; Potts, C.; Manning, C.D. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Lisbon, Portugal, 17–21 September 2015; Association for Computational Linguistics: Stroudsburg, PA, USA, 2015.

95. Speer, R.; Chin, J.; Havasi, C. Conceptnet 5.5: An open multilingual graph of general knowledge. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.

96. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Lawrence Zitnick, C.; Parikh, D. Vqa: Visual question answering. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.

97. Hart, M. Project Gutenberg. 1971. Available online: https://www.gutenberg.org (accessed on 24 October 2020).

98. NIST. The Text REtrieval Conference (TREC). Available online: https://trec.nist.gov/ (accessed on 24 October 2020).

99.   Raison, M.; Mazaré, P.E.; Das, R.; Bordes, A. Weaver: Deep co-encoding of questions and documents for machine reading. *arXiv* **2018**, arXiv:1804.10490.

100.  Sampath, R.; Ma, P. QA with Wiki: Improving Information Retrieval and Machine Comprehension. Available online: https://web.stanford.edu/class/cs224n/reports/custom/15737727.pdf (accessed on 24 October 2020).

101.  Callan, J.; Hoy, M.; Yoo, C.; Zhao, L. The ClueWeb09 Dataset. 2009. Available online: https://lemurproject.org/clueweb09/ (accessed on 24 October 2020).

102.  Graff, D.; Kong, J.; Chen, K.; Maeda, K. English gigaword. *Linguist. Data Consortium Phila* **2003**, *4*, 34.

103.  Kemmerer, D. *Cognitive Neuroscience of Language*; Psychology Press: London, UK, 2014.

104.  Barsalou, L.W. Grounded cognition. *Annu. Rev. Psychol.* **2008**, *59*, 617–645. [CrossRef]

105.  Pecher, D.; Zwaan, R.A. *Grounding Cognition: The Role of Perception and Action in Memory, Language, and Thinking*; Cambridge University Press: Cambridge, UK, 2005.

106.  Semin, G.R.; Smith, E.R. *Embodied Grounding: Social, Cognitive, Affective, and Neuroscientific Approaches*; Cambridge University Press: Cambridge, UK, 2008.

107.  Gibbs, R.W., Jr. *Embodiment and Cognitive Science*; Cambridge University Press: Cambridge, UK, 2005.

108.  Barsalou, L.W. Perceptual symbol systems. *Behav. Brain Sci.* **1999**, *22*, 577–660. [CrossRef]

109.  Shapiro, L. *Embodied Cognition*; Routledge: London, UK, 2019.

110.  Storks, S.; Gao, Q.; Chai, J.Y. Commonsense reasoning for natural language understanding: A survey of benchmarks, resources, and approaches. *arXiv* **2019**, arXiv:1904.01172.

111.  Kincaid, J.P.; Fishburne, R.P., Jr.; Rogers, R.L.; Chissom, B.S. *Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel*; Butterworth: Ann Arbor, MI, USA, 1975.

112.  Benzahra, M.; Yvon, F. Measuring Text Readability with Machine Comprehension: A Pilot Study. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 28 July–2 August 2019.

113.  Hough, A.R.; Gluck, K. The understanding problem in cognitive science. In Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems, Cambridge, MA, USA, 2–5 August 2019; pp. 55–74.

114.  Jackson, P.C., Jr. *Toward Human-Level Knowledge Representation with a Natural Language of Thought*; Dover Publications: Mineola, NY, USA, 2020.

115.  Fodor, J.A. *The Language of Thought*; Harvard University Press: Cambridge, MA, USA, 1975; Volume 5.

116.  Smith, E.E. Theories of semantic memory. *Handbook Learn. Cogn. Process.* **1978**, *6*, 1–56.

117.  Pylyshyn, Z.W. *Computation and Cognition*; MIT Press: Cambridge, MA, USA, 1984.

118.  Wang, X.; Men, W.; Gao, J.; Caramazza, A.; Bi, Y. Two Forms of Knowledge Representations in the Human Brain. *Neuron* **2020**. [CrossRef]