Natural Language Engineering (2022), **28**, pp. 683–732
doi:10.1017/S1351324921000395

CAMBRIDGE UNIVERSITY PRESS

SURVEY PAPER

A Survey on Machine Reading Comprehension Systems

Razieh Baradaran**, Razieh Ghiasi** and Hossein Amirkhani*

Computer and Information Technology Department, University of Qom, Qom, Iran
*Corresponding author. E-mail: amirkhani@qom.ac.ir
**Equal contribution

(Received 20 July 2019; revised 27 October 2021; accepted 8 November 2021; first published online 19 January 2022)

Abstract

Machine Reading Comprehension (MRC) is a challenging task and hot topic in Natural Language Processing. The goal of this field is to develop systems for answering the questions regarding a given context. In this paper, we present a comprehensive survey on diverse aspects of MRC systems, including their approaches, structures, input/outputs, and research novelties. We illustrate the recent trends in this field based on a review of 241 papers published during 2016–2020. Our investigation demonstrated that the focus of research has changed in recent years from answer extraction to answer generation, from single- to multi-document reading comprehension, and from learning from scratch to using pre-trained word vectors. Moreover, we discuss the popular datasets and the evaluation metrics in this field. The paper ends with an investigation of the most-cited papers and their contributions.

Keywords: Natural Language Processing; question answering; Machine Reading Comprehension; deep learning; literature review

1. Introduction

Machine Reading Comprehension (MRC) is a challenging task in Natural Language Processing (NLP) aimed to evaluate the extent to which machines achieve the goal of natural language understanding. In order to assess the comprehension of a machine of a piece of natural language text, a set of questions about the text is given to the machine, and the responses of the machine are evaluated against the gold standard. Nowadays, MRC is known as the research area of reading comprehension for machines based on question answering (QA). Each instance in MRC datasets contains a context C, a related question Q, and an answer A. Figure 1 shows some examples of the SQuAD (Rajpurkar *et al.* 2016) and CNN/Daily Mail (Hermann *et al.* 2015) datasets. The goal of MRC systems is to learn the predictive function f that extracts/generates the appropriate answer A based on the context C and the related question Q:

$$f: (C, Q) \rightarrow A$$

Furthermore, MRC systems have important applications in distinct areas, such as conversational agents (Hewlett, Jones and Lacoste 2017; Reddy, Chen and Manning 2019) and customer service support (Cui *et al.* 2017a).

Although MRC is referred to as QA in some studies, these two concepts are different in the following ways:

- The main objective of QA systems is to answer the input questions, while the main goal of an MRC system, as the name suggests, is to demonstrate the machine's ability in understanding natural languages through answering questions about specific context that it reads.

SQuAD dataset	CNN/Daily Mail dataset
Context In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called \showers"	**Context** (@entity4) if you feel a ripple in the force today, it may be the news that the official @entity6 is getting its first gay character. according to the sci-fi website @entity9, the upcoming novel "@entity11" will feature a capable but flawed @entity13 official named @entity14 who " also happens to be a lesbian." The character is the first gay figure in the official @entity6 – the movies, television shows, comics and books approved by @entity6 franchise owner @entity22 -- according to @entity24, editor of "@entity6" books at @entity28 imprint @entity26.
Question Where do water droplets collide with ice crystals to form precipitation?	**Question** characters in "@placeholder" movies have gradually become more diverse.
Answer within a cloud	**Answer** @entity6

Figure 1. Samples from SQuAD (Rajpurkar *et al.* 2016) and CNN/Daily Mail (Chen, Bolton and Manning 2016) datasets. The original article of the CNN/Daily Mail example can be found at https://edition.cnn.com/2015/03/10/entertainment/feat-star-wars-gay-character

- The only input to QA systems is the question, while the inputs to MRC systems entail the question and the corresponding context, which should be used to answer the question. As a result, sometimes MRC is referred to as QA from the text (Deng and Liu 2018).
- The main information source used to answer questions in MRC systems is natural language texts, while in QA systems, the structured and semi-structured data sources, such as knowledge-based ones, are commonly applied, in addition to the non-structured data like texts.

1.1. History

The history of reading comprehension for machines dates back to the 1970s when researchers identified it as a convenient way to test computer comprehension ability. One of the most prominent early studies was the QUALM system (Lehnert 1977). This system was limited to handwritten scripts and could not be easily generalized to larger domains. Due to the complexity of this task, research in this area was reduced in 1980s and 1990s. In the late 1990s, Hirschman *et al.* (1999) revived the field of MRC by creating a new dataset, including 120 stories and questions from 3rd- to 6th-grade material, followed by a workshop on comprehension test as a tool for assessing machine comprehension at ANLP/NAACL 2000.

Another revolution in this field occurred between 2013 and 2015 by introducing labeled training datasets mapping (context, question) pairs to the answer. This transformed the MRC problem into a supervised learning task (Chen 2018). Two prominent datasets in this period were the MCTest dataset (Richardson, Burges and Renshaw 2013) with 500 stories and 2000 questions and the ProcessBank dataset (Berant *et al.* 2014) with 585 questions over 200 paragraphs related to biological processes. In 2015, the introduction of large datasets such as CNN/Daily Mail (Hermann *et al.* 2015) and SQuAD (Rajpurkar *et al.* 2016) opened a new window in the MRC field by allowing the development of deep models.

In recent years, with the success of machine learning techniques, especially the neural networks, and the usage of recurrent neural networks to process sequential data such as texts, MRC has become an active area in the field of NLP. The goal of this paper is to categorize these studies, provide related statistics, and show the trends in this field. Some recent surveys focused on QA systems (Bouziane *et al.* 2015; Kodra and Meçe 2017). Some papers presented a partial survey on some MRC systems but did not provide a comprehensive classification of different aspects

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 1. Number of reviewed papers over different years.

Year	Number of papers
2016	13
2017	34
2018	62
2019	83
2020	49
Total	241

and different statistics in this field (Arivuchelvan and Lakahmi 2017; Zhang *et al.* 2019a). Ingale and Singh only provided a review on MRC datasets (Ingale and Singh 2019). Liu *et al.* (2019c) provide a review on different aspects of neural MRC models including definitions, differences, popular datasets, architectures, and new trends based on 85 papers. In our study, we present a comprehensive review on 241 papers, analyzing and categorizing MRC studies from different aspects including problem-solving approaches, input/output, model structures, research novelties, and datasets. We also provide statistics on the amount of research attention to these aspects in different years, which are not provided in previous reviews.

1.2. Outline

In order to select papers, the queries "reading comprehension," "machine reading," "machine reading comprehension," and "machine comprehension" were submitted to the Google Scholar service[a]. Also, the ACL Anthology website[b], which includes top related NLP conferences such as *ACL*, *EMNLP*, *NAACL*, and *CoNLL,* was searched with the same queries to extract the remaining related papers. We excluded the retrieved papers that were published only on *arXiv* as well as the QA papers with no novelty in the MRC phase. We also excluded the papers with the *conversational* or *dialogue MRC* as subjects because these papers focus on multi-turn QA in conversational contexts with different challenges (Gupta, Rawat and Yu 2020). We limited our study to the papers published in recent years, that is, from 2016 to September 2020. Table 1 shows the number of reviewed papers over different years.

The contributions of this paper are as follows:

- Investigating recently published MRC papers from different perspectives including problem-solving approaches, system input/outputs, contributions of these studies, and evaluation metrics.
- Providing statistics for each category over different years and highlighting the trends in this field.
- Reviewing available datasets and classifying them based on important factors.
- Discussing specific aspects of the most-cited papers such as their contribution, their citation number, and year and venue of their publication.

The rest of this paper is organized as follows. Section 2 focuses on the main problem-solving approaches for the MRC task. The review of the papers based on the basic phases of an MRC

[a] https://scholar.google.com/
[b] https://www.aclweb.org/anthology/

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

system is presented in Section 3. Section 4 provides an analysis of the type of input/outputs of MRC systems. The recent datasets and evaluation measures are reviewed in Sections 5 and 6, respectively. In Section 7, the MRC studies are categorized based on their contributions and novelties. The most-cited papers are investigated in Section 8. Section 9 provides future trends and opportunities. Finally, the paper is concluded in Section 10.

2. Problem-solving approaches

The approaches used for developing MRC systems can be grouped into three categories: rule-based methods, classical machine learning-based methods, and deep learning-based methods.

The traditional rule-based methods use the rules handcrafted by linguistic experts. These methods suffer from the problem of the incompleteness of the rules. Also, this approach is domain-specific where for any new domain a new set of rules should be handcrafted. As an example, Riloff and Thelen (2000) present a rule-based MRC system called Quarc, which reads a short story and answers the input question by extracting the most relevant sentences. Quarc uses a separate set of rules for each question type (WHO, WHAT, WHEN, WHERE, and WHY). In this system, several NLP tools are used for parsing, part of speech tagging, morphological analysis, entity recognition, and semantic class tagging. As another example, Akour *et al.* (2011) introduce the QArabPro system, which is a system for answering reading comprehension questions in the Arabic language. It is also developed using a set of rules for each type of question and uses multiple NLP components, including question classification, query reformulation, stemming, and root extraction.

The second approach is based on the classical machine learning. These methods rely on a set of human-defined features and train a model for mapping input features to the output. Note that in classical machine learning-based methods, even though the handcrafted rules are not necessary, feature engineering is a critical necessity.

For example, Ng, Teo and Kwan (2000) have developed a machine learning-based MRC system and introduced some of features to be extracted from a context sentence like "the number of matching words/verb types between the question and the sentence," "the number of matching words/verb types between the question and the previous/next sentence," "co-reference information," and binary features like "sentence-contain-person," "sentence-contain-time," "sentence-contain-location," "sentence-is-title," and so on

The third approach uses deep learning methods to learn features from raw input data automatically. These methods require a large amount of training data to create high accuracy models. Because of the growth of available data and computational power in recent years, deep learning methods have gained state-of-the-art results in many tasks. In the MRC task, most of the recent research falls into this category. Two main deep learning architectures used by MRC researchers are the *Recurrent Neural Network* (RNN) and *Convolutional Neural Network* (CNN).

RNNs are often used for modeling sequential data by iterating through the sequence elements and maintaining a state containing information relative to what have seen so far. Two common types of RNNs are Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) and Gated Recurrent Unit (GRU) (Cho *et al.* 2014) in uni-directional and bi-directional versions (Chen *et al.* 2016; Kobayashi *et al.* 2016; Chen *et al.* 2017; Dhingra *et al.* 2017; Seo *et al.* 2017; Clark and Gardner 2018; Ghaeini *et al.* 2018; Hoang, Wiseman and Rush 2018; Hu *et al.* 2018a; Liu *et al.* 2018a). In MRC systems, like other NLP tasks, these architectures have been commonly used in different parts of the pipeline, such as for representing questions and contexts (Chen *et al.* 2016; Kobayashi *et al.* 2016; Chen *et al.* 2017; Dhingra *et al.* 2017; Seo *et al.* 2017; Clark and Gardner 2018; Ghaeini *et al.* 2018; Hoang *et al.* 2018; Hu *et al.* 2018a; Liu *et al.* 2018a) or in higher levels of the MRC system such as the modeling layer (Choi *et al.* 2017b; Seo *et al.* 2017; Li, Li and Lv 2018). In recent years, the attention-based Transformer (Vaswani *et al.* 2017) has emerged as a powerful alternative to the RNN architecture. For more detailed information, refer to Section 3.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 2. Statistics of different embedding methods used by reviewed papers.

Year	Word embedding	Hybrid (word-char embedding)	Sentence embedding
2016	100%	0%	5%
2017	56%	40%	4%
2018	45%	54%	6%
2019	67%	33%	3%
2020	86%	14%	18%
All	64%	35%	7%

CNN is a type of deep learning model that is universally used in computer vision applications. It utilizes layers with convolution filters that are applied to local spots of their inputs (LeCun *et al*. 1998). CNN models have subsequently been shown to be effective for NLP and have achieved excellent results in various NLP tasks (Kim 2014). In MRC systems, CNN is used in the embedding phase (especially, character embedding) (Seo *et al*. 2017; Indurthi *et al*. 2018) as well as in higher-level phases (introduced in Section 3) for modeling interactions between the question and passage like in the QANet (Yu *et al*. 2018). QANet uses CNN and self-attention blocks instead of the RNN, which results in faster answer span detection on the SQuAD dataset (Rajpurkar *et al*. 2016).

3. MRC phases

Most of the recent deep learning-based MRC systems have the following phases: *embedding phase*, *reasoning phase,* and *prediction phase*. Many of the reviewed papers focus on developing new structures for these phases, especially the reasoning phase.

3.1. Embedding phase

In this phase, input characters, words, or sentences are represented by real-valued dense vectors in a meaningful space. The goal of this phase is to provide question and context embedding. Different levels of embedding are used in MRC systems. Character-level and word-level embeddings can capture the properties of words, subwords, and characters, and higher-level representations can represent syntactic and semantic information of input text. Table 2 shows the statistics of various embedding methods used in the reviewed papers. Since there is not any paper that uses the character embedding as the only embedding method, there is no character embedding column in this table. For a complete list of papers categorized based on their embedding methods, refer to Table A1.

3.1.1. Character embedding

Some papers use character embedding as part of their embedding phase. This type of embedding is useful to overcome unknown and rare words problems (Dhingra *et al*. 2016; Seo *et al*. 2017). To generate the input representation, deep neural network models are commonly used. Inspired by Kim's work (Kim 2014), some papers have used CNN models to embed the input characters (Seo *et al*. 2017; Zhang *et al*. 2017; Gong and Bowman 2018; Kundu and Ng 2018a). Some other papers have used character level information captured from the final state of an RNN model like LSTM (or Bi-LSTM) and GRU (or Bi-GRU) (Wang, Yuan and Trischler 2017a; Yang *et al*. 2017a; Du and Cardie 2018; Hu *et al*. 2018a; Wang *et al*. 2018b). As another approach which uses both CNN and LSTM to embed input characters, LSTM-char CNN (Kim *et al*. 2016) is also used in

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

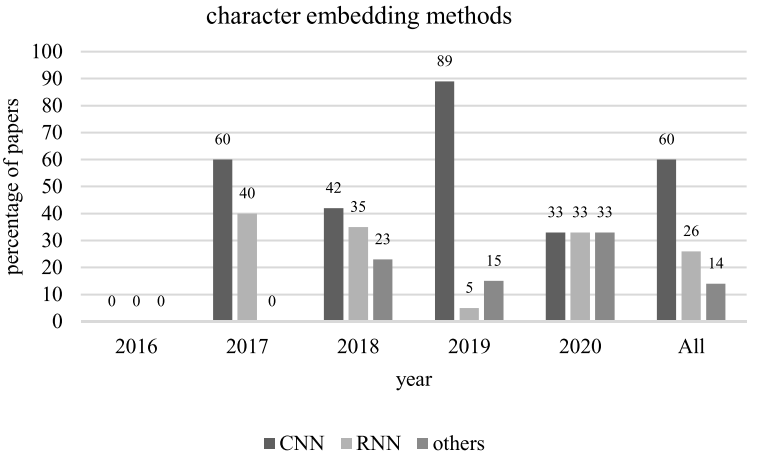

Figure 2. The percentage of different character embedding methods over different years

MRC literature (Prakash, Tripathy and Banu 2018). We classify these papers in two categories, CNN and RNN, and so the sum of percentages is greater than 100% in Figure 2. This figure shows the percentage of different character embedding methods over different years. Other methods include skip-gram, n-grams, and more recent methods like ELMo (Peters *et al.* 2018). It is clear that the use of CNN has been consistently higher than the use of RNN for character embedding in different years.

3.1.2. Word embedding

Word embedding is to represent the words or subwords in a numeric vector space, which is performed by two main approaches: 1) non-contextual embedding and 2) contextual embedding.

Non-contextual word embedding

Non-contextual embeddings present a single general representation for each word, regardless of its context. There are three main non-contextual embeddings: 1) one hot encoding, 2) learning word vectors jointly with the main task, and 3) using pre-trained word vectors (fixed or fine-tuned).

One hot encoding is the most basic way to turn a token into a vector. These are binary, sparse, and very high dimensional vectors with the size of the vocabulary (the number of unique words in the corpus). To represent a word *w*, all the vector elements are set to zero, except the one which identifies *w*. This approach has been less popular than other approaches in recent papers (Cui *et al.* 2016; Liu and Perez 2017).

Another popular way to represent words is learned word vectors, which delivers dense real-valued representations. In the presence of a large amount of training data, it is advised to learn the word vectors from scratch jointly with the main task (Chollet 2017).

Some studies have shown that initializing word vectors with pre-trained values results in better accuracies than random initialization (Dhingra *et al.* 2017; Ren, Cheng and Su 2020; Wang *et al.* 2020a; Zhou, Luo and Wu 2020b). This approach is especially useful in low-data scenarios (Chollet 2017; Dhingra *et al.* 2017). GloVe embedding (Pennington, Socher and Manning 2014) is the most common pre-trained word representation among non-contextual representations, used in the reviewed papers (Chen *et al.* 2016; Yin, Ebert and Schütze 2016; Chen *et al.* 2017; Liu *et al.* 2017; Wang *et al.* 2017a; Xiong, Zhong and Socher 2017; Gong and Bowman 2018; Wang *et al.* 2018b). Word2Vec (Mikolov *et al.* 2013) is another word embedding used in this task (Kobayashi *et al.* 2016; Chaturvedi, Pandit and Garain 2018; Šuster and Daelemans 2018). These pre-trained

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 3. Statistics of different word representation methods in the reviewed papers.

Year	Non-contextual				Contextual			
	One hot encoding	Learned word embedding	Fixed pre-trained	Fine-tuned	Learned word embedding		Fixed pre-trained	Fine-tuned
					RNN	CNN		
2016	11%	56%	22%	22%	67%	11%	0%	0%
2017	4%	12%	36%	40%	41%	0%	0%	0%
2018	4%	8%	29%	18%	33%	6%	4%	0%
2019	0%	0%	38%	21%	0%	0%	5%	45%
2020	9%	14%	27%	14%	50%	0%	23%	36%
All	4%	9%	38%	27%	46%	2%	4%	21%

word vectors are fine-tuned (Chen *et al.* 2016; Kobayashi *et al.* 2016; Zhang *et al.* 2017; Clark and Gardner 2018; Liu *et al.* 2018c; Šuster and Daelemans 2018) or left as fixed vectors (Seo *et al.* 2017; Shen *et al.* 2017; Weissenborn, Wiese and Seiffe 2017; Gong and Bowman 2018). Fine-tuning some keywords such as "what," "how," "which," and "many" could be crucial for QA systems, while most of the pre-trained word vectors can be kept fixed (Chen *et al.* 2017). Table 3 shows the statistics of these approaches in different years. Finally, it is worth noting that some papers use hand-designed word features such as named entity (NE) tag and part of speech (POS) tag along with embedding of words (Huang *et al.* 2018).

Contextual word embedding

Despite the relative success of non-contextual embeddings, they are static, so all meanings of a word are represented with a fixed vector (Ethayarajh 2019). Different from static word embeddings, contextual embeddings move beyond word-level semantics and represent each word considering its context (surrounding words). To obtain the context-based representation of the words, two approaches can be adopted: 1) learning the word vectors jointly with the main task and 2) using pre-trained contextual word vectors (fixed or fine-tuned).

For learning the contextual word vectors, a sequence modeling method, usually an RNN, is used. For example, Chen *et al.* (2017) used a multi-layer Bi-LSTM model for this purpose. In Yang, Kang and Seo (2020) study, forward and backward GRU hidden states are combined to generate contextual representations of query and context words. Bajgar, Kadlec and Kleindienst (2017) used different approaches for the query and context words, where the combination of forward and backward GRU hidden states is exploited for representing the context words, while the final hidden state of GRU is used for the query. On the other hand, pre-trained contextualized embeddings such as ELMo (Peters *et al.* 2018), BERT (Devlin *et al.* 2018), and GPT (Radford *et al.* 2018) are deep neural language models that are trained on large unlabeled corpuses. The ELMo method (Peters *et al.* 2018) obtains the contextualized embeddings by a 2-layer Bi-LSTM, while BERT (Devlin *et al.* 2018) and GPT (Radford *et al.* 2018) are bi-directional and uni-directional Transformer-based (Vaswani *et al.* 2017) language models, respectively. These embeddings are used either besides other embeddings (Hu *et al.* 2018a; Hu *et al.* 2018b; Seo *et al.* 2018; Lee and Kim 2020; Ren *et al.* 2020) or alone (Bauer, Wang and Bansal 2018; Zheng *et al.* 2020).

Due to the success of the contextual word embeddings in many NLP tasks, there is a clear trend toward using these embeddings in recent years (Bauer *et al.* 2018; Hu *et al.* 2018a; Hu *et al.* 2018b; Wang and Bansal 2018). This is obvious in Table 3, where the use of fixed pre-trained

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

and fine-tuned contextual embeddings has increased from 0% in 2016 to 23% and 36% in 2020, respectively. Note that some papers use multiple methods, so the sum of percentages in the tables may be greater than 100%.

3.1.3. Hybrid word-character embedding

The combination of word embedding and character embedding is used in some reviewed papers (Seo *et al.* 2017; Yang *et al.* 2017a; Zhang *et al.* 2017; Gong and Bowman 2018). Hybrid embedding tries to use the strengths of both word and character embeddings. A simple approach is to concatenate the word and character embeddings. As an example, Lee and Kim (2020) used GloVe as the word embedding and the output of the CNN model as the character embedding.

This approach suffers from a potential problem. Word embedding has better performance for frequent words (subwords), while it can have negative effects for representing rare words (subwords). The reverse is true for character embedding (Yang *et al.* 2017a). To solve this problem, some researchers introduced a gating mechanism which regulates the flow of information. Yang *et al.* (2017a) used a fine-grained gating mechanism for dynamic concatenation of word and characters embedding. This mechanism uses a gate vector, which is a linear multiplication of word features (POS and NE), to control the flow of information of word and character embeddings. Seo *et al.* (2017) used highway networks (Srivastava, Greff and Schmidhuber 2015) for embedding concatenation. These networks use the gating mechanism learned by the LSTM network. According to Table 2, the use of hybrid embedding in reviewed papers has increased from 0% in 2016 to 54% in 2018. However, with the success of language model-based contextual embeddings, the direct combination of character and word embeddings has decreased thereafter.

3.1.4. Sentence embedding

Sentence embedding is a high-level representation in which the entire sentence is encoded in a single vector. It is often used along with other embeddings (Yin *et al.* 2016). However, sentence embedding is not so popular in MRC systems, because the answer is often a sentence part, not the whole sentence.

3.2. Reasoning phase

The goal of this phase is to match the input query (question) with the input document (context). In other words, this phase determines the related parts of the context for answering the question by calculating the relevance between question and context parts. Recently, Phrase Indexed Question Answering (PIQA) model (Seo *et al.* 2018) enforces complete independence between document encoder and question encoder and does not include any cross attention between question and document. In this model, each document is processed beforehand, and its *phrase index vectors* are generated. Then, at inference time, the answer is obtained by retrieving the nearest indexed phrase vector to the query vector.

The attention mechanism (Bahdanau, Cho and Bengio 2015), originally introduced for machine translation, is used for this phase. In recent years, with the advent of attention-based Transformer architecture (Vaswani *et al.* 2017) as an alternative to common sequential structures like RNN, new Transformer-based language models, such as BERT (Devlin *et al.* 2018) and XLNet (Yang *et al.* 2019b), have been introduced. They are used as the basis for new state-of-the-art results in MRC task (Sharma and Roychowdhury 2019; Su *et al.* 2019; Yang *et al.* 2019a; Tu *et al.* 2020; Zhang *et al.* 2020a; Zhang *et al.* 2020b) by adding or modifying final layers and fine-tuning them on the target task.

The attention mechanism used in MRC systems can be explored in three perspectives: *direction*, *dimension,* and *number of steps.* For the statistics, refer to Table 4.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 4. Statistics of different attention mechanisms used in the reasoning phase of MRC systems.

Year	Direction		Dimension		Number of steps		
	One direction	Two direction	One dimension	Two dimension	Single	Multi-fixed	Multi-dynamic
2016	89%	11%	78%	22%	89%	22%	0%
2017	58%	42%	10%	90%	83%	12%	5%
2018	51%	49%	21%	83%	76%	22%	2%
2019	12%	89%	2%	98%	44%	56%	4%
2020	23%	77%	18%	82%	55%	46%	0%
All	35%	65%	15%	86%	61%	36%	3%

3.2.1. Direction

Some research only uses the context-to-query (C2Q) attention vector (Cui *et al*. 2016; Wang and Jiang 2017; Weissenborn *et al*. 2017; Huang *et al*. 2018) called *one-directional* attention mechanism. It signifies which query words are relevant to each context word (Cui *et al*. 2017b; Seo *et al*. 2017).

In *bi-directional* attention mechanism, query-to-context (Q2C) attention weights are also calculated (Cui *et al*. 2017b; Liu *et al*. 2017; Min, Seo and Hajishirzi 2017; Seo *et al*. 2017; Xiong *et al*. 2017; Clark and Gardner 2018) along with C2Q. It signifies which context words have the closest similarity to one of the query words and are hence critical for answering the question (Cui *et al*. 2017b; Seo *et al*. 2017). In Transformer-based MRC models like BERT-based models, the question and context are processed as one sequence, so the attention mechanism can be considered as bi-directional attention. As shown in Table 4, the ratio of bi-directional attention usage has increased in recent years.

3.2.2. Dimension

There are two attention dimensions in the reviewed papers: *one-dimensional* and *two-dimensional* attentions. In one-dimensional attention, the whole question is represented by one embedding vector, which is usually the last hidden state of the contextual embedding (Chen *et al*. 2016; Kadlec *et al*. 2016b; Dhingra *et al*. 2017; Shen *et al*. 2017; Weissenborn *et al*. 2017). It does not pay more attention to important question words. On the contrary, in two-dimensional attention, every word in the query has its own embedding vector (Chen *et al*. 2017; Cui *et al*. 2017b; Seo *et al*. 2017; Yang *et al*. 2017b; Clark and Gardner 2018).

According to Table 4, 86% of all reviewed papers use two-dimensional attention. Also, the use of two-dimensional attention has increased over recent years.

3.2.3. Number of steps

According to the number of reasoning steps, three types of MRC systems can be seen: *single-step reasoning*, *multi-step reasoning with a fixed number of steps,* and *dynamic multi-step reasoning*.

In the single-step reasoning, question and passage matching is done in a single step. However, the obtained representation can be processed through multiple layers to extract or generate the answer (Chen *et al*. 2016; Seo *et al*. 2017; Clark and Gardner 2018). In multi-step reasoning, question and passage matching is done in multiple steps such that the question-aware context representation is updated by integrating the intermediate information in each step. The number of steps can be static (Yang *et al*. 2017b; Hu *et al*. 2018a) or dynamic (Dhingra *et al*. 2017; Shen *et al*. 2017; Song *et al*. 2018). Dynamic multi-step reasoning uses a termination module to

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 5. Statistics of different prediction phase categories in the reviewed papers.

Year	Extraction mode		Generation mode	
	Span detection	Candidate ranking	Answer generation	Candidate ranking
2016	40%	50%	0%	10%
2017	70%	26%	4%	0%
2018	66%	20%	8%	8%
2019	69%	5%	16%	19%
2020	59%	23%	32%	23%
All	65%	18%	13%	13%

decide whether the inferred information is sufficient for answering or more reasoning steps are still needed. Therefore, the number of reasoning steps in this model depends on the complexity of the passage and question. It is obvious that in multi-step reasoning, the model complexity increases by the number of reasoning steps. In the Transformer-based MRC models, the number of steps is fixed and depends on the number of layers.

According to Table 4, about 61% of reviewed papers use single-step reasoning, but the popularity of multi-step reasoning has increased over recent years. For a detailed list of the used reasoning methods in different papers, refer to Table A2.

3.3. Prediction phase

The final output of an MRC system is specified in the prediction phase. The output can be extracted from context or generated according to context. In generation mode, a decoder module generates answer words one by one (Hewlett *et al.* 2017). In some cases, multiple choices are presented to the system, and it must select the best answer according to the question and passage(s) (Greco *et al.* 2016). These multi-choice systems can be seen in both extractive and generative models based on whether the answer choices occur in the passage or not.

The extraction mode is implemented in different forms. If the answer is a span of context, the start and end indices of the span are predicted in many studies by estimating the probability distribution of indices over the entire context (Chen *et al.* 2017; Wang and Jiang 2017; Xiong *et al.* 2017; Yang *et al.* 2017a; Yang *et al.* 2017b; Clark and Gardner 2018).

In some studies, the candidate chunks (answers) are extracted first, which are ranked by a trained model. These chunks can be sentences (Duan *et al.* 2017; Min *et al.* 2017) or entities (Sachan and Xing 2018). In the Ren *et al.* (2020) study, after extracting the candidate chunks from various contexts, a linear transformation is used along with a sigmoid function to compute the score of the answer candidates.

Table 5 shows the statistics of these categories in the reviewed papers. It is clear that most papers (65%) extract the answer span in the passage(s). It seems that developing rich span-based datasets like SQuAD (Rajpurkar *et al.* 2016) is the reason for this popularity. Also, the generation-based papers have increased from 10% in 2016 to 55% in 2020 (sum of the *answer generation* and *candidate ranking* columns in the *generation mode*). For more details, refer to Table A3.

4. Input/Output-based analysis

4.1. MRC systems input

The inputs to an MRC system are question and passage texts. The passage is often referred to as context. Moreover, in some systems, the candidate answer list is part of the input.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

4.1.1. Question

Input questions can be grouped into three categories: f*actoid* questions, n*on-factoid* questions, and y*es/no* questions.

Factoid questions are questions that can be answered with simple facts expressed in short text answers like a personal name, temporal expression, or location (Jurafsky and Martin 2019). For example, the answer to the question "Who founded Virgin Airlines?" is a personal name; or questions "What is the average age of the onset of autism?" and "Where is Apple Computer based?" have number and location as an answer, respectively. In other words, the answers to factoid questions are one or more entities or a short expression. Because of its simplicity compared to other types, most research in MRC literature has focused on this type of question (Chen *et al*. 2017; Seo *et al*. 2017; Clark and Gardner 2018; Huang *et al*. 2018).

Non-factoid questions, on the other hand, are open-ended questions that usually require long and complex passage-level answers, such as descriptions, opinions, and explanations (Hashemi *et al*. 2020). For example, "Why does Queen Elizabeth sign her name Elizabeth?" "What is the difference between MRC and QA?" and "What do you think about MRC?" are instances of non-factoid questions. In our reviewed papers, 32% of works focus on non-factoid questions. Because of their difficulty, the systems dealing with non-factoid questions have often lower accuracies (Wang *et al*. 2016; Tan *et al*. 2018a; Wang *et al*. 2018a).

Yes/No questions, as indicated by their name, have yes or no as answers. According to our investigation, the papers which deal with this type of question consider other types of questions as well (Li *et al*. 2018; Liu *et al*. 2018b; Zhang *et al*. 2018).

Refer to Table 6 for the statistics of input/output types in MRC systems. It is clear from the table that the popularity of non-factoid and yes/no questions is increased. Note that since some papers focus on multiple question types, the sum of percentages is greater than 100% in this table.

4.1.2. Context

The input context can be a *single* passage or *multiple* passages. It is obvious that as the context gets longer, finding the answer becomes harder and more time-consuming. Until now, most of the papers have focused on a single passage (Seo *et al*. 2017; Wang *et al*. 2017b; Yang *et al*. 2017a; Yang *et al*. 2017b; Zhang *et al*. 2017). But multiple passages MRC systems are becoming more popular (Xie and Xing 2017; Huang *et al*. 2018; Liu *et al*. 2018b; Wang *et al*. 2018b). According to Table 6, only 4% of the reviewed papers focused on multiple passages in 2016, but this ratio has increased in recent years reaching 52% and 32% in 2019 and 2020, respectively.

4.2. MRC systems output

The output of MRC systems can be classified into two categories: *abstractive (generative)* output and *extractive (selective)* output.

In the abstractive mode, the answer is not necessarily an exact span in the context and is generated according to the question and context. This output type is especially suitable for non-factoid questions (Greco *et al*. 2016; Choi *et al*. 2017b; Tan *et al*. 2018a).

In the extractive mode, the answer is a specific span of the context (Liu *et al*. 2017; Min *et al*. 2017; Seo *et al*. 2017; Yang *et al*. 2017b; Liu *et al*. 2018c). This output type is appropriate for factoid questions; however, it is possible that the answer to a factoid question may be generative or the answer to a non-factoid question may be extractive. For example, the answer to a non-factoid question may be a whole sentence which is extracted from the context.

There has generally been more focus on extractive MRC systems, but according to Table 6, the popularity of abstractive MRC systems has increased over recent years. From another point of view, MRC outputs can be categorized as *multiple-choice style, cloze style*, and d*etail style*:

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 6. Statistics of input/output types in MRC systems.

Year	Input					Output				
	Question			Context						
	Factoid	Non-factoid	Yes/No	Single passage	Multi-passage	Extractive	Abstractive	Multi-choice	Cloze	Detail
2016	100%	8%	0%	85%	15%	91%	9%	38%	54%	8%
2017	100%	14%	0%	100%	0%	100%	9%	5%	27%	95%
2018	100%	26%	10%	76%	34%	84%	12%	18%	14%	74%
2019	100%	43%	2%	50%	52%	76%	36%	21%	3%	79%
2020	100%	32%	23%	68%	32%	82%	50%	36%	18%	59%
All	100%	30%	7%	70%	42%	82%	25%	21%	16%	72%

- In the multiple-choice style mode, the answer is one of the multiple candidate answers that must be selected from a predefined set A containing k candidate answers:

$$A = \{a_1, \ldots, a_k\},$$

where a_j can be a word, a phrase, or a sentence with length l_j:

$$a_j = \left(a_{j,1}, a_{j,2}, \ldots, a_{j,l_j}\right)$$

- In the cloze style mode, the question includes a blank that must be filled as an answer according to the context. In the available cloze style datasets, the answer is an entity from the context.
- In the detail style mode, there is no candidate or blank, so the answer must be extracted or generated according to the context. In extractive mode, the answer must be a definite range in the context, so it can be shown as (a_{start}, a_{end}), where a_{start} and a_{end} are, respectively, the start and end indices of the answer in the context. In generative mode, on the other hand, the answer can be generated from any custom vocabulary V, and it is not limited to the context words.

As shown in Table 6, most of the reviewed papers (72%) have focused on the detail style mode. Also, about 100%, 70%, and 82% of the reviewed papers have focused on factoid questions, multi-passage context, and extractive answers, respectively, due to their lower complexity and existence of rich datasets. For a more detailed categorization of papers based on their input/outputs, refer to Table A4.

5. MRC datasets

Rich datasets are the first prerequisite for having accurate machine learning models. Especially, deep neural network models require high volumes of training data to achieve good results. For this reason, in recent years, many researchers have focused on collecting big datasets. For example, Stanford Question Answering Dataset (SQuAD) (Rajpurkar *et al.* 2016), which is a popular MRC dataset used in many studies, includes over 100,000 training samples.

MRC datasets can be categorized according to their volume, domain, question type, answer type, context type, data collection method, and language.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

In terms of domain, MRC datasets can be classified into two categories: *open domain* and *closed domain*. Open domain datasets contain diverse subjects, while closed domain datasets focus on specific areas such as the medical domain. For example, the SQuAD (Rajpurkar *et al*. 2016) dataset, which contains Wikipedia articles, is an open domain dataset, and emrQA (Pampari *et al*. 2018), BIOMRC (Pappas *et al*. 2020), LUPARCQ (Horbach *et al*. 2020) are a closed domain dataset with biology as its subject.

There are two data collection approaches in MRC datasets, *automatic* approach, and *crowdsourcing* approach. The former generates questions/answers without direct human interventions. For instance, datasets that contain cloze style questions, such as Children's Book Test dataset (Hill *et al*. 2016), are generated by removing important entities from text. Also, in some datasets, questions are automatically extracted from the search engine's user logs (Nguyen *et al*. 2016) or real reading comprehension tests (Lai *et al*. 2017).

On the other hand, in the crowdsourcing approach, humans generate questions, answers, or select related paragraphs. Of course, a dataset can be generated by a combination of these two approaches. For instance, in MS MARCO (Nguyen *et al*. 2016), questions have been generated automatically, while these questions have been answered and evaluated by crowdsourcing.

Table 7 shows a detailed list of the datasets proposed from 2016-2020 in chronological order. In this table, the datasets with a link address are publicly available. Finally, Figure 3 shows the progress made on two representative datasets, SQuAD1.1 and RACE, as representatives of the advancements in the field. Note that only the articles available in our reviewed papers are reported in this figure. According to this figure, the state-of-the-art model's performance on SQuAD1.1 dataset increased from about 0.7 in 2017 to super human level of 0.95 in 2019. For the RACE dataset, despite the progress made in the accuracy from around 0.4 in 2017 to around 0.8 in 2020, it is still under the human-level performance which shows that this dataset is more challenging.

6. MRC evaluation measures

Based on the system output type, different evaluation metrics are introduced. We classify these measures into two categories: *extractive* metrics and g*enerative* metrics.

6.1. Extractive metrics

These metrics are used for the extractive outputs. Table 8 shows the statistics of these measures in the reviewed papers.

- ***F1 score:*** The harmonic mean of precision and recall is a common extractive metric for evaluating MRC systems. It takes into account the system output and the ground truth answer as bag-of-tokens (words). Precision is calculated as the number of correctly predicted tokens divided by the number of all predicted tokens. The recall is also the number of correctly predicted tokens divided by the number of ground truth tokens. The final F1 score is then obtained by averaging over all question-answer pairs.
- ***Exact Match (EM)***. This is the percentage of answers that exactly match with the correct answers. If there are multiple answers to a question in a dataset, a match with at least one of the answers is considered as an exact match. Some QA systems such as multiple-choice QA systems (Zhang *et al*. 2020c) or sentence selection QA systems (Min *et al*. 2017) call this measure as accuracy (ACC) instead of EM.
- ***Mean Average Precision (MAP).*** This measure is used when the system returns several answers along with their ratings. The MAP for a set of question-answer pairs is the mean of Average Precision scores (AveP) for each one. AveP is an evaluation measure used in information retrieval systems which evaluates a ranked list of documents in response to a given query.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 7. MRC datasets proposed from 2016 to 2020. (A: Answer, P: passage, Q: Question)

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2016	Children's Book Test (Hill *et al*. 2016)	Open	English	Factoid	Single paragraph	Extractive (Cloze Style)	687343	108 book	Automatic	No	https://research.fb.com/downloads/babi/
2016	People Daily news dataset (Cui *et al*. 2016)	Open	Chinese	Factoid	Single paragraph	Extractive (Cloze Style)	876K	60k Articles	Automatic	No	–
2016	Children's Fairy Tale (CFT) (Cui *et al*. 2016)	Open	Chinese	Factoid	Single paragraph	Extractive (Cloze Style)	3.5K	60 k Passages	Automatic	No	–
2016	Who did what (Onishi *et al*. 2016)	Open	English	Factoid	Single paragraph	Extractive (Multiple Choice Style)	330K	200 k Passages	Automatic	No	https://tticnlp.github.io/who_did_what/
2016	SQuAD (Rajpurkar *et al*. 2016)	Open	English	Factoid	Single paragraph	Extractive (Detail)	100K	536 Articles	Crowdsourced	No	https://stanford-qa.com
2016	MS MARCO (Nguyen *et al*. 2016)	Open	English	Factoid	Multi-document	Abstractive	100K	1M Passage +200K Document	Q: Automatic A:Crowdsourced	Yes	http://www.msmarco.org
2017	BookTest (NE, CN) (Bajgar *et al*. 2017)	Open	English	Factoid	Single paragraph	Extractive (Cloze Style)	14M	13.5K Books	Automatic	No	–
2017	NewsQA (Trischler *et al*. 2017)	Open	English	Factoid	Single paragraph	Extractive (Detail)	100K	10K Articles	Crowdsourced	No	https://www.microsoft.com/en-us/research/project/newsqa-dataset/
2017	TriviaQA (Joshi *et al*. 2017)	Open	English	Factoid	Multi-document	Extractive (Detail)	95K	650k Passages	Automatic	No	http://nlp.cs.washington.edu/triviaqa/
2017	RACE (Lai *et al*. 2017)	Open	English	Factoid and non-Factoid	Multi-paragraph	Abstractive (Multiple Choice style)	97K	27K Passages	Automatic	No	https://www.cs.cmu.edu/glai1/data/race
2017	SciQ (Welbl, Liu and Gardner 2017)	Close (science)	English	Factoid	Single document	Extractive (Multiple Choice style)	13.7K	13.7K Passages	Crowdsourced	No	https://allenai.org/data/sciq

Table 7. Continued

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2018	Wikihop (Welbl, Stenetorp and Riedel 2018)	Open	English	Factoid	Multi-paragraph	Extractive (Multiple Choice Style)	51318	3-63 for each Q	Q: Automatic A: Crowdsourced	No	http://qangaroo.cs.ucl.ac.uk/
2018	Medhop (Welbl *et al*. 2018)	Close (molecular biology)	English	Factoid	Multi-paragraph	Extractive (Multiple Choice Style)	2508	5-64 for each Q	Q: Automatic A: Crowdsourced	No	http://qangaroo.cs.ucl.ac.uk/
2018	NarrativeQA (Kočiský *et al*. 2018)	Open	English	Factoid and non-Factoid	Multi-paragraph	Abstractive	46,765	1,572 stories (books, movie scripts)	Crowdsourced	No	https://github.com/deepmind/narrativeqa
2018	MCScript (Ostermann *et al*. 2018)	Open	English	Factoid and non-Factoid	Single paragraph	Abstractive	32K	2K Passages	Crowdsourced	No	http://www.sfb1102.uni-saarland.de/?page_id=2582
2018	CliCl (Šuster and Daelemans 2018)	Close (medical)	English	Factoid	Single paragraph	Extractive (Cloze Style)	105K	12K Passages	Automatic	No	https://github.com/clips/clicr
2018	DRCD (Shao *et al*. 2018)	Open	Chinese	Factoid	Single paragraph	Extractive (Detail)	30K	10K Paragraphs from 2K articles	Crowdsourced	Yes	https://github.com/DRCSolutionService/DRCD
2018	MultiRC (Khashabi *et al*. 2018)	Open	English	Factoid and non-Factoid	Single paragraph	Abstractive (Multiple Choice style)	6K	+800 Passages	Crowdsourced	No	http://cogcomp.org/multirc/
2018	DuReader (He *et al*. 2018)	Open	Chinese	Factoid and non-Factoid	Multi-document	Abstractive	200K Questions, 420K answers	1M Passages	Q & P: Automatic A: Crowdsourced	Yes	http://ai.baidu.com/broad/download?dataset=dureader
2018	DuoRC (Saha *et al*. 2018)	Open	English	Factoid and non-Factoid	Multi-paragraph	Abstractive	186K	7.5K Passages	Crowdsourced	No	https://duorc.github.io/

Table 7. Continued

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2018	SQuAD 2.0 (Rajpurkar, Jia and Liang 2018)	Open	English	Factoid	Single paragraph	Extractive (Detail)	150K	505 Articles	Crowdsourced	No	https://rajpurkar.github.io/SQuAD-explorer
2018	SQuAD-T (Tan *et al*. 2018b)	Open	English	Factoid	Single Paragraph	Extractive (Detail)	100K	536 Articles	Crowdsourced	No	https://github.com/Chuanqi1992/SQuAD-T
2018	CLOTH (Xie *et al*. 2018)	Open	English	Factoid	Single paragraph	Abstractive (Multiple Choice style)	99K	7K Passage	Automatic	Yes	–
2018	emrQA (Pampari *et al*. 2018)	Close (electronic medical records)	English	Factoid	Single paragraph	Extractive (Detail)	455K	2K Passages	Automatic	No	https://www.i2b2.org/NLP/DataSets/
2018	HotpotQA (Yang *et al*. 2018)	Open	English	Factoid	Multi-paragraph	Extractive (Detail, yes/no)	112779	–	Crowdsourced	No	https://hotpotqa.github.io/
2018	QBLink (Elgohary, Zhao and Boyd-Graber 2018)	Open	English	Factoid	Multi-paragraph	Extractive (Detail)	56K	Context is extracted before reading.	Automatic	No	https://sites.google.com/view/qanta/projects/qblink
2019	HindiRC (Anuranjana, Rao and Mamidi 2019)	Open	Hindi	Factoid and non-Factoid	Single paragraph	Extractive (Detail)	127	24	Crowdsourced	Yes	https://github.com/erzaliator/HindiRC-Data
2019	DROP (Dua *et al*. 2019)	Open	English	Factoid	Single paragraph	Extractive and Abstractive	96.6K	6.6K	Crowdsourced	No	https://allennlp.org/drop
2019	MCScript2.0 (Ostermann, Roth and Pinkal 2019)	Close (narrative)	English	Factoid	Single paragraph	Abstractive (Multiple Choice)	20K	3.5K	Crowdsourced	No	http://www.sfb1102.uni-saarland.de/?page_id=2582

Table 7. Continued

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2019	ARCD (Mozannar *et al*. 2019)	Open	Arabic	Factoid	Single paragraph	Extractive (Detail)	1.4K	155 Article	Crowdsourced	No	https://github.com/husseinmozannar/SOQAL
2019	Natural Question (Kwiatkowski *et al*. 2019)	Open	English	Factoid and non-Factoid	Multi-paragraph	Abstractive	323045	–	Q:Automatic A:Crowdsourced	No	https://ai.google.com/research/NaturalQuestions
2019	AmazonQA (Gupta *et al*. 2019)	Open	English	Factoid and non-Factoid	Multi-paragraph	Extractive (Detail)	923k	14M	Crowdsourced	Yes	https://github.com/amazonqa/amazonqa
2019	XQA (Liu *et al*. 2019a)	Open	9 languages	Factoid	Multi-paragraph	Extractive (Detail)	90k	top-10 Wikipedia articles for each question	Automatic	No	http://github.com/thunlp/XQA
2019	(Hardalov, Koychev and Nakov 2019)	Close (history, biology, geography, and philosophy)	Bulgarian	Factoid and non-Factoid	–	Abstractive (Multiple Choice)	2.6K	–	Automatic	Yes	https://github.com/mhardalov/bg-reason-BERT
2019	FaQuAD (Sayama, Araujo and Fernandes 2019)	Close (higher education)	Brazilian	Factoid	Single paragraph	Extractive	900	249	Crowdsourced	No	https://github.com/liafacom/faquad
2019	SMART (Yao *et al*. 2019)	Open	Chinese	Factoid	Single paragraph	Extractive (Detail)	39.4K	564 Article	Crowdsourced	Yes	–
2019	CMRC 2018 (Cui *et al*. 2019b)	Open	Chinese	Factoid and non-Factoid	Single paragraph	Extractive (Detail)	20K	–	Crowdsourced	No	https://bit.ly/2ZdS8Ct
2019	COSMOS QA (Huang	Open	English	Factoid and	Single	Abstractive (Multiple Choice)	35.6K	21.9K	Crowdsourced	No	https://wilburone.github.io/cosmos

Table 7. Continued

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2019	QUOREF (Dasigi *et al*. 2019)	Open	English	Factoid	Single paragraph	Extractive (Detail)	24K	4.7K	Crowdsourced	No	https://allennlp.org/quoref
2019	ROPES (Lin *et al*. 2019)	Open	English	Factoid	Single paragraph	Extractive (Detail)	14.3K	1.4K	Crowdsourced	No	https://allennlp.org/ropes
2019	BiPaR (Jing, Xiong and Zhen 2019)	Close (novels)	Chinese English	Factoid and non-Factoid	Single paragraph	Extractive (Detail)	14.7K	3.7K	Crowdsourced	No	https://multinlp.github.io/BiPaR/
2019	RACE-C (Liang, Li and Yin 2019)	Open	English	Factoid and non-Factoid	Multi-paragraph	Abstractive (Multiple Choice)	14K	4K	Automatic	No	–
2019	(Li, Li and Liu 2019c)	Open	Chinese	Factoid and non-Factoid	Single paragraph	Extractive (Detail)	242K	85K	Crowdsourced	No	–
2020	ReCo (Wang *et al*. 2020b)	Open	Chinese	Factoid	Single paragraph	Generative (Multiple Choice)	300K	300K	Q: Automatic P, A: Crowdsourced	No	https://github.com/benywon/ReCO
2020	(Watarai and Tsuchiya 2020)	Open	Japanese	Factoid	Single Paragraph	Extractive (Multiple Choice and Cloze)	22790	64040	Automatic	No	–
2020	SQuAD2-CR (Lee, Hwang and Cho 2020)	Open	English	Factoid	Single paragraph	Extractive (Detail)	150K (16403 annotated)	500 article	Crowdsourced	No	https://antest1.github.io/SQuAD2-CR/
2020	LUPARCQ (Horbach *et al*. 2020)	Close (Biology)	English(E), German(G), Basque(B)	Factoid	Multi-paragraph	Extractive (Detail)	E: 299 G:95 E: 152	E: 24 G:21 B:18	Crowdsourced Automatic	No	–

Table 7. Continued

Year	Dataset	Open/ Closed Domain	Language	Question Type	Context Type	Answer Type	#Question	#Context	Collect Data	Question Classification	Link Address
2020	C^3(Sun *et al.* 2020)	Open	Chinese	Factoid	Multi-paragraph	Extractive (Multiple Choice)	19577	13369	Automatic	No	https://dataset.org/c3/
2020	R^4C(Inoue, Stenetorp and Inui 2020)	Open	English	Factoid	Multi-paragraph	Extractive (Detail)	4588	45880 (10 Paragraphs per Question)	Crowdsourced	No	https://naoya-i.github.io/r4c/
2020	OneStepQA (Berzak, Malmaud and Levy 2020)	Open	English	non-Factoid	Single paragraph	Abstractive (Multiple Choice)	1458 (486 × three version)	30 article (162 paragraph)	Crowdsourced	No	https://github.com/berzak/onestop-qa
2020	BIOMRC (Large) (Pappas *et al.* 2020)	Close (Biomedical)	English	Factoid	Single Paragraph	Extractive (Cloze)	812707	812707	Automatic	No	https://archive.org/details/biomrc_dataset
2020	ReClor (Yu *et al.* 2020)	Open	English	non-factoid	Singl paragraph	Abstractive (Multiple Choice)	6138	6138	Automatic	No	https://whyu.me/reclor

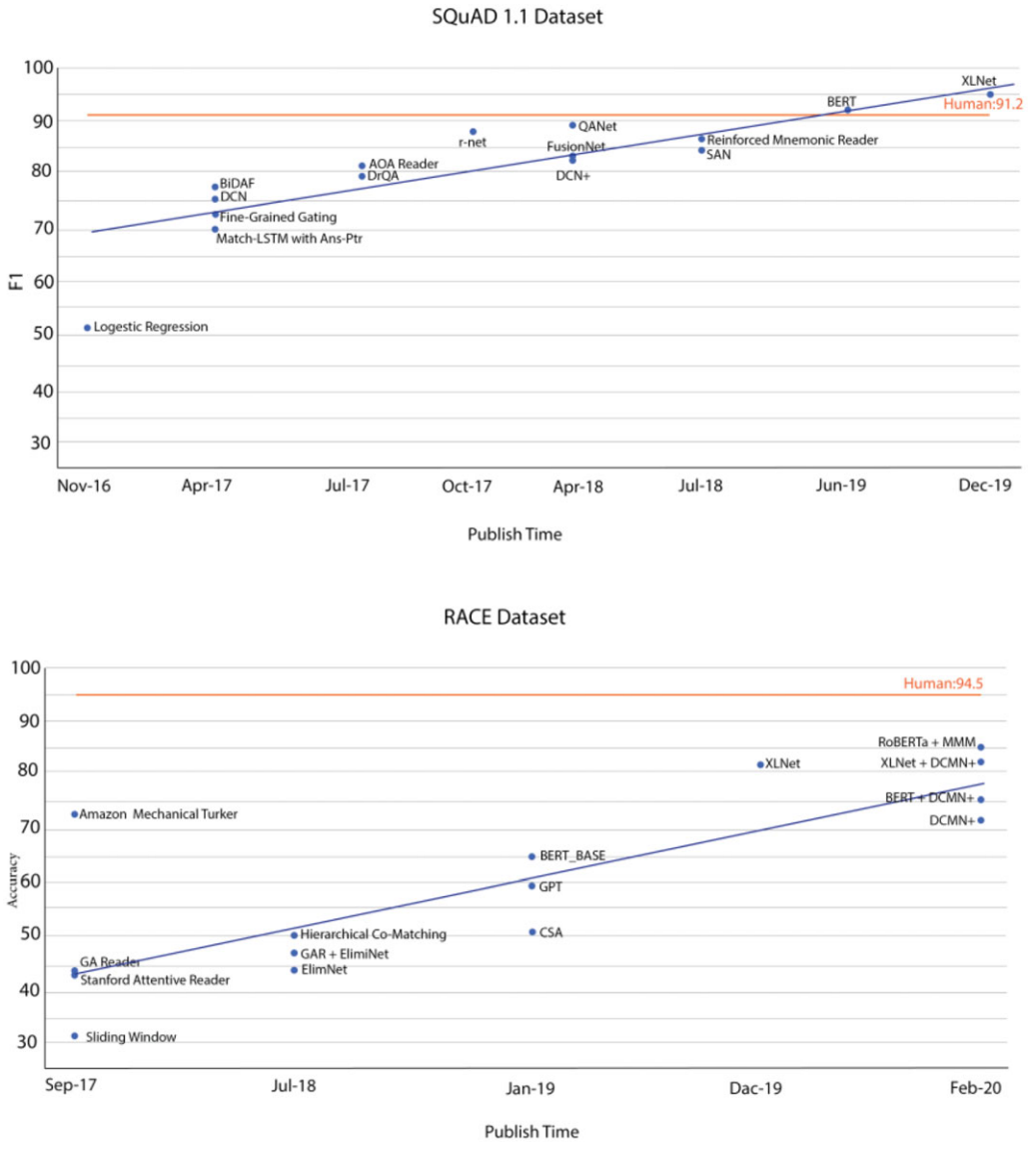

Figure 3. The progress made on two datasets: SQuAD1.1 (top) and RACE (down). The data points are taken from https://rajpurkar.github.io/SQuAD-explorer and http://www.qizhexie.com/data/RACE_leaderboard.html, respectively. Only the articles available in our reviewed papers are reported.

AveP for a single query is calculated by taking the mean of the precision scores obtained after each relevant document is retrieved, with relevant documents that are not retrieved receiving a precision score of zero (Turpin and Scholer 2006). In MRC literature, the ranked list of answers for a given question is evaluated.

- ***Mean Reciprocal Rank (MRR).*** This is a common evaluation metric for factoid QA systems introduced in TREC QA track 1999. According to the definition presented in the "Evaluation of Factoid Answers" Section of the "Speech and Language Processing" book (Jurafsky and Martin 2019), MRR evaluates a ranked list of answers based on the inverse of the rank of the correct answer. For example, if the rank of the correct answer in the output list of a system is

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 8. Statistics of evaluation measures used in reviewed papers.

Year	Extractive Metrics								Generative Metrics			
	EM	F1	MAP	MRR	P@1	R@1	ACC	Hit@k/ Top-k	ROUGE_L	BLEU	METEOR	CIDEr
2016	0%	0%	13%	13%	0%	0%	87%	13%	5%	0%	0%	0%
2017	59%	63%	7%	7%	4%	0%	41%	0%	3%	7%	0%	3%
2018	58%	80%	4%	2%	7%	7%	36%	0%	22%	18%	7%	2%
2019	64%	64%	2%	2%	3%	2%	2%	0%	19%	17%	5%	0%
2020	46%	55%	0%	0%	23%	18%	50%	0%	9%	14%	0%	0%
All	56%	60%	4%	3%	7%	5%	29%	1%	16%	14%	4%	1%

4, the reciprocal rank score for that question would be 1/4. This measure is then averaged for all questions in the test set.

- ***Precision@K.*** This measure is also borrowed from information retrieval literature. It is the number of correct answers in the first k returned answers without considering the position of these correct ones (Manning, Raghavan and Schütze 2008).
- ***Hit@K or Top-K.*** Hit@K, which is equivalent to the Top-K accuracy, counts the number of samples where their first k returned answers include the correct answer.

6.2. Generative metrics

The metrics used for evaluating the performance of generative MRC systems are the same metrics used for machine translation and summarization evaluation. Table 8 shows the statistics of these measures in the reviewed papers.

- ***Recall-Oriented Understudy for Gisting Evaluation (ROUGE).*** This measure compares a system-generated answer with the human-generated one (Lin 2004). It is defined as the recall of the system based on the n-grams, that is, the number of correctly generated n-grams divided by the total number of n-grams in the human-generated answer.
- ***BiLingual Evaluation Understudy (BLEU).*** This metric was first introduced for evaluating the output of machine translation task. It is defined as the precision of the system based on the n-grams, that is, the number of correctly generated n-grams divided by the total number of n-grams in the system-generated answer (Papineni *et al*. 2002).
- ***Metric for Evaluation of Translation with Explicit Ordering (METEOR).*** This measure is designed to fix some weaknesses of the popular BLEU measure. METEOR is based on an alignment between the system output and reference output. It introduces a penalty for adjacent tokens that cannot be mapped between the reference and system output. For this, unigrams are grouped into longer chunks if they are adjacent in both reference and system output. The more adjacent unigrams, the fewer chunks and the fewer penalty will be (Banerjee and Lavie 2005).
- ***Consensus-based Image Description Evaluation (CIDEr).*** This measure is initially introduced for evaluating the image description generation task (Vedantam, Lawrence Zitnick and Parikh 2015). It is based on the n-gram matching of the system output and reference output in the stem or root forms. According to this measure, the n-grams that are not in the reference output should not be in the system output. Also, the common n-grams in the dataset are less informative and have lower weights.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 9. Statistics of different research contributions to MRC task in the reviewed papers.

Year	Model structure	Dataset	Other tasks	Evaluation measure
2016	50%	50%	21%	7%
2017	54%	14%	23%	6%
2018	71%	31%	14%	5%
2019	68%	20%	24%	11%
2020	57%	20%	29%	31%
All	61%	23%	21%	12%

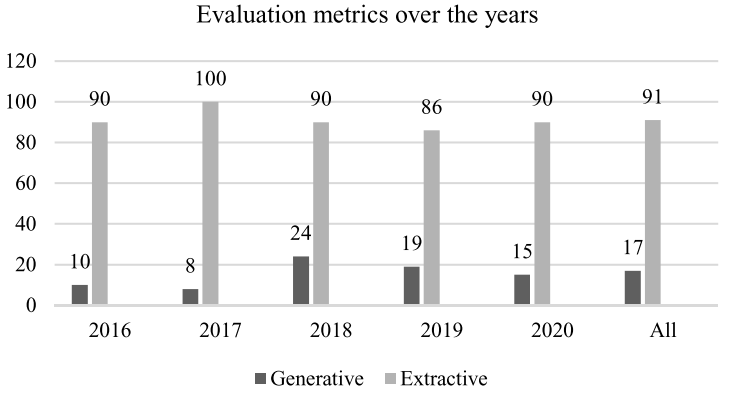

Figure 4. Ratio of reviewed papers (%) for extractive/generative evaluation metrics

Figure 4 shows the ratio of the used extractive/generative measures in the reviewed papers. According to this figure, the generative measures have been more popular in recent years than in 2016 and 2017. The obvious reason for this is the trend toward developing abstractive MRC systems. For more details, refer to Table A5.

7. Research contribution

The contribution of MRC research can be grouped into four categories: developing new model structures, creating new datasets, combining with other tasks and improvement, and introducing new evaluation measures. Table 9 shows the statistics of these categories. Note that some studies have more than one contribution type, so the sum of some rows is greater than 100%. For example, Ma, Jurczyk and Choi (2018) introduced a new dataset from the "Friends" sitcom transcripts and developed a new model architecture as well. For more details, refer to Table A6.

7.1. Developing new model structures

Many MRC papers have focused on developing new model structures to address the weaknesses of previous models. Most of them developed new internal structures (Cui *et al.* 2016; Kadlec *et al.* 2016b; Kobayashi *et al.* 2016; Cui *et al.* 2017b; Dhingra *et al.* 2017; Seo *et al.* 2017; Shen *et al.* 2017; Xiong *et al.* 2017; Hu *et al.* 2018a; Huang *et al.* 2018). Some others changed the system inputs. For example, in Chen *et al.* (2017) study, in addition to word vectors, linguistic features such as NE and POS vectors have also been used as the input to the model. Also, some papers

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

introduced a new way of entering the input into the system. For example, Hewlett *et al.* (2017) proposed breaking the context into overlapping windows and entering each window as an input to the system.

7.2. Creating new datasets

One of the main reasons for advancing the MRC research in recent years is the introduction of rich datasets. Many studies have focused on creating new datasets with new features in recent years (Nguyen *et al.* 2016; Rajpurkar *et al.* 2016; Joshi *et al.* 2017; Lai *et al.* 2017; Trischler *et al.* 2017; He *et al.* 2018; Šuster and Daelemans 2018). The main trend is to develop multi-document datasets, abstractive style outputs, and more complex questions that require more advanced reasoning. Also, some papers focus on customizing the available datasets instead of creating new ones. For example, Horbach *et al.* (2020) proposed to turn existing datasets, such as SQuAD and NewsQA, into interactive datasets. Some other papers added annotations to the existing datasets to provide interpretable clues for investigating the models' behavior as well as to prevent models from exploiting biases and annotation artifacts. For example, SQuAD 2.0 (Rajpurkar *et al.* 2018) was created by adding unanswerable questions to SQuAD, and SQuAD2-CR (Lee *et al.* 2020) was developed by adding causal and relational annotations to unanswerable questions in SQuAD 2.0. Similarly, R^4C (Inoue *et al.* 2020) was created by adding derivations to questions in hotpotQA.

7.3. Combining with other tasks

Simultaneous learning of multiple tasks (multi-task learning) (Collobert and Weston 2008) and exploiting the learned knowledge from one task in another task (transfer learning) (Ruder 2019) have been promising directions for obtaining better results, especially in the data-poor setting. As an example, Wang *et al.* (2017a) trained their MRC task with a question generation task and achieved better results. Besides these approaches, some papers exploit other task solutions as sub-modules in their MRC system. As an example, Yin *et al.* (2016) used a question classifier and a natural language inference (NLI) system as two sub-modules in their MRC system.

7.4. Introducing new evaluation measures

Reliable assessment of an MRC system is still a challenging topic. While some systems go beyond human performance in specific datasets such as SQuAD by the current measures (Rajpurkar *et al.* 2016), further investigation shows that these systems fail to achieve a thorough and true understanding of human language (Jia and Liang 2017; Wang and Bansal 2018). In these papers, the passage is successfully edited to mislead the model. These papers can be seen as a measure to evaluate the true comprehension of systems. Also, some papers have evaluated the required comprehension and reasoning capabilities for solving the MRC problem in available datasets (Chen *et al.* 2016; Sugawara *et al.* 2018).

8. Hot MRC papers

Table 10 shows the top papers in different years from 2016-2020 based on the number of citations in the Google Scholar service until September 2020. For all years but 2020, ten papers have been selected; while for the year 2020, just five papers have been chosen because the number of citations in this year is not enough in the time of writing the paper. According to this table, hot papers often introduce a new successful model structure or a new dataset.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 10. Hot papers based on the number of citations in the Google Scholar service until September 2020.

Title	Publication venue	Year	Contribution	Citations
SQuAD: 100,000+ Questions for Machine Comprehension of Text (Rajpurkar *et al*. 2016)	EMLP	2016	Dataset	2068
The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations (Hill *et al*. 2016)	ICLR	2016	Dataset	426
A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task (Chen *et al*. 2016)	ACL	2016	Evaluation dataset	420
MS MARCO: A Human Generated Machine Reading Comprehension Dataset (Nguyen *et al*. 2016)	CoCo@NIPS	2016	Dataset	407
Text Understanding with the Attention Sum Reader Network (Kadlec *et al*. 2016b)	ACL	2016	Model Structure (AS)	241
Who Did What: A Large-Scale Person-centered Cloze Dataset (Onishi *et al*. 2016)	EMNLP	2016	Dataset	97
Attention-Based Convolutional Neural Network for Machine Comprehension (Yin *et al*. 2016)	NAACL	2016	Model Structure (HABCNN), Knowledge Transfer	78
Consensus Attention-Based Neural Networks for Chinese Reading Comprehension (Cui *et al*. 2016)	COLING	2016	Model structure (CAS Reader), Dataset	58
Dynamic Entity Representation with Max-Pooling Improves Machine Reading (Kobayashi *et al*. 2016)	ACL	2016	Model Structure (DER Network)	40
Employing External Rich Knowledge for Machine Comprehension (Wang *et al*. 2016)	IJCAI	2016	Model Structure, Knowledge Transfer	12
Bidirectional Attention Flow for Machine Comprehension (Seo *et al*. 2017)	ICLR	2017	Model structure (BiDAF)	1091
Reading Wikipedia to Answer Open-Domain Questions (Chen *et al*. 2017)	ACL	2017	Model structure (DrQA)	681
Adversarial Examples for Evaluating Reading Comprehension Systems (Jia and Liang 2017)	EMNLP	2017	Evaluation measure	547
Dynamic Coattention Networks for Question Answering (Xiong *et al*. 2017)	ICLR	2017	Model structure (DCN)	448
Gated Self-Matching Networks for Reading Comprehension and Question Answering (Wang *et al*. 2017b)	ACL	2017	Model structure (R-NET)	430
TriviaQA: a Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension (Joshi *et al*. 2017)	ACL	2017	Dataset	418
Machine Comprehension using Match-LSTM and Answer Pointer (Wang and Jiang 2017)	ICLR	2017	Model structure (Match-LSTM and Answer Pointer)	392
NewsQA: A Machine Comprehension Dataset (Trischler *et al*. 2017)	RepL4NLP	2017	Dataset	302

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 10. Continued

Title	Publication venue	Year	Contribution	Citations
Attention-Over-Attention Neural Networks for Reading Comprehension (Cui *et al*. 2017b)	ACL	2017	Model structure (AoA Reader)	291
Gated-Attention Readers for Text Comprehension (Dhingra *et al*. 2017)	ACL	2017	Model structure (GA-Reader)	288
Know What You Don't Know: Unanswerable Questions for SQuAD (Rajpurkar *et al*. 2018)	ACL	2018	Dataset	529
QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension (Yu *et al*. 2018)	ICLR	2018	Model structure (QANet)	407
HotpotQA: A Dataset for Diverse, Explainable Multi-Hop Question Answering (Yang *et al*. 2018)	EMNLP	2018	Dataset	260
Simple and Effective Multi-Paragraph Reading Comprehension (Clark and Gardner 2018)	ACL	2018	Model structure	212
Reinforced Mnemonic Reader for Machine Reading Comprehension (Hu *et al*. 2018a)	IJCAI	2018	Model structure (R.M-Reader)	184
The Narrativeqa Reading Comprehension Challenge (KoČiskÝ *et al*. 2018)	ACL	2018	Dataset	172
Constructing Datasets for Multi-Hop Reading Comprehension Across Documents (Welbl *et al*. 2018)	ACL	2018	Dataset	170
R3: Reinforced Ranker-Reader for Open-Domain Question Answering (Wang *et al*. 2018b)	AAAI	2018	Model structure (R3)	150
Stochastic Answer Networks for Machine Reading Comprehension (Liu *et al*. 2018c)	ACL	2018	Model structure (SAN)	130
Natural Questions: a Benchmark for Question Answering Research (Kwiatkowski *et al*. 2019)	ACL	2019	Dataset	192
DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs (Dua *et al*. 2019)	NAACL-HLT	2019	Model Structure (NAQANET), Dataset	108
Read + Verify: Machine Reading Comprehension with Unanswerable Questions (Hu *et al*. 2019c)	AAAI	2019	Model Structure (RMR)	69
Improving Machine Reading Comprehension with General Reading Strategies (Sun *et al*. 2019)	NAACL-HLT	2019	Model Structure, Knowledge Transfer	54
MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension (Talmor and Berant 2019)	ACL	2019	Evaluation Measure	49
Multi-Step Retriever-Reader Interaction for Scalable Open-Domain Question Answering (Das *et al*. 2019)	ICLR	2019	Model Structure	47
Cognitive Graph for Multi-Hop Reading Comprehension at Scale (Ding *et al*. 2019)	ACL	2019	Model Structure (CogQA)	43

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table 10. Continued

Title	Publication venue	Year	Contribution	Citations
Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning (Huang *et al*. 2019b)	EMNLP-IJCNLP	2019	Dataset, Model Structure	32
Multi-Hop Reading Comprehension Through Question Decomposition and Rescoring (Min *et al*. 2019)	ACL	2019	Model Structure	28
MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension (Fisch *et al*. 2019)	ACL	2019	Evaluation Measure	27
SG-Net: Syntax-Guided Machine Reading Comprehension (Zhang *et al*. 2020c)	AAAI	2020	Model Structure (SG-Net), Knowledge Transfer	17
Select, Answer and Explain: Interpretable Multi-Hop Reading Comprehension Over Multiple Documents (Tu *et al*. 2020)	AAAI	2020	Model Structure (SAE), Knowledge Transfer	10
DCMN+: Dual Co-Matching Network for Multi-Choice Reading Comprehension (Zhang *et al*. 2020b)	AAAI	2020	Model Structure (DCMN+)	9
Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets (Sugawara *et al*. 2020)	AAAI	2020	Evaluation Measure	8
MMM: Multi-Stage Multi-Task Learning for Multi-Choice Reading Comprehension (Jin *et al*. 2020)	AAAI	2020	Model Structure (MMM), Knowledge Transfer	5

9. Future trends and opportunities

The MRC task has witnessed a great progress in recent years. Especially, like other NLP tasks, fine-tuning the pre-trained language models like BERT (Devlin *et al*. 2018) and XLNet (Yang *et al*. 2019b) on the target task has achieved an impressive success in MRC task, such that many state-of-the-art systems use these language models. However, they suffer from some shortcomings, which make them far from the real reading comprehension. In the following, we list some of these challenges and new trends in the MRC field:

- *Out-of-domain distributions*: Despite the high accuracy of the current MRC models on test samples from their training distribution, they are too fragile for out-of-domain distributed data. To address this shortcoming, some recent papers focused on improving the generalization capability of MRC models (Fisch *et al*. 2019; Wang *et al*. 2019e; Nishida *et al*. 2020).
- *Multi-document MRC:* One important challenge in the MRC task is the multi-hop reasoning, which is to infer the answer from multiple texts. These texts can be either several paragraphs of a document (Frermann 2019; Tay *et al*. 2019) or heterogeneous paragraphs from multiple documents (Tu *et al*. 2020). One of the new trends is to use the graph structures, such as graph neural networks, for multi-hop reasoning (Ding *et al*. 2019; Tu *et al*. 2019; Song *et al*. 2020; Tu *et al*. 2020).
- *Numerical reasoning:* Many questions in real-world applications require numerical inference including addition, subtraction, comparison, and so on. For example, consider the following

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

question from DROP dataset (Dua *et al*. 2019) which needs a subtraction: "How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?" Developing MRC models capable of numerical reasoning has become more popular in recent years, especially after the creation of numerical datasets such as DROP (Dua *et al*. 2019; Ran *et al*. 2019; Chen *et al*. 2020).

- *No-answer questions:* One of the new trends that makes the MRC systems more usable in real-world applications is to enable models to identify the questions which cannot be answered using the given context. With the development of datasets containing this kind of questions, such as SQuAD 2.0 (Rajpurkar *et al*. 2018) and Natural Questions (Kwiatkowski *et al*. 2019), more attention has been paid to this issue (Back *et al*. 2020; Liu *et al*. 2020a; Zhang *et al*. 2020c).
- *Non-factoid questions:* Answering non-factoid questions, such as *why* and *opinion* questions, often requires generating answers rather than selecting a span of context. The accuracy of the existing models in answering these questions is still far from the desired level. In recent years, several datasets which contain non-factoid questions have attracted more attention to this kind of questions (Saha *et al*. 2018; Gupta *et al*. 2019; Berzak *et al*. 2020).
- *Low-resource language datasets and models:* It is worth noting that most available datasets are in resource-rich languages, such as English and Chinese. Creating new datasets and models for low-resource languages and developing them in the multi-lingual or multi-task setting can also be seen as a new trend in this field (Jing *et al*. 2019; Amirkhani *et al*. 2020; Gupta and Khade 2020; Yuan *et al*. 2020a).

10. Conclusion

The MRC, as a hot research topic in NLP, focuses on reading the document(s) and answering the corresponding questions. The main goal of an MRC system is to gain a comprehensive understanding of text documents to be able to justify and respond to related questions. In this paper, we presented an overview of the variable aspects of recent MRC studies, including approaches, internal architecture, input/output type, research contributions, and evaluation measures. We reviewed 241 papers published during 2016-2020 to investigate recent studies and find new trends.

Based on the question type, MRC papers are categorized into factoid, non-factoid, and yes/no questions. In addition, the input context is categorized into single or multiple passages. According to the results of statistics, a trend toward non-factoid questions and multiple passages was obvious in recent years. The output types are classified into extractive and abstractive outputs. From another point of view, the output types are categorized as multiple choice, cloze, and detail styles. The statistics showed that although the extractive outputs have been more popular, the abstractive outputs are becoming more popular in recent years.

Furthermore, we reviewed the developed datasets along with their features, such as data volume, domain, question type, answer type, context type, collection method, and data language. The number of developed datasets has increased in recent years, and they are in general more challenging than previous datasets. Regarding research contributions, some papers develop new model structures, some introduce new datasets, some combine MRC tasks with other tasks, and others introduce new evaluation measures. The majority of papers developed novel model structures or introduced new datasets. Moreover, we presented the most-cited papers, which indicate the most popular datasets and models in the MRC literature. Finally, we mentioned the possible future trends and important challenges of available models, including the issues related to out-of-domain distributions, multi-document MRC, numerical reasoning, no-answer questions, non-factoid questions, and low-resource languages.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

References

Aghaebrahimian A. (2018). Linguistically-Based Deep Unstructured Question Answering. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 433-43.

Akour M., **Abufardeh S.**, **Magel K.** and **Al-Radaideh Q.** (2011). QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic. *American Journal of Applied Sciences* **8**, 652–661.

Amirkhani H., **Jafari M.A.**, **Amirak A.**, **Pourjafari Z.**, **Jahromi S.F.** and **Kouhkan Z**. (2020). FarsTail: A Persian Natural Language Inference Dataset. *arXiv preprint arXiv:2009.08820*.

Andor D., **He L.**, **Lee K.** and **Pitler E.** (2019). Giving BERT a Calculator: Finding Operations and Arguments with Reading Comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5947-52. Hong Kong, China.

Angelidis S., **Frermann L.**, **Marcheggiani D.**, **Blanco R.** and **MÀrquez L.** (2019). Book QA: Stories of Challenges and Opportunities. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 78-85. Hong Kong, China.

Anuranjana K., **Rao V.** and **Mamidi R**. (2019). HindiRC: A Dataset for Reading Comprehension in Hindi. *0th International Conference on Computational Linguistics and Intelligent Text*, La Rochelle, France.

Arivuchelvan K.M. and **Lakahmi K.** (2017). Reading Comprehension System-A Review. *Indian Journal of Scientific Research (IJSR)* **14**, 83–90.

Asai A. and **Hajishirzi H.** (2020). Logic-Guided Data Augmentation and Regularization for Consistent Question Answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5642–50.

Back S., **Chinthakindi S.C.**, **Kedia A.**, **Lee H.** and **Choo J**. (2020). NeurQuRI: Neural Question Requirement Inspector for Answerability Prediction in Machine Reading Comprehension. *International Conference on Learning Representations*.

Back S., **Yu S.**, **Indurthi S.R.**, **Kim J.** and **Choo J**. (2018). MemoReader: Large-Scale Reading Comprehension through Neural Memory Controller. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2131-40.

Bahdanau D., **Cho K.** and **Bengio Y**. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Bajgar O., **Kadlec R.** and **Kleindienst J.** (2017). Embracing data abundance: BookTest Dataset for Reading Comprehension. *International Conference on Learning Representations (ICLR) Workshop*.

Banerjee S. and **Lavie A**. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65-72.

Bao H., **Dong L.**, **Wei F.**, **Wang W.**, **Yang N.**, **Cui L.**, **Piao S.** and **Zhou M.** (2019). Inspecting Unification of Encoding and Matching with Transformer: A Case Study of Machine Reading Comprehension. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 14-18.

Bauer L., **Wang Y.** and **Bansal M**. (2018). Commonsense for Generative Multi-Hop Question Answering Tasks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4220–30

Berant J., **Srikumar V.**, **Chen P.-C.**, **Vander Linden A.**, **Harding B.**, **Huang B.**, **Clark P.** and **Manning C.D.** (2014). Modeling Biological Processes for Reading Comprehension. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1499-510.

Berzak Y., **Malmaud J.** and **Levy R**. (2020). STARC: Structured Annotations for Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5726–35.

Bi B., **Wu C.**, **Yan M.**, **Wang W.**, **Xia J.** and **Li C**. (2019). Incorporating External Knowledge into Machine Reading for Generative Question Answering. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2521-30. Hong Kong, China.

Bouziane A., **Bouchiha D.**, **Doumi N.** and **Malki M.** (2015). Question answering systems: survey and trends. *Procedia Computer Science* **73**, 366–375.

Cao Y., **Fang M.**, **Yu B.** and **Zhou J.T**. (2020). Unsupervised Domain Adaptation on Reading Comprehension. *AAAI*, pp. 7480-87.

Charlet D., **Damnati G.**, **BÉchet F.** and **Heinecke J**. (2020). Cross-lingual and cross-domain evaluation of Machine Reading Comprehension with Squad and CALOR-Quest corpora. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5491-97.

Chaturvedi A., **Pandit O.** and **Garain U**. (2018). CNN for Text-Based Multiple Choice Question Answering. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 272-77.

Chen A., **Stanovsky G.**, **Singh S.** and **Gardner M**. (2019a). Evaluating question answering evaluation. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 119-24.

Chen D. (2018). *Neural Reading Comprehension and Beyond*. Stanford University.

Chen D., **Bolton J.** and **Manning C.D.** (2016). A Thorough Examination of the Cnn/Daily Mail Reading Comprehension Task. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 2358–67.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Chen D., **Fisch A.**, **Weston J.** and **Bordes A**. (2017). Reading wikipedia to answer open-domain questions. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pp. 1870-79.

Chen X., **Liang C.**, **Yu A.W.**, **Zhou D.**, **Song D.** and **Le Q.V.** (2020). Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. *International Conference on Learning Representations (ICLR2020)*.

Chen Z., **Cui Y.**, **Ma W.**, **Wang S.** and **Hu G**. (2019b). Convolutional spatial attention model for reading comprehension with multiple-choice questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6276-83.

Cho K., **Van MerriËnboer B.**, **Gulcehre C.**, **Bahdanau D.**, **Bougares F.**, **Schwenk H.** and **Bengio Y.** (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–34. Doha, Qatar.

Choi E., **Hewlett D.**, **Lacoste A.**, **Polosukhin I.**, **Uszkoreit J.** and **Berant J.** (2017a). Hierarchical Question Answering for Long Documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 209-20. Vancouver, Canada.

Choi E., **Hewlett D.**, **Uszkoreit J.**, **Polosukhin I.**, **Lacoste A.** and **Berant J**. (2017b). Coarse-to-fine question answering for long documents. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 209-20.

Chollet F. (2017). *Deep learning with python*. Manning Publications Co.

Clark C. and **Gardner M**. (2018). Simple and effective multi-paragraph reading comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers*, pp. 845–55

Collobert R. and **Weston J**. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. *Proceedings of the 25th international conference on Machine learning*, pp. 160-67. ACM.

Cui L., **Huang S.**, **Wei F.**, **Tan C.**, **Duan C.** and **Zhou M**. (2017a). Superagent: a customer service chatbot for e-commerce websites. *Proceedings of ACL, System Demonstrations*, pp. 97-102. Association for Computational Linguistics.

Cui Y., **Che W.**, **Liu T.**, **Qin B.**, **Wang S.** and **Hu G.** (2019a). Cross-Lingual Machine Reading Comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1586-95. Hong Kong, China.

Cui Y., **Chen Z.**, **Wei S.**, **Wang S.**, **Liu T.** and **Hu G.** (2017b). Attention-over-attention neural networks for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* pp. 593–602.

Cui Y., **Liu T.**, **Che W.**, **Xiao L.**, **Chen Z.**, **Ma W.**, **Wang S.** and **Hu G.** (2019b). A Span-Extraction Dataset for Chinese Machine Reading Comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5883-89. Hong Kong, China.

Cui Y., **Liu T.**, **Chen Z.**, **Wang S.** and **Hu G**. (2016). Consensus Attention-Based Neural Networks for Chinese Reading Comprehension. *Proceedings of the COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1777–86.

Das R., **Dhuliawala S.**, **Zaheer M.** and **McCallum A.** (2019). Multi-step Retriever-Reader Interaction for Scalable Open-domain Question Answering. *International Conference on Learning Representations*.

Dasigi P., **Liu N.F.**, **MarasoviĆ A.**, **Smith N.A.** and **Gardner M.** (2019). Quoref: A Reading Comprehension Dataset with Questions Requiring Coreferential Reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5925-32. Hong Kong, China.

Dehghani M., **Azarbonyad H.**, **Kamps J.** and **de Rijke M.** (2019). Learning to transform, combine, and reason in open-domain question answering. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pp. 681-89.

Deng L and **Liu Y**. (2018). *Deep Learning in Natural Language Processing*. Springer Singapore.

Devlin J., **Chang M.-W.**, **Lee K.** and **Toutanova K.** (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171-86.

Dhingra B., **Liu H.**, **Yang Z.**, **Cohen W.W.** and **Salakhutdinov R.** (2017). Gated-attention readers for text comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* pp. 1832–46.

Dhingra B., **Pruthi D.** and **Rajagopal D.** (2018). Simple and Effective Semi-Supervised Question Answering. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pp. 582–87.

Dhingra B., **Zhou Z.**, **Fitzpatrick D.**, **Muehl M.** and **Cohen W.W.** (2016). Tweet2vec: Character-based distributed representations for social media. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 269–74.

Ding M., **Zhou C.**, **Chen Q.**, **Yang H.** and **Tang J.** (2019). Cognitive Graph for Multi-Hop Reading Comprehension at Scale. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2694-703. Florence, Italy.

Du X. and **Cardie C.** (2018). Harvesting Paragraph-Level Question-Answer Pairs from Wikipedia. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1907–17.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Dua D., **Singh S.** and **Gardner M.** (2020). Benefits of Intermediate Annotations in Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5627-34.

Dua D., **Wang Y.**, **Dasigi P.**, **Stanovsky G.**, **Singh S.** and **Gardner M.** (2019). DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs. *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)*, pp. 2368-78. Minneapolis, Minnesota.

Duan N., **Tang D.**, **Chen P.** and **Zhou M.** (2017). Question generation for question answering. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 866-74.

Dunietz J., **Burnham G.**, **Bharadwaj A.**, **Chu-Carroll J.**, **Rambow O.** and **Ferrucci D.** (2020). To Test Machine Comprehension, Start by Defining Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7839–59.

Elgohary A., **Zhao C.** and **Boyd-Graber J.** (2018). A dataset and baselines for sequential open-domain question answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1077-83.

Ethayarajh K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 55–65. Hong Kong, China.

Fisch A., **Talmor A.**, **Jia R.**, **Seo M.**, **Choi E.** and **Chen D.** (2019). MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 1-13. Hong Kong, China.

Frermann L. (2019). Extractive NarrativeQA with Heuristic Pre-Training. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 172-82.

Gardner M., **Berant J.**, **Hajishirzi H.**, **Talmor A.** and **Min S.** (2019). On Making Reading Comprehension More Comprehensive. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 105-12.

Ghaeini R., **Fern X.Z.**, **Shahbazi**, **H.** and **Tadepalli**, **P.** (2018). Dependent Gated Reading for Cloze-Style Question Answering. *Proceedings of the 27th International Conference on Computational Linguistics* pp. 3330–45.

Giuseppe A. (2017). Question Dependent Recurrent Entity Network for Question Answering. *NL4AI: 1st Workshop on Natural Language for Artificial Intelligence*, pp. 69-80. CEUR.

Golub D., **Huang P.-S.**, **He X.** and **Deng L.** (2017). Two-stage synthesis networks for transfer learning in machine comprehension. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* pp. 835–44.

Gong Y. and **Bowman S.R.** (2018). Ruminating reader: Reasoning with gated multi-hop attention. *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 1–11 Association for Computational Linguistics.

Greco C., **Suglia A.**, **Basile P.**, **Rossiello G.** and **Semeraro G.** (2016). Iterative multi-document neural attention for multiple answer prediction. *2016 AI∗ IA Workshop on Deep Understanding and Reasoning: A Challenge for Next-Generation Intelligent Agents (URANIA)*.

Guo S., **Li R.**, **Tan H.**, **Li X.**, **Guan Y.**, **Zhao H.** and **Zhang Y**. (2020). A Frame-based Sentence Representation for Machine Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 891-96.

Gupta D., **Lenka P.**, **Ekbal A.** and **Bhattacharyya P.** (2018). Uncovering Code-Mixed Challenges: A Framework for Linguistically Driven Question Generation and Neural based Question Answering. *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pp. 119-30.

Gupta M., **Kulkarni N.**, **Chanda R.**, **Rayasam A.** and **Lipton Z.C.** (2019). AmazonQA: A Review-Based Question Answering Task. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*, pp. 4996-5002.

Gupta S. and **Khade N**. (2020). BERT Based Multilingual Machine Comprehension in English and Hindi. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.s* 19.

Gupta S., **Rawat B.P.S.** and **Yu H.** (2020). Conversational Machine Comprehension: a Literature Review. *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 2739-53.

Hardalov M., **Koychev I.** and **Nakov P.** (2019). Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian. *Proceedings of Recent Advances in Natural Language Processing*, pp. 447-59. Varna, Bulgaria.

Hashemi H., **Aliannejadi M.**, **Zamani H.** and **Croft W.B.** (2020). ANTIQUE: A non-factoid question answering benchmark. *European Conference on Information Retrieval*, pp. 166-73. Springer.

He W., **Liu K.**, **Liu J.**, **Lyu Y.**, **Zhao S.**, **Xiao X.**, **Liu Y.**, **Wang Y.**, **Wu H.** and **She Q**. (2018). Dureader: a Chinese Machine Reading Comprehension Dataset from Real-World Applications. *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. 37–46. Association for Computational Linguistics.

Hermann K.M., **Kocisky T.**, **Grefenstette E.**, **Espeholt L.**, **Kay W.**, **Suleyman M.** and **Blunsom P.** (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, pp. 1693-701.

Hewlett D., **Jones L.** and **Lacoste A.** (2017). Accurate supervised and semi-supervised machine reading for long documents. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2011-20.

Hill F., **Bordes A.**, **Chopra S.** and **Weston J.** (2016). The goldilocks principle: Reading children's books with explicit memory representations. *Proceedings of the International Conference on Learning Representations (ICLR)*.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Hirschman L., **Light M.**, **Breck E.** and **Burger J.D.** (1999). Deep Read: A Reading Comprehension System. *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, pp. 325-32.

Hoang L., **Wiseman S.** and **Rush A.M.** (2018). Entity Tracking Improves Cloze-style Reading Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1049–55

Hochreiter S. and **Schmidhuber J**. (1997). Long short-term memory. *Neural Computation* **9**, 1735–1780.

Horbach A., **Aldabe I.**, **Bexte M.**, **de Lacalle O.L.** and **Maritxalar M.** (2020). Linguistic Appropriateness and Pedagogic Usefulness of Reading Comprehension Questions. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 1753-62.

Htut P.M., **Bowman S.R.** and **Cho K.** (2018). Training a Ranking Function for Open-Domain Question Answering. *Proceedings of NAACL-HLT 2018: Student Research Workshop*, pp. 120–27.

Hu M., **Peng Y.**, **Huang Z.** and **Li D.** (2019a). A Multi-Type Multi-Span Network for Reading Comprehension that Requires Discrete Reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1596-606. Hong Kong, China.

Hu M., **Peng Y.**, **Huang Z.** and **Li D.** (2019b). Retrieve, Read, Rerank: Towards End-to-End Multi-Document Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy.

Hu M., **Peng Y.**, **Huang Z.**, **Qiu X.**, **Wei F.** and **Zhou M.** (2018a). Reinforced Mnemonic Reader for Machine Reading Comprehension. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pp. 4099-106.

Hu M., **Peng Y.**, **Wei F.**, **Huang Z.**, **Li D.**, **Yang N.** and **Zhou M.** (2018b). Attention-Guided Answer Distillation for Machine Reading Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4243–52.

Hu M., **Wei F.**, **Peng Y.**, **Huang Z.**, **Yang N.** and **Li D**. (2019c). Read+ Verify: Machine Reading Comprehension with Unanswerable Questions. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6529-37.

Huang H.-Y., **Zhu C.**, **Shen Y.** and **Chen W.** (2018). Fusionnet: Fusing via fully-aware attention with application to machine comprehension. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Huang K., **Tang Y.**, **Huang J.**, **He X.** and **Zhou B.** (2019a). Relation Module for Non-Answerable Predictions on Reading Comprehension. *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 747-56.

Huang L., **Le Bras R.**, **Bhagavatula C.** and **Choi Y.** (2019b). Cosmos QA: Machine Reading Comprehension with Contextual Commonsense Reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2391-401. Hong Kong, China.

Indurthi S., **Yu S.**, **Back S.** and **CuayÁhuitl H.** (2018). Cut to the Chase: A Context Zoom-in Network for Reading Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 570–75. Association for Computational Linguistics.

Ingale V. and **Singh P**. (2019). Datasets for Machine Reading Comprehension: A Literature Review. *Available at SSRN 3454037*.

Inoue N., **Stenetorp P.** and **Inui K.** (2020). R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6740-50.

Jia R. and **Liang P**. (2017). Adversarial examples for evaluating reading comprehension systems. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* pp. 2021–31.

Jiang Y., **Joshi N.**, **Chen Y.-C.** and **Bansal M.** (2019). Explore, Propose, and Assemble: An Interpretable Model for Multi-Hop Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2714–25. Florence, Italy.

Jin D., **Gao S.**, **Kao J.-Y.**, **Chung T.** and **Hakkani-tur D.** (2020). Mmm: Multi-stage multi-task learning for multi-choice reading comprehension. *The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*.

Jin W., **Yang G.** and **Zhu H.** (2019). An Efficient Machine Reading Comprehension Method Based on Attention Mechanism. *2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 1297-302. IEEE.

Jing Y., **Xiong D.** and **Zhen Y.** (2019). BiPaR: A Bilingual Parallel Dataset for Multilingual and Cross-lingual Reading Comprehension on Novels. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2452–62. Hong Kong, China.

Joshi M., **Choi E.**, **Weld D.S.** and **Zettlemoyer L.** (2017). Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601-11.

Jurafsky D. and **Martin J.H.** (2019). *Speech and language processing*. Pearson London.

Kadlec R., **Bajgar O.** and **Kleindienst J.** (2016a). From Particular to General: A Preliminary Case Study of Transfer Learning in Reading Comprehension. *NIPS Machine Intelligence Workshop*.

Kadlec R., **Schmid M.**, **Bajgar O.** and **Kleindienst J.** (2016b). Text understanding with the attention sum reader network. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* pp. 908–18.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Ke N.R., **Zolna K.**, **Sordoni A.**, **Lin Z.**, **Trischler A.**, **Bengio Y.**, **Pineau J.**, **Charlin L.** and **Pal C.** (2018). Focused Hierarchical RNNs for Conditional Sequence Processing. *Proceedings of the 35 th International Conference on Machine Learning*.

Khashabi D., **Chaturvedi S.**, **Roth M.**, **Upadhyay S.** and **Roth D.** (2018). Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 252-62.

Kim Y. (2014). Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1746-51. Doha, Qatar.

Kim Y., **Jernite Y.**, **Sontag D.** and **Rush A.M.** (2016). Character-aware neural language models. *Thirtieth AAAI Conference on Artificial Intelligence*.

Kobayashi S., **Tian R.**, **Okazaki N.** and **Inui K.** (2016). Dynamic entity representation with max-pooling improves machine reading. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 850-55.

KoČiskÝ T., **Schwarz J.**, **Blunsom P.**, **Dyer C.**, **Hermann K.M.**, **Melis G.** and **Grefenstette E.** (2018). The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics* **6**, 317–328.

Kodra L. and **Meçe E.K.** (2017). Question Answering Systems: A Review on Present Developments, Challenges and Trends. *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE* **8**, 217–224.

Kundu S. and **Ng H.T.** (2018a). A Nil-Aware Answer Extraction Framework for Question Answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4243-52.

Kundu S. and **Ng H.T.** (2018b). A Question-Focused Multi-Factor Attention Network for Question Answering. *Association for the Advancement of Artificial Intelligence (AAAI2018)*.

Kwiatkowski T., **Palomaki J.**, **Redfield O.**, **Collins M.**, **Parikh A.**, **Alberti C.**, **Epstein D.**, **Polosukhin I.**, **Devlin J.** and **Lee K.** (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466.

Lai G., **Xie Q.**, **Liu H.**, **Yang Y.** and **Hovy E.** (2017). Race: Large-scale reading comprehension dataset from examinations. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–94

LeCun Y., **Bottou L.**, **Bengio Y.** and **Haffner P.** (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE* **86**, 2278–2324.

Lee G., **Hwang S.-w.** and **Cho H.** (2020). SQuAD2-CR: Semi-supervised Annotation for Cause and Rationales for Unanswerability in SQuAD 2.0. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5425-32.

Lee H.-g. and **Kim H.** (2020). GF-Net: Improving Machine Reading Comprehension with Feature Gates. *Pattern Recognition Letters* **129**, 8–15.

Lee K., **Park S.**, **Han H.**, **Yeo J.**, **Hwang S.-w.** and **Lee J.** (2019a). Learning with limited data for multilingual reading comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2833-43.

Lee S., **Kim D.** and **Park J.** (2019b). Domain-agnostic Question-Answering with Adversarial Training. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 196-202. Hong Kong, China.

Lehnert W.G. (1977). The process of question answering. Yale Univ New Haven Conn Dept Of Computer Science.

Li H., **Zhang X.**, **Liu Y.**, **Zhang Y.**, **Wang Q.**, **Zhou X.**, **Liu J.**, **Wu H.** and **Wang H**. (2019a). D-NET: A Simple Framework for Improving the Generalization of Machine Reading Comprehension. *Proceedings of 2nd Machine Reading for Reading Comprehension Workshop at EMNLP*.

Li J., **Li B.** and **Lv X.** (2018). Machine Reading Comprehension Based on the Combination of BIDAF Model and Word Vectors. *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, pp. 89. ACM.

Li X., **Zhang Z.**, **Zhu W.**, **Li Z.**, **Ni Y.**, **Gao P.**, **Yan J.** and **Xie G.** (2019b). Pingan Smart Health and SJTU at COIN-Shared Task: utilizing Pre-trained Language Models and Common-sense Knowledge in Machine Reading Tasks. *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pp. 93-98.

Li Y., **Li H.** and **Liu J.** (2019c). Towards Robust Neural Machine Reading Comprehension via Question Paraphrases. *2019 International Conference on Asian Language Processing (IALP)*, pp. 290-95. IEEE.

Liang Y., **Li J.** and **Yin J.** (2019). A New Multi-Choice Reading Comprehension Dataset for Curriculum Learning. *Asian Conference on Machine Learning*, pp. 742-57.

Lin C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. *Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*: Association for Computational Linguistics.

Lin K., **Tafjord O.**, **Clark P.** and **Gardner M.** (2019). Reasoning Over Paragraph Effects in Situations. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 58-62. Hong Kong, China.

Lin X., **Liu R.** and **Li Y.** (2018). An Option Gate Module for Sentence Inference on Machine Reading Comprehension. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1743-46. ACM.

Liu C., **Zhao Y.**, **Si Q.**, **Zhang H.**, **Li B.** and **Yu D.** (2018a). Multi-Perspective Fusion Network for Commonsense Reading Comprehension. *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pp. 262-74. Springer.

Liu D., **Gong Y.**, **Fu J.**, **Yan Y.**, **Chen J.**, **Jiang D.**, **Lv J.** and **Duan N.** (2020a). RikiNet: Reading Wikipedia Pages for Natural Question Answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6762–71.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Liu F. and **Perez J**. (2017). Gated end-to-end memory networks. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 1-10.

Liu J., **Lin Y.**, **Liu Z.** and **Sun M.** (2019a). XQA: A cross-lingual open-domain question answering dataset. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2358-68.

Liu J., **Wei W.**, **Sun M.**, **Chen H.**, **Du Y.** and **Lin D.** (2018b). A Multi-answer Multi-task Framework for Real-world Machine Reading Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2109-18.

Liu K., **Liu X.**, **Yang A.**, **Liu J.**, **Su J.**, **Li S.** and **She Q.** (2020b). A Robust Adversarial Training Approach to Machine Reading Comprehension. *AAAI*, pp. 8392-400.

Liu R., **Hu J.**, **Wei W.**, **Yang Z.** and **Nyberg E**. (2017). Structural embedding of syntactic trees for machine comprehension. *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* pp. 815–24

Liu S., **Zhang S.**, **Zhang X.** and **Wang H.** (2019b). R-trans: RNN Transformer Network for Chinese Machine Reading Comprehension. *IEEE Access* **7**, 27736–27745.

Liu S., **Zhang X.**, **Zhang S.**, **Wang H.** and **Zhang W.** (2019c). Review Neural Machine Reading Comprehension: Methods and Trends. *Applied Sciences* **9**, 3698.

Liu X., **Shen Y.**, **Duh K.** and **Gao J.** (2018c). Stochastic Answer Networks for Machine Reading Comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*: 1694–704

Liu Y., **Huang Z.**, **Hu M.**, **Du S.**, **Peng Y.**, **Li D.** and **Wang X.** (2018d). MFM: A Multi-level Fused Sequence Matching Model for Candidates Filtering in Multi-paragraphs Question-Answering. *Pacific Rim Conference on Multimedia*, pp. 449-58. Springer.

Longpre S., **Lu Y.**, **Tu Z.** and **DuBois C**. (2019). An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 220-27.

Ma K., **Jurczyk T.** and **Choi J.D.** (2018). Challenging Reading Comprehension on Daily Conversation: Passage Completion on Multiparty Dialog. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 2039-48.

Manning C.D., **Raghavan P.** and **Schütze H.** (2008). Chapter 8: Evaluation in information retrieval. *Introduction to information retrieval*: Cambridge University Press.

Miao H., **Liu R.** and **Gao S.** (2019a). A Multiple Granularity Co-Reasoning Model for Multi-choice Reading Comprehension. *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-7. IEEE.

Miao H., **Liu R.** and **Gao S.** (2019b). Option Attentive Capsule Network for Multi-choice Reading Comprehension. *International Conference on Neural Information Processing*, pp. 306-18. Springer.

Mihaylov T. and **Frank A**. (2018). Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 821–32.

Mihaylov T. and **Frank A**. (2019). Discourse-Aware Semantic Self-Attention for Narrative Reading Comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2541-52. Hong Kong, China.

Mihaylov T., **Kozareva Z.** and **Frank A.** (2017). Neural Skill Transfer from Supervised Language Tasks to Reading Comprehension. *Workshop on Learning with Limited Labeled Data: Weak Supervision and Beyond at NIPS*.

Mikolov T., **Sutskever I.**, **Chen K.**, **Corrado G.S.** and **Dean J.** (2013). Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, pp. 3111-19.

Min S., **Seo M.** and **Hajishirzi H.** (2017). Question answering through transfer learning from large fine-grained supervision data. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)* pp. 510–17.

Min S., **Zhong V.**, **Socher R.** and **Xiong C.** (2018). Efficient and Robust Question Answering from Minimal Context over Documents. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1725–35.

Min S., **Zhong V.**, **Zettlemoyer L.** and **Hajishirzi H.** (2019). Multi-hop Reading Comprehension through Question Decomposition and Rescoring. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6097-109. Florence, Italy.

Mozannar H., **Maamary E.**, **El Hajal K.** and **Hajj H.** (2019). Neural Arabic Question Answering. *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pp. 108-18. Florence, Italy.

Munkhdalai T. and **Yu H.** (2017). Reasoning with memory augmented neural networks for language comprehension. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Nakatsuji M. and **Okui S.** (2020). Answer Generation through Unified Memories over Multiple Passages. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI-20)*.

Ng H.T., **Teo L.H.** and **Kwan J.L.P.** (2000). A Machine Learning Approach to Answering Questions for Reading Comprehension Tests. *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 124-32.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Nguyen T., **Rosenberg M.**, **Song X.**, **Gao J.**, **Tiwary S.**, **Majumder R.** and **Deng L.** (2016). MS MARCO: A Human Generated Machine Reading Comprehension Dataset. *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (CoCo@NIPS)*, Barcelona, Spain.

Nie Y., **Wang S.** and **Bansal M.** (2019). Revealing the Importance of Semantic Retrieval for Machine Reading at Scale. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2553-66. Hong Kong, China.

Nishida K., **Nishida K.**, **Nagata M.**, **Otsuka A.**, **Saito I.**, **Asano H.** and **Tomita J.** (2019a). Answering while Summarizing: Multi-task Learning for Multi-hop QA with Evidence Extraction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2335-45. Florence, Italy.

Nishida K., **Nishida K.**, **Saito I.**, **Asano H.** and **Tomita J.** (2020). Unsupervised Domain Adaptation of Language Models for Reading Comprehension. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pp. 5392–99.

Nishida K., **Saito I.**, **Nishida K.**, **Shinoda K.**, **Otsuka A.**, **Asano H.** and **Tomita J.** (2019b). Multi-style Generative Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2273-84. Florence, Italy.

Nishida K., **Saito I.**, **Otsuka A.**, **Asano H.** and **Tomita J.** (2018). Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 647-56. ACM.

Niu Y., **Jiao F.**, **Zhou M.**, **Yao T.**, **Xu J.** and **Huang M.** (2020). A Self-Training Method for Machine Reading Comprehension with Soft Evidence Extraction. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3916–27.

Onishi T., **Wang H.**, **Bansal M.**, **Gimpel K.** and **McAllester D.** (2016). Who did what: A large-scale person-centered cloze dataset. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2230–35.

Osama R., **El-Makky N.M.** and **Torki M.** (2019). Question Answering Using Hierarchical Attention on Top of BERT Features. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 191-95.

Ostermann S., **Modi A.**, **Roth M.**, **Thater S.** and **Pinkal M.** (2018). MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Ostermann S., **Roth M.** and **Pinkal M.** (2019). MCScript2.0: A Machine Comprehension Corpus Focused on Script Events and Participants. *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (∗SEM 2019)*, pp. 103-17. Minneapolis, Minnesota.

Pampari A., **Raghavan P.**, **Liang J.** and **Peng J.** (2018). emrQA: A Large Corpus for Question Answering on Electronic Medical Records. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2357–68

Pang L., **Lan Y.**, **Guo J.**, **Xu J.**, **Su L.** and **Cheng X.** (2019). Has-qa: Hierarchical answer spans model for open-domain question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 6875-82.

Papineni K., **Roukos S.**, **Ward T.** and **Zhu W.-J.** (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311-18. Association for Computational Linguistics.

Pappas D., **Stavropoulos P.**, **Androutsopoulos I.** and **McDonald R.** (2020). BioMRC: A Dataset for Biomedical Machine Reading Comprehension. *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pp. 140-49.

Park C., **Lee C.** and **Song H.** (2019). VS3-NET: Neural Variational Inference Model for Machine-Reading Comprehension. *ETRI Journal* **41**, 771–781.

Pennington J., **Socher R.** and **Manning C.** (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-43.

Peters M.E., **Neumann M.**, **Iyyer M.**, **Gardner M.**, **Clark C.**, **Lee K.** and **Zettlemoyer L.** (2018). Deep contextualized word representations. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2227-37. New Orleans, Louisiana.

Prakash T., **Tripathy B.K.** and **Banu K.S.** (2018). ALICE: A Natural Language Question Answering System Using Dynamic Attention and Memory. *International Conference on Soft Computing Systems*, pp. 274-82. Springer.

Pugaliya H., **Route J.**, **Ma K.**, **Geng Y.** and **Nyberg E.** (2019). Bend but Don't Break? Multi-Challenge Stress Test for QA Models. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 125-36.

Qiu D., **Zhang Y.**, **Feng X.**, **Liao X.**, **Jiang W.**, **Lyu Y.**, **Liu K.** and **Zhao J.** (2019). Machine Reading Comprehension Using Structural Knowledge Graph-aware Network. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5898-903.

Radford A., **Narasimhan K.**, **Salimans T.** and **Sutskever I.** (2018). Improving language understanding by generative pre-training.

Rajpurkar P., **Jia R.** and **Liang P.** (2018). Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pp. 784–89.

Rajpurkar P., **Zhang J.**, **Lopyrev K.** and **Liang P.** (2016). Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–92.

Ran Q., **Lin Y.**, **Li P.**, **Zhou J.** and **Liu Z.** (2019). NumNet: Machine Reading Comprehension with Numerical Reasoning. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2474-84. Hong Kong, China.

Reddy S., **Chen D.** and **Manning C.D.** (2019). Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics* **7**, 249–266.

Ren M., **Huang H.**, **Wei R.**, **Liu H.**, **Bai Y.**, **Wang Y.** and **Gao Y.** (2019). Multiple Perspective Answer Reranking for Multi-passage Reading Comprehension. *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 736-47. Springer.

Ren Q., **Cheng X.** and **Su S.** (2020). Multi-Task Learning with Generative Adversarial Training for Multi-Passage Machine Reading Comprehension. *AAAI*, pp. 8705-12.

Richardson M., **Burges C.J.C** and **Renshaw E.** (2013). Mctest: A challenge dataset for the open-domain machine comprehension of text. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193-203.

Riloff E. and **Thelen M**. (2000). A Rule-Based Question Answering System for Reading Comprehension Tests. *Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding sytems*, pp. 13-19. Association for Computational Linguistics.

Ruder S. (2019). *Neural Transfer Learning for Natural Language Processing*. NATIONAL UNIVERSITY OF IRELAND, GALWAY.

Sachan M. and **Xing E**. (2018). Self-Training for Jointly Learning to Ask and Answer Questions. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 629-40.

Saha A., **Aralikatte R.**, **Khapra M.M.** and **Sankaranarayanan K.** (2018). DuoRC: Towards Complex Language Understanding with Paraphrased Reading Comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1683–93.

Salant S. and **Berant J**. (2018). Contextualized Word Representations for Reading Comprehension. *Proceedings of NAACL-HLT 2018*, pp. 554–59.

Sayama H.F., **Araujo A.V.** and **Fernandes E.R.** (2019). FaQuAD: Reading Comprehension Dataset in the Domain of Brazilian Higher Education. *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 443-48. IEEE.

Schlegel V., **Valentino M.**, **Freitas A.**, **Nenadic G.** and **Batista-Navarro R.** (2020). A framework for evaluation of Machine Reading Comprehension Gold Standards. *Proceedings of the 12th Conference on Language Resources and Evaluation*, pp. 5359–69.

Seo M., **Kembhavi A.**, **Farhadi A.** and **Hajishirzi H.** (2017). Bidirectional attention flow for machine comprehension. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Seo M., **Kwiatkowski T.**, **Parikh A.P.**, **Farhadi A.** and **Hajishirzi H.** (2018). Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* pp. 559–64.

Shao C.C., **Liu T.**, **Lai Y.**, **Tseng Y.** and **Tsai S.** (2018). DRCD: a Chinese Machine Reading Comprehension Dataset. *Proceedings of the Workshop on Machine Reading for Question Answering*, pp. pages 37–46. Association for Computational Linguistics.

Sharma P. and **Roychowdhury S.** (2019). IIT-KGP at COIN 2019: Using pre-trained Language Models for modeling Machine Comprehension. *Proceedings of the First Workshop on Commonsense Inference in Natural Language Processing*, pp. 80-84.

Shen Y., **Huang P.-S.**, **Gao J.** and **Chen W.** (2017). Reasonet: Learning to Stop Reading in Machine Comprehension. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*, pp. 1047-55. ACM.

Sheng Y., **Lan M.** and **Wu Y.** (2018). ECNU at SemEval-2018 Task 11: Using Deep Learning Method to Address Machine Comprehension Task. *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 1048-52. Association for Computational Linguistics.

Song J., **Tang S.**, **Qian T.**, **Zhu W.** and **Wu F.** (2018). Reading Document and Answering Question via Global Attentional Inference. *Pacific Rim Conference on Multimedia (PCM 2018)*, pp. 335-45. Springer.

Song L., **Wang Z.**, **Yu M.**, **Zhang Y.**, **Florian R.** and **Gildea D.** (2020). Evidence Integration for Multi-hop Reading Comprehension with Graph Neural Networks. *IEEE Transactions on knowledge data engineering*.

Soni S. and **Roberts K**. (2020). Evaluation of Dataset Selection for Pre-Training and Fine-Tuning Transformer Language Models for Clinical Question Answering. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 5532-38.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Srivastava R.K., **Greff K.** and **Schmidhuber J.** (2015). Training very deep networks. *Advances in neural information processing systems*, pp. 2377-85.

Su D., **Xu Y.**, **Winata G.I.**, **Xu P.**, **Kim H.**, **Liu Z.** and **Fung P.** (2019). Generalizing Question Answering System with Pre-trained Language Model Fine-tuning. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 203-11.

Sugawara S., **Inui K.**, **Sekine S.** and **Aizawa A.** (2018). What Makes Reading Comprehension Questions Easier? *Proceedings of 2018 conference on empirical methods in natural language processing*, pp. 4208-40219.

Sugawara S., **Kido Y.**, **Yokono H.** and **Aizawa A.** (2017). Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 806-17.

Sugawara S., **Stenetorp P.**, **Inui K.** and **Aizawa A.** (2020). Assessing the Benchmarking Capacity of Machine Reading Comprehension Datasets. *AAAI*, pp. 8918-27.

Sun K., **Yu D.**, **Yu D.** and **Cardie C.** (2019). Improving Machine Reading Comprehension with General Reading Strategies. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2633-43.

Sun K., **Yu D.**, **Yu D.** and **Cardie C.** (2020). Investigating Prior Knowledge for Challenging Chinese Machine Reading Comprehension. *Transactions of the Association for Computational Linguistics* **8**, 141–155.

Šuster S. and **Daelemans W**. (2018). CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension. *Proceedings of NAACL-HLT 2018* pp. 1551–63.

Swayamdipta S., **Parikh A.P.** and **Kwiatkowski T.** (2018). Multi-mention learning for reading comprehension with neural cascades. *Proceeding of the International Conference on Learning Representations (ICLR)*.

Takahashi T., **Taniguchi M.**, **Taniguchi T.** and **Ohkuma T.** (2019). CLER: Cross-task learning with expert representation to generalize reading and understanding. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 183-90.

Talmor A. and **Berant J**. (2019). MultiQA: An Empirical Investigation of Generalization and Transfer in Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4911-21. Florence, Italy.

Tan C., **Wei F.**, **Yang N.**, **Du B.**, **Lv W.** and **Zhou M.** (2018a). S-Net: From Answer Extraction to Answer Synthesis for Machine Reading Comprehension. *Association for the Advancement of Artificial Intelligence (AAAI)*.

Tan C., **Wei F.**, **Zhou Q.**, **Yang N.**, **Lv W.** and **Zhou M.** (2018b). I Know There Is No Answer: Modeling Answer Validation for Machine Reading Comprehension. *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 85-97. Springer.

Tang H., **Hong Y.**, **Chen X.**, **Wu K.** and **Zhang M.** (2019a). How to Answer Comparison Questions. *2019 International Conference on Asian Language Processing (IALP)*, pp. 331-36. IEEE.

Tang M., **Cai J.** and **Zhuo H.H.** (2019b). Multi-matching network for multiple choice reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7088-95.

Tay Y., **Luu A.T.** and **Hui S.C.** (2018). Multi-Granular Sequence Encoding via Dilated Compositional Units for Reading Comprehension. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2141-51.

Tay Y., **Wang S.**, **Luu A.T.**, **Fu J.**, **Phan M.C.**, **Yuan X.**, **Rao J.**, **Hui S.C.** and **Zhang A.** (2019). Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4922-31. Florence, Italy.

Trischler A., **Wang T.**, **Yuan X.**, **Harris J.**, **Sordoni A.**, **Bachman P.** and **Suleman K.** (2017). Newsqa: A machine comprehension dataset. *Proceedings of the 2nd Workshop on Representation Learning for NLP*.

Tu M., **Huang K.**, **Wang G.**, **Huang J.**, **He X.** and **Zhou B.** (2020). Select, Answer and Explain: Interpretable Multi-Hop Reading Comprehension over Multiple Documents. *AAAI*, pp. 9073-80.

Tu M., **Wang G.**, **Huang J.**, **Tang Y.**, **He X.** and **Zhou B.** (2019). Multi-hop Reading Comprehension across Multiple Documents by Reasoning over Heterogeneous Graphs. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2704-13. Florence, Italy.

Turpin A. and **Scholer F**. (2006). User performance versus precision measures for simple search tasks. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 11-18. ACM.

Vaswani A., **Shazeer N.**, **Parmar N.**, **Uszkoreit J.**, **Jones L.**, **Gomez A.N.**, **Kaiser Ł.** and **Polosukhin I.** (2017). Attention is all you need. *Advances in neural information processing systems*, pp. 5998-6008.

Vedantam R., **Lawrence Zitnick C.** and **Parikh D.** (2015). Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566-75.

Wang B., **Guo S.**, **Liu K.**, **He S.** and **Zhao J.** (2016). Employing External Rich Knowledge for Machine Comprehension. *International Joint Conferences on Artificial Intelligence (IJCAI)*, pp. 2929-25.

Wang B., **Yao T.**, **Zhang Q.**, **Xu J.** and **Wang X.** (2020b). ReCO: A Large Scale Chinese Reading Comprehension Dataset on Opinion. *AAAI*, pp. 9146-53.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Wang B., **Yao T.**, **Zhang Q.**, **Xu J.**, **Liu K.**, **Tian Z.** and **Zhao J.** (2019a). Unsupervised Story Comprehension with Hierarchical Encoder-Decoder. *Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval*, pp. 93-100.

Wang B., **Zhang X.**, **Zhou X.** and **Li J.** (2020a). A Gated Dilated Convolution with Attention Model for Clinical Cloze-Style Reading Comprehension. *International Journal of Environmental Research Public Health* **17**, 1323.

Wang C. and **Jiang H.** (2019). Explicit Utilization of General Knowledge in Machine Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2263–72. Florence, Italy.

Wang H., **Gan Z.**, **Liu X.**, **Liu J.**, **Gao J.** and **Wang H.** (2019e). Adversarial Domain Adaptation for Machine Reading Comprehension. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2510-20. Hong Kong, China.

Wang H., **Lu W.** and **Tang Z.** (2019d). Incorporating External Knowledge to Boost Machine Comprehension Based Question Answering. *European Conference on Information Retrieval*, pp. 819-27. Springer.

Wang H., **Yu D.**, **Sun K.**, **Chen J.**, **Yu D.**, **McAllester D.** and **Roth D.** (2019b). Evidence Sentence Extraction for Machine Reading Comprehension. *Proceedings of the 23rd Conference on Computational Natural Language Learning*, pp. 696–707. Hong Kong, China.

Wang H., **Yu M.**, **Guo X.**, **Das R.**, **Xiong W.** and **Gao T.** (2019c). Do Multi-hop Readers Dream of Reasoning Chains? *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 91-97. Hong Kong, China.

Wang S. and **Jiang J**. (2017). Machine comprehension using match-lstm and answer pointer. *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1-15. Toulon, France.

Wang S., **Yu M.**, **Chang S.** and **Jiang J.** (2018a). A Co-Matching Model for Multi-choice Reading Comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Short Papers)*: 746–51.

Wang S., **Yu M.**, **Guo X.**, **Wang Z.**, **Klinger T.**, **Zhang W.**, **Chang S.**, **Tesauro G.**, **Zhou B.** and **Jiang J.** (2018b). R3: Reinforced ranker-reader for open-domain question answering. *Association for the Advancement of Artificial Intelligence (AAAI 2018)*.

Wang S., **Yu M.**, **Jiang J.**, **Zhang W.**, **Guo X.**, **Chang S.**, **Wang Z.**, **Klinger T.**, **Tesauro G.** and **Campbell M**. (2018c). Evidence Aggregation for Answer Re-Ranking in Open-Domain Question Answering. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wang T., **Yuan X.** and **Trischler A.** (2017a). A joint model for question answering and question generation. *Learning to generate natural language workshop, ICML 2017*.

Wang W., **Yan M.** and **Wu C.** (2018d). Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1705-14.

Wang W., **Yang N.**, **Wei F.**, **Chang B.** and **Zhou M.** (2017b). Gated self-matching networks for reading comprehension and question answering. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 189-98.

Wang Y. and **Bansal M**. (2018). Robust Machine Comprehension Models via Adversarial Training. *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018)*, pp. 575–81. Association for Computational Linguistics.

Wang Y., **Liu K.**, **Liu J.**, **He W.**, **Lyu Y.**, **Wu H.**, **Li S.** and **Wang H.** (2018e). Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1918-27.

Wang Z., **Liu J.**, **Xiao X.**, **Lyu Y.** and **Wu T.** (2018f). Joint Training of Candidate Extraction and Answer Selection for Reading Comprehension. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pp. 1715–24.

Watarai T. and **Tsuchiya M**. (2020). Developing Dataset of Japanese Slot Filling Quizzes Designed for Evaluation of Machine Reading Comprehension. *Proceedings of The 12th Language Resources and Evaluation Conference*, pp. 6895-901.

Weissenborn D., **Wiese G.** and **Seiffe L.** (2017). Making Neural QA as Simple as Possible but not Simpler. *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL)*, pp. 271–80. Vancouver, Canada.

Welbl J., **Liu N.F.** and **Gardner M.** (2017). Crowdsourcing multiple choice science questions. *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106. Association for Computational Linguistics.

Welbl J., **Minervini P.**, **Bartolo M.**, **Stenetorp P.** and **Riedel S.** (2020). Undersensitivity in neural reading comprehension. *International Conference on Learning Representations (ICLR2020)*.

Welbl J., **Stenetorp P.** and **Riedel S.** (2018). Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics* **6**, 287–302.

Wu B., **Huang H.**, **Wang Z.**, **Feng Q.**, **Yu J.** and **Wang B.** (2019). Improving the Robustness of Deep Reading Comprehension Models by Leveraging Syntax Prior. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 53-57.

Wu Z. and **Xu H**. (2020). Improving the Robustness of Machine Reading Comprehension Model with Hierarchical Knowledge and Auxiliary Unanswerability Prediction. *Knowledge-Based Systems*: 106075.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Xia J., **Wu C.** and **Yan M.** (2019). Incorporating Relation Knowledge into Commonsense Reading Comprehension with Multi-task Learning. *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 2393–96. Beijing, China.

Xie P. and **Xing E**. (2017). A constituent-centric neural architecture for reading comprehension. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1405-14.

Xie Q., **Lai G.**, **Dai Z.** and **Hovy E.** (2018). Large-scale Cloze Test Dataset Created by Teachers. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2344-56.

Xiong C., **Zhong V.** and **Socher R.** (2017). Dynamic coattention networks for question answering. *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Xiong C., **Zhong V.** and **Socher R.** (2018). Dcn+: Mixed objective and deep residual coattention for question answering. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Xiong W., **Yu M.**, **Guo X.**, **Wang H.**, **Chang S.**, **Campbell M.** and **Wang W.Y.** (2019). Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering. *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*, pp. 48-52. Hong Kong, China.

Xu Y., **Liu W.**, **Chen G.**, **Ren B.**, **Zhang S.**, **Gao S.** and **Guo J.** (2019a). Enhancing Machine Reading Comprehension With Position Information. *IEEE Access* **7**, 141602–141611.

Xu Y., **Liu X.**, **Shen Y.**, **Liu J.** and **Gao J.** (2019b). Multi-task Learning with Sample Re-weighting for Machine Reading Comprehension. *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)*, pp. 2644-55. Minneapolis, Minnesota.

Yadav M., **Vig L.** and **Shroff G.** (2017). Learning and Knowledge Transfer with Memory Networks for Machine Comprehension. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pp. 850-59.

Yan M., **Xia J.**, **Wu C.**, **Bi B.**, **Zhao Z.**, **Zhang J.**, **Si L.**, **Wang R.**, **Wang W.** and **Chen H.** (2019). A deep cascade model for multi-document reading comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7354-61.

Yan M., **Zhang H.**, **Jin D.** and **Zhou J.T.** (2020). Multi-source Meta Transfer for Low Resource Multiple-Choice Question Answering. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7331-41.

Yang A., **Wang Q.**, **Liu J.**, **Liu K.**, **Lyu Y.**, **Wu H.**, **She Q.** and **Li S.** (2019a). Enhancing Pre-Trained Language Representations with Rich Knowledge for Machine Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2346-57.

Yang Y., **Kang S.** and **Seo J.** (2020). Improved Machine Reading Comprehension Using Data Validation for Weakly Labeled Data. *IEEE Access* **8**, 5667–5677.

Yang Z., **Dai Z.**, **Yang Y.**, **Carbonell J.**, **Salakhutdinov R.** and **Le Q.V.** (2019b). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Advances in neural information processing systems*, pp. 5753-63.

Yang Z., **Dhingra B.**, **Yuan Y.**, **Hu J.**, **Cohen W.W.** and **Salakhutdinov R.** (2017a). Words or characters? fine-grained gating for reading comprehension. *Proceedings of the 5th International Conference on Learning Representations, (ICLR)*, Toulon, France.

Yang Z., **Hu J.**, **Salakhutdinov R.** and **Cohen W.W.** (2017b). Semi-supervised QA with generative domain-adaptive nets. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1040-50.

Yang Z., **Qi P.**, **Zhang S.**, **Bengio Y.**, **Cohen W.**, **Salakhutdinov R.** and **Manning C.D.** (2018). HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369-80.

Yao J., **Feng M.**, **Feng H.**, **Wang Z.**, **Zhang Y.** and **Xue N.** (2019). Smart: A stratified machine reading test. *CCF International Conference on Natural Language Processing and Chinese Computing*, pp. 67-79. Springer.

Yin W., **Ebert S.** and **Schütze H.** (2016). Attention-based convolutional neural network for machine comprehension. *Proceedings of 2016 NAACL Human-Computer Question Answering Workshop*, pp. 15–21.

Yu A.W., **Dohan D.**, **Luong M.-T.**, **Zhao R.**, **Chen K.**, **Norouzi M.** and **Le Q.V.** (2018). QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Yu J., **Zha Z.** and **Yin J.** (2019). Inferential machine comprehension: Answering questions by recursively deducing the evidence chain from text. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2241-51.

Yu W., **Jiang Z.**, **Dong Y.** and **Feng J.** (2020). ReClor: A Reading Comprehension Dataset Requiring Logical Reasoning. *International Conference on Learning Representations (ICLR2020)*.

Yuan F., **Shou L.**, **Bai X.**, **Gong M.**, **Liang Y.**, **Duan N.**, **Fu Y.** and **Jiang D.** (2020a). Enhancing Answer Boundary Detection for Multilingual Machine Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 925–34.

Yuan F., **Xu Y.**, **Lin Z.**, **Wang W.** and **Shi G.** (2019). Multi-perspective Denoising Reader for Multi-paragraph Reading Comprehension. *International Conference on Neural Information Processing*, pp. 222-34. Springer.

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Yuan X., **Fu J.**, **Cote M.-A.**, **Tay Y.**, **Pal C.** and **Trischler A.** (2020b). Interactive machine comprehension with information seeking agents. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*: 2325–38.

Yue X., **Gutierrez B.J.** and **Sun H.** (2020). Clinical Reading Comprehension: A Thorough Analysis of the emrQA Dataset. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Zhang C., **Luo C.**, **Lu J.**, **Liu A.**, **Bai B.**, **Bai K.** and **Xu Z.** (2020a). Read, Attend, and Exclude: Multi-Choice Reading Comprehension by Mimicking Human Reasoning Process. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1945-48.

Zhang J., **Zhu X.**, **Chen Q.**, **Ling Z.**, **Dai L.**, **Wei S.** and **Jiang H.** (2017). Exploring question representation and adaptation with neural networks. *Computer and Communications (ICCC), 2017 3rd IEEE International Conference on*, pp. 1975-84. IEEE.

Zhang S., **Zhao H.**, **Wu Y.**, **Zhang Z.**, **Zhou X.** and **Zhou X.** (2020b). DCMN+: Dual co-matching network for multi-choice reading comprehension. *AAAI*

Zhang X. and **Wang Z**. (2020). Reception: Wide and Deep Interaction Networks for Machine Reading Comprehension (Student Abstract). *AAAI*, pp. 13987-88.

Zhang X., **Wu J.**, **He Z.**, **Liu X.** and **Su Y.** (2018). Medical Exam Question Answering with Large-scale Reading Comprehension. *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*.

Zhang X., **Yang A.**, **Li S.** and **Wang Y.** (2019a). Machine Reading Comprehension: a Literature Review. *arXiv preprint arXiv:1907.01686*.

Zhang Z., **Wu Y.**, **Zhou J.**, **Duan S.**, **Zhao H.** and **Wang R.** (2020c). SG-Net: Syntax-Guided Machine Reading Comprehension. *AAAI*, pp. 9636-43.

Zhang Z., **Zhao H.**, **Ling K.**, **Li J.**, **Li Z.**, **He S.** and **Fu G.** (2019b). Effective subword segmentation for text comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **27**, 1664–1674.

Zheng B., **Wen H.**, **Liang Y.**, **Duan N.**, **Che W.**, **Jiang D.**, **Zhou M.** and **Liu T.** (2020). Document Modeling with Graph Attention Networks for Multi-grained Machine Reading Comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6708–18.

Zhou M., **Huang M.** and **Zhu X.** (2020a). Robust reading comprehension with linguistic constraints via posterior regularization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28.

Zhou X., **Luo S.** and **Wu Y.** (2020b). Co-Attention Hierarchical Network: Generating Coherent Long Distractors for Reading Comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9725-32.

Zhu H., **Dong L.**, **Wei F.**, **Wang W.**, **Qin B.** and **Liu T.** (2019). Learning to Ask Unanswerable Questions for Machine Reading Comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4238-48. Florence, Italy.

Zhuang Y. and **Wang H**. (2019). Token-level dynamic self-attention network for multi-passage reading comprehension. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2252-62.

Appendix

Table A1. Reviewed papers categorized based on their embedding phase

Character embedding	CNN			(Clark and Gardner 2018), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Shen *et al*. 2017), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Salant and Berant 2018), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Kundu and Ng 2018a), (Back *et al*. 2018), (Seo *et al*. 2018), (Seo *et al*. 2017), (Indurthi *et al*. 2018), (Sugawara *et al*. 2018), (Wang and Bansal 2018), (Yan *et al*. 2019), (Mihaylov and Frank 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Tang, Cai and Zhuo 2019b), (Yuan *et al*. 2019), (Ran *et al*. 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Xiong *et al*. 2019), (Zhuang and Wang 2019), (Park, Lee and Song 2019), (Nishida *et al*. 2019a), (Lee and Kim 2020),
	RNN			(Hu *et al*. 2018a),(Wang *et al*. 2017b), (Tan *et al*. 2018a), (Wang *et al*. 2017a), (Dhingra *et al*. 2017), (Prakash *et al*. 2018), (Hu *et al*. 2018b), (Dhingra, Pruthi and Rajagopal 2018), (Yang *et al*. 2017a), (Ghaeini *et al*. 2018), (Gupta *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Hu *et al*. 2019c), (Wu and Xu 2020)
	other			(Dua *et al*. 2019), (Xiong, Zhong and Socher 2018), (Li *et al*. 2018),(Wang, Yan and Wu 2018d), (Wang *et al*. 2018e), (Min *et al*. 2018), (Yu *et al*. 2018), (Yang *et al*. 2020),
Word embedding	Non-contextual	One hot		(Cui *et al*. 2016),(Liu and Perez 2017), (Tay, Luu and Hui 2018), (Nishida *et al*. 2018), (Lee and Kim 2020), (Chen *et al*. 2020)
		Learned		(Choi *et al*. 2017b), (Greco *et al*. 2016), (Cui *et al*. 2017b), (Liu *et al*. 2018b), (Bauer *et al*. 2018), (Yang *et al*. 2017a), (Cui *et al*. 2016), (Kadlec, Bajgar and Kleindienst 2016a), (Kadlec *et al*. 2016b), (Ke *et al*. 2018), (Du and Cardie 2018), (Guo *et al*. 2020), (Chen *et al*. 2020),
		Fixed Pre-train		(Yan *et al*. 2019), (Angelidis *et al*. 2019), (Chen *et al*. 2019b), (Mihaylov and Frank 2019), (Dua *et al*. 2019), (Pang *et al*. 2019), (Wu *et al*. 2019), (Yu, Zha and Yin 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Hu *et al*. 2019c), (Huang *et al*. 2019a), (Tay *et al*. 2019), (Xiong *et al*. 2019), (Zhuang and Wang 2019), (Wang *et al*. 2019a), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Das *et al*. 2019), (Nakatsuji and Okui 2020), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Song *et al*. 2020), (Back *et al*. 2020)
		Fine-tune		(Miao, Liu and Gao 2019a), (Wang *et al*. 2019e), (Jin, Yang and Zhu 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Tang *et al*. 2019a), (Bi *et al*. 2019), (Nishida *et al*. 2019b), (Miao, Liu and Gao 2019b), (Liu *et al*. 2019b), (Wang *et al*. 2020a), (Zhou *et al*. 2020b), (Ren *et al*. 2020)
	Contextual	Learned	RNN	(Chen *et al*. 2017), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Liu *et al*. 2017), (Hu *et al*. 2018a), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Wang *et al*. 2018b), (Wang *et al*. 2018c), (Golub *et al*. 2017), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Mihaylov, Kozareva and Frank 2017), (Wang *et al*. 2018a), (Kundu and Ng 2018b), (Liu *et al*. 2018b), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Aghaebrahimian 2018), (Seo *et al*. 2018), (Li *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Liu *et al*. 2018a), (Yang *et al*. 2017a),

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A1. Continued

			(Chen *et al*. 2017), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Liu *et al*. 2017), (Hu *et al*. 2018a), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Wang *et al*. 2018b), (Wang *et al*. 2018c), (Golub *et al*. 2017), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Mihaylov, Kozareva and Frank 2017), (Wang *et al*. 2018a), (Kundu and Ng 2018b), (Liu *et al*. 2018b), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Aghaebrahimian 2018), (Seo *et al*. 2018), (Li *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Liu *et al*. 2018a), (Yang *et al*. 2017a), (Lin, Liu and Li 2018), (Hoang *et al*. 2018), (Back *et al*. 2018), (Dhingra *et al*. 2018), (Chen *et al*. 2016), (Trischler *et al*. 2017), (Cui *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng, Lan and Wu 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Tan *et al*. 2018b), (Mihaylov and Frank 2018), (Song *et al*. 2018), (Lee and Kim 2020), (Yang *et al*. 2020), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Wu and Xu 2020), (Ren *et al*. 2020), (Chen *et al*. 2020), (Song *et al*. 2020), (Zhang *et al*. 2020a), (Zhang and Wang 2020)
		CNN	(Yin *et al*. 2016), (Ma *et al*. 2018), (Gupta *et al*. 2018), (Yu *et al*. 2018)
		Fixed Pre-train	(Wang *et al*. 2019b), (Xu *et al*. 2019b), (Hu *et al*. 2019c), (Lee and Kim 2020), (Wu and Xu 2020), (Back *et al*. 2020), (Zhang and Wang 2020), (Nakatsuji and Okui 2020)
		Fine-tune	(Hu *et al*. 2019a), (Ding *et al*. 2019), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Yang *et al*. 2019a), (Su *et al*. 2019), (Andor *et al*. 2019), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Xia, Wu and Yan 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Ren *et al*. 2019), (Nishida *et al*. 2019b), (Li *et al*. 2019b), (Osama, El-Makky and Torki 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie, Wang and Bansal 2019), (Frermann 2019), (Li *et al*. 2019c), (Lee, Kim and Park 2019b), (Liu *et al*. 2020a), (Back *et al*. 2020), (Zhang *et al*. 2020c), (Zhang *et al*. 2020b), (Zhang *et al*. 2020a), (Tu *et al*. 2020), (Chen *et al*. 2020), (Zheng *et al*. 2020)
Hybrid	–		(Clark and Gardner 2018), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Liu *et al*. 2017), (Hu *et al*. 2018a), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Tan *et al*. 2018a), (Zhang *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Xiong *et al*. 2018), (Salant and Berant 2018), (Dhingra *et al*. 2017), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Kundu and Ng 2018a),(Hu *et al*. 2018b), (Back *et al*. 2018), (Seo *et al*. 2018), (Li *et al*. 2018), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Dhingra *et al*. 2018), (Dhingra *et al*. 2017), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Gupta *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Yu *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018; Yan *et al*. 2019), (Mihaylov and Frank 2019), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Ran *et al*. 2019), (Liu *et al*. 2019b), (Hu *et al*. 2019c), (Huang *et al*. 2019a), (Zhuang and Wang 2019), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Xiong *et al*. 2019), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020)
Sentence embedding			(Yin *et al*. 2016), (Liu and Perez 2017), (Chaturvedi *et al*. 2018), (Min *et al*. 2018), (Htut, Bowman and Cho 2018), (Wang *et al*. 2019a), (Park *et al*. 2019), (Guo *et al*. 2020), (Zhou *et al*. 2020b), (Tu *et al*. 2020), (Jin *et al*. 2020)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A2. Reviewed papers categorized based on their reasoning phase

Direction	One direction	(Chen *et al*. 2017), (Yang *et al*. 2017b), (Weissenborn *et al*. 2017), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Huang *et al*. 2018), (Tan *et al*. 2018a), (Wang *et al*. 2018b), (Hewlett *et al*. 2017), (Xie and Xing 2017), (Salant and Berant 2018), (Wang *et al*. 2018c), (Yadav, Vig and Shroff 2017), (Dhingra *et al*. 2017), (Wang *et al*. 2018a), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Lin *et al*. 2018), (Liu *et al*. 2018d), (Liu *et al*. 2018a), (Wang and Jiang 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Cui *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Choi *et al*. 2017a), (Chaturvedi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Tan *et al*. 2018b), (Wang *et al*. 2018f), (Mihaylov and Frank 2018), (Song *et al*. 2018), (Sachan and Xing 2018), (Htut *et al*. 2018), (Liu and Perez 2017), (Yan *et al*. 2019), (Wang *et al*. 2019e), (Angelidis *et al*. 2019), (Chen *et al*. 2019b), (Zhang *et al*. 2019b), (Xu *et al*. 2019b), (Tay *et al*. 2019), (Wang *et al*. 2020a), (Zhang and Wang 2020), (Ren *et al*. 2020), (Song *et al*. 2020)
	Two direction	(Clark and Gardner 2018), (Min *et al*. 2017), (Liu *et al*. 2017), (Golub *et al*. 2017), (Swayamdipta, Parikh and Kwiatkowski 2018), (Duan *et al*. 2017), (Cui *et al*. 2017b), (Liu *et al*. 2018b), (Zhang *et al*. 2018), (Aghaebrahimian 2018), (Back *et al*. 2018), (Seo *et al*. 2018), (Wang *et al*. 2018e), (Hu *et al*. 2018a), (Greco *et al*. 2016), (Zhang *et al*. 2017), (Gong and Bowman 2018), (Xiong *et al*. 2018), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Li *et al*. 2018), (Wang *et al*. 2018d), (Dhingra *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Xiong *et al*. 2017), (Indurthi *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Yu *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Miao *et al*. 2019a), (Hu *et al*. 2019a), (Jin *et al*. 2019), (Ding *et al*. 2019), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang *et al*. 2019b), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Su *et al*. 2019), (Andor *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Wu *et al*. 2019), (Bi *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Das *et al*. 2019), (Nishida *et al*. 2019b), (Ran *et al*. 2019), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Hu *et al*. 2019c), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Guo *et al*. 2020), (Niu *et al*. 2020), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Jin *et al*. 2020), (Pappas *et al*. 2020), (Zheng *et al*. 2020), (Chen *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c), (Zhang *et al*. 2020a)
Dimension	One dimension	(Weissenborn *et al*. 2017), (Shen *et al*. 2017), (Tan *et al*. 2018a), (Hewlett *et al*. 2017), (Lin *et al*. 2018), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Chaturvedi *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Ke *et al*. 2018), (Tan *et al*. 2018b), (Mihaylov and Frank 2018), (Sachan and Xing 2018), (Liu and Perez 2017), (Angelidis *et al*. 2019), (Tu *et al*. 2020), (Niu *et al*. 2020), (Ren *et al*. 2020), (Song *et al*. 2020)
	Two dimension	(Chen *et al*. 2017), (Yang *et al*. 2017b), (Clark and Gardner 2018), (Min *et al*. 2017), (Liu *et al*. 2017), (Golub *et al*. 2017), (Swayamdipta *et al*. 2018), (Duan *et al*. 2017), (Cui *et al*. 2017b), (Liu *et al*. 2018b), (Zhang *et al*. 2018), (Aghaebrahimian 2018),(Back *et al*. 2018), (Seo *et al*. 2018), (Wang *et al*. 2018e), (Wang *et al*. 2017b), (Wang *et al*. 2018b), (Salant and Berant 2018), (Wang *et al*. 2018c), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Liu *et al*. 2018d), (Liu *et al*. 2018a), (Hu *et al*. 2018a), (Lai *et al*. 2017), (Greco *et al*. 2016), (Dhingra *et al*. 2017), (Zhang *et al*. 2017), (Gong and Bowman 2018), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Li *et al*. 2018), (Wang *et al*. 2018d), (Dhingra *et al*. 2018), (Huang *et al*. 2018), (Wang *et al*. 2018a), (Xiong *et al*. 2018), (Xie and Xing 2017), (Yadav *et al*. 2017), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Wang *et al*. 2018f), (Yu *et al*. 2018), (Song *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Htut *et al*. 2018), (Miao *et al*. 2019a), (Yan *et al*. 2019), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Ding *et al*. 2019), (Chen *et al*. 2019b), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b) (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang *et al*. 2019b), (Wang and Jiang 2019) (Jiang *et al*. 2019),

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A2. Reviewed papers categorized based on their reasoning phase

		(Su *et al*. 2019), (Andor *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Wu *et al*. 2019), (Bi *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Das *et al*. 2019), (Nishida *et al*. 2019b), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Hu *et al*. 2019c), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Tay *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Zhang *et al*. 2020a), (Liu *et al*. 2020a), (Jin *et al*. 2020), (Pappas *et al*. 2020), (Zheng *et al*. 2020), (Chen *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)
Number of steps	Single	(Chen *et al*. 2017), (Clark and Gardner 2018), (Min *et al*. 2017), (Huang *et al*. 2018), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Golub *et al*. 2017), (Swayamdipta *et al*. 2018), (Duan *et al*. 2017), (Cui *et al*. 2017b), (Liu *et al*. 2018b), (Zhang *et al*. 2018), (Aghaebrahimian 2018), (Back *et al*. 2018), (Seo *et al*. 2018), (Wang *et al*. 2018e), (Weissenborn *et al*. 2017), (Tan *et al*. 2018a), (Hewlett *et al*. 2017), (Wang *et al*. 2017b), (Wang *et al*. 2018b), (Salant and Berant 2018), (Wang *et al*. 2018c), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Liu *et al*. 2018d), (Liu *et al*. 2018a), (Xiong *et al*. 2018), (Xie and Xing 2017), (Yadav *et al*. 2017), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Choi *et al*. 2017a), (Chaturvedi *et al*. 2018), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Tan *et al*. 2018b), (Wang *et al*. 2018f), (Mihaylov and Frank 2018), (Yu *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Sachan and Xing 2018), (Htut *et al*. 2018), (Miao *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Chen *et al*. 2019b), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Wu *et al*. 2019), (Bi *et al*. 2019), (Tu *et al*. 2019), (Yuan *et al*. 2019), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Miao *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), [8], (Tay *et al*. 2019), (Xiong *et al*. 2019), (Zhuang and Wang 2019), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Niu *et al*. 2020), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Tu *et al*. 2020), (Ren *et al*. 2020), (Song *et al*. 2020), (Chen *et al*. 2020)
	Multi-fixed	(Yang *et al*. 2017b), (Lin *et al*. 2018), (Hu *et al*. 2018a), (Greco *et al*. 2016), (Dhingra *et al*. 2017), (Gong and Bowman 2018), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Li *et al*. 2018), (Wang *et al*. 2018d), (Dhingra *et al*. 2018), (Wang *et al*. 2018a), (Liu and Perez 2017), (Yan *et al*. 2019), (Hu *et al*. 2019a), (Angelidis *et al*. 2019), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Yang *et al*. 2019a), (Wang *et al*. 2019b), (Su *et al*. 2019), (Andor *et al*. 2019), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Xia *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tang *et al*. 2019b), (Ren *et al*. 2019), (Das *et al*. 2019), (Nishida *et al*. 2019b), (Li *et al*. 2019b), (Hu *et al*. 2019c), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Frermann 2019), (Li *et al*. 2019c), (Lee *et al*. 2019b), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Pappas *et al*. 2020), (Zheng *et al*. 2020), (Jin *et al*. 2020), (Back *et al*. 2020), (Liu *et al*. 2020a), (Zhang *et al*. 2020c), (Zhang *et al*. 2020a)
	Multi-dynamic	(Shen *et al*. 2017), (Song *et al*. 2018), (Ding *et al*. 2019), (Yu *et al*. 2019)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A3. Reviewed papers categorized based on their prediction phase

Extraction mode	Boundary identification	(Chen *et al*. 2017), (Yang *et al*. 2017b), (Clark and Gardner 2018), (Weissenborn *et al*. 2017), (Hu *et al*. 2018a), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Wang *et al*. 2018b), (Golub *et al*. 2017), (Swayamdipta *et al*. 2018), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang *et al*. 2018c), (Mihaylov *et al*. 2017), (Cui *et al*. 2017b), (Dhingra *et al*. 2017), (Kundu and Ng 2018b), (Liu *et al*. 2018b), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Hoang *et al*. 2018), (Back *et al*. 2018), (Seo *et al*. 2018), (Khashabi *et al*. 2018), (Li *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Dhingra *et al*. 2018), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Trischler *et al*. 2017), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Munkhdalai and Yu 2017), (Ghaeini *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Tan *et al*. 2018b), (Mihaylov and Frank 2018), (Yu *et al*. 2018), (Nishida *et al*. 2018), (Htut *et al*. 2018), (Yan *et al*. 2019), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Ding *et al*. 2019), (Takahashi *et al*. 2019), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang and Jiang 2019), (Su *et al*. 2019), (Andor *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Wu *et al*. 2019), (Min *et al*. 2019), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Das *et al*. 2019), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Hu *et al*. 2019c), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Ren *et al*. 2020), (Niu *et al*. 2020), (Zheng *et al*. 2020), (Chen *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)
	Candidate ranking	(Min *et al*. 2017), (Choi *et al*. 2017b), (Shen *et al*. 2017), (Golub *et al*. 2017), (Duan *et al*. 2017), (Yadav *et al*. 2017), (Prakash *et al*. 2018), (Aghaebrahimian 2018), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Liu and Perez 2017), (Ma *et al*. 2018), (Chaturvedi *et al*. 2018), (Sheng *et al*. 2018), (Tay *et al*. 2018), (Du and Cardie 2018), (Wang *et al*. 2018f), (Song *et al*. 2018), (Sachan and Xing 2018), (Angelidis *et al*. 2019), (Jiang *et al*. 2019), (Tu *et al*. 2019), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Ren *et al*. 2020), (Pappas *et al*. 2020), (Song *et al*. 2020)
Generation mode	Answer generation	(Hewlett *et al*. 2017), (Bauer *et al*. 2018), (Tan *et al*. 2018a), (Indurthi *et al*. 2018), (Saha *et al*. 2018), (Ke *et al*. 2018), (Hu *et al*. 2019a), (Dua *et al*. 2019), (Andor *et al*. 2019), (Bi *et al*. 2019), (Dehghani *et al*. 2019), (Nishida *et al*. 2019b), (Ran *et al*. 2019), (Tay *et al*. 2019), (Wang *et al*. 2019a), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Zheng *et al*. 2020), (Chen *et al*. 2020), (Zhang *et al*. 2020c)
	Candidate ranking	(Greco *et al*. 2016), (Wang *et al*. 2018a), (Lin *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018a), (Miao *et al*. 2019a), (Chen *et al*. 2019b), (Huang *et al*. 2019b), (Wang *et al*. 2019b), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Tang *et al*. 2019b), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Zhang *et al*. 2020b), (Zhang *et al*. 2020a), (Jin *et al*. 2020), (Niu *et al*. 2020), (Zhang *et al*. 2020c)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A4. Reviewed papers categorized based on their input/output

Input	Question	Factoid	(Chen *et al*. 2017), (Yang *et al*. 2017b), (Clark and Gardner 2018), (Weissenborn *et al*. 2017), (Choi *et al*. 2017b), (Hu *et al*. 2018a), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Tan *et al*. 2018a), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Wang *et al*. 2018b), (Hewlett *et al*. 2017), (Golub *et al*. 2017), (Swayamdipta *et al*. 2018), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang *et al*. 2018c), (Mihaylov *et al*. 2017), (Greco *et al*. 2016), (Cui *et al*. 2017b), (Dhingra *et al*. 2017), (Wang *et al*. 2018a), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Lin *et al*. 2018), (Liu *et al*. 2018b), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Bauer *et al*. 2018), (Hoang *et al*. 2018), (Aghaebrahimian 2018), (Seo *et al*. 2018), (Back *et al*. 2018), (Li *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Liu *et al*. 2018a), (Dhingra *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Trischler *et al*. 2017), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Liu and Perez 2017)„ (Munkhdalai and Yu 2017), (Ma *et al*. 2018), (Chaturvedi *et al*. 2018), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Wang *et al*. 2018f), (Mihaylov and Frank 2018), (Yu *et al*. 2018), (Song *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Sachan and Xing 2018), (Htut *et al*. 2018), (Miao *et al*. 2019a), (Yan *et al*. 2019), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Angelidis *et al*. 2019), (Ding *et al*. 2019), (Chen *et al*. 2019b), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang *et al*. 2019b), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Pang *et al*. 2019),(Su *et al*. 2019), (Andor *et al*. 2019), (Tang *et al*. 2019a), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Wu *et al*. 2019), (Bi *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Nishida *et al*. 2019b), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Hu *et al*. 2019c), (Hu *et al*. 2019b), (Huang *et al*. 2019a), (Nie *et al*. 2019), (Tay *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Wang *et al*. 2019a), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Das *et al*. 2019), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Zhang *et al*. 2020a), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Jin *et al*. 2020), (Chen *et al*. 2020), (Niu *et al*. 2020), (Pappas *et al*. 2020), (Zheng *et al*. 2020), (Song *et al*. 2020), (Ren *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)
		Non-factoid	(Min *et al*. 2017), (Choi *et al*. 2017b), (Tan *et al*. 2018a)„ (Hewlett *et al*. 2017), (Wang *et al*. 2018c), (Wang *et al*. 2018a),(Lin *et al*. 2018), (Liu *et al*. 2018b), (Bauer *et al*. 2018), (Aghaebrahimian 2018), (Li *et al*. 2018), (Zhang *et al*. 2018), (Wang *et al*. 2018e), (Liu *et al*. 2018a),(Dhingra *et al*. 2018), (Wang *et al*. 2016), (Tay *et al*. 2018), (Miao *et al*. 2019a), (Yan *et al*. 2019), (Jin *et al*. 2019), (Chen *et al*. 2019b), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Wang *et al*. 2019b), (Su *et al*. 2019), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Tang *et al*. 2019b), (Ren *et al*. 2019), (Nishida *et al*. 2019b), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Osama *et al*. 2019), (Tay *et al*. 2019), (Frermann 2019), (Lee *et al*. 2019b), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Liu *et al*. 2020a), (Jin *et al*. 2020), (Zheng *et al*. 2020), (Zhang *et al*. 2020a), (Zhang *et al*. 2020c)
		Yes/No	(Liu *et al*. 2018b), (Li *et al*. 2018), (Wang *et al*. 2018e), (Zhang *et al*. 2018), (Liu *et al*. 2018a), (Hu *et al*. 2019a), (Nakatsuji and Okui 2020), (Niu *et al*. 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Zheng *et al*. 2020)
	Context	Single paragraph	(Chen *et al*. 2017),(Yang *et al*. 2017b), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Choi *et al*. 2017b), (Hu *et al*. 2018a), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Hewlett *et al*. 2017), (Golub *et al*. 2017), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang *et al*. 2018c), (Mihaylov *et al*. 2017), (Cui *et al*. 2017b), (Dhingra *et al*. 2017),(Wang *et al*. 2018a), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Lin *et al*. 2018), (Kundu and Ng 2018a), (Hu *et al*. 2018b),(Bauer *et al*. 2018), (Hoang *et al*. 2018), (Aghaebrahimian 2018), (Back *et al*. 2018), (Seo *et al*. 2018), (Wang *et al*. 2018d), (Liu *et al*. 2018a), (Dhingra *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016),

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A4. Continued

			(Trischler *et al*. 2017), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Liu and Perez 2017),, (Munkhdalai and Yu 2017), (Ma *et al*. 2018), (Chaturvedi *et al*. 2018), (Indurthi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Mihaylov and Frank 2018), (Yu *et al*. 2018), (Song *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Sachan and Xing 2018), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Chen *et al*. 2019b), (Takahashi *et al*. 2019), (Huang *et al*. 2019b), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang and Jiang 2019), (Su *et al*. 2019), (Andor *et al*. 2019), (Sharma and Roychowdhury 2019), (Wu *et al*. 2019), (Xia *et al*. 2019), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Hu *et al*. 2019c), (Li *et al*. 2019c), (Park *et al*. 2019), (Lee *et al*. 2019b), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Zhang *et al*. 2020a), (Jin *et al*. 2020), (Chen *et al*. 2020), (Niu *et al*. 2020), (Pappas *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)
		Multi-paragraph	(Clark and Gardner 2018), (Tan *et al*. 2018a), (Wang *et al*. 2018b), (Swayamdipta *et al*. 2018), (Wang *et al*. 2018c),(Greco *et al*. 2016), (Liu *et al*. 2018b), (Aghaebrahimian 2018), (Li *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018e), (Nguyen *et al*. 2016), (Tay *et al*. 2018), (Ke *et al*. 2018), (Wang *et al*. 2018f), (Yu *et al*. 2018), (Sachan and Xing 2018), (Htut *et al*. 2018), (Miao *et al*. 2019a), (Yan *et al*. 2019), (Wang *et al*. 2019e), (Angelidis *et al*. 2019), (Ding *et al*. 2019), (Wang *et al*. 2019c), (Wang *et al*. 2019b), (Jiang *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Sun *et al*. 2019), (Bi *et al*. 2019), (Yu *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Nishida *et al*. 2019b), (Miao *et al*. 2019b), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Tay *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Zhuang and Wang 2019), (Wang *et al*. 2019a), (Nishida *et al*. 2019a), (Das *et al*. 2019), (Nakatsuji and Okui 2020), (Yang *et al*. 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Ren *et al*. 2020), (Zheng *et al*. 2020), (Song *et al*. 2020)
Output	–	Extractive	(Yang *et al*. 2017b), (Yang *et al*. 2017b), (Clark and Gardner 2018), (Min *et al*. 2017), (Weissenborn *et al*. 2017), (Hu *et al*. 2018a),(Shen *et al*. 2017), (Wang *et al*. 2017b), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Wang *et al*. 2018b), (Golub *et al*. 2017), (Swayamdipta *et al*. 2018), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang *et al*. 2018c), (Mihaylov *et al*. 2017), (Cui *et al*. 2017b), (Dhingra *et al*. 2017), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Liu *et al*. 2018b), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Hoang *et al*. 2018), (Aghaebrahimian 2018), (Back *et al*. 2018), (Seo *et al*. 2018), (Li *et al*. 2018), (Liu *et al*. 2018d), (Wang *et al*. 2018d), (Wang *et al*. 2018e), (Dhingra *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Trischler *et al*. 2017), (Xiong *et al*. 2017), (Cui *et al*. 2016), (Yin *et al*. 2016), (Wang *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Liu and Perez 2017), (Munkhdalai and Yu 2017), (Ma *et al*. 2018), (Chaturvedi *et al*. 2018), (Ghaeini *et al*. 2018), (Sheng *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Wang *et al*. 2018f), (Mihaylov and Frank 2018), (Yu *et al*. 2018), (Song *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Sachan and Xing 2018), (Htut *et al*. 2018),(Yan *et al*. 2019), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Angelidis *et al*. 2019), (Ding *et al*. 2019), (Takahashi *et al*. 2019), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Su *et al*. 2019), (Andor *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Wu *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Tu *et al*. 2019), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Hu *et al*. 2019c), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Das *et al*. 2019), (Guo *et al*. 2020), (Wang *et al*. 2020a), (Zhang *et al*. 2020b), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Ren *et al*. 2020), (Niu *et al*. 2020), (Pappas *et al*. 2020), (Zheng *et al*. 2020), (Song *et al*. 2020), (Chen *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A4. Continued

Abstractive	(Choi *et al*. 2017b), (Tan *et al*. 2018a), (Hewlett *et al*. 2017), (Greco *et al*. 2016),(Wang *et al*. 2018a),(Lin *et al*. 2018), (Bauer *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018a), (Indurthi *et al*. 2018), (Ke *et al*. 2018), (Miao *et al*. 2019a), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Chen *et al*. 2019b), (Huang *et al*. 2019b), (Dua *et al*. 2019), (Wang *et al*. 2019b), (Andor *et al*. 2019), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Bi *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Tang *et al*. 2019b), (Nishida *et al*. 2019b), (Ran *et al*. 2019), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Liu *et al*. 2019b), (Tay *et al*. 2019), (Wang *et al*. 2019a), (Niu *et al*. 2020), (Nakatsuji and Okui 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Zhang *et al*. 2020a), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Jin *et al*. 2020), (Zheng *et al*. 2020), (Chen *et al*. 2020), (Zhang *et al*. 2020c)
Multiple Choice	(Wang *et al*. 2018c), (Greco *et al*. 2016), (Wang *et al*. 2018a), (Lin *et al*. 2018), (Zhang *et al*. 2018), (Liu *et al*. 2018a), (Yang *et al*. 2017a), (Yin *et al*. 2016), (Wang *et al*. 2016), (Chaturvedi *et al*. 2018), (Sheng *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Miao *et al*. 2019a), (Chen *et al*. 2019b), (Huang *et al*. 2019b), (Wang *et al*. 2019b), (Sharma and Roychowdhury 2019), (Sun *et al*. 2019), (Xia *et al*. 2019), (Yu *et al*. 2019), (Tu *et al*. 2019), (Tang *et al*. 2019b), (Miao *et al*. 2019b), (Li *et al*. 2019b), (Guo *et al*. 2020), (Niu *et al*. 2020), (Zhou *et al*. 2020b), (Zhang *et al*. 2020b), (Zhang *et al*. 2020a), (Jin *et al*. 2020), (Song *et al*. 2020), (Zhang *et al*. 2020c)
Cloze	(Yadav *et al*. 2017), (Cui *et al*. 2017b), (Dhingra *et al*. 2017), (Hoang *et al*. 2018), (Dhingra *et al*. 2018), (Yang *et al*. 2017a), (Seo *et al*. 2017), (Chen *et al*. 2016), (Kobayashi *et al*. 2016), (Cui *et al*. 2016), (Kadlec *et al*. 2016a), (Kadlec *et al*. 2016b), (Liu and Perez 2017), (Munkhdalai and Yu 2017), (Ma *et al*. 2018), (Ghaeini *et al*. 2018), (Sugawara *et al*. 2018), (Mihaylov and Frank 2018), (Song *et al*. 2018), (Zhang *et al*. 2019b), (Wang *et al*. 2019a), (Wang *et al*. 2020a), (Niu *et al*. 2020), (Pappas *et al*. 2020)
Detail	(Chen *et al*. 2017), (Yang *et al*. 2017b), (Weissenborn *et al*. 2017), (Hu *et al*. 2018a), (Shen *et al*. 2017), (Wang *et al*. 2017b), (Liu *et al*. 2018c), (Huang *et al*. 2018), (Zhang *et al*. 2017), (Liu *et al*. 2017), (Gong and Bowman 2018), (Wang *et al*. 2017a), (Golub *et al*. 2017), (Xiong *et al*. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Mihaylov *et al*. 2017), (Kundu and Ng 2018b), (Prakash *et al*. 2018), (Kundu and Ng 2018a), (Hu *et al*. 2018b), (Seo *et al*. 2018), (Back *et al*. 2018), (Wang *et al*. 2018d), (Clark and Gardner 2018), (Wang *et al*. 2018b), (Swayamdipta *et al*. 2018), (Wang *et al*. 2018c), (Liu *et al*. 2018d), (Min *et al*. 2017), (Choi *et al*. 2017b), (Hewlett *et al*. 2017), (Bauer *et al*. 2018), (Tan *et al*. 2018a), (Liu *et al*. 2018b),(Li *et al*. 2018), (Wang *et al*. 2018e), (Aghaebrahimian 2018), (Dhingra *et al*. 2018),(Yang *et al*. 2017a), (Seo *et al*. 2017), (Wang and Jiang 2017), (Trischler *et al*. 2017), (Xiong *et al*. 2017), (Wang *et al*. 2016), (Liu and Perez 2017),, (Indurthi *et al*. 2018), (Min *et al*. 2018), (Sugawara *et al*. 2018), (Tay *et al*. 2018), (Gupta *et al*. 2018), (Ke *et al*. 2018), (Du and Cardie 2018), (Tan *et al*. 2018b), (Wang *et al*. 2018f), (Yu *et al*. 2018), (Nishida *et al*. 2018), (Wang and Bansal 2018), (Sachan and Xing 2018), (Htut *et al*. 2018), (Yan *et al*. 2019), (Hu *et al*. 2019a), (Wang *et al*. 2019e), (Jin *et al*. 2019), (Angelidis *et al*. 2019), (Ding *et al*. 2019), (Takahashi *et al*. 2019), (Cui *et al*. 2019a), (Mihaylov and Frank 2019), (Li *et al*. 2019a), (Wang *et al*. 2019c), (Dua *et al*. 2019), (Zhang *et al*. 2019b), (Xu *et al*. 2019a), (Yang *et al*. 2019a), (Wang and Jiang 2019), (Jiang *et al*. 2019), (Su *et al*. 2019), (Andor *et al*. 2019), (Pang *et al*. 2019), (Tang *et al*. 2019a), (Wu *et al*. 2019), (Bi *et al*. 2019), (Dehghani *et al*. 2019), (Min *et al*. 2019), (Yuan *et al*. 2019), (Ren *et al*. 2019), (Nishida *et al*. 2019b), (Xu *et al*. 2019b), (Ran *et al*. 2019), (Li *et al*. 2019b), (Osama *et al*. 2019), (Liu *et al*. 2019b), (Huang *et al*. 2019a), (Hu *et al*. 2019b), (Nie *et al*. 2019), (Tay *et al*. 2019), (Xiong *et al*. 2019), (Frermann 2019), (Hu *et al*. 2019c), (Zhuang and Wang 2019), (Li *et al*. 2019c), (Park *et al*. 2019), (Nishida *et al*. 2019a), (Lee *et al*. 2019b), (Das *et al*. 2019), (Nakatsuji and Okui 2020), (Lee and Kim 2020), (Yang *et al*. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Liu *et al*. 2020a), (Tu *et al*. 2020), (Ren *et al*. 2020), (Chen *et al*. 2020), (Niu *et al*. 2020), (Zheng *et al*. 2020), (Back *et al*. 2020), (Zhang *et al*. 2020c)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A5. Reviewed papers categorized based on their evaluation metric

Extractive Metric	EM	(Chen et al. 2017), (Yang et al. 2017b), (Clark and Gardner 2018), (Joshi et al. 2017), (Weissenborn et al. 2017), (Hu et al. 2018a), (Shen et al. 2017), (Wang et al. 2017b), (Liu et al. 2018c), (Huang et al. 2018), (Liu et al. 2017), (Gong and Bowman 2018), (Wang et al. 2017a), (Wang et al. 2018b), (Golub et al. 2017), (Swayamdipta et al. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang et al. 2018c), (Mihaylov et al. 2017), (Kundu and Ng 2018b), (Kundu and Ng 2018a), (Hu et al. 2018b), (Back et al. 2018), (Seo et al. 2018), (Wang et al. 2018d), (Dhingra et al. 2018), (Prakash et al. 2018), (Aghaebrahimian 2018), (Yang et al. 2017a), (Seo et al. 2017), (Wang and Jiang 2017), (Trischler et al. 2017), (Xiong et al. 2017), (Min et al. 2018), (Gupta et al. 2018), (Ke et al. 2018), (Du and Cardie 2018), (Wang et al. 2018f), (Yu et al. 2018), (Nishida et al. 2018), (Htut et al. 2018), (Yan et al. 2019), (Hu et al. 2019a), (Wang et al. 2019e), (Ding et al. 2019), (Takahashi et al. 2019), (Cui et al. 2019a), (Li et al. 2019a), (Wang et al. 2019c), (Dua et al. 2019), (Zhang et al. 2019b), (Xu et al. 2019a), (Yang et al. 2019a), (Wang et al. 2019b), (Wang and Jiang 2019), (Su et al. 2019), (Andor et al. 2019), (Pang et al. 2019), (Tang et al. 2019a), (Wu et al. 2019), (Yu et al. 2019), (Dehghani et al. 2019), (Yuan et al. 2019), (Das et al. 2019), (Xu et al. 2019b), (Ran et al. 2019), (Li et al. 2019b), (Osama et al. 2019), (Hu et al. 2019c), (Huang et al. 2019a), (Nie et al. 2019), (Xiong et al. 2019), (Zhuang and Wang 2019), (Li et al. 2019c), (Park et al. 2019), (Nishida et al. 2019a), (Lee et al. 2019b), (Wang et al. 2020a), (Niu et al. 2020), (Lee and Kim 2020), (Yang et al. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Tu et al. 2020), (Ren et al. 2020), (Chen et al. 2020), (Zhang et al. 2020c)
	F1	(Chen et al. 2017), (Yang et al. 2017b), (Clark and Gardner 2018), (Joshi et al. 2017), (Weissenborn et al. 2017), (Hu et al. 2018a), (Shen et al. 2017), (Wang et al. 2017b), (Liu et al. 2018c), (Huang et al. 2018), (Liu et al. 2017), (Gong and Bowman 2018), (Wang et al. 2017a), (Wang et al. 2018b), (Hewlett et al. 2017), (Golub et al. 2017), (Swayamdipta et al. 2018), (Xie and Xing 2017), (Salant and Berant 2018), (Wang et al. 2018c), (Mihaylov et al. 2017), (Kundu and Ng 2018b), (Kundu and Ng 2018a), (Hu et al. 2018b), (Back et al. 2018), (Seo et al. 2018), (Wang et al. 2018d), (Dhingra et al. 2018), (Prakash et al. 2018), (Aghaebrahimian 2018), (Yang et al. 2017a), (Seo et al. 2017), (Wang and Jiang 2017), (Trischler et al. 2017), (Xiong et al. 2017), (Min et al. 2018), (Sugawara et al. 2018), (Tay et al. 2018), (Gupta et al. 2018), (Ke et al. 2018), (Du and Cardie 2018), (Tan et al. 2018b), (Wang et al. 2018f), (Yu et al. 2018), (Nishida et al. 2018), (Wang and Bansal 2018), (Htut et al. 2018), (Yan et al. 2019), (Hu et al. 2019a), (Wang et al. 2019e), (Ding et al. 2019), (Takahashi et al. 2019), (Cui et al. 2019a), (Li et al. 2019a), (Wang et al. 2019c), (Dua et al. 2019), (Zhang et al. 2019b), (Xu et al. 2019a), (Yang et al. 2019a), (Wang et al. 2019b), (Wang and Jiang 2019), (Su et al. 2019), (Andor et al. 2019), (Pang et al. 2019), (Tang et al. 2019a), (Wu et al. 2019), (Yu et al. 2019), (Dehghani et al. 2019), (Min et al. 2019), (Yuan et al. 2019), (Das et al. 2019), (Xu et al. 2019b), (Ran et al. 2019), (Li et al. 2019b), (Osama et al. 2019), (Hu et al. 2019c), (Huang et al. 2019a), (Nie et al. 2019), (Xiong et al. 2019), (Zhuang and Wang 2019), (Park et al. 2019), (Nishida et al. 2019a), (Lee et al. 2019b), (Wang et al. 2020a), (Niu et al. 2020), (Lee and Kim 2020), (Yang et al. 2020), (Wu and Xu 2020), (Zhang and Wang 2020), (Liu et al. 2020a), (Tu et al. 2020), (Ren et al. 2020), (Chen et al. 2020), (Zheng et al. 2020), (Zhang et al. 2020c)
	MAP	(Min et al. 2017), (Duan et al. 2017), (Wang et al. 2016), (Min et al. 2018), (Sachan and Xing 2018)
	MRR	(Min et al. 2017), (Duan et al. 2017), (Wang et al. 2016), (Sachan and Xing 2018), (Angelidis et al. 2019)
	P@1	(Min et al. 2017), (Du and Cardie 2018), (Tan et al. 2018b), (Sachan and Xing 2018), (Angelidis et al. 2019), (Ding et al. 2019), (Wang et al. 2020a), (Niu et al. 2020), (Liu et al. 2020a), (Zheng et al. 2020), (Tu et al. 2020)
	R@1	(Liu et al. 2018d), (Du and Cardie 2018), (Tan et al. 2018b), (Ding et al. 2019), (Wang et al. 2020a), (Niu et al. 2020), (Liu et al. 2020a), (Zheng et al. 2020)
	ACC	(Choi et al. 2017b), (Lai et al. 2017), (Yadav et al. 2017), (Cui et al. 2017b), (Dhingra et al. 2017), (Wang et al. 2018a), (Lin et al. 2018), (Hoang et al. 2018), (Zhang et al. 2018), (Liu et al. 2018a), (Prakash et al. 2018), (Duan et al. 2017), (Yang et al. 2017a), (Seo et al. 2017), (Chen et al. 2016), (Kobayashi et al. 2016), (Cui et al. 2016), (Yin et al. 2016), (Kadlec et al. 2016a), (Kadlec et al. 2016b), (Liu and Perez 2017), (Choi et al. 2017a), (Munkhdalai and Yu 2017), (Ma et al. 2018), (Chaturvedi et al. 2018), (Ghaeini et al. 2018), (Sheng et al. 2018), (Min et al. 2018), (Sugawara et al. 2018), (Tay et al. 2018), (Tan et al. 2018b), (Mihaylov and Frank 2018), (Song et al. 2018), (Miao et al. 2019a), (Chen et al. 2019b), (Huang et al. 2019b), (Zhang et al. 2019b), (Wang et al. 2019b), (Jiang et al. 2019), (Andor et al. 2019), (Sharma and Roychowdhury 2019), (Sun et al. 2019), (Xia et al. 2019), (Yu et al. 2019), (Tu et al. 2019), (Tang et al. 2019b), (Miao et al. 2019b), (Li et al. 2019b), (Wang et al. 2019a), (Guo et al. 2020), (Niu et al. 2020), (Pappas et al. 2020), (Zhang et al. 2020b), (Zhang et al. 2020a), (Jin et al. 2020), (Pappas et al. 2020), (Song et al. 2020), (Tu et al. 2020), (Zhang et al. 2020c)

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A5. Continued

	Hit@k/Top@k	(Greco et al. 2016)
Generative Metric	ROUGE_L	(Weissenborn et al. 2017), (Aghaebrahimian 2018), (Liu et al. 2018b), (Tan et al. 2018a), (Li et al. 2018), (Wang et al. 2018e), (Bauer et al. 2018), (Nguyen et al. 2016), (Indurthi et al. 2018), (Sugawara et al. 2018), (Tay et al. 2018), (Ke et al. 2018), (Yan et al. 2019), (Wang et al. 2019e), (Jin et al. 2019), (Mihaylov and Frank 2019), (Bi et al. 2019), (Ren et al. 2019), (Nishida et al. 2019b), (Liu et al. 2019b), (Tay et al. 2019), (Frermann 2019), (Li et al. 2019c), (Nakatsuji and Okui 2020), (Zhou et al. 2020b)
	BLEU	(Weissenborn et al. 2017), (Aghaebrahimian 2018), (Tan et al. 2018a), (Li et al. 2018), (Wang et al. 2018e), (Bauer et al. 2018), (Trischler et al. 2017), (Indurthi et al. 2018), (Tay et al. 2018), (Ke et al. 2018), (Yan et al. 2019), (Wang et al. 2019e), (Jin et al. 2019), (Mihaylov and Frank 2019), (Bi et al. 2019), (Ren et al. 2019), (Nishida et al. 2019b), (Liu et al. 2019b), (Tay et al. 2019), (Frermann 2019), (Wang et al. 2020a), (Nakatsuji and Okui 2020), (Zhou et al. 2020b)
	METEOR	(Bauer et al. 2018), (Indurthi et al. 2018), (Tay et al. 2018), (Tay et al. 2019), (Frermann 2019)
	CIDEr	(Bauer et al. 2018), (Trischler et al. 2017)

Table A6. Reviewed papers categorized based on their novelties

Model Structure	Input/output	(Choi *et al.* 2017b), (Liu *et al.* 2018c), (Tan *et al.* 2018a), (Liu *et al.* 2017), (Hewlett *et al.* 2017), (Swayamdipta *et al.* 2018), (Salant and Berant 2018), (Wang *et al.* 2018c), (Liu *et al.* 2018d), (Longpre *et al.* 2019), (Wang *et al.* 2019b), (Zhu *et al.* 2019), (Guo *et al.* 2020), (Lee and Kim 2020), (Zheng *et al.* 2020)
	Internal	(Chen *et al.* 2017), (Clark and Gardner 2018), (Weissenborn *et al.* 2017), (Hu *et al.* 2018a), (Shen *et al.* 2017), (Wang *et al.* 2017b), (Lai *et al.* 2017), (Tan *et al.* 2018a), (Huang *et al.* 2018), (Gong and Bowman 2018), (Xiong *et al.* 2018), (Xie and Xing 2017), (Yadav *et al.* 2017), (Cui *et al.* 2017b), (Dhingra *et al.* 2017), (Wang *et al.* 2018a), (Kundu and Ng 2018b), (Prakash *et al.* 2018), (Lin *et al.* 2018), (Liu *et al.* 2018b), (Kundu and Ng 2018a), (Bauer *et al.* 2018), (Aghaebrahimian 2018), (Back *et al.* 2018), (Seo *et al.* 2018), (Wang *et al.* 2018d), (Wang *et al.* 2018e), (Liu *et al.* 2018a), (Zhang *et al.* 2018), (Giuseppe 2017), (Yang *et al.* 2017a), (Seo *et al.* 2017), (Wang and Jiang 2017), (Kobayashi *et al.* 2016), (Trischler *et al.* 2017), (Xiong *et al.* 2017), (Cui *et al.* 2016), (Yin *et al.* 2016), (Wang *et al.* 2016), (Kadlec *et al.* 2016b), (Liu and Perez 2017), (Munkhdalai and Yu 2017), (Ma *et al.* 2018), (Chaturvedi *et al.* 2018), (Indurthi *et al.* 2018), (Ghaeini *et al.* 2018), (Sheng *et al.* 2018), (Min *et al.* 2018), (Tay *et al.* 2018), (Gupta *et al.* 2018), (Ke *et al.* 2018), (Tan *et al.* 2018b), (Wang *et al.* 2018f), (Mihaylov and Frank 2018), (Yu *et al.* 2018), (Song *et al.* 2018), (Htut *et al.* 2018), (Miao *et al.* 2019a), (Yan *et al.* 2019), (Hu *et al.* 2019a), (Wang *et al.* 2019e), (Jin *et al.* 2019), (Angelidis *et al.* 2019), (Ding *et al.* 2019), (Chen *et al.* 2019b), (Takahashi *et al.* 2019), (Huang *et al.* 2019b), (Cui *et al.* 2019a), (Mihaylov and Frank 2019), (Wang *et al.* 2019c), (Dua *et al.* 2019), (Zhang *et al.* 2019b), (Xu *et al.* 2019a), (Yang *et al.* 2019a), (Jiang *et al.* 2019), (Su *et al.* 2019), (Andor *et al.* 2019), (Pang *et al.* 2019), (Tang *et al.* 2019a), (Sharma and Roychowdhury 2019), (Sun *et al.* 2019), (Wu *et al.* 2019), (Bi *et al.* 2019), (Xia *et al.* 2019), (Yu *et al.* 2019), (Dehghani *et al.* 2019), (Min *et al.* 2019), (Tu *et al.* 2019), (Tang *et al.* 2019b), (Yuan *et al.* 2019), (Ren *et al.* 2019), (Das *et al.* 2019), (Nishida *et al.* 2019b), (Xu *et al.* 2019b), (Ran *et al.* 2019), (Miao *et al.* 2019b), (Li *et al.* 2019b), (Osama *et al.* 2019), (Liu *et al.* 2019b), (Hu *et al.* 2019c), (Huang *et al.* 2019a), (Hu *et al.* 2019b), (Nie *et al.* 2019), (Tay *et al.* 2019), (Xiong *et al.* 2019), (Frermann 2019), (Zhuang and Wang 2019), (Wang *et al.* 2019a), (Park *et al.* 2019), (Nishida *et al.* 2019a), (Lee *et al.* 2019b), (Wang *et al.* 2020a), (Yang *et al.* 2020), (Niu *et al.* 2020), (Nakatsuji and Okui 2020),

https://doi.org/10.1017/S1351324921000395 Published online by Cambridge University Press

Table A6. Continued

	(Zhou *et al.* 2020b), (Zhang *et al.* 2020b), (Wu and Xu 2020), (Zhang and Wang 2020), (Zhang *et al.* 2020a), (Liu *et al.* 2020a), (Tu *et al.* 2020), (Welbl *et al.* 2020), (Jin *et al.* 2020), (Ren *et al.* 2020), (Chen *et al.* 2020), (Pappas *et al.* 2020), (Zheng *et al.* 2020), (Song *et al.* 2020), (Asai and Hajishirzi 2020), (Yan *et al.* 2020), (Back *et al.* 2020), (Zhang *et al.* 2020c), (Cao *et al.* 2020) (Zhou, Huang and Zhu 2020a),
Dataset	(Joshi *et al.* 2017), (Lai *et al.* 2017), (He *et al.* 2018), (Welbl *et al.* 2017), (Xie *et al.* 2018), (Pampari *et al.* 2018), (Khashabi *et al.* 2018), (Ostermann *et al.* 2018), (Nguyen *et al.* 2016), (Trischler *et al.* 2017), (Bajgar *et al.* 2017), (Cui *et al.* 2016), (Rajpurkar *et al.* 2016), (Onishi *et al.* 2016), (Ma *et al.* 2018), (Šuster and Daelemans 2018), (Shao *et al.* 2018), (Saha *et al.* 2018), (Elgohary *et al.* 2018), (Tan *et al.* 2018b), (Rajpurkar *et al.* 2018), (Hill *et al.* 2016), (KoČiskÝ *et al.* 2018), (Kwiatkowski *et al.* 2019), (Welbl *et al.* 2018), (Yang *et al.* 2018), (Liang *et al.* 2019), (Cui *et al.* 2019b), (Gupta *et al.* 2019), (Hardalov *et al.* 2019), (Jing *et al.* 2019), (Huang *et al.* 2019b), (Dua *et al.* 2019), (Sayama *et al.* 2019), (Ostermann *et al.* 2019), (Mozannar *et al.* 2019), (Dasigi *et al.* 2019), (Yao *et al.* 2019), (Li *et al.* 2019c), (Liu *et al.* 2019a), (Anuranjana *et al.* 2019), (Lin *et al.* 2019), (Pappas *et al.* 2020), (Watarai and Tsuchiya 2020), (Sun *et al.* 2020), (Inoue *et al.* 2020), (Yu *et al.* 2020), (Wang *et al.* 2020b), (Lee *et al.* 2020), (Berzak *et al.* 2020), (Yuan *et al.* 2020b), (Horbach *et al.* 2020)
Knowledge transfer	(Chen *et al.* 2017), (Yang *et al.* 2017b), (Min *et al.* 2017), (Wang *et al.* 2017a), (Golub *et al.* 2017), (Duan *et al.* 2017), (Yadav *et al.* 2017), (Mihaylov *et al.* 2017), (Hu *et al.* 2018b), (Hoang *et al.* 2018), (Wang *et al.* 2018e), (Dhingra *et al.* 2018), (Liu *et al.* 2018d), (Yin *et al.* 2016), (Wang *et al.* 2016), (Kadlec *et al.* 2016b), (Du and Cardie 2018), (Nishida *et al.* 2018), (Sachan and Xing 2018), (Longpre *et al.* 2019), (Hardalov *et al.* 2019), (Cui *et al.* 2019a), (Mihaylov and Frank 2019), (Li *et al.* 2019a), (Yang *et al.* 2019a), (Wang and Jiang 2019), (Su *et al.* 2019), (Sun *et al.* 2019), (Wu *et al.* 2019), (Bi *et al.* 2019), (Xia *et al.* 2019), (Yu *et al.* 2019), (Qiu *et al.* 2019), (Xu *et al.* 2019b), (Nishida *et al.* 2019a), (Wang, Lu and Tang 2019d), (Lee *et al.* 2019b), (2019), (Soni and Roberts 2020), (Wu and Xu 2020), (Tu *et al.* 2020), (Nishida *et al.* 2020), (Cao *et al.* 2020), (Zhou *et al.* 2020a), (Asai and Hajishirzi 2020), (Jin *et al.* 2020), (Yan *et al.* 2020), (Ren *et al.* 2020), (Dua, Singh and Gardner 2020), (Yuan *et al.* 2020a), (Zhang *et al.* 2020c), (Gupta and Khade 2020)
Evaluation measure	(Jia and Liang 2017), (Sugawara *et al.* 2017), (Chen *et al.* 2016), (Sugawara *et al.* 2018), (Tan *et al.* 2018b), (Wang and Bansal 2018), (Pugaliya *et al.* 2019), (Wang *et al.* 2019c), (Chen *et al.* 2019a), (Bao *et al.* 2019), (Lee *et al.* 2019a), (Fisch *et al.* 2019), (Talmor and Berant 2019), (Gardner *et al.* 2019), (Li *et al.* 2019c), (Wu and Xu 2020), (Schlegel *et al.* 2020), (Liu *et al.* 2020b), (Sugawara *et al.* 2020), (Yue, Gutierrez and Sun 2020), (Charlet *et al.* 2020), (Soni and Roberts 2020), (Inoue *et al.* 2020), (Berzak *et al.* 2020), (Welbl *et al.* 2020), (Zhou *et al.* 2020a), (Ren *et al.* 2020), (Cao *et al.* 2020), (Dunietz *et al.* 2020), (Horbach *et al.* 2020)

Cite this article: Baradaran R, Ghiasi R and Amirkhani H (2022). A Survey on Machine Reading Comprehension Systems. *Natural Language Engineering* **28**, 683–732. https://doi.org/10.1017/S1351324921000395