Journal of Intelligent Information Systems (2022) 58:453–480
https://doi.org/10.1007/s10844-021-00655-8

A systematic review of question answering systems for non-factoid questions

Eduardo Gabriel Cortes[1] · Vinicius Woloszyn[2] · Dante Barone[1] · Sebastian Möller[2] · Renata Vieira[3]

Received: 29 March 2021 / Revised: 6 July 2021 / Accepted: 6 July 2021 /
Published online: 25 September 2021

Abstract
Question Answering (QA) is a field of study addressed to develop automatic methods for answering questions expressed in natural language. Recently, the emergence of the new generation of intelligent assistants, such as Siri, Alexa, and Google Assistant, has intensified the importance of an effective and efficient QA system able to handle questions with different complexities. Regarding the type of question to be answered, QA systems have been divided into two sub-areas: (i) factoid questions that require a single fact – e.g., a name of a person or a date, and (ii) non-factoid questions that need a more complex answer – e.g., descriptions, opinions, or explanations. While factoid QA systems have overcome human performance on some benchmarks, automatic systems for answering non-factoid questions remain a challenge and an open research problem. This work provides an overview of recent research addressing non-factoid questions. It focuses on which methods have been applied in each task, the data sets available, challenges and limitations, and possible research directions. From a total of 455 recent studies, we selected 75 papers based on our quality control system and exclusion criteria for an in-depth analysis. This systematic review helped to answer what are the tasks and methods involved in non-factoid, what are the data sets available, what the limitations are, and what is the recommendations for future research.

Keywords Systematic review · Non-factoid question · Question answering

1 Introduction

Question Answering (QA) system aims to automatically answer questions posed by users, normally, using natural language. Those systems generally consist of three main components: (i) Question Processing, responsible for extracting relevant information about what

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

✉ Eduardo Gabriel Cortes
egcortes@inf.ufrgs.br

Extended author information available on the last page of the article.

is being asked; (ii) Information Retrieval, which recovers data from a knowledge base, and (iii) Answering Processing, which selects the most pertinent pieces of information to provide the final answer. Despite the computational tasks involved in answering a question automatically, recent works have reported a very high performance on particular tasks. For example, since 2018, new methods addressing factoid questions have outperformed humans on the Stanford Question Answering Dataset (SQuAD)[1] achieving new state-of-the-art results (Joshi et al., 2020; Young et al., 2018; Shen et al., 2020; Wang et al., 2017). Nevertheless, non-factoid questions are considered the most challenging task, and they are still an open research problem.

Non-factoid questions are defined as open-ended questions that require complex answers, like descriptions, opinions, or explanations. For example, questions like "*How should I treat measles in a 12-year-old boy?*" are non-factoid. The answer for this class of questions is typically found in different textual segments, such as phrases, sentences, or paragraphs (Yang et al., 2016); however, the automatic process of creating coherent and concise answers based on multiple textual segments is challenging and still an open research problem (Bau et al., 2020). Conversely, factoid questions are the simplest form of a question that require a single fact as the answer – as a name of a person or a date, for example, "*What football team does the pope support?*".

Once the QA area is vast and presents several studies over the years, there are many QA reviews in the literature. The vast majority are broadly focused, encompassing the area of QA as a whole. Therefore, there is importance in QA studies going deep in a specifying QA subarea. Recently, the emergence of the new generation of search interfaces and intelligent assistant services, such as Siri, Alexa, and Google Assistant, has intensified the importance of an effective and efficient QA system able to handle questions with different complexities. Also, there are recent studies in literature observing an increase in publication regarding QA for non-factoid questions (Kodra & Kajo, 2017; Malviya & Soni, 2020).

This work proposes a systematic review of the state-of-the-art for non-factoid QA systems, focusing on different tasks and methods, as well as the available databases, evaluation strategies, outcomes and recommendations for future research. Therefore, the objective of the review is to answer the following research questions:

1. What are the methods and tasks involved in non-factoid Question Answering systems?
2. What are the data sets available for non-factoid Question Answering systems?
3. What are the limitations of non-factoid Question Answering?

We describe the main concepts and general architecture of non-factoid QA systems in Section 2. The methodology of the systematic review is present in Section 3. The results and discussion are described in Section 4. Finally, we present our conclusions and final remarks in Section 5.

2 Background

In this Chapter, we first describe the field of non-factoid Question Answering Systems by introducing key concepts and general architecture. Next, we present the previously works

[1]Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset (https://rajpurkar.github.io/SQuAD-explorer/)

addressing the non-factoid questions as well as highlighting the importance of a systematic review of the literature about this problem.

2.1 General architecture of non-factoid question answering systems

As observed in recent works (Kodra & Kajo, 2017; Calijorne Soares & Parreiras, 2020; Dimitrakis et al., 2019; Noraset et al., 2021), factoid and non-factoid QA systems present a similar architecture, which is based on three key components: 1) Question Processing, 2) Information Retrieval, and 3) Answer Processing. While Question Processing and Information Retrieval have similar behavior for both factoid and non-factoid QA systems, Answer Processing presents the most significant difference. Figure 1 depicts the general architecture of QA systems and is explained as detailed in the following paragraphs:

1. Question Processing is the first component responsible for understanding a user's questions. Questions posed by users using a reduced set of words increases the chances for misinterpretation by the system, therefore extra information about the question is generally incorporated. The extra information is used to help the system to reduce the search space for the corresponding answer. For example, QA systems that work with different types of questions, like factoid and non-factoid, usually need to understand what is the focus of the question the user is asking (*sports*, *health*, *weather*, etc.) and the expected type of answer (*person name*, *date*, *location*, etc.). Therefore, the detection of the focus of the question and the expected answer type is performed in *a) Question Classification* in order to apply the appropriate strategies during the next steps (Wu et al., 2015; Ben Abacha & Zweigenbaum, 2015; Bondarenko et al., 2020; Cortes et al., 2020). *b) Question Reformulation*, also known as Question Expansion, aims to enhancing the question by transforming it into a semantically equivalent one. The objective of these reformulations is to increase the likelihood of finding the correct answers in the text (Hermjakob et al., 2002). For example, the question *"Who invented the telephone?"* would be reformulated to "$<$ *who* $>$ *received a patent for the telephone?"* matching the following passage in the corpus *"Alexander Graham Bell received a patent for the telephone"*.

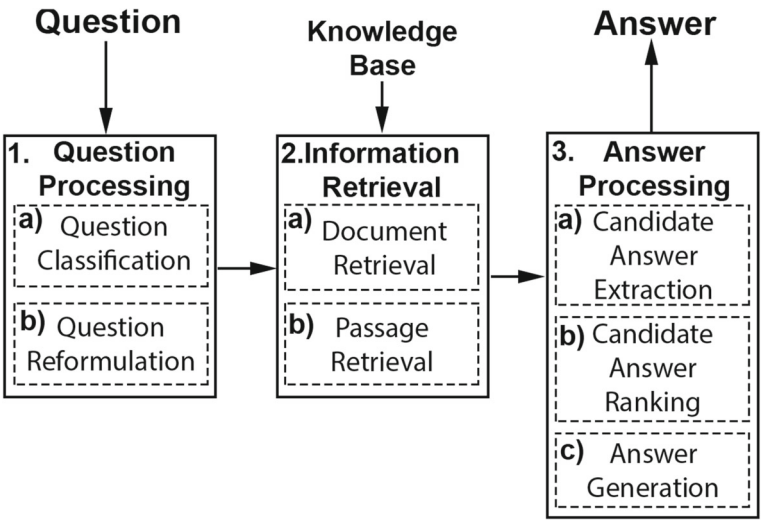

Fig. 1 General Architecture of Question-Answering Systems

2. Information Retrieval is responsible for searching and retrieving relevant information necessary for composing a final answer. The tasks in this component are determined by the type of knowledge available for the system. For instance, systems that use textual documents as a knowledge source (e.g., books or Wikipedia articles) generally at this component identify the relevant document(s) for answering the question. Normally, there are two important tasks involved: *a) Document Retrieval* and *b) Passage Extraction*. The *Document Retrieval* searches for a subset of documents and sort them by relevance. *Passage Extraction* extracts passages from the retrieved documents to reduce the amount of text to be analyzed, reducing the answers' search space into a few passages (Calijorne Soares & Parreiras, 2020). The passages have different processing pipelines depending on whether they refer to a factoid or non-factoid question. For example, in factoid question, only a small part of a sentence is normally used to generate the final answer – e.g., a Named Entity such as person names, organizations, locations, etc. On the other hand, non-factoid QA systems can consider individual or multiples passages for answering a single question.

3. Answer Processing is the last component responsible for generating a final answer to the user. Among all components, Answer Processing is the most challenging for non-factoid questions since it needs to provide an answer – usually long – using the extracted passages from the previous components. There are three tasks associated with this components (Dimitrakis et al., 2019; Yogish et al., 2018; Papadakis & Tzitzikas, 2015; Liu et al., 2016): *a) Candidate Answer Extraction*, *b) Candidate Answer Ranking*, and *c) Answers Generation*. *Candidate Answer Extraction* aims to identify answer candidates in the list of passages. We consider *Passage Extraction* and *Candidate Answer Extraction* being two distinct tasks since the first concerns techniques for finding passage/paragraph limits, while the second sorts and selects the top-k passages related to the question. *Candidate Answer Ranking* aims to select the best answers candidates ranking them based on how associated they are to the input question. Finally, *Answer Generation* is responsible for composing a final answer using the information extracted from previous tasks.

However, the studies do not always design their systems following precisely this architecture. The architecture here presented shows the main components and tasks of a general architecture following previously works (Kodra & Kajo, 2017; Calijorne Soares & Parreiras, 2020; Dimitrakis et al., 2019; Yogish et al., 2018; Liu et al., 2016; Chali et al., 2015). Regarding the terminology used to describe the components and tasks, we tried to simplify and cover as many problems as possible since QA systems' development involves several challenges.

2.2 Previous reviews

We have found many studies making a review of the current development of QA systems. Most of these address the area of QA broadly or focus only on factoid methods. There is still no study addressing only non-factoid systems, but some reviews have included it in their analyses (Kodra & Kajo, 2017; Malviya & Soni, 2020; Calijorne Soares & Parreiras, 2020; Dimitrakis et al., 2019; Mishra & Jain, 2016; Shah et al., 2019).

Some studies conducted by (Kodra & Kajo, 2017; Calijorne Soares & Parreiras, 2020; Dimitrakis et al., 2019) observed that most of the methods are concerned with characteristic as open domain, factoid, non-community QA, and documents as World Wide Web for the information source. Also, (Calijorne Soares & Parreiras, 2020) shows that medicine subject was the most used among the restrict-domain systems.

Regarding the techniques employed in QA systems, (Dimitrakis et al., 2019) observe that most of the current methods are based on neural networks and exploit structured data,

as knowledge graphs, somehow. The study conducted by (Shah et al., 2019) shows that it is possible to achieve significant results without relying on complex approaches. According to the studies analyzed by (Mishra & Jain, 2016), non-factoid questions can compromise the performance and increase QA systems' complexity. Also, the authors conclude that the accuracy of QA systems had a high dependency on the quality of the source data and the users' questions.

Other studies have concentrated on analyzing the paradigms, stages, and metrics of a QA system. (Kodra & Kajo, 2017) identifies that most methods concern tasks from the question processing stage, as question classification and question reformulation. (Sultana & Badugu, 2020), on the other hand, observed a considerable number of studies employing information retrieval approaches. (Calijorne Soares & Parreiras, 2020) shows that most analyzed studies have concentrated efforts in the natural language paradigm, and it had the best average evaluation performance over the years. Also, the review concludes that most of the studies employed *Precision* and *Recall* as evaluation metrics.

The analyzed reviews suggest that the research area in QA should explore challenges related to the word sense disambiguation (Hazrina et al., 2017), distributed exploit across different data bases (Chali et al., 2015), multilingualism, and complex non-factoid questions (Malviya & Soni, 2020). Although most analyzed studies have focused on factoid questions, some reviews noted an increase in publication regarding systems for non-factoid questions (Kodra & Kajo, 2017; Malviya & Soni, 2020).

Even though some studies presented the differences between factoid and non-factoid questions, there still no research particularly addressing only the problems, resources, and solutions for non-factoid questions. Hence, this work provides a survey of the recent works addressing non-factoid questions, the main problems, data sets, methods, and possible future research directions.

3 Systematic review

A Systematic Review differs from traditional narrative reviews by adopting a replicable, scientific, and transparent process to minimize bias through exhaustive literature searches (Tranfield et al., 2003; Denyer & Tranfield, 2009; Higgins et al., 2019). Despite the relative maturity of systematic reviews, there is no firm agreement about the number of stages for conducting a systematic review. For example, while the Cochrane Reviewers' Handbook (Higgins et al., 2019) and National Health Service Dissemination (2001) agreed in 9 Stages for a systematic review, recent studies have used a simplified approach (Khan et al., 2003; Dybå & Dingsøyr, 2008). We have adopted the same method proposed by Dybå (Dybå & Dingsøyr, 2008) by breaking down the study into (i) development of a protocol, (ii) definition of inclusion and exclusion criteria, (iii) selection of data sources to be scrutinized, as well as, definition of search string, (iv) definition of quality control for selection of the works, (v) definition of data extraction strategy, and (vi) synthesis of findings. In the remainder of this section, we describe the details of these stages and the methods employed.

3.1 Protocol

The protocol is a document that gives a general overview of how the review is performed. Typically, it specifies the research questions, search strategy, inclusion, exclusion and quality criteria, data extraction, synthesis method, etc. In previous studies, we found different guidelines for designing a protocol, which complements each other. Therefore, we have

relied on guidelines presented in (Tranfield et al., 2003; Denyer & Tranfield, 2009; Higgins et al., 2019) for designing our protocol.

3.2 Systematic search

A systematic search begins with the definition of search terms and electronic databases to be scrutinized. The search terms are used in databases of scientific articles (e.g., Web of Science or Google Scholar) for retrieving only related work, in our case on non-factoid QA systems. This is the first systematic review particularly addressed to non-factoid QA systems; Consequently, we had to create our own set of terms and electronic databases built based on related studies (Dimitrakis et al., 2019; Kolomiyets & Moens, 2011), and fine-tuned during discussions with the review team composed by the authors of this study. The final set of search terms is presented below:

1. non-factoid
2. definition question
3. confirmation question
4. causal question
5. comparative question
6. opinionated question

In order to select studies that respect both criteria, the keywords were combined through the Boolean "OR" and "AND" operators to formulate a query string, as following:

QUERY STRING = (1 OR 2 OR 3 OR 4 OR 5 OR 6) AND "question answering"

Only papers that contain the patterns described in the query string were considered in this review. The following electronic databases were employed in this study:

DATABASES = [*"ACM Digital Library","Web of Science", "IEEE Xplore", "Science Direct - Elsevier", "Springer Link"*]

3.3 Inclusion and exclusion criteria

We have developed different inclusion and exclusion criteria to reduce the number of unrelated and less significant papers for the manual review. For instance, to retrieve only updated and relevant works, we have only considered studies written in English since 2010 published in international conferences. Additionally, we have performed a semi-automatized analyzes using a tool to make sure only papers focusing on the QA topic were considered. We employed Rayyan QCRI (Ouzzani et al., 2016), which is a tool used to highlight a set of keywords in the papers and facilitate the process of inclusion and exclusion. When a paper present several keywords highlights, it is assumed that the paper covers the review's topics and should be included. Furthermore, once our research focused on non-factoid questions, we excluded studies that employ only factoid questions in their experiments or do not propose any method for non-factoid questions. We also excluded surveys and reviews from our study. In sum, the paper is included in our study only if it fulfills all the following criteria:

– Written in the English language.
– Published between 2010 and 2020.
– Focus on QA systems, as indicated by the search terms.
– Consider the challenges of non-factoid questions in the solutions.
– Employ non-factoid questions in the experiments.

3.4 Work eligibility and quality control

In order to ensure that only relevant studies are included in this systematic review, we have created a quality control system, which consists of questionnaire. To answer the questionnaire we used three annotators – the authors of this review – to read the papers and answer a questionnaire regarding a paper's eligibility and quality. The questionnaire contains questions with three possible options: "Yes", "Partial" and "No". Only studies which have "Yes" for most of the questions and none "No" were considered. The questions are:

- Does the study address empirical research, or is it a merely "lessons learned" report based on an expert's opinion?
- Were the objectives and conclusions clearly reported?
- Is not the study an example of editorials, prefaces, article summary, interview, new, or review?
- Does the study provide an understandable description of the proposed methods?
- Was the research methodology suitable for the aims of the research?
- Was there a description of the data sets employed, and whether they contain non-factoid questions?
- Does the data set present quality data, enough instances, and was it adequate for the experiments?
- Does the study employ adequate methods to analyze the data?
- Were the results calculated using metrics appropriate to the experiment?

3.5 Data extraction & annotation process

To answer the research questions posed in this study, we have employed the software Rayyan QCR for a semi-automatic annotation of papers. The software enables the extraction of metadata, such as author, institution, year, etc. Simultaneously, the annotator were responsible for extracting in-depth information which was not explicit on the paper, such as the name of the methods or the data sets employed. We used three annotators, being at least two annotators per work, and a third one doing disambiguation for the cases where the two annotations did not agree on a particular label. In sum, we have annotated the following set of features from the papers:

1. **Year of publication:** the year that the study was published.
2. **Language:** these are the languages of the data used in the analyzed study.
3. **Question type:** the types of question used in the study experiments. We consider a taxonomy of question types derived from (Dimitrakis et al., 2019) that consist of a) *Definition*: questions requiring a definition; b) *How*: questions requiring an instruction; c) *Why*: questions requiring a reason; d) *Opinion*: questions requiring an opinion; e) *Comparison*: questions requiring a comparison between entities; f) *Conformation*: questions checking a fact; g) *Factoid Included*: the data set include factoid questions.
4. **Domain of knowledge:** The knowledge is the proposed method has focused on. We first classify the study into the open-domain or the closed-domain. When classified in the closed-domain, we also identify which knowledge area it is focusing on.
5. **Knowledge source type:** this feature corresponds to the characteristics of the information source used by the QA system. We propose four categories derived from the reviews (Calijorne Soares & Parreiras, 2020; Dimitrakis et al., 2019): *Documents*:

the system uses a collection of raw text documents; *Web*: the system uses information retrieved from the web, such as pages resulting from a search engine; *Knowledge Graph*: uses a structured knowledge base; *Answer List*: usually used by community QA systems, in which for each question, the system consults a list of possible candidate answers.

6. **Metrics used for Evaluation:** The evaluation metrics used by the study to evaluate the proposed methods.
7. **Research problem:** the QA tasks the study has focused on. We classify each problem focused by the study on one of the tasks presented in Section 2.1. Thus, the following tasks are associated with the study analyzed when it follows the criteria:
   - *Question Classification*: assigned to studies with methods related to text classification that somehow classify the input question. It includes tasks like answer type classification and topic classification.
   - *Question Reformulation*: assigned to studies that enhance the question by transforming it into semantically equivalent, like improving the question text with data from the WordNet.
   - *Document Retrieval*: assigned to studies that aim to retrieve documents with relevant information, like text documents and web pages.
   - *Passage Extraction*: assigned to studies that aim to provide methods to extract passages from the retrieved documents.
   - *Candidate Answer Extraction*: assigned to studies that propose methods to identify answer candidates in a list of text passages.
   - *Candidate Answer Ranking*: assigned to studies that propose methods to rank or select the candidate answers most likely to be correct.
   - *Answer Generation*: assigned to studies that propose composing a final answer using the information extracted from previous tasks. It includes methods like natural language generation and summarization.
8. **Method employed:** the methods proposed by the study to solve each task.
9. **Dataset employed:** the data collections used in the experiments.
10. **Results:** results and conclusion of the study.

3.6 Synthesis of the findings

The information extracted from the papers in this review were summarized in a tabular format, where each row represents a study and each column represents an extracted feature. The tabular organization enables comparison across works, and reciprocal translation of findings into a higher-order of interpretation, as well as it is a well-employed method and highly recommended for qualitative data analysis (Seers, 2012; Corbin & Strauss, 2014). Table 1 presented the metadata and the features extracted from all works considered in this review.

3.7 Limitations of this review

The main limitation of this review is related to a possible bias in the selection of the studies. Nevertheless, we aimed at reducing this bias by querying the digital databases following the standards and keywords used by previously systematic reviews. Another possible limitation

Table 1 Analyzed Studies

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
P1	2020	English	-	Open-Domain	Documents	P@k, MAP, MRR, NDCG	Candidate Answer Ranking	ANTIQUE
P2	2020	English	Confirmation, Definition, Factoid	Open-Domain	Knowledge Graph	Accuracy	Candidate Answer Extraction	SimpleQuestions (v2)
P3	2020	Russian	Comparison, Opinion, Factoid	Open-Domain	-	F1	Question Classification	Created by Authors
P4	2020	English	Confirmation, Factoid	Health	Documents	F1, Accuracy, MRR	Question Classification, Document Retrieval, Passage Extraction, Candidate Answer Extraction	BioASQ
P5	2020	Japanese	How	Open-Domain	Documents	Accuracy	Answer Generation	Created by Authors
P6	2019	English	-	Open-Domain	Documents	P@1, MRR	Candidate Answer Ranking	L6 - Yahoo! Answers QA
P7	2019	English	Why, Definition, Confirmation, How, Factoid	Health	Documents	MRR, Recall@K	Document Retrieval	HealthQA
P8	2019	Chinese	-	Geographical	Documents, Knowledge Graph	F1, MRR	Question Reformulation	Created by Authors
P9	2019	English	-	Open-Domain	Documents, Answer List	F1, Accuracy	Candidate Answer Ranking	SemEval-2015, 2016 and 2017
P10	2019	English	How	Open-Domain	Answer List	P@k, MRR	Candidate Answer Ranking	L5 - Yahoo! Answers QA
P11	2019	English	Comparison	Open-Domain	Web, Answer List	P@k, MAP	Candidate Answer Ranking	Yahoo! Answers
P12	2019	English	Comparison	E-Commerce	Documents, Web,	P@k		

Table 1 (continued)

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
Candidate Answer Extraction	Created by Authors							
					Knowledge Graph			
P13	2018	English	-	Open-Domain	Documents	MAP, MRR, P@k, NDCG	Candidate Answer Extraction	WikiPassageQA
P14	2018	English	Factoid	Open-Domain	Documents	Accuracy, Rouge	Candidate Answer Extraction	WebAP, MSMARCO
P15	2018	English	Factoid	Open-Domain	Web, Knowledge Graph	judged by LiveQA	Question Reformulation, Candidate Answer Extraction, Candidate Answer Ranking	LiveQA TREC
P16	2018	English	-	Open-Domain	Web, Answer List	Rouge	Candidate Answer Extraction	Yahoo! Answers
P17	2018	English	Opinion	Open-Domain, Insurance	Documents, Answer List	P@k, MRR, NDCG	Candidate Answer Extraction, Candidate Answer Ranking	FiQA, InsuranceQA
P18	2018	English	Opinion	Open-Domain, Financial, Insurance	Documents, Answer List	MAP, MRR	Candidate Answer Ranking	InsuranceQA, FiQA
P19	2018	English	Factoid	Open-Domain	Web, Answer List	judged by LiveQA	Question Reformulation, Document Retrieval, Candidate Answer Extraction, Candidate Answer Ranking	LiveQA TREC
P20	2018	English	Factoid	Open-Domain	Web	MRR	Candidate Answer Ranking	Created by Authors
P21	2018	English	-	Open-Domain	Answer List	MRR,	Candidate Answer Ranking	WebAP, nfl6,

Table 1 (continued)

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
						NDCG		WikiPassageQA
P22	2018	English	-	Open-Domain	Answer List	MAP, MRR	Candidate Answer Ranking	LiveQA TREC, InsuranceQA
P23	2018	English	-	Open-Domain	Documents	P@k, MAP, MRR, NDCG	Candidate Answer Extraction, Candidate Answer Ranking	GOV2, ClueWeb09B
P24	2017	English	-	Open-Domain	Web	MAP, MRR, NDCG	Question Reformulation, Candidate Answer Extraction, Candidate Answer Ranking	L6 - Yahoo! Answers QA, LiveQA TREC
P25	2017	English	-	Open-Domain	Web	P@1, MRR, Manual	Candidate Answer Ranking	Yahoo! Answers, LiveQA TREC
P26	2017	English, Chinese	-	Open-Domain	-	Accuracy	Candidate Answer Ranking	Created by Authors
P27	2017	English, Chinese	- Agriculture	Insurance,	Documents	Accuracy	Candidate Answer Ranking AgricultureQA	InsuranceQA,
P28	2017	English	-	Open-Domain	Web	MAP, MRR, NDCG	Question Reformulation, Document Retrieval, Passage Extraction, Candidate Answer Extraction, Candidate Answer Ranking	L6 - Yahoo! Answers QA, TREC-QA
P29	2017	English	-	Open-Domain	Answer List	Rouge	Candidate Answer Extraction	Yahoo! Answers
P30	2017	English	-	Open-Domain	Documents	P@k, MRR, NDCG	Question Reformulation	WebAP
P31	2016	English	How	Open-Domain	Documents	P@1, MRR	Candidate Answer Ranking	L6 - Yahoo! Answers QA
P32	2016	English	-	Open-Domain	Web	MRR, NDCG,	Candidate Answer Extraction,	WebAP

Table 1 (continued)

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
						P@10	Candidate Answer Ranking	
P33	2016	English	Definition, Factoid	Open-Domain	Documents, Web	Accuracy	Candidate Answer Ranking	TREC-QA, AQUAINT, AQUAINT-2
P34	2016	Indonesian	Comparison	Open-Domain	Web	Accuracy	Question Classification	Created by Authors
P35	2016	Arabic	Why	Open-Domain	Documents	c@1	Candidate Answer Extraction	Created by Authors
P36	2016	Chinese	-	Tourism	Documents	P@k, MRR	Candidate Answer Ranking	Created by Authors
P37	2016	English	Definition	Open-Domain	Documents	P@k, NDCG, Rouge	Candidate Answer Extraction	WebAP
P38	2016	English	-	Open-Domain	Answer List	MAP, MRR	Candidate Answer Ranking	SemEval-2016
P39	2016	English	-	Open-Domain	Answer List	P@k, MRR	Candidate Answer Ranking	Yahoo! Answers
P40	2016	English	-	Open-Domain	Documents	MAP, MRR, Accuracy	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	MCTest
P41	2016	English	-	Open-Domain	Documents	F1, Exact match	Candidate Answer Extraction, Candidate Answer Ranking, Answer Generation	SQuAD
P42	2015	Chinese	-	Open-Domain	Documents, Web	F1	Question Classification	NTCIR
P43	2015	English	-	Insurance	Documents	Accuracy	Candidate Answer Ranking	Created by Authors
P44	2015	English	-	Open-Domain	Answer List	MAP, MRR	Candidate Answer Ranking	Yahoo! Answers
P45	2015	English	-	Open-Domain	Documents	P@k, MAP, MRR	Candidate Answer Ranking	TREC-QA, AQUAINT
P46	2015	English	-	Open-Domain	Web	P@k, MRR, NDCG	Passage Ranking	WebAP
P47	2014	English	Why, How	Biology	Answer List	P@k, MRR	Candidate Answer Ranking	Yahoo! Answers, Biology Textbook Corpus
P48	2014	Arabic, Chinese,	Opinion	Open-Domain	Answer List	F1	Question Classification, Candidate Answer Ranking	BOLT, TAC

Table 1 (continued)

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
		English						
P49	2014	Korean	Definition	Open-Domain	Web	F1, P@k, MRR, Recall@K	Candidate Answer Extraction, Candidate Answer Ranking	Created by Authors
P50	2013	Japanese	Why	Open-Domain	Documents	P@1, MAP	Candidate Answer Ranking	NTCIR-6 QAC
P51	2013	Arabic	Opinion	Political	Documents	Precision	Question Classification	Created by Authors
P52	2013	English	Comparison	Open-Domain	-	F1	Question Classification	Yahoo! Answers
P53	2013	English	How	Open-Domain	Answer List	MRR	Candidate Answer Ranking	Yahoo! Answers
P54	2013	English, Italian	-	Open-Domain	Documents	Accuracy, MRR	Candidate Answer Ranking	ResPubliQA (CLEF 2010)
P55	2012	Japanese	Why	Open-Domain	Documents, Web	P@k, MAP	Candidate Answer Extraction, Candidate Answer Ranking	Yahoo! Answers, Created by Authors
P56	2012	Japanese	How	Open-Domain	Web	F1	Candidate Answer Extraction	Created by Authors
P57	2012	Arabic	Why, How	Open-Domain	Web	Accuracy	Candidate Answer Extraction	Created by Authors
P58	2012	English	-	Open-Domain	Knowledge Graph	F1, Rouge	Question Classification, Question Reformulation, Candidate Answer Ranking	Created by Authors
P59	2012	Japanese	Why	Open-Domain	Web	F1, Accuracy	Candidate Answer Extraction	Yahoo! Answers
P60	2011	English	-	Health	Documents	Manual	Question Classification, Question Reformulation, Passage Extraction	Clinical Collection
P61	2011	English	Why, Definition, Confirmation, How, Factoid	Open-Domain	Documents	P@1, Precision	Question Classification	Created by Authors

Table 1 (continued)

Study	Year of Publication	Language	Question Type	Domain of Knowledge	Knowledge Source Type	Evaluation Metrics	Research Problem	Datasets
P62	2011	English	Definition, Factoid	Open-Domain	Documents, Knowledge Graph	F1	Question Classification	Jeopardy!
P63	2011	English	-	Open-Domain	Answer List	Rouge	Candidate Answer Extraction	Yahoo! Answers
P64	2011	English	How	Open-Domain	Answer List	P@k, MRR	Question Reformulation, Candidate Answer Ranking	Yahoo! Answers
P65	2011	Chinese	Why, How	Open-Domain	Documents	P@k, MRR	Candidate Answer Extraction, Candidate Answer Ranking	Created by Authors
P66	2011	Chinese	Why, Definition	Open-Domain	Documents, Web	F1	Candidate Answer Extraction, Candidate Answer Ranking	NTCIR-8 CCLQA
P67	2011	English	How	Open-Domain	Web, Answer List	P@k, MRR, Recall@K	Candidate Answer Extraction, Candidate Answer Ranking	Yahoo! Answers
P68	2010	English	Why, How, Definition	Open-Domain	Documents, Web	F1	Candidate Answer Ranking	WEB-QA, TREC-QA
P69	2010	English	Why	Open-Domain	Documents	F1	Candidate Answer Extraction	Yahoo! Answers
P70	2010	English	Definition, Opinion	Open-Domain	Documents, Web	F1	Candidate Answer Ranking	Yahoo! Answers, TREC-QA
P71	2010	English	Opinion	Open-Domain	Documents, Web	F1	Question Reformulation	MPQA
P72	2010	Arabic	Definition	Open-Domain	Documents, Web	P@k	Candidate Answer Extraction, Candidate Answer Ranking	Created by Authors
P73	2010	Chinese	-	Open-Domain	Web	P@k, MRR	Candidate Answer Ranking	Created by Authors
P74	2010	English	-	Health	Documents	F1	Question Classification, Question Reformulation	Created by Authors
P75	2010	English	Definition	Open-Domain	Web	F1, MAP	Candidate Answer Ranking	TREC-QA, Created by Authors

is related to coverage since we only covered studies focused on QA systems that explicitly performed experiments addressing non-factoid questions. Consequently, related problems such as automatic text summarization were not included because they were not evaluated using QA benchmarks and standards.

4 Results and discussion

Our systematic review initially covered a total of 455 studies; nevertheless, after carefully removing duplicates, this number reduced to 422 papers. The inclusion and exclusion criteria were divided into two steps: while the first step removed 165 unrelated studies based on titles and keywords, the second step excluded 154 studies based on a manual reading of the abstract. During the eligibility and quality control, we employed semi-automatic annotation to remove 28 studies that did not fulfill the quality criteria. Therefore, we considered only 75 studies for a manual analysis listed in Appendix. Figure 2 summarises the systematic review process.

4.1 Publication over the years

In our study, we have observed a recent interest in non-factoid Question Answering. Figure 3 presents the works' distribution over the years, where the red line represents the accumulated amount of studies and shows how accentuated the evolution in the number of studies is. It shows a reduction of publications in 2010 and a rise in interest since 2015. We believe

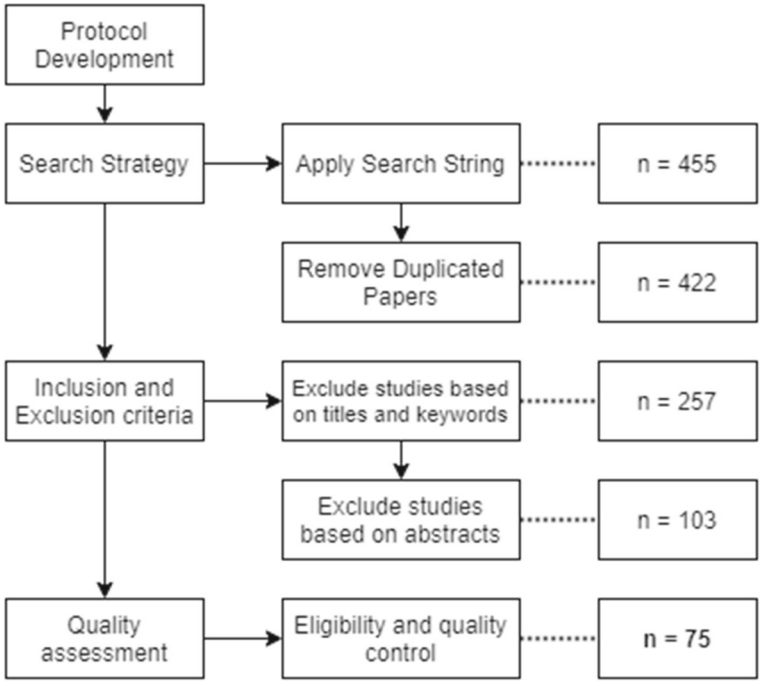

Fig. 2 Systematic review processes

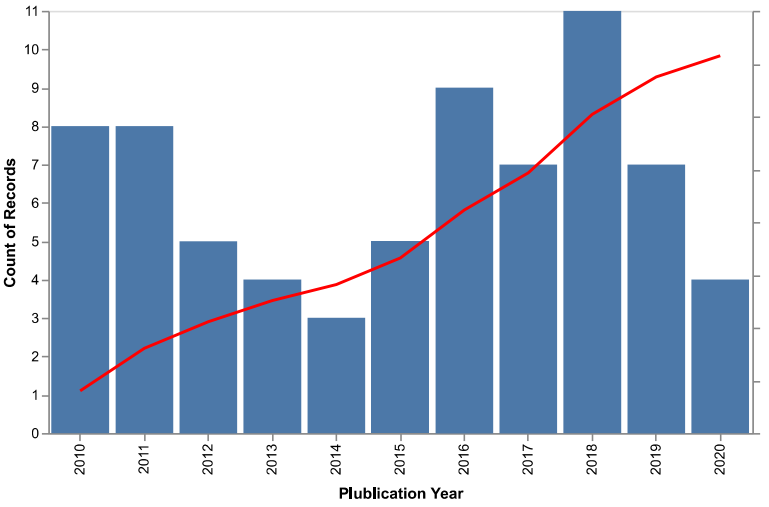

Fig. 3 Number of publication over the years

that this fact is related to the development of new and efficient language models, such as Word2vec and Bidirectional Encoder Representations from Transformers (BERT), and the use of models based on Deep Learning methods that proved to be suitable for solving QA problems.

Also, about 20% of the studies were published in journals, while the leftover was published in the conference literature. We observe that most of the analyzed studies focus on specific tasks and techniques of a QA system and not on the system as a whole. Thus, these studies often do not have enough content to justify their publication in a journal, fitting better into the conference literature.

4.2 Language focus

Concerning Language, as expected, non-factoid QA systems have mainly addressed English documents. Among all studies, 74.7% (56 studies) used English as a primary language, and only 5.3% (4) use a multi-language strategy (P26, P27, P48, P54). Among non-English works, Chinese is the most addressed Language representing a total 12.6% (9) (P8, P36, P42), followed by Arabic with 7% (5) (P35, P48, P51, P57, P72), and Japanese with 5.3% (4) (P5, P55, P56, P59). Indonesian (P34), Italian (P54), Korean (P49), and Russian (P3) had only one study each, representing 5.3% of the works.

4.3 Types of non-factoid questions

Most of the studies (53.4%) addressed definition (what), casualty (how), and reasoning (why) questions (P1, P5, P8, P9, P10), followed by studies focused on confirmation and comparison questions (14.1%) (P2, P3, P4, P7, P11). Table 2 gives an overview of the different types of questions addressed by non-factoid QA works. We observed that studies addressing definition, casualty, or reasoning usually address multiple types of questions because of their similarity. Nevertheless, we found many studies have addressed a single type of question, such as confirmation or comparison. Although most of the studies have

Table 2 Distribution of studies in relation to the type of question

Question Type	Studies
Definition	14
How	13
Why	11
Opinion	7
Comparison	6
Confirmation	4
Factoid Included	12

*The sum of studies is less than the total of analyzed paper because we just included papers that specified the question type

focused on non-factoid questions, few works (16%) also included factoid questions in their experiments (P3, P4, P19, P20, P33).

4.4 Application domain

Many studies focused on an open-domain, such as the class of questions where the answer is found on *Wikipedia* or *Yahoo! answers*. In total, 85.3% (62) of studies focused on an open-domain, while only 17.3% (13) focused on a closed-domain, such as health and insurance. Although, the main areas of application among the closed-domains are health and insurance, we have also observed that non-factoid QA systems have been used for answering questions in the domain of agriculture (P27), biology (P47), E-commerce (P12), financial (P18), geography (P8), political (P51), and tourism (P36).

4.5 Knowledge source

Most of the works use unstructured data such as textual documents (P5, P7, P23) and webpages (P16, P28, P59) for answering questions. Table 3 presents the knowledge source type used by the analyzed studies. In total, 50.7% (36) of works use textual documents, 36% (27) use webpages, 25.3% (19) use a list of possible answers, and 8% (6) use knowledge graphs as a source of knowledge. Few studies employ a combination of different knowledge sources, such as knowledge graphs and documents (P8, P12, P62).

4.6 Most employed metrics for quality assessment

Before, we present a brief explanation of the most common metrics and evaluation methods used by the analyzed studies.

– **Mean Reciprocal Rank (MRR)** is a statistic measure for evaluating any process that produces a list of possible responses to a sample of queries, ordered by probability of

Table 3 Studies distribution over the type of Knowledge Source

Knowledge Source	Studies
Documents	38
Web	27
Answer List	19
Knowledge Graph	6

correctness. Regarding QA systems, giving a set of questions Q, the Mean Reciprocal Rank is definied as:

$$MRR = \frac{1}{Q}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

where $rank_i$ is the rank position of the first relevant answer for the i-th question (Dimitrakis et al., 2019).

- **Precision@k (P@k)** corresponds to the number of relevant results among the top-k-answers. For example, the precision of a QA model that return k possible answers for a question q is given by:

$$P@k = \frac{|Found(q)|}{k}$$

where $Found(q)$ is the list of correct answers returned by the model.

- **Mean Average Precision (MAP)** evaluates the mean of the average precision for a set of queries. It is mainly used to evaluate ranked-results. For example, giving a set of questions Q, the MAP is calculated by:

$$\frac{MAP = \sum_{q=1}^{|Q|} AveP(q)}{|Q|}$$

where $AveP(q)$ consider the order in which the returned result are presented by computing a precision and recall at every position in the ranked sequence of results. Therefore, it is the average value of the precision as a function of the recall of the question q.

- **Accuracy** is the fraction of the questions that are answered correctly. For example, for a set of questions Q, the Accuracy is calculated as:

$$Accuracy = \frac{|CQ|}{|Q|}$$

where CQ are those questions that were answered correctly.

- **F-Score** is a weighted harmonic mean between precision and recall. The F-Score is computed as:

$$F_\beta = \frac{(1+\beta^2)\cdot(precision\cdot recall)}{\beta^2\cdot precision + recall}$$

.

- **Normalized Discounted Cumulative Gain (nDCG)** measures the usefulness, or gain, of a document based on its position in the result list. It is normally employed to measure of ranking quality. The nDCG is computed as:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

where p is the particular rank position. The DCG_p is the discounted cumulative gain and penalizes highly relevant answers that appear lower in the rank of answers candidates. It is computed as:

$$DCG_p = \sum_{i=1}^{p}\frac{2^{rel_i}-1}{log_2(i_1)}$$

where rel_i is the graded relevance of the result at position i. The $IDCG_p$ is computed as:

$$IDCG_p = \sum_{i=1}^{|REL_p|} \frac{2^{rel_i} - 1}{log_2(i_1)}$$

where REL_p is the the list of relevant answers ordered by their relevance.

– **Recall-Oriented Understudy for Gisting Evaluation (ROUGE)** evaluates the answers returned by the QA system, comparing them against correct answers. It works by comparing the produced answer against a set of reference answers. The $ROUGE_{Recall}$ is computed as:

$$ROUGE_{Recall} = \frac{overlaps}{correct\_words}$$

where $overlaps$ is the number of overlapping words between the produced answer and the correct one. $correct\_words$ is the total of words in the correct answer. The $ROUGE_{Precision}$ is computed as:

$$ROUGE_{Precision} = \frac{overlaps}{produced\_words}$$

where $produced\_words$ is the total of words in the produced answer.

There are different variants of ROUGE. For example, ROUGE-N, which considers the overlap of N-grams, and ROUGE-L evaluates the longest common sub-sequence between the predicted and correct answer.

– **Human Assessment** is an evaluation strategy mainly used for real-time competitions, such as LiveQA Trec (Agichtein et al., 2015). It uses humans to evaluate QA systems through a manual judgment of answers.

Table 4 presents the most used metrics to evaluate QA systems. The most employed metrics are MRR (45.3%) and P@K (37.3%), used mainly for assessment of Candidate Answer Extraction (P4, P13, P17) and Candidate Answer Ranking (P20, P21, P22). Besides, few studies have employed Accuracy (18.7%) and F-score (17.3%) to assess Answer Generation (P5, P40, P41) and Question Classification (P48, P62, P74). The method called "Human Assessment" (5.3%) represents a manual evaluations – mainly used for competitions, such as LiveQA Trec –, where experts assess the systems in real-time (Agichtein et al., 2015).

4.7 Tasks involved and methods used for non-factoid question answering systems

We divided our analysis according to the task present in Section 2. We observed that 81.3% (61) of the works have focused on Answer Processing, followed by 26.7% (20) on Question Processing, and 6.7% (5) on Information Retrieval. Many studies have focused on Answer Processing since this component is different from the conventional factoid QA system. Many studies have also focused on Question Processing, which shows concern for extracting pertinent information for non-factoid questions. Table 5 summarizes the distribution of the tasks over the works.

Candidate Ranking is the most addressed task by the analyzed studies. The principal strategy employed in these studies is supervised learning (P27, P36), where the goal is to rank a list of potential answers. There are two main approaches, namely ranking and classification. Regarding ranking, most of the works try to estimate the distance between question input and the answer candidate. While some works set a score for each candidate based on lexical and semantic features related to the difference between the input question and the

Table 4 Evaluation strategy used by studies

Metric	Associated Task	Studies
MRR	Question Reformulation	34
	Document Retrieval	
	Passage Extraction	
	Candidate Answer Extraction	
	Candidate Answer Ranking	
P@K	Candidate Answer Extraction	28
	Candidate Answer Ranking	
MAP	Candidate Answer Ranking	14
Accuracy	Question Classification	14
	Candidate Answer Extraction	
	Candidate Answer Ranking	
	Answer Generation	
F-score	Question Classification	13
	Question Reformulation	
	Answer Generation	
NDCG	Candidate Answer Extraction	12
	Candidate Answer Ranking	
ROUGE	Candidate Answer Extraction	6
Human Assessment	Candidate Answer Ranking	4
Others		9

answer candidate (P23, P39, P49), other studies propose learning to rank strategies based on lexical, semantic, and other textual features extracted from the text using machine learning models (P19, P20, P54). Conversely, some studies treated this task as a binary classification problem, where the system classifies the candidate answer as correct or incorrect (P15, P27, P67). Not least of all, some studies presented considerable results using pre-trained models, and attention mechanisms, such as BERT (P1, P5). These studies show that these methods based on neural model significantly outperform other models, such as BM25, an effective term-matching retrieval model.

Table 5 Studies distribution over QA architecture stages and tasks

Architecture Stage	Task	Studies
Question Processing	Question Classification	12
(20)	Question Reformulation	11
Information Retrieval	Document Retrieval	4
(5)	Passage Extraction	3
Answer Processing	Candidate Answer Extraction	29
(61)	Candidate Answer Ranking	46
	Answer Generation	3

*The sum of studies go over the total of analyzed papers, once it is possible to assign more than one stage or task by study

The second most addressed task in non-factoid QA systems is *Candidate Answer Extraction*. The most employed approaches are based on traditional information retrieval methods, such as BM25, which estimates the relevance of a document giving a query (P17, P19), to deep neural models (P13, P16, P23). These studies show that deep neural models, such as Long short-term memory (LSTM) and Convolutional Neural network (CNN), present better results than traditional ones. Few studies employ summarization methods to create candidate answers (P16, P29, P63), such as deep auto-encoder and LSTM auto-encoder for sentence representation. Some other studies also use answer lists from community QA websites selecting text fragments that are more likely to bear answers to the query. Experiments show a positive impact on the performance optimization-based summaries (P23). Furthermore, several studies have used knowledge graphs to support Candidate Answer Extraction (P2, P4, P15, P66) by harness the unique properties of knowledge graphs to treat data redundancy, access the links between data objects, run efficient queries against the knowledge graph and explore the updated nature of the knowledge. Also, some studies employ metathesaurus and sentiment analysis for answer extraction (P4).

Works addressing *Question Classification* have combined different features, such as lexico-syntactic (P42, P62) and sentiment analysis (P58). Handcrafted rules usually perform well due to experts' effort to create manual rules (Cortes et al., 2020), however only few works have proposed a combination of handcrafted rules with lexico-syntactic patterns to classify questions (P3, P4). We have also observed studies proposing new taxonomy for non-factoid questions that best fit specific domains (P51, P61). The results suggest that new taxonomy with multi-label classification is better than a single-label, once it helps to reduce the search space for answers.

Studies on *Question Reformulation* have proposed methods based on the extraction of different information from the questions. For example, (P8, P15) decompose the question into sub-queries and resolve each of them individually to create a final answer. On the other hand, some studies have expanded the question using external knowledge bases, such as Wikipedia or a knowledge graph (P64, P71).

The studies on *Document Retrieval* have directly looked for the answers on web pages using search engines (P19, P28, P64). Usually, these studies use commercial search engines, such as Google Search API and Bing, to mine answers candidates from the top web pages retrieved. Also, some studies use the search engines themselves to expand the questions using the snippets of the top search results (P64). Few studies have proposed methods for dealing with technical terminology of particular domains through special encoders (P4, P7) applying metathesaurus, synonyms, and cross-attention mechanisms between the query and document words to discover the important terms.

In spite of *Answer Generation*, we observed that most studies preferred to use multiple passages extracted from the original document instead of generating a single response using Natural Language Generation. The few works addressing *Answer Generation* proposed end-to-end methods based on neural reading comprehension models to extract and generate the answer from documents (P5, P40, P41). Some of them also employed external knowledge to capture deep semantic relationships between sentences and questions to acquire question-aware representations for the document (P40, P41).

4.8 Data sets

We have observed that many works do not use standard benchmarks for evaluating their systems. While some of them use a subset of an existing data set, others build a new data set from scratch. During the analysis process, we have noticed that the works tend to modify the

data sets according to their experiments' needs. Therefore, we have found several different versions of the same data set and overlapping of data. Table 6 shows those data sets and some of their relevant features for the QA research. Table 7 has references that help to access these data sets.

Among the data sets used, several can be classified as community QA data sets. They are collections composed of questions and answers created by users from community QA web portals. The majority of a community QA data set questions are not trivial to be answered with a simple web search. Therefore these questions can be classified as complex and, most of the time, as non-factoid. Also, different from conventional collections that are created

Table 6 Data sets used by the studies

Collection	Questions	Documents	Language	Domain
AgricultureQA	3,000	-	Chinese	Agriculture
ANTIQUE	2,626	-	English	Open
BioASQ	3,243	-	English	Biomedical
Biology Textbook Corpus (Bio)	378	-	English	Biology
BOLT	455	62,000	Arabic, Chinese and English	Open
Clinical Questions Collection	4,654	-	English	Health
FiQA	6,646	57,641	English	Financial
HealthQA	7,517	7,355	English	Health
InsuranceQA	16,889	-	English	Insurance
L5 - Yahoo! Manner Questions	142,627	-	English	Open
L6 - Yahoo! Comprehensive QA	4,483,032	-	English	Open
LC-QuAD	5,000	-	English	Open
MPQA	30	98	English	Open
MS MARCO	100,000	200,000	English	Open
NTCIR 2008	30	-	Chinese	Open
ResPubliQA (CLEF 2010)	200	10,700	Bulgarian, Dutch, English, French, German, Italian, Portuguese, Romanian and Spanish	Open
SemEval-2015, 2016 and 2017	2,942	-	Arabic and English	Open
SimpleQuestions (v2)	108,442	-	English	Open
SQuAD	107,785	536	English	Open
TAC 2008 Opinion QA track	89	100,649	English	Open
TREC LiveQA 2015	1,087	-	English	Open & Health
TREC LiveQA 2016	1,015	-	English	Open & Health
TREC LiveQA 2017	1,182	-	English	Open & Health
TREC-QA	2,256	-	English	Open
WEB-QA	1,309	-	English	Open
WebAP	82	710	English	Open
WikiPassageQA	4,165	244,136	English	Open

requiring the user to invent a question for given information, this type of collection has the advantage of being naturally created by the user in natural conditions of questioning.

The main difference between the Community QA data set and the conventional ones is that the Community QA collections' answers should be considered candidate answers (Bae & Ko, 2019; Khushhal et al., 2020). It is usually not validated by certified experts and has a score or a ranked order from users' votes. Also, (Surdeanu et al., 2008; Yan & Zhou, 2015) describe that these candidate answers have a high variance of quality like answers range from exceptionally informative to completely irrelevant, and someones can be even abusive.

Unlike conventional End-to-End QA, the task involving Community QA data sets usually relies on selecting the most appropriate answers from a given list of answers candidates for the target questions. The author usually picked the most voted answer as the correct candidate to elaborate on the list of candidate answers during these data sets' construction. The rest of the irrelevant one is picked from answers candidates of other questions (Cohen

Table 7 Data sets reference (links accessed in June 2021)

Collection	Reference
AgricultureQA	Cited by http://dx.doi.org/10.1109/IALP.2017.8300620
ANTIQUE	https://ciir.cs.umass.edu/downloads/Antique/
BioASQ	http://participants-area.bioasq.org/datasets/
Biology Textbook Corpus (Bio)	Cited by http://dx.doi.org/10.3115/v1/p14-1092
BOLT	Cited by https://doi.org/10.1145/2566486.2567999
Clinical Questions Collection	https://www.nlm.nih.gov/databases/download/CQC.html
FiQA	https://sites.google.com/view/fiqa/home
HealthQA	https://github.com/mingzhu0527/HAR
InsuranceQA	https://github.com/shuzi/insuranceQA
L5 - Yahoo! Answers Manner Questions	https://webscope.sandbox.yahoo.com/
L6 - Yahoo! Answers Comprehensive QA	https://webscope.sandbox.yahoo.com/
LC-QuAD	https://figshare.com/projects/LC-QuAD/21812
MPQA	http://mpqa.cs.pitt.edu/corpora/mpqa_corpus/
MS MARCO	https://microsoft.github.io/msmarco/
NTCIR 2008	Cited by http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/NTCIR/toc_eval.html
ResPubliQA (CLEF 2010)	Cited by https://doi.org/10.1145/2484028.2484233
SemEval-2015, 2016 and 2017	https://alt.qcri.org/semeval2017/task3/index.php?id=data-and-tools
SimpleQuestions (v2)	https://research.fb.com/downloads/babi/
SQuAD	https://rajpurkar.github.io/SQuAD-explorer/
TAC 2008 Opinion QA track	https://tac.nist.gov/data/
TREC LiveQA 2015	https://trec.nist.gov/data/qa/2015_LiveQA.html
TREC LiveQA 2016	https://trec.nist.gov/data/qa/2016_LiveQA.html
TREC LiveQA 2017	https://trec.nist.gov/data/qa/2017_LiveQA.html
TREC-QA	http://disi.unitn.it/~silviaq/resources.html
WEB-QA	http://disi.unitn.it/~silviaq/resources.html
WebAP	http://ciir.cs.umass.edu/downloads/WebAP/
WikiPassageQA	https://ciir.cs.umass.edu/downloads/wikipassageqa/

et al., 2018). Therefore, most of the candidate answers list may not have any semantic relationship to the target question (Table 7).

4.9 Limitations of non-factoid question answering

The limitation of the analyzed studies is related to how they provide the answer. The majority of studies focus on selecting few passages from different documents and ranking them according to their usefulness to answer a question. However, it is common for non-factoid questions to have several restrictions, narrowing the search space down to a specific answer. For instance, for the question "How should I treat measles in a 12-year-old boy?" the ideal passage to be used as an answer should cover "treatment", "measles", "12-year-old" and "boy", which is very unlikely and there may not be a ready-made passage in the knowledge base containing all the information needed. In this case, the ideal system must search for different information pieces in different documents and merge them to compose a single answer. However, this is challenging and still an open research problem.

Some works have tried to overcome this limitation by presenting a set of sentences grouped by the terms (P4, P60, P70). However, this approach still requires a great interpretation effort from the user.

Regarding evaluation, non-factoid QA requires a great deal of manual effort to verify the system's correctness. Unlike factoid QA systems, where a question usually has one of few correct alternatives, answers for non-factoid questions can be expressed in infinitive manners. Therefore, it is challenging for humans to assess end-to-end non-factoid QA systems.

5 Conclusion

In this paper, we presented a systematic review of the literature addressing non-factoid QA systems. From a total of 455 recent studies, we selected 75 papers based on our quality control system and exclusion criteria for an in-depth analysis. This work aims to explain the particular aspects of non-factoid QA systems, such as the distinct tasks and methods, the available benchmarks, and the different types of questions addressed in recent works. This systematic review helped to answer the following questions:

What are the tasks and methods involved in non-factoid Question Answering systems? We observed that the general architecture of non-factoid QA systems does not differ from factoids. Nevertheless, the methods employed in each task of the non-factoid QA system vary to some extent. For example, our empirical analysis showed that many studies on non-factoid questions have focused on *Candidate Answer Extraction*. While in factoid question, the *Candidate Answer Extraction* is responsible for extracting entities as a possible answer candidate, non-factoid QA systems extract multiples passages from a document(s) to compose a single answer. The methods used in this task vary from BM25 based methods (P1, P4, P28, P60) to Deep Neural models (P12, P17, P22, P43). Our review also revealed that although the composition of an answer based on multiple passages is one of the most distinct characteristics of non-factoid questions, only a few works have addressed this problem so far. The only few works that have addressed this issue have used Automatic Text Summarization to digest multiple passages retrieved by previous steps and generate the user's final answer (P16, P29, P63).

What are the data sets available for non-factoid Question Answering systems? We have found an increasing number of available data sets for non-factoid questions. As expected,

most data sets address the English Language; however, we have found data sets for non-English Languages such as Chinese, Arabic, and Japanese. Regarding area of application, most of the available data sets addresses health and insurance, followed by agriculture (P27), biology (P47), E-commerce (P12), financial (P18), geography (P8), political (P51), and tourism (P36). We have also noticed that many works tend to modify the data sets according to their needs. Consequently, different versions of the same data set are used for evaluation, which makes a fair comparison between the studies difficult.

What are the limitations of non-factoid Question Answering? Automatic generation of answers based on multiple passages is a critical issue for developing full end-to-end non-factoid question-answer systems. The problem emerges from the fact that automatic generation of coherent and cohesive text – especially for long passages – is still an open research question (Bau et al., 2020). Broadly speaking, coherence and cohesion refer to how a text is organized so that it can hold together. In a coherent answer, concepts are connected meaningfully and logically by using grammatical and lexical cohesive devices. Furthermore, evaluation of an end-to-end non-factoid QA system seems to be a challenging issue. Although quality estimation is a critical component for developing better systems, this kind of problem is not exclusively of QA systems but also for all text-to-text applications, such as machine translation, text simplification, text summarization, grammatical error correction, and natural language generation (Specia et al., 2018).

The future directions in the non-factoid QA should concern methods that generate natural language and use several information sources to compose complex answers instead of using a simple extracted sentence. Also, there are challenges in other QA pipeline stages to compose answers through structured knowledge bases and extract relevant information from complex questions. Finally, researchers must seek to share the same data set versions in their experiments to compare results between proposed methods.

Author Contributions The idea of the systematic review on Non-factoid Question Answering came from the author Eduardo G. Cortes. All authors contributed to the study design. The survey of the analyzed studies and their analysis was carried out by Eduardo G. Cortes and Vinicius Woloszyn with Dante Barone's supervision, Renata Vieira and Sebastian Möller. The first draft of the manuscript was written, evaluated, and corrected by all authors Eduardo G. Cortes, Vinicius Woloszyn, Dante Barone, Renata Vieira, and Sebastian Möller. All authors read and approved the final manuscript.

Funding This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Consent to participate There is the consent of all authors.

Consent for Publication There is the consent of all authors.

Conflicts of Interest/Competing Interests There are no conflicts or competing interests. We ensure that this manuscript has a novel contribution and has not been published or submitted to any other publisher before.

References

Agichtein, E., Carmel, D., Pelleg, D., Pinter, Y., & Harman, D. (2015). Overview of the trec 2015 liveqa track. In *TREC*.

Bae, K., & Ko, Y. (2019). Efficient question classification and retrieval using category information and word embedding on cQA services. *Journal of Intelligent Information Systems*, *53*(1), 27–49. https://doi.org/10.1007/s10844-019-00556-x.

Bau, D., Liu, S., Wang, T., Zhu, J.-Y., & Torralba, A. (2020). Rewriting a deep generative model. In A. Vedaldi, H. Bischof, T. Brox, & J.-M. Frahm (Eds.) *Computer Vision – ECCV 2020, pp 351–369. Springer International Publishing. Cham*.

Ben Abacha, A., & Zweigenbaum, P. (2015). Means: A medical question-answering system combining nlp techniques and semantic web technologies. *Information Processing & Management*, *51*(5), 570–594. https://doi.org/10.1016/j.ipm.2015.04.006. https://www.sciencedirect.com/science/article/pii/S0306457315000515.

Bondarenko, A., Braslavski, P., Völske, M., Aly, R., Fröbe, M., Panchenko, A., Biemann, C., Stein, B., & Hagen, M. (2020). Comparative web search questions. In *Proceedings of the 13th International Conference on Web Search and Data Mining, WSDM '20, pp 52–60. Association for Computing Machinery, New York, NY, USA*. https://doi.org/10.1145/3336191.3371848.

Calijorne Soares, M. A., & Parreiras, F. S. (2020). A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, *32*(6), 635–646. https://doi.org/10.1016/j.jksuci.2018.08.005.

Chali, Y., Hasan, S. A., & Mojahid, M. (2015). A reinforcement learning formulation to the complex question answering problem. *Information Processing & Management*, *51*(3), 252–272. https://doi.org/10.1016/j.ipm.2015.01.002, https://www.sciencedirect.com/science/article/pii/S0306457315000035.

Cohen, D., Yang, L., & Croft, W.B. (2018). WikiPassageQA: A benchmark collection for research on non-factoid answer passage retrieval. 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018, pp 1165–1168. https://doi.org/10.1145/3209978.3210118.

Corbin, J., & Strauss, A. (2014). Basics of qualitative research: Techniques and procedures for developing grounded theory. Sage publications.

Cortes, E., Woloszyn, V., Binder, A., Himmelsbach, T., Barone, D., & Möller, S (2020). An empirical comparison of question classification methods for question answering systems. In *Proceedings of the 12th Language Resources and Evaluation Conference, pp 5408–5416. European Language Resources Association, Marseille, France*. https://www.aclweb.org/anthology/2020.lrec-1.665.

Denyer, D., & Tranfield, D. (2009). Producing a systematic review. Sage Publications Ltd, 671–689.

Dimitrakis, E., Sgontzos, K., & Tzitzikas, Y. (2019). A survey on question answering systems over linked data and documents. *Journal of Intelligent Information Systems*, *55*, 233–259.

Dybå, T., & Dingsøyr, T. (2008). Empirical studies of agile software development: A systematic review. *Information and software technology*, *50*(9-10), 833–859.

Hazrina, S., Sharef, N. M., Ibrahim, H., Murad, M. A. A., & Noah, S.A.M. (2017). Review on the advancements of disambiguation in semantic question answering system. *Information Processing & Management*, *53*(1), 52–69. https://doi.org/10.1016/j.ipm.2016.06.006, https://www.sciencedirect.com/science/article/pii/S0306457316302102.

Hermjakob, U., Echihabi, A., & Marcu, D. (2002). Natural language based reformulation resource and web exploitation for question answering. In *Proceedings of TREC, 11. Citeseer*.

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V.A. (2019). *Cochrane handbook for systematic reviews of interventions*. New York: John Wiley & Sons.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). SpanBERT: Improving Pre-training by Representing and Predicting Spans. *Transactions of the Association for Computational Linguistics*, *8*, 64–77. https://doi.org/10.1162/tacl_a_00300.

Khan, K. S., Kunz, R., Kleijnen, J., & Antes, G. (2003). Five steps to conducting a systematic review. *Journal of the royal society of medicine*, *96*(3), 118–121.

Khushhal, S., Majid, A., Abbas, S. A., Nadeem, M. S. A., & Shah, S. (2020). Question retrieval using combined queries in community question answering. *Journal of Intelligent Information Systems*, *55*, 307–327. https://doi.org/10.1007/s10844-020-00612-x.

Kodra, L., & Kajo, E. (2017). Question Answering Systems: A Review on Present Developments, Challenges and Trends. *International Journal of Advanced Computer Science and Applications*, *8*(9), 217–224. https://doi.org/10.14569/ijacsa.2017.080931.

Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. *Information Sciences*, *181*(24), 5412–5434. https://doi.org/10.1016/j.ins.2011.07.047.

Liu, Y., Yi, X., Chen, R., & Song, Y. (2016). A Survey on Frameworks and Methods of Question Answering. Proceedings - 2016 3rd International Conference on Information Science and Control Engineering, ICISCE 2016, pp 115–119. https://doi.org/10.1109/ICISCE.2016.35.

Malviya, M., & Soni, M. (2020). Question answering schemes: A review. *International Journal of Scientific Research & Engineering Trends*, *6*(4), 2641–2648.

Mishra, A., & Jain, S. K. (2016). A survey on question answering systems with classification. *Journal of King Saud University - Computer and Information Sciences*, *28*(3), 345–361. https://doi.org/10.1016/j.jksuci.2014.10.007.

Noraset, T., Lowphansirikul, L., & Tuarob, S. (2021). Wabiqa: A wikipedia-based thai question-answering system. *Information Processing & Management*, *58*(1), 102431. https://doi.org/10.1016/j.ipm.2020.102431.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic reviews*, *5*(1), 210.

Papadakis, M., & Tzitzikas, Y. (2015). Answering keyword queries through cached subqueries in best match retrieval models. *Journal of Intelligent Information System*, *44*(1), 67–106. https://doi.org/10.1007/s10844-014-0330-7.

Seers, K. (2012). Qualitative data analysis. *Evidence-based nursing*, *15*(1), 2–2.

Shah, A. A., Ravana, S. D., Hamid, S., & Ismail, M.A. (2019). Accuracy evaluation of methods and techniques in Web-based question answering systems: a survey. *Knowledge and Information Systems*, *58*(3), 611–650. https://doi.org/10.1007/s10115-018-1203-0.

Shen, S., Dong, Z., Ye, J., Ma, L., Yao, Z., Gholami, A., Mahoney, M. W., & Keutzer, K. (2020). Q-bert: Hessian based ultra low precision quantization of bert. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8815–8821. https://doi.org/10.1609/aaai.v34i05.6409, https://ojs.aaai.org/index.php/AAAI/article/view/6409.

Specia, L., Scarton, C., & Paetzold, G.H. (2018). Quality estimation for machine translation. *Synthesis Lectures on Human Language Technologies*, *11*(1), 1–162.

Sultana, T., & Badugu, S. (2020). A review on different question answering system approaches. In S. C. Satapathy, K. S. Raju, K. Shyamala, D. R. Krishna, & M. N. Favorskaya (Eds.) *Advances in Decision Sciences, Image Processing, Security and Computer Vision, pp 579–586. Springer International Publishing, Cham*.

Surdeanu, M., Ciaramita, M., & Zaragoza, H. (2008). Learning to rank answers on large online QA collections. In *ACL-08: HLT - 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp 719–727*.

Tranfield, D., Denyer, D., & Smart, P. (2003). Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *British journal of management*, *14*(3), 207–222.

Wang, W., Yang, N., Wei, F., Chang, B., & Zhou, M. (2017). Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp 189–198. Association for Computational Linguistics, Vancouver, Canada*. https://www.aclweb.org/anthology/P17-1018.

Wu, Y., Hori, C., Kashioka, H., & Kawai, H. (2015). Leveraging social Q&A collections for improving complex question answering. *Computer Speech and Language*, *29*(1), 1–19. https://doi.org/10.1016/j.csl.2014.06.001.

Yan, Z., & Zhou, J. (2015). Optimal answerer ranking for new questions in community question answering. *Information Processing & Management*, *51*(1), 163–178. https://doi.org/10.1016/j.ipm.2014.07.009.

Yang, L., Ai, Q., Spina, D., Chen, R. C., Pang, L., Bruce Croft, W., Guo, J., & Scholer, F. (2016). Beyond factoid QA: Effective methods for non-factoid answer sentence retrieval. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 9626, pp 115–128. Springer Verlag*.

Yogish, D., Manjunath, T. N., & Hegadi, R.S. (2018). Survey on trends and methods of an intelligent answering system. International Conference on Electrical, Electronics, Communication Computer Technologies and Optimization Techniques, ICEECCOT 2017, 2018-Janua:346–353. https://doi.org/10.1109/ICEECCOT.2017.8284526.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738.

Affiliations

Eduardo Gabriel Cortes[1] · Vinicius Woloszyn[2] · Dante Barone[1] · Sebastian Möller[2] · Renata Vieira[3]

Vinicius Woloszyn
woloszyn@tu-berlin.de

Dante Barone
barone@inf.ufrgs.br

Sebastian Möller
sebastian.moeller@tu-berlin.de

Renata Vieira
renatav@uevora.pt

1 Federal University of Rio Grande do Sul, Porto Alegre, Brazil

2 Technische Universität Berlin, Berlin, Germany

3 CIDEHUS, Évora University, Évora, Portugal