Knowledge and Information Systems (2024) 66:2235–2268
https://doi.org/10.1007/s10115-023-02019-w

REVIEW

Techniques, datasets, evaluation metrics and future directions of a question answering system

Faiza Qamar[1] · **Seemab Latif**[1] · **Asad Shah**[1]

Received: 20 July 2022 / Revised: 15 July 2023 / Accepted: 31 October 2023 /
Published online: 22 December 2023

Abstract
Question answering has been around for more than half a century. The problem was addressed with different solutions in the eras of different technologies. Some proved more helpful and accurate than the other. Different studies are available online which list and summarize the work done in this domain. This SLR adds up to that list with answers to some questions which will assist the researchers in this field to comprehend the existing knowledge, quickly analyze the available facts and determine some research gaps and future directions. In this article, we investigate different solution domains applied to question answering systems, their results, and methodologies. We also list and discuss different datasets provided to the community for experiments along with their availability status. In the light of this study, we analyze different solution domains and the areas where they produce promising results. Moreover, we focused on different evaluation metrices used in the papers that were included in this study and shed light on some metrices which should be included in the results if the community wants to achieve greater results. Lastly, we also looked into an interesting possibility of a question answering system where answer could be generated using multiple sources. And for that we suggested a domain based on the Quran, Tafseer and Ahadith data sources as the Quran and Ahadith contribute collectively in the Islamic legislation. We hope this article will help the new researchers in the field of question answering to start their research.

Keywords Question answering system · Deep learning · Knowledge graphs · Quran · Tafseer · Ahadith

✉ Faiza Qamar
fqamar.dphd18seecs@seecs.edu.pk

Seemab Latif
seemab.latif@seecs.edu.pk

Asad Shah
asad.shah@seecs.edu.pk

[1] School of Electrical Engineering and Computer Science (SEECS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

1 Introduction

In recent times when everyone seeks the instant results of everything after putting in their input, the same trend exists in consuming information. People want the answer readily available against their query. This quest lead humans to automatic question answering systems (QAS). QAS is a problem which is addressed multiple times in the literature. It is defined as to find the answer to a given natural language question instead of some relevant documents, automatically. Questions can be of different types with two main categories, i.e., factoid and non-factoid. Both categories are further divided into subcategories, i.e., multiple choice questions (MCQ's), list, definition, description, casual, opinion, hypothetical, procedural, etc. [1, 2].

Numerous studies have been conducted focusing on the automatic question answering problem and provided the solutions. The solutions were based on structured knowledge and free text belonging to domains like information retrieval, machine learning/deep learning, natural language processing (NLP) (statistical approaches), knowledge graph (KG) and combination of these. Each technique has its own pros and cons and some restricted environment and requirements to produce the results [2]. As the techniques to solve the question answering problem grew, so did the datasets related to the problem. Researchers worked with existing datasets and curated new ones to throw a new challenge and to address a new aspect of this problem.

To analyze the available solutions, datasets and to summarize the work done, a systematic literature review (SLR) is presented in this paper. In addition to general question answering system reviews, this study also investigates the implication of such systems on the Quran, Tafseer and Ahadith domain. The Quran and Ahadith are the main source of legislation for Muslims with some others. And Muslims consist of 24% of the world population [3] which makes these religious scriptures, that are followed by masses important and interesting to explore. Many questions are asked to get the guidance on the daily life matters in the light of these scriptures which take Muslim scholars days to answer (keeping in view the large number of questions they receive and the amount of time they need to put in to answer those question). Hence, it can prove to be a very interesting application of an automatic question answering system. Main contributions of this paper are fivefold:

- To summarize the techniques, their merits, demerits and limitations.
- To analyze all the datasets which have been curated over time.
- To analyze the evaluation techniques and highlight the requirements to improve that.
- To investigate the question answering system for the Quran, Tafseer and Hadith domain.
- To list the open research directions to be explored in future.

This study is divided into five sections. Section 1 briefly discusses the QAS and gives a really short introduction of what this study investigated. In Sect. 2, we describe the protocol for this SLR including research questions, selection of sources, inclusion and exclusion criterion. In Sect. 3, we analyze the studies that are included in this review and answered the research questions which are defined in Sect. 2. Section 3 also has the summary of all the papers in tabular form to get a quick glance on individual papers and their contributions (see Table 4). It also lists all the introduced datasets with their brief descriptions (see Table 5). Section 4 presents the results and discussions followed by conclusion in Sect. 5.

2 Systematic mapping

This SLR is prepared on the guidelines provided by Kitchenham [4]. A well thought problem has led to certain questions which will be answered in this study. The formulated questions are

Table 1 Research questions

	Research questions (RQ)
RQ1	Solution based on which area (information retrieval, machine learning/deep learning, knowledge base, natural language processing) is producing promising results in question answering system?
RQ2	Does the results of the method get affected by the type of domain they are applied to, i.e., Open or closed domain?
RQ3	Which datasets have been used in studies from 2016 to 2022?
RQ4	Does literature provide any guideline to produce the answer from different sources?
RQ5	What evaluation metrics are used w.r.t answer type, i.e., factoid or non-factoid?
RQ6	What efforts have been made to apply QAS on the Quran, Tafseer and Ahadith domain?

stated in Table 1. This study is targeting the facts about question answering in general and its application to the domain of Quran, Tafseer and Ahadith. So, the search string was structured keeping that in mind to retrieve all the papers which have addressed question answering systems. Whether those were from open or closed domain or were targeting English or Arabic language. We used Google Scholar used as a search engine. When passed the search string with the date range from 2016 to 2022, 4497 results were returned. Those results were filtered through inclusion and exclusion criteria described in Table 2. The filtered results were passed through a quality assessment checklist which consisted of several questions. We have stated this process in detail, below.

2.1 Research question formalization

Since the purpose of this study is to summarize and analyze the latest work being done in the question answering domain which includes details about; what methods are being used? Which techniques and methods are proving to be the best according to which type of questions? Which datasets resemble the most real-life use cases? What progress has been made in the Quran, Tafseer and Ahadith domain? Does the type of domain create any difference in performance of algorithm/method/technique? All these aspects are covered into well-defined research questions presented in Table 1.

2.2 Search string

One thing would be common in all the papers that would be selected/retrieved from the sources, i.e., the term question answering. Other than that, all the related terms can/cannot be present in the paper. Keeping that in mind, the following search string has been designed which will return the results that must have a "Question answering System" phrase anywhere in the article and all other phrases and terms will be optional.

Search String- "Question answering system" AND (English OR Arabic OR Quran OR Hadith OR Tafseer)

Table 2 Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Only those studies which are published in A*, A and B ranked journal and conference according to CORE Ranking Portal [5]	Studies other than in English Language
Studies from 2016 to Feb 2022	Studies not providing the discrete findings
Studies which have applied QA systems on textual data	Publications where full text is not available
Studies targeting complete question answering system	Books, thesis, editorials, prefaces, article summaries, interviews, news, reviews, correspondences, discussions, comments, reader's letters and panels, and poster sessions and patents, Survey papers
Studies targeting any phase of QAS and investigating its effect on overall system	Studies where a new dataset is introduced without applying any QAS

Words/phrase in double quotes represent the string which should be searched from the search engine "as it is" and it must be there. "OR" represents the concept that this could and could not be present in returned results.

2.3 Selection of sources

The identified search query was executed in Google Scholar to get related papers. The search process also covered journal articles and conference papers available in four of the most reliable electronic databases that are scientifically and technically peer reviewed: ACM Digital Library, IEEE Xplore, Springer Link, Science Direct and other Journals/conferences. The reason for the selection of the Google Scholar was the accessibility of high-quality proceedings of key conferences and journals with reference to computer science and engineering at one place rather searching each digital library separately. Since we were interested in recent articles in this research, we confined our search to articles published in the 2016–2022.

2.4 Inclusion and exclusion criteria

Initial criteria are applied on the papers that were extracted from the Google Scholar after executing the search string. We excluded the literature published in any language other than English. Any study targeting question answering on any type of data other than textual was also excluded. Detailed criteria are mentioned in Table 2.

2.5 Quality assessment checklist (QAC)

A quality assessment checklist was developed to assess the individual studies, based on Kitchenham [4] guidelines. This checklist included the following questions: (a) Does the research paper clearly specify the research methodology? (b) Is the research methodology appropriate for the problem under consideration? (c) Is the analysis of the study properly done? If the study met the assessment criteria, then it was given a "yes."

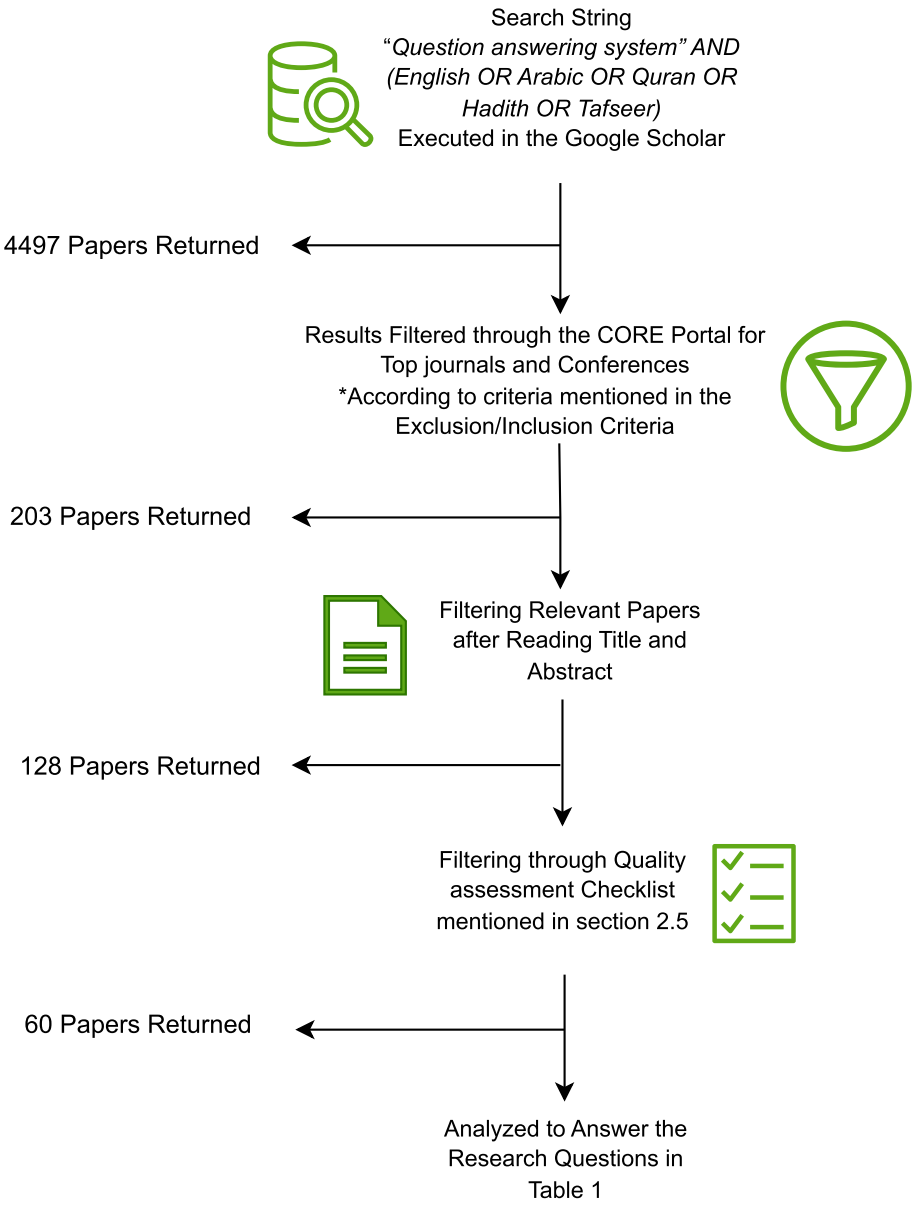

Fig. 1 Research paper retrieval and selection process

2.6 Research paper retrieval

We executed the designed search string on Google Scholar and passed the returned results through an inclusion/exclusion criteria. Studies which passed that criteria were then filtered according to quality assessment checklist. And the studies which passed all the criterion were then critically analyzed to answer the formulated research questions aforementioned in Table

Table 3 Included journal and conferences

Conference	Paper count	Conference	Paper count
AAAI	9	ESWC	1
ACL	14	WWW	5
SIGKDD	2	CIKM	1
PKDD	1	WSDM	4
SIGIR	5	ICLR	1
BioNLP	1	COLING	1
ICWE	1	ICML	1
EMNLP	4	ICANN	1

Journal	Paper count	Journal	Paper count
VLDB	1	Intl. Journal of Speech and Technology	1
Cluster Computing	1	BMC Informatics	1
IEEE Transactions on Knowledge and Data Engg	1	Information Processing and Management	3
Computer Speech and Language	1		

1. The process and number of papers after each step are presented in the form of a flow chart in Fig. 1.

After applying the paper selection process, the resulting papers belonged to different prestigious journals and conferences. Their details are provided in Table 3.

3 Analysis

This section provides the answers to the research question through the analysis of the included research papers. A short summary of those papers is presented in Table 4. It includes details about papers such as domain, type and language in which the QAS is implemented along with the dataset that was used in the paper. The table also explains the methodology implemented by the authors and results achieved by those methods. Answers to research questions with the discussion according to the available literature is available in later part of this section.

RQ1: Solutions based on which area (machine learning/deep learning, knowledge base, statistical approaches in natural language processing) are producing promising results in question answering system?

More than 50 papers, studied and critically analyzed for this SLR, used techniques based on all the research areas mentioned in the question. Authors [6–20] worked on factoid type question answering system using machine or deep learning techniques. All of these studies belonged to open domain. Authors [21, 22] also worked with deep learning models but they targeted the conversational system in open domain. Authors [9, 16, 23, 24] targeted non-factoid question answering system using deep learning approaches. Authors [25–28] applied the combination of deep learning and knowledge graphs. Authors [1, 29–35] used two knowledge graphs to solve the problem; Freebase and DBpedia.

We also noticed an interesting trend that is; In 59% of the studies, where machine or deep learning techniques were applied, it was used in combination with Knowledge base.

Table 4 Summary of papers included in this literature review

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[38]	Open	English	TREC	Factoid	Bi-LSTM (long short-term memory) to sequentially read words from question and answer to output the relevance score	Mean Average Precision (MAP): 71.34%, Mean Reciprocal Rank (MRR): 79.13%
[39]	IT and movies subtitles	English	IT Helpdesk Troubleshooting dataset movies subtitle dataset	Conversation Based	Recurrent Neural Network (RNN)	They didn't provide the quantitative measures for their system
[40]	Open	English	TREC	Factoid	Convolution Neural Network (CNN) with Logistic Regression and Softmax at the output layer	MAP: 74.59%, MAP: 80.78%
[41]	Medical	English	i2b2 corpus, scientific articles clinical reports	Factoid , yes/no	Lexical patterns for Question classification semantic search for answers using SPARQL	Settings: BIO-CRF-H Precision (P): 72.18%, Recall (R): 83.78%, F-score (F): 77.55%"
[42]	Open	English	TREC	Factoid	They collected candidate answers from web resources and then mapped those Candidates to freebase entities to select the top answers	P: 57.92% R: 57.92% F1-score (F1): 57.92%, MRR: 65.32%
[43]	Open	English	bAbi	Factoid	It consists of two layers (1) Statistical NLP layer applies Abstract Meaning Representation Parser (2) Formal reasoning layer uses; answer set programming as knowledge representation. applies inductive programming algorithm for reasoning	100% accuracy on 19 tasks, 93.6% On Basic Induction
[22]	Open	English	MovieTriples, Movie-Dic	Conversation Based	hierarchical recurrent encoder-decoder (HRED)	Perplexity: 26.31 $\pm$ 0.19, Error-rate: 63.91% $\pm$ 0.09

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[44]	Open	English	NY Regents Science Exam (MCQs)	Factoid	They implemented the combination of three techniques (1) IR-based search (2) Support Vector Machine (SVM) based solution with RNN-based embeddings (3) Rule-based answer extraction	71.3% Accuracy (Acc)
[16]	Open	English	InsuranceQA, TREC-QA	Factoid	They used a number of hybrid methods. The best results were produced by Attentive LSTM	(max-pooling K=50) MAP: 0.753, MRR: 0.830
[25]	Open	English	Freebase	Factoid	They used constraints conditional based Logistic Regression, Neural Network (NN) and CNN	CNN produced the best results. Accuracy: 65.19%
[10]	Open	English	CNN and Daily Mail the Children's Book Test (CBT)	Factoid	They used RNN as reader and attention to extract the answer	Single model on CNN; accuracy: 69.5% Single model on CBT; Accuracy: 68.6%
[35]	Open	English	WebQuestions	Factoid	They designed a system named as Knowledge Base Question Answering (KBQA) which extracts the answer from the knowledge graph by incorporating the external text related to question obtained from a web resource say bing.com	Average (Avg) R: 63.5%, Avg P: 50.6%, F1 with avg R and P: 56.3%, Avg F1: 52.2%
[26]	Open	English	WebQuestions, Bing search logs	Factoid	Convolutional Deep Structured Semantic Model (C-DSSM) TabCell and ParaSempre	WebQuestions; P: 77.02, R: 67.65, F1: 69.98 Bing.com P: 61.86, R: 61.86, F1: 61.86"
[1]	Open	English	Question Answering over Linked Data (QALD) Challenge	factoid	Qanary available @ https://github.com/WDAqua/Qanary	P: 0.90, R: 0.78, F1: 0.68

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[28]	Open	English	(1) WebQSP Questions (2) SimpleQuestions	Factoid	Hier-Res-BiLSTM	SimpleQuestions; Accuracy: 93.3% WebQSP; Accuracy: 82.5%"
[6]	Open	English	SQuAD, CuratedTREC, WebQuestions, WikiMovies	Factoid	bigram hashing and Term Frequency-Inverse Document Frequency (TF-IDF) for pages retrieval multilayer RNN for answer detection/extraction	Reader only; EM: 70%, F1: 7%
[45]	Medical	English	Biomedical Semantic Indexing and Question Answering (BioASQ) Challenge	Factoid	They used Systemized Nomenclature of Medicine—Clinical Terms (SNOMED-CT) as knowledge base and multilayer, nested structure of templates to map the question into the semantic relationships of SNOMED-CT and to find the answer from that graph they used ontology-based lexical reference inference	F1: 86.41%
[33]	Open	English	DBpedia Resource Description Framework (RDF) repository, Freebase QALD	Factoid	They proposed a semantic query graph to convert the natural language query into a graph and converting the problem into a sub-graph matching problem. The graph of query can be generated using two different approaches, i.e., edge-first or node-first	Node first framework; on QALD-6; P: 0.89, R: 070, F1: 0.78 On WebQuestions; Avg F1: 49.6"

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[12]	Open	English	NewsQA, TriviaQA, SearchQA	Factoid	Layers in model: word embedding using Glove BiLSTM Cartesian similarity-based attention layer Question dependent passage encoding Multi-factor attentive encoding (Main contribution) to collect the facts from multiple sentences to infer the answer Question-focused Attentional Pointing	On NewsQA; EM: 48.4, F1: 63.3
[46]	Open	English	Complex Sequential QA (CSQA)	Conversation based	They presented a model for CSQA which is a combination of (i) the HRED model for dialogue system and (ii) the key value memory network model for the QA system	R: 15.83%, P: 6.7% F1 score range for different question types: 11.6–40.2% Bilingual Evaluation Understudy Score (BLEU)-4 for clarification type: 15.58"
[18]	Open	English	Quasar-T SQuAD WikiMovies CuratedTREC WebQuestion	Factoid	BM25 and LSTM with 3 layers	Quasar-T; F1: 40.9, EM: 34.2 SQuAD; F1: 37.5, EM: 19.1 wikiMoives; F1:39.9, EM:38.8 CuratedTREC; F1:34.3, EM:28.4 WebQuestion; F1: 24.6, EM: 17.1"
[24]	E-Commerce	English	Data of conversations from the E-Commerce website	Conversation based	Hybrid model combining sentence encoding-based and sentence interaction-based methods	No quantitative measures provided

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[14]	Open	English	Quasar-T SearchQA TriviaQA CuratedTREC WebQuestions	Factoid	(1) Paragraph selector: Multi-layer perceptron (MLP), RNN (2) Paragraph reader: LSTM available @ https://github.com/thunlp/OpenQA	Quasar-T EM: 42.2, F1: 49.3 SearchQA EM: 58.8, F1: 64.5 TriviaQA EM: 48.7, F1: 56.3 CuratedTREC EM: 29.1 WebQuestion EM: 18.5, F1: 25.6”
[47]	Multi	English, Hindi	MMQA (Multi-domain, Multi-lingual dataset)	Factoid	(1) Question Classification based on CNN and RNN (2) Retrieval of passage using Boolean and BM25 vector space model (3) Candidate Answer selection (4) Answer scoring and ranking	Factoid; EM: 30.24, MRR: 49.10 Descriptive; BLEU: 41.37, ROUGE-L:40.19”
[15]	Open	English	SQuAD, Japanese Dataset	Factoid	They implemented Bi-directional Attention Flow (BiDAF) for Reading Comprehension with Stanford Question Answering Dataset (SQuAD) but retrieved the passage from Wikipedia pages and then applied the RC to find the answers instead of the passage available with SQuAD which surely contains the answer. And for retrieval, they used multi- and single-task learning	Reading Comprehension (RC): On SQuAD; EM: 35.6, F1: 42.6 Information Retrieval part: (multi-task learning) MTL; S@1: 81.1, M@5: 86.3
[11]	Open	English	SQuAD, Curated TREC, WikiMovies, WebQuestions	Factoid	They extended DrQA to adapt the retrieved number of documents according to query and size of corpus in retrieval module	different settings for threshold, produced better results on different datasets

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[48]	Medical	English	BioASQ	Factoid	They used Named Entity Recognition (NER) based method to answer the factoid question and lexical chaining technique for extractive summarization	The study also discussed the lack of effective evaluation techniques. They achieved highest Recall-Oriented Understudy for Gisting Evaluation (ROGUE) and ROGUE-SU4 for all the test batches of the challenge
[27]	Open	Chinese		Non-Factoid	seq2seq based generative model combining raw text and knowledge base (1) Candidate facts retriever (2) Question Encoder (3) Reply Decoder (4) Response words predictor classifier (5) Universal Schema (5) Knowledge enquirer (6) State Update	BLEU: model (text) 0.45 human evaluation- model (model: text+KB); Fluency: 4.12, correctness: 4.42, Grammar: 4.19
[29]	Open	English	Web Questions, Simple Questions	Factoid	They created a model which trains offline to map the sentences to a knowledge base and then querying that knowledge base to get the answer. If the user query does not match to already learned query templates, then it incorporates a non-expert user's feedback and learns the query representation for that question	WebQuestions; Avg P: 40.6 Avg R: 49.5 Avg F1: 40.8
[13]	Open	English	Curated TREC, SQuAD-open	Factoid	SPARC available @ https://github.com/jhyuklee/sparc. (They used and fine-tuned Bidirectional Encoder Representations from Transformers (BERT)-Large for their encoders)	DENSPI + Sparco on C.TREC: EM: 35.7 on SUQAD; EM: 40.7, F1: 49.0
[20]	Open	English	SQuAD	Factoid	BERTserini: Finetuned BERT with Anserini information retrieval toolkit available @ http://anserini.io/	EM: 38.6, F1: 46.1, R: 85.8

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[8]	Open	English	SQuAD, Quasar-T	Factoid	Transformer for input encoding, universal transformer for multihop reasoning	SearchQA: EM:52.9, F1:65.1 Quasar-T: EM:43.2, F1:54.0
[49]	Open	English	Natural Questions, Web Questions, CuratedTREC, TriviaQA, SQuAD	Factoid	Open retrieval question answering system (ORQA) which establishes that it is possible to retrieve and read jointly from question answer pair without separate IR system	EM on different datasets: Natural Questions: 33.3 WebQuestions: 36.4 CuratedTREC: 30.1 TriviaQA: 47.1 SQuAD: 33.2 Not the best for last two
[34]	Open	English	Complex questions from WikiAnswers (CQ-W), Complex questions from Trends (CQ-T)	Factoid	QUEST (for "QUEstion answering with Steiner Trees") QUEST dynamically creates an adhoc knowledge graph of the documents related to question, retrieved from the documents and they name that resulting graph Quasi Knowledge graph. The method is completely unsupervised	For CQ-W; MRR: 0.355 P@1: 0.268 Hit@5: 0.376 For CQ-T; MRR: 0.467 P@1: 0.394 Hit@5: 0.53
[50]	Open	English			They provided a novel risk control framework to evaluate the risk involved. and drawbacks of evaluating any Machine Reading Comprehension (MRC) by F1 and EM	Improvement in BiDAF: 10.8% using PROBE-CNN Improvement in BERT: 14.16% using PROBE-CNN
[19]	Open	English	SQuAD WikiMovies SearchQA Web Questions	Factoid	CNN based document selection LSTM based answer generation	SQuAD: EM:33.9, F1:43.6 SearchQA: EM:61.4, F1:65.7 WebQuestions: EM:22.6, F1:27.3 WikiMovies: EM:36.2, F1:42.5

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[23]	Open and Medical	English	TREC, MedQuaD https://github.com/abachaa/MedQuAD	Non-Factoid	They curated the Medical Question Answering Dataset (MedQuAD) (47,457 pairs of Question Answers). Hybrid method (Combining Logistic regression and Information Retrieval (IR) methods)	Accuracy: 80.57 Precision: 70.29 Recall: 72.10 F1: 71.19 MAP: 77.47 MRR: 83.79"
[7]	Open	English	Trivia-QA unfiltered, SearchQA, SQuAD-open, QUASAR-T	Factoid	They introduced a framework in which retriever and reader interact with each other, in iterative manner.	Quasar-T: EM:40.63, F1:46.97 SearchQA: EM:56.26, F1:61.36 TriviaQA-unfiltered: EM:61.56, F1:68.03 SQuAD-open: EM:31.93, F1:39.22
[51]	Open	English	20 Newsgroup	Factoid	(1) Question Preprocessing a. Question type through "Tagger based Question Pattern Analysis" (T-QPA) (2) grouping of documents for retrieval based on DKS-KBC algorithm (3) Answer generation through Semantic Word based Answer Generation (SWAG)	No comparison with other systems
[52]	Open	English	Telecom company FAQs, retail FAQs	Conversation Based	(1) Extractive Summarization (2) Set of paraphrasing techniques to generate paraphrased title and selective sentences (3) Candidate selection algorithm to select the best paraphrases. At each step they used multiple algorithms and compared their results	Overall, their approach work well. Results are expanded on multiple stages with multiple algorithms making it unable to summarize effectively

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[53]	Open	English	NaturalQuestions (NQ), TriviaQA, WebQuestion, CuratedTREC, SQuADv1.1	Factoid	The study used English Wikipedia dumps as knowledge source to answer. FAISS library for searching similarity and clustering. BERT for passage and question encoding	NQ: Acc:79.4, EM:41.5 TriviaQA: Acc:79.9,EM:57.9 WebQuestion: Acc:75.0, EM:42.4 TREC: Acc:89.1, EM:50.6 SQuAD: Acc:66.2, EM:35.8
[54]	Medical (Covid)	English	Covid-19 dataset available @ https://covid-19-infobot.org/data/	FAQ based	Poly encoder trained from scratch on data extracted from subreddit	JHU-COVID-QA@20 Acc:79.4,F1: 83.1, BLEU:79.4, MRR:87.5 JHU-COVID-QA@10 Acc:98.5,F1:98.9, BLEU-4:98.5,MRR:99.2"
[32]	Open	English, German, French, Italian, Spanish, Portuguese, Arabic, Chinese		Factoid	Their system is based upon following steps: query expansion, Retrieval, Answer selection, Feedback	Cocktails Dataset; F1:0.37 HR; F1:0.52 EU; F1:0.70"
[55]	Open	Arabic	Arabic Question Answering (AQA)-WebCorp	Non-Factoid	They convert the Arabic text into conceptual graphs and then transform it into the logical representation to extract the answer	Accuracy: 74% (test set just contained 250 questions)
[56]	Medical	English	Covid Dataset	Factoid	Web Understanding and Learning with AI-Question Answering (WULAI-QA): (1) Feature Engineering using NLP tools, i.e., TF-IDF, Linear Discriminant Analysis (LDA), Word Mover's Distance (WMD) (2) Reteriver using BERT (3) Reader using BERT (4) User feedback using entropy loss	Rouge-L : 0.74

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[56]	Open	English	Simple Questions	Factoid	They used the Bayesian Neural Network (BNN) to estimate the uncertainties of the model's predictions. They integrated the entity and relation detection and employed the uncertainty estimation through BNN (Bayesian LSTM) in an end-to-end model	75.1% accuracy
[57]	Open	English	Complex Question Answering (CQA)	Non-Factoid	They implemented the idea that if phrases are introduced in the information network, they can produce better results. They introduced the algorithm to create the graph for the phrases and the selected the best answer amongst the already available one. They did not gave much information about the case where no answer exists, already	On stack-overflow: Precision@K: 0.65 Recall@ K:0.51 Mean Absolute Error (MAE)@K: 0.76 RMSE@K:0.8546
[17]	Open	English	SQuAD	Factoid	Gated self-matching networks consisting of 4 layers: (1) Gated Recurrent Unit (GRU) layer for Question and Passage (2) Gated-attention-based recurrent network to match passage and question (3) Passage self-matching layer using self-matching attention mechanism (4) Pointer networks to locate the answer in the passage in output layer"	Exact Match (EM): 71.3%, with ensemble EM: 75.9%
[58]	Open	English	WikiPassage MS Marco WikiQA	Factoid	They proposed a self-matching attention pooling mechanism to highlight the important terms in question-passage pair to lessen the effect on passage ranking of irrelevant matching terms in the passage	WikiPassageQA: MAP: 0.769 MRR: 0.838 On the other two datasets, their results are not the best

Table 4 continued

Paper	Domain	Language	Dataset	Type	Methodology	Result(s)
[36]	Open	English	GrailQA	Factoid	They released a new dataset with 64,331 questions. They suggested three levels of generalization for KBQA models, i.e., iid, compositional and zero-shot and also provided the evaluation of generalization of all three levels	EM: 50.6 F1: 58.0
[59]	Open	English	ConvQuestions Largescale Complex Question Answering Dataset (LC-QuAD)	Factoid	Contracting answer spaces with scored Lists and top-k Operators for Complex QA (CLOCQ): Their main contribution is that they presented an algorithm to reduce the search space for the QAS. They used top k query processor for the relevant results returned. To rank the final answer they used lexical matching, similarity between question and candidate answer, coherence and connectivity in the knowledge graph.	Answer was present in the final search space by CLOCQ: LC-QuAD: 82.6% ConvQuestions: 84.7%
[60]	Medical	English Chinese Japanese	Curated the Dataset with 164125 QA pairs	Non-Factoid	They used the translation module at both question and answer ends. A question in language L can be answered in language L and translated in language L‘. The same question could be translated in L‘ and then answered in language L‘	With parallel data: BLEU-3: 35.5 Without parallel data: BLEU-3: 35.3

Gu et al. [36] used Zero-Shot learning with Knowledge Base in continuation of this trend. Some pre-processing steps were common in all because every system had to deal with natural language.

Every solution area has its pros and cons. Knowledge graph-based solutions are found to be more efficient for closed domain QA systems because of the ontologies that exist to represent the knowledge/information of that domain. But, again, if any domain specific ontology does not exist then one has to work on that.

Any machine or deep learning solution needs large, labeled data to train and produce promising results. Considerable number of resources, which are both general and domain specific, are produced for that. But these resources have some defined constraints or somehow a controlled environment. For example, most of these resources represent factoid type questions where the answer is one or some words or a sentence span which is assumed to have the answer, see Fig. 2. They have very limited text (context) to search the answer from. Existing systems, mostly based on deep learning0 are using high computation power and memory to produce those answers. Though the knowledge graph system relies on inference and can answer factoid data well, it has problems while inference on multi-level relations in the graph [37]. A system trained to produce the knowledge graph of any text and then running structured queries on it with inference extended to multiple levels and entities could be a solution.

RQ2: Does the results of the method get affected by the type of domain they are applied to, i.e., open or closed domain?

Vinyals and Le [39] designed a seq-to-seq model which they tested on open and closed domain. For open domain they applied it to a movie's subtitles dataset and in closed domain they applied it to a dataset of IT helpdesk which used to solve the problems that users faced. One thing was common in both domains and that was conversation and some common inference. According to the results that they reported, their model worked quite well. It had better perplexity as compared to n-grams models though there was a lack of consistency. But they did not report any difference of performance on open or closed domain. McElvain et al. [61] applied an open domain system to the legal domain. The system was trained and judged by the domain expert attorney editors. Though they applied the question classification to map them according to legal taxonomy. Their answers got 52% clickthrough rate to see the full case against the questions of the attorneys.

Certain domains are rich in resources to train on large, labeled data. For the domains which do not have sufficient resources such as the E-commerce domain, Yu et al. [24] applied transfer learning to paraphrase identification and natural language inference and proposed a general framework to efficiently adapt the shared knowledge between domains.

Where some have reported no difference in results, Zhang et al. [62] has pointed out that in some cases general information is not enough and efficient. As they studied the e-commerce domain and discovered that some reviews and questions would be about certain products from certain producers, we cannot generalize them.

Domains like these or some others where the information provided to users should be inferred from a particular piece of text and rules, need domain specific systems. If a method is trained on an open domain, there is a big chance that it will miss a certain pattern in the closed domain and produce inefficient results w.r.t that domain.

RQ3: Which datasets have been used in studies from 2016 to 2022?

After studying sixty papers, a list of all the datasets was created which were used between 2016 to 2022 for the question answering system. Their names, statistics and where they are available are all summarized in Table 5. They are divided into two main categories; (1) open domain, (2) closed domain. A large section belongs to the open domain, almost 80% to be

Table 5 Datasets used in studies which are included in this article from 2016–2022

Sr.	Dataset	Year	Description
1	Web Questions	2013	WEBQUESTIONS, a dataset collected by [66]. It contains 5810 questions crawled from the Google Suggest service, with answers annotated on Amazon Mechanical Turk. All questions contain at least one answer from Freebase. It has factoid questions
2	IT Helpdesk Troubleshooting dataset	2015	In this service, costumers face computer related issues, and a specialist help them by conversing and walking through a solution. Typical interactions (or threads) are 400 words long, and turn taking is clearly signaled. The training set contains 30M tokens, and 3M tokens were used as validation. Some amount of clean up was performed, such as removing common names, numbers, and full URLs. (Not available publicly)
3	OpenSubtitles dataset	2018	OpenSubtitles is collection of multilingual parallel corpora. The dataset is compiled from a large database of movie and TV subtitles and includes a total of 1689 bitexts spanning 2.6 billion sentences across 60 languages. Available at https://opus.nlpl.eu/OpenSubtitles2018.php
4	Movie-Triples Corpus (MTC)	2012	The MovieTriples dataset has been developed by expanding and preprocessing the Movie-DiC dataset by [67] to make it fit the generative dialogue modeling framework. The dataset is available upon request
5	NY Regents 14th Grade Science exams	2016	Elementary Science Corpus: 80k sentences about elementary science, consisting of a Regents study guide, CK12 textbooks, 3 and automatically collected Web sentences of similar style and content to that material. 2. Web Corpus: 5 × 10 10 tokens (280 GB of plain text) extracted from Web pages. Available at www.allenai.org
6	InsuranceQA	2017	git clone https://github.com/shuzi/insuranceQA.git this dataset contains question answer from the website https://www.insurancelibrary.com/. It has questions from real users and answers are collected from experts of the field. It contains 16,889 Questions, 27987 Answers
7	TREC	1992 and onward	TREC stands for The Text Retrieval Conference. TREC question classification dataset contains 5500 labeled questions for training and 500 more for testing purpose. It is available at https://cogcomp.seas.upenn.edu/Data/QA/QC/, https://trec.nist.gov/data.html
8	CNN and Daily Mail	2016	It is a dataset for summarization task. It has about 300K articles by journalists in CNN and DailyMail. Which fall into categories of both extractive and abstractive summaries. Available at https://github.com/deepmind/rc-data

Table 5 continued

Sr.	Dataset	Year	Description
9	Children's Book Test (CBT)	2016	http://www.thespermwhale.com/jaseweston/babi/CBTest.tgz It is created form the children's books that are freely available. It has a context, a question and multiple answers to choose from. 20 sentences of ay part of the book are considered the context. The very next 21st sentence is then edited by removing a word that turns it into the query to which answer should be generated
10	MovieQA	2016	Available at http://movieqa.cs.toronto.edu This dataset is mainly for comprehension tasks and is collected based upon both text and videos. It is collected from 408 movies and consists of 14,944 Q/A pairs
11	TimeML/ TempEval-3		TimeML is a corpus annotated with: (a) time expressions; (b) events and (c) links between them. The TempEval-3 Platinum TimeML annotations consists of twenty English newswire documents, each annotated for events, temporal expressions and temporal relations by multiple experts and an adjudicator. This is the corpus used to rank participant systems in the TempEval-3 evaluation exercise. Annotations are in TimeML-strict, a subset of TimeML. Available at https://aclweb.org/aclwiki/TempEval-3_Platinum_TimeML_annotations_(Repository)
12	SQuAD 1.1	2016	It contains 100K+ question answer pairs from more than 500 articles. Format is to have a contextual paragraph which contains the answer to the question. Available at https://rajpurkar.github.io/SQuAD-explorer/
13	SQuAD 2.0	2018	SQuAD 2.0 adds 50K Questions to the SQuAD 1.1 which are unanswerable but looks very similar to the ones present in the dataset. Available at https://rajpurkar.github.io/SQuAD-explorer/
14	MS-MARCO	2016	This dataset is released by Microsoft in 2016. It has about 1M question answer pairs. Dataset is used for question answering, natural language generation, ranking and extraction
15	WikiMovies	2017	Contains 100K questions from movies available at https://metatext.io/datasets/the-wikimovies-dataset
16	CuratedTREC		It is factoid type dataset created by combining irc/ and trec/ available at https://github.com/brmson/dataset-factoid-curated
17	BioASQ	2015	BioASQ is an entity which organizes competition/challenges for question answering and indexing of medical domain resources, available at http://bioasq.org/
18	bAbi	2015	Facebook has started an initiative regarding machine comprehension tasks which also includes question answering, dataset available at https://github.com/facebook/bAbI-tasks

Table 5 continued

Sr.	Dataset	Year	Description
19	NewsQA	2016	This dataset is collected by crowdworkers from 1000 articles from CNN. It contains 100K question answer sets manually written by the crowdworkers. Available at https://www.microsoft.com/en-us/research/project/newsqa-dataset/
20	TriviaQA	2017	TriviaQA consists of question–answer pairs authored by trivia enthusiasts and independently gathered evidence documents from Wikipedia as well as Bing Web search. It contains 650K question answer pairs, available at https://nlp.cs.washington.edu/triviaqa/
21	SearchQA	2017	SearchQA is a dataset which is created using a question answer pair with snippet from google. It also includes the metadata of that snippet, i.e., the URL of that google snippet which is being used as context. It has 140K question answer pairs. It is available at https://github.com/nyu-dl/dl4ir-searchQA
22	Quasar-T/S	2017	Quasar-T and Quasar-S are two datasets which consist of open domain trivia questions and answer pairs. Quasar-T consists of 43K Q/A pairs. Both are available at http://curtis.ml.cmu.edu/datasets/quasar/
23	MMQA	2018	MMQA is a multilingual, multidomain question answering dataset. It has data in English, Hindi language and from multiple domains, i.e., History, Tourism, Geography, Environment, Disease and economics. It is available at http://www.iitp.ac.in/~ai-nlp-ml/resources.html
24	CoQA	2018	Conversational Question Answering Challenge. It consists of 127k conversation turns collected from 8k conversations over text passages. The average conversation length is 15 turns, and each turn consists of a question and an answer. It contains free-form answers, and each answer has a span-based rationale highlighted in the passage. Its text passages are collected from seven diverse domains: five are used for in-domain evaluation and two are used for out of domain evaluation. It is available at https://stanfordnlp.github.io/coqa/
25	CSQA	2018	This dataset contains complex questions which need to refer to more than one tuple to answer the question. Same is the case with dialogues which need the reference to previous/current conversations as reference to answer. It is available at https://amritasaha1812.github.io/CSQA/download/
26	CQA	2019	This dataset consists of 22M questions about 113K images released by researchers from Stanford to be used for visual question answering. Available at https://cs.stanford.edu/people/dorarad/gqa/download.html

Table 5 continued

Sr.	Dataset	Year	Description
27	SQA	2016	SQA (sequential question answering) was created to answer sequence of questions. It has 17,553 questions with more than 6000 sequences. Available at https://www.microsoft.com/en-us/download/details.aspx?id=54253
28	NarrativeQA	2017	Created by deepmind for the machine comprehension purpose, it has summaries from Wikipedia pages, full article and question and answer pairs related to those articles. Available at https://github.com/deepmind/narrativeqa
29	HotpotQA	2018	This dataset contains questions which need more than one Wikipedia document to answer any question. It also introduced comparison-based factoid questions. It consists of 113K question answering pairs based in Wikipedia. Available at https://hotpotqa.github.io/
30	SimpleQuestion	2015	Known as SimpleQA consists of more than 108K instances in the form of Subject, predicate and object. It has question answer pairs with some facts to support the answer. Available at https://github.com/davidgolub/SimpleQA/tree/master/datasets/SimpleQuestions
31	Natural Questions	2019	It contains question and answer pairs. The challenge is to find the answer to a question from a Wikipedia document which may or may not have it. It has challenge for both factoid and long answer. Available at https://ai.google.com/research/NaturalQuestions/dataset
32	WikiPassageQA	2018	Crowd workers were asked to create non-factoid questions based on a Wikipedia article, and indicate the location of their respective answer passages within the document
33	ANTIQUE	2019	It consists of more than 2.6K non-factoid question answer pairs which were crowdsourced by real users from the communities like yahoo answers etc. Available at https://ir-datasets.com/antique.html
34	SogouRC	2018	SogouRC is a large-scale Web QA dataset released by the Chinese commercial search engine Sogou. It includes 30000 queries selected from search logs that can be satisfied by short answers, and the corresponding top ranked passages from the search result
35	MedQuAD	2019	It consists of more than 46K question answer pairs, related to medical domain, annotated in a way that it can be used for several NLP tasks including Question Answering. Available at https://github.com/abachaa/MedQuAD
36	SNLI	2014	It consists of 570K open-domain sentences used for inference. Dataset contains a sentence, hypothesis and label for each data point. Available at https://nlp.stanford.edu/projects/snli/

Table 5 continued

Sr.	Dataset	Year	Description
37	MultiNLI	2018	It is on the same format as SNLI, but it has more diverse range of sentences than SNLI. Available at https://cims.nyu.edu/~sbowman/multinli/
38	SemEval-cQA	2017	An annotated dataset for answer selection from community question answer where a single question has hundreds of answers and users find it overwhelming to get to the accurate one. Available at https://alt.qcri.org/semeval2016/task3/
39	STS Benchmark	2018	Semantic text Similarity dataset hub. Available at https://github.com/brmson/dataset-sts
40	MRPC corpus	2005	This corpus consists of a single document which contains 5800 pairs of sentences that are extracted from different news articles. Only one sentence is extracted from one article. Corpus is available at https://deepai.org/dataset/mrpc
41	Event-QA	2020	First dataset which covers the events rather entities using knowledge graph. It covers almost 970K events. They used EventKG to answer the questions related to events. Available at http://eventcqa.l3s.uni-hannover.de/
42	Covid-19 dataset	2020	Available at https://covid-19-infobot.org/data/. Not released yet [10th May 2021] It has queries/information from some official sources, e.g., World Health Organization (WHO) and some unofficial and informal sources, i.e., twitter to collect the queries from public
43	ComQA	2021	Available at https://github.com/benywon/ComQA/tree/main/data. It has 120K human labeled Q/A pairs where an answer is
44	wikiQA	2016	Available at: https://huggingface.co/datasets/wiki_qa. This dataset consists of 3047 questions. Each question has multiple sentences as candidate answers and only one sentence marked as true, at most. The candidate sentences were selected from the summary section of the Wikipedia pages assuming that they contain the most valuable information
45	GrailQA	2021	Available at: https://dl.orangedox.com/WyaCpL/ GrailQA consists of 64,331 annotated QA pairs for question answering on knowledge bases using freebase. They are logically mapped on different syntax, e.g., SPARQL, S-expression, etc.
46	ConvQuestions	2019	Available at: https://convex.mpi-inf.mpg.de/ ConvQuestions was presented in 2019 along with the model called CONVEX. It was created for knowledge bases and it consists of 21,000 conversations which can be evaluated on the wikiQA

precise. It is mainly because open domain QAS are considered more challenging and to cover other domains within itself. But that does not deny the fact that some domains are so delicate that they cannot be left to be managed within open domain systems, e.g., medical domain. Question answering solutions for these domains need domain specific resources to be efficient.

There are more categories in these datasets. Some of these are based on Knowledge graphs, i.e., FreeBase, DBpedia etc. while others have free text sources. Visual data-based datasets also exist for visual question answering but that is out of the scope of this SLR [63, 64]. Some of the existing datasets are context based. Which means they have context in their annotated data from which they extract the answer. Though most of the datasets targeted the factoid question answers or where answer span consisted of few words or MCQ's type. Facebook collected a dataset known as ELI5; collected from a reddit thread (Explain Like I am Five). It consists of both extractive and abstractive answers in non-factoid type targeting seq length more than 512 [65].

RQ4: Does literature provide any guideline to produce the answer from different sources?

The literature has reported many studies where the system had to extract or generate the answer form a contextual paragraph. The trend was followed specifically after the introduction of SQuAD by Stanford [68] There are some studies in which documents related to a particular query were retrieved from hundreds and thousands of documents. The general pattern is to retrieve related paragraphs from related documents. Extract the sentences with the highest scores and present them as answer. Cesario et al. [69] presented an incremental approach to eliminate the duplicate entries through clustering. This helps in accommodating multiple facts and aspects in answer generation rather focusing on the one redundant aspect. Other than that, there is no study reported where an answer was generated by consulting multiple sources and inferring and combining the facts provided in those sources.

RQ5: What evaluation metrics are used with respect to Q/A type, i.e., factoid and non-factoid?

Following evaluation metrics are reported in the literature for both factoid and non-factoid type question answering systems.

There are some techniques from Table 6 which favor correct answers over no or wrong answers, for example, accuracy; it only considers the correct number of answers.

$$\text{Accuracy} = \frac{\text{Number of questions correctly Answered}}{\text{Total number of questions}} \tag{1}$$

On the other hand, MRR as described in table above is used when multiple answers are generated against a single question. So, they are ranked according to their position on the answer list. Though it incentivizes the system for returning the right answer in upper ranks, it does not differentiate between wrong answers only or no answer.

All these automatic evaluation techniques are quantitative measures. None of these checks whether the text, which is generated by the system, mostly in non-factoid systems, makes sense or not. Is it grammatically correct? Does it cover all aspects to answer the question? Only one study measured these aspects, but the evaluator was human.

An open question whether systems should be rewarded for not any answer against a wrong one? According to [77], yes. This will improve the confidence level of the system.

Distribution of types of question answering systems is visualized in a chart in Fig. 2. Most of the systems are factoid based which include single- or multi-word answers or span of sentence which has the answer to the question. And most of the evaluation metrics are designed for the factoid type QAS to quantify the response, whereas the world is moving

Table 6 Evaluation metrics with their descriptions

Evaluation metric	Description
Precision	Precision is calculated as percentage of shared words between actual and system provided answer [70]
Recall	Recall is the percentage of possible correct words for the answer according to the ground truth [70]
F1	F1 is the harmonic mean of Precision and Recall. Precision is biased toward words that should/should not be in the generated texts, i.e., true positives and true negatives. While recall is a quantitative measure of how many words that should have been in the generated text, are there. F-measure balances out that biasness [70]
Mean Average Precision (MAP)	Average precision is the average of precision at each recall level (all those points at which a relevant word is retrieved). And mean average precision is the average of AP over different queries [71]
Mean reciprocal rank (MRR)	Reciprocal of the rank at which the first relevant word is retrieved is called reciprocal rank (RR). Average of all the RRs over multiple queries is called MRR [72]
Accuracy	It is defined as ratio of total number of words which are correct to the total number of words retrieved [73]
Exact Match (EM)	Exact match has a binary value. If the ground truth and model prediction match exactly excluding the punctuation marks or articles, etc., then its value is 1; otherwise, it is 0 [74]
Perplexity	In language models, perplexity is interpreted as a branching factor, e.g., if there are 50 words from which the model has to pick the next word then the perplexity is 50. More formally it can be described relating to cross-entropy. Cross-entropy defines the N bits that will make a sentence and perplexity is the number of words that can fill those bits [75]
Word Error-Rate	It is described as the total number of words that are wrong. Mathematically it is the ratio between sum of substituted, inserted and deleted words to the complete number of words. It is mostly related to speech recognition and transcription [75]
BLEU	Bilingual evaluation understudy, as the name suggests, is often used for evaluating the translation. But it can also be used to evaluate other generated texts such as answers, stories etc. It is used very often mainly because it correlates with human evaluation. In general, it is referred to as BLEU-N where n represents n-gram. Uni, bi or multi-grams, i.e., words are compared in reference and generated texts [76]
Fluency, Correctness, Grammar	Fluency, correctness and grammar, all three evaluations were done by human beings rating between 1–5. Only one study used these measures [27]

toward non-factoid QAS. As the generated answers will be from 'summary', 'detailed' and 'list', etc., categories, the evaluation techniques also need to be updated to match that and to progress in the field. Though we have BLEU, ROGUE, Perplexity and METEOR, etc., they are also n-grams-based techniques that failed to evaluate correctness, scope and fluency of the answer.

RQ6: What efforts have been made to apply QAS on the Quran, Tafseer and Ahadith domain?

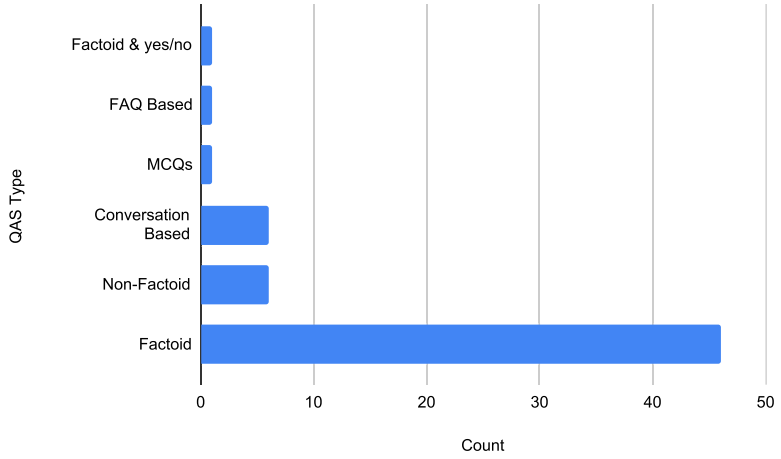

Fig. 2 Q/A type distribution

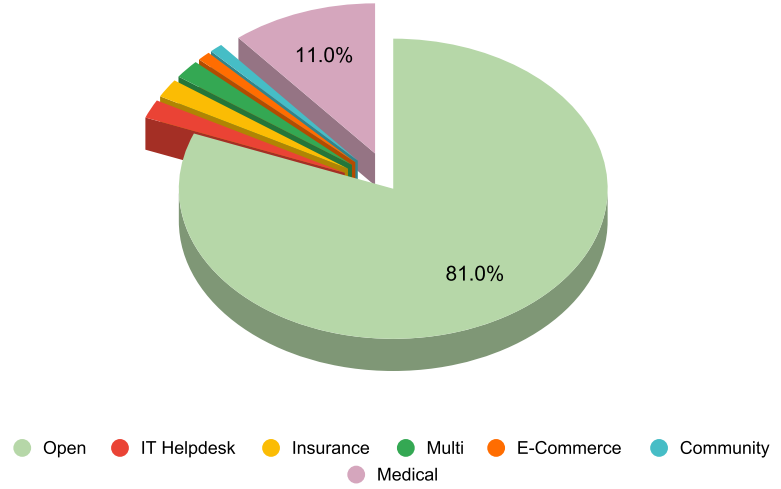

Fig. 3 Domain distribution for question answering systems

Quran is the holy book of Muslims which was revealed upon Holy Prophet Muhammad (S.A.W) to guide Muslims through this journey called life. Ahadith are the words, actions or silence of Holy Prophet Muhammad (S.A.W) on any matter. Quran and Ahadith are primary sources for Islamic legislation. Tafseer is the explanation of the Quranic verses by different scholars. It helps Muslims understand the Holy book [78].

Several domains were targeted in the time range of 2016–2022 for automatic question answering problems. Majority of the studies were performed on an open domain. Second most used domain was Medical and then some other domains were targeted very infrequently (see Fig. 3). Abdi et al. [79] worked on Ahadith in Arabic language. This study proposed a framework to answer the Arabic questions from Ahadith corpus. To extract the relevant Ahadith to user's query they retrieved semantically and syntactically similar Ahadith to user's query and similar Ahadith to the returned Ahadith. They name their method "Question Answering System in Al-Hadith using Linguistic Knowledge" (ASHLK).

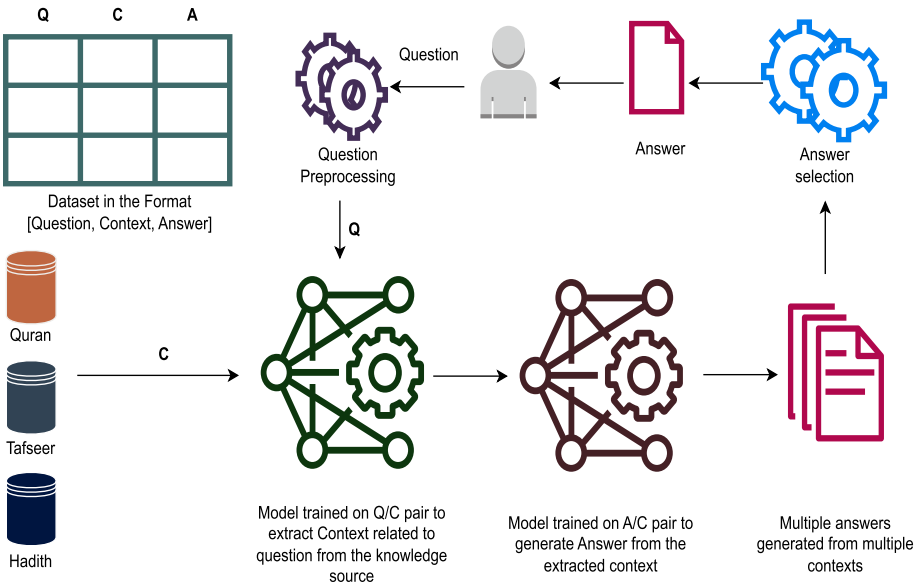

Fig. 4 Proposed integration of components for a QAS system for the Quran, Tafseer and Ahadith domain

It was excluded from the list because of the CORE ranking clause in inclusion/exclusion criteria (see Table 2). Two main reasons for lesser contributions in this area are (1) lack of publicly available resources and (2) unavailability of work in well reputed venues. No study which worked on the said domain, which has great challenges in it, passed the criteria. For example, it is a very interesting challenge to extract or generate the answers from multiple sources. It will involve the subchallenges of removing the redundancies, ranking and prioritizing different sources. Another interesting challenge is to link the entities on multiple levels weighing the source and exploiting the references in text. All these tasks make this domain interesting to explore and to experiment on.

4 Discussion

This study was conducted in a systematic manner to answer some research questions related to question answering systems in natural language processing. Detailed analysis and answers to those research questions pointed toward some research gap in the literature. Statistical and machine learning-based techniques were used in the past to solve the question answering problem and then the process kicked off with the addition of deep learning. Deep learning-based models were able to beat human accuracy level in factoid question answering for some datasets, e.g., SQuAD. Detailed results are mentioned in Table 4. But same models were not proven much efficient when it came to non-factoid type. The inefficiency was caused by multiple reasons, e.g., longer contextual information also known as max sequence length. Moreover, deep learning models highly depend upon the amount of data they are trained on. And there were not many datasets for the non-factoid question answering systems until recently. After 2020, more studies focused on non-factoid systems and with the introduction of longformer and BART the max sequence length also increased.

This study helped us identify some gaps that we have discussed earlier. Hence, considering the absence of a specific question answering system (QAS) targeting multiple answer streams of the Quran, Tafseer and Ahadith, one can apply these progressing models to achieve good results. A pipeline is proposed for this problem presented in Fig. 4. There are many sources available online where user ask question about any matter that concern them. And scholars/expert of these scripture answer those questions referring to the original text as facts. A dataset could be collected from these sources comprising of the question, answer and context. Context is the text(s) that has the answer to user's question. Large enough dataset would be sufficient to train the deep learning models on it.

It has text of the Quran, Tafseer and Ahadith in English as data source from which it will search for the answer to user's questions. First module of the system would train on the pair of question and context to learn the connection and which part of the texts from our data source can have the answer to the question. Next module will train on the answer and context pair to learn, how to extract the answer if the system has the context (text that contains the information to answer the question). The system will extract multiple ranked answers and the top answer will be returned to the user. It is a generic framework in which we can later add different sub tasks and stages to improve it, e.g., different ranking of the text based on if it belongs to the Quran or Ahadith. Different categories of Ahadith such as Sahih, Hasan, Daif and Mawdu will also have different ranking.

Another aspect that we considered in this study is evaluation metrics. When we come to the evaluation of these models the metrics are based on the mere comparison of the words present or not present in the answer. They do not take the grammar, cohesion, and conciseness of the answer into the account. Some of these points could be addressed through our proposed model.

5 Conclusion

We conducted a systematic literature review in this paper which covered the literature in question answering from different perspectives. This review is mainly for the beginners in the field of natural language processing or for those who are just starting with the question answering system. Others could also benefit from the different information structured in this article. NLP got its boost with the deep learning models and from then there is no going back. Deep learning is producing the most promising results in the different tasks of natural language processing (NLP). Question answering is also one of them. Solutions for both open and closed domain systems have been proposed and a vast variety of the datasets has been created to take this research forward. Models which were producing results close to human performance for factoid type questions are not working that well for long form answers. One reason is the inability to store the context for the longer texts. Transformers can save the relatively longer context, but it also has the limit of 512 tokens. We also have reformers and longformers which can save longer contexts than transformers by using less memory, i.e., 16GB to be exact. Another promising research aspect is to generate answers by inferring from multiple context paragraphs. For that we also proposed a the Quran, Tafseer and Ahadith domain along with a proposed framework. Furthermore, the evaluation measures that are used for non-factoid answers are not sufficient and efficient enough. We need automatic evaluation

techniques which can evaluate the fluency, grammar and correctness of the answer generated or extracted.

Author Contributions FQ wrote the main manuscript and prepared all the figures, SL contributed to the formalization of research questions and reviewed the paper. AAS reviewed the paper.

Data availability All data generated or analyzed during this study are included in this article (see Table 5).

Declarations

Competing interests The authors declare no competing interests.

References

1. Diefenbach D, Singh K, Both A, Cherix D, Lange C, Auer S (2017) The Qanary ecosystem: getting new insights by composing question answering pipelines. In: International conference in web engineering (ICWE), pp 171–189. https://doi.org/10.1007/978-3-319-60131-1
2. Shah AA, Ravana SD, Hamid S, Ismail MA (2019) Accuracy evaluation of methods and techniques in Web-based question answering systems: a survey. Knowl Inf Syst 58(3):611–650. https://doi.org/10.1007/s10115-018-1203-0
3. Kettani H (2010) 2010 World Muslim population. In: Proceedings of the 8th Hawaii international conference on arts and humanities, pp 1–61
4. Kitchenham B, Pearl Brereton O, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering: a systematic literature review. Inf Softw Technol 51(1):7–15. https://doi.org/10.1016/j.infsof.2008.09.009
5. Computing Research and Education (CORE) rankings portal. https://www.core.edu.au/conference-portal
6. Chen D, Fisch A, Weston J, Bordes A (2017) Reading Wikipedia to answer open-domain questions. In: ACL 2017—55th annual meeting of the association for computational linguistics, proceedings of the conference (long papers), vol 1, pp 1870–1879. https://doi.org/10.18653/v1/P17-1171
7. Das R, Dhuliawala S, Zaheer M, McCallum A (2019) Multi-step retriever-reader interaction for scalable open-domain question answering. In: 7th international conference on learning representations, ICLR 2019, pp 1–13
8. Dehghani M, Azarbonyad H, Kamps J, De Rijke M (2019) Learning to transform, combine, and reason in open domain question answering. In: WSDM 2019—Proceedings of the 12th ACM international conference on web search and data mining, vol 2491, pp 681–689. https://doi.org/10.1145/3289600.3291012
9. Gupta D, Ekbal A, Bhattacharyya P (2019) A deep neural network framework for English Hindi question answering. In: ACM transactions on Asian and low-resource language information processing, vol 19. https://doi.org/10.1145/3359988
10. Kadlec R, Schmid M, Bajgar O, Kleindienst J (2016) Text understanding with the attention sum reader network. In: 54th Annual meeting of the association for computational linguistics, ACL 2016—long papers, vol 2, pp 908–918. https://doi.org/10.18653/v1/p16-1086
11. Kratzwald B, Feuerriegel S (2020) Adaptive document retrieval for deep question answering. In: Proceedings of the 2018 conference on empirical methods in natural language processing, EMNLP 2018, pp 576–581. https://doi.org/10.18653/v1/d18-1055
12. Kundu S, Ng HT (2018) A question-focused multi-factor attention network for question answering. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 5828–5835
13. Lee J, Seo M, Hajishirzi H, Kang J (2020) Contextualized sparse representations for real-time open-domain question answering. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 912–919. https://doi.org/10.18653/v1/2020.acl-main.85
14. Lin Y, Ji H, Liu Z, Sun M (2018) Denoising distantly supervised open-domain question answering. In: ACL 2018—56th annual meeting of the association for computational linguistics, proceedings of the conference (long papers), vol 1, pp 1736–1745. https://doi.org/10.18653/v1/p18-1161
15. Nishida K, Saito I, Otsuka A, Asano H, Tomita J (2018) Retrieve-and-read: multi-task learning of information retrieval and reading comprehension. In: International conference on information and knowledge management, proceedings, pp 647–656. https://doi.org/10.1145/3269206.3271702

16. Tan M, Santos CD, Xiang B, Zhou B (2016) Improved representation learning for question answer matching. In: 54th annual meeting of the association for computational linguistics, ACL 2016—long papers, vol 1, pp 464–473. https://doi.org/10.18653/v1/p16-1044
17. Wang W, Yang N, Wei F, Chang B, Zhou M (2017) Gated self-matching networks for MC and QA. Association for Computational Linguistic, pp 189–198
18. Wang S, Yu M, Guo X, Wang Z, Klinger T, Zhang W, Chang S, Tesauro G, Zhou B, Jiang J (2018) R 3: reinforced ranker-reader for open-domain question answering. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 5981–5988
19. Wang B, Yao T, Zhang Q, Xu J, Tian Z, Liu K, Zhao J (2019) Document gated reader for open-domain question answering. In: SIGIR 2019—proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 85–94. https://doi.org/10.1145/3331184.3331190
20. Yang W, Xie Y, Lin A, Li X, Tan L, Xiong K, Li M, Lin J (2019) End-to-end open-domain question answering with BERTserini. In: NAACL HLT 2019—2019 conference of the North American chapter of the association for computational linguistics: human language technologies—proceedings of the demonstrations session, pp 72–77. https://doi.org/10.18653/v1/N19-4013
21. Ghazvininejad M, Brockett C, Chang MW, Dolan B, Gao J, Yih WT, Galley M (2018) A knowledge-grounded neural conversation model. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 5110–5117
22. Serban IV, Sordoni A, Bengio Y, Courville A, Pineau J (2016) Building end-To-end dialogue systems using generative hierarchical neural network models. In: 30th AAAI conference on artificial intelligence, AAAI 2016, pp 3776–3783
23. Ben Abacha A, Demner-Fushman D (2019) A question-entailment approach to question answering. BMC Bioinform 20(1):1–23. https://doi.org/10.1186/s12859-019-3119-4. arXiv:1901.08079ZD
24. Yu J, Qiu M, Jiang J, Huang J, Song S, Chu W, Chen H (2018) Modelling domain relationships for transfer learning on retrieval-based question answering systems in E-commerce. In: WSDM 2018—proceedings of the 11th ACM international conference on web search and data mining, pp 682–690. https://doi.org/10.1145/3159652.3159685
25. Aghaebrahimian A, Jurčíček F (2016) Open-domain factoid question answering via knowledge graph search. In: Proceedings Of2016 NAACL human-computer question answering workshop, pp 22–28. https://doi.org/10.18653/v1/w16-0104
26. Sun H, Ma H, He X, Yih WT, Su Y, Yan X (2016) Table cell search for question answering. In: 25th international world wide web conference, WWW 2016, pp 771–782. https://doi.org/10.1145/2872427.2883080
27. Ye Z, Cai R, Liao Z, Hao Z, Li J (2018) Generating natural answers on knowledge bases and text by sequence-to-sequence learning. In: International conference on artificial neural networks. Springer, pp 447–455. https://doi.org/10.1007/978-3-030-01418-6-44
28. Yu M, Yin W, Hasan KS, dos Santos C, Xiang B, Zhou B (2017) Improved neural relation detection for knowledge base question answering. In: ACL 2017—55th annual meeting of the association for computational linguistics, proceedings of the conference (long papers), vol 1, pp 571–581. https://doi.org/10.18653/v1/P17-1053
29. Abujabal A, Saha Roy R, Yahya M, Weikum G (2018) Never-ending learning for open-domain question answering over knowledge bases. In: The web conference 2018—proceedings of the world wide web conference, WWW 2018, pp 1053–1062. https://doi.org/10.1145/3178876.3186004
30. Bakari W, Neji M (2020) A novel semantic and logical-based approach integrating RTE technique in the Arabic question-answering. Int J Speech Technol. https://doi.org/10.1007/s10772-020-09684-0
31. Cui W, Xiao Y, Wang H, Song Y, Hwang SW, Wang W (2017) KBQA: learning question answering over QA corpora and knowledge bases. In: Proceedings of the VLDB endowment, vol 10, pp 565–576. https://doi.org/10.14778/3055540.3055549
32. Diefenbach D, Giménez-García J, Both A, Singh K, Maret P (2020) QAnswer KG: designing a portable question answering system over RDF data. In: Extended semantic web conference (ESWC) 12123 LNCS, pp 429–445. https://doi.org/10.1007/978-3-030-49461-2_25
33. Hu S, Zou L, Yu JX, Wang H, Zhao D (2018) Answering natural language questions by subgraph matching over knowledge graphs. IEEE Trans. Knowl. Data Eng. 30:824–837. https://doi.org/10.1109/TKDE.2017.2766634
34. Lu X, Abujabal A, Pramanik S, Wang Y, Roy RS, Weikum G (2019) Answering complex questions by joining multi-document evidence with quasi knowledge graphs. In: SIGIR 2019—proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 105–114. https://doi.org/10.1145/3331184.3331252
35. Savenkov D, Agichtein E (2016) When a knowledge base is not enough: question answering over knowledge bases with external text data. In: SIGIR 2016—proceedings of the 39th international ACM SIGIR

conference on research and development in information retrieval, pp 235–244. https://doi.org/10.1145/2911451.2911536
36. Gu Y, Kase S, Vanni M, Sadler B, Liang P, Yan X, Su Y (2021) Beyond I.I.D.: three levels of generalization for question answering on knowledge bases. In: The web conference 2021—proceedings of the world wide web conference, WWW 2021, pp 3477–3488. https://doi.org/10.1145/3442381.3449992
37. Angeli G, Premkumar MJ, Manning CD (2015) Leveraging linguistic structure for open domain information extraction. In: ACL-IJCNLP 2015—53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the Asian federation of natural language processing, proceedings of the conference, vol 1, pp 344–354. https://doi.org/10.3115/v1/p15-1034
38. Wang D, Nyberg E (2015) A long short-term memory model for answer sentence selection in question answering. In: ACL-IJCNLP 2015—53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing of the Asian federation of natural language processing, proceedings of the conference, vol 2, pp 707–712. https://doi.org/10.3115/v1/p15-2116
39. Vinyals O, Le Q (2015) A neural conversational model. In: Proceedings of the 31st international conference on machine learning, vol 37. arXiv:1506.05869
40. Severyn A, Moschittiy A (2015) Learning to rank short text pairs with convolutional deep neural networks. In: SIGIR 2015—proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval, pp 373–382. https://doi.org/10.1145/2766462.2767738
41. Ben Abacha A, Zweigenbaum P (2015) MEANS: a medical question-answering system combining NLP techniques and semantic Web technologies. Inf Process Manag 51(5):570–594. https://doi.org/10.1016/j.ipm.2015.04.006
42. Sun H, Ma H, Yih WT, Tsai CT, Liu J, Chang MW (2015) Open domain question answering via semantic enrichment. In: WWW 2015—proceedings of the 24th international conference on world wide web, pp 1045–1055. https://doi.org/10.1145/2736277.2741651
43. Mitra A, Baral C (2016) Addressing a question answering challenge by combining statistical methods with inductive rule learning and reasoning. In: 30th AAAI conference on artificial intelligence, AAAI 2016, pp 2779–2785
44. Clark P, Etzioni O, Khot T, Sabharwal A, Tafjord O, Turney P, Khashabi D (2016) Combining retrieval, statistics, and inference to answer elementary science questions. In: 30th AAAI conference on artificial intelligence, AAAI 2016, pp 2580–2586
45. Sarrouti M, Ouatik El Alaoui S (2017) A biomedical question answering system in BioASQ 2017. In: BioNLP 2017, pp 296–301. https://doi.org/10.18653/v1/w17-2337
46. Saha A, Pahuja V, Khapra MM, Sankaranarayanan K, Chandar S (2018) Complex sequential question answering: towards learning to converse over linked question answer pairs with a knowledge graph. In: 32nd AAAI conference on artificial intelligence, AAAI 2018, pp 705–713. https://doi.org/10.5281/zenodo.3268649
47. Gupta D, Kumari S, Ekbal A, Bhattacharyya P (2019) MMQA: a multi-domain multi-lingual question-answering framework for English and Hindi. In: LREC 2018—11th international conference on language resources and evaluation, pp 2777–2784
48. Bhandwaldar A, Zadrozny W (2019) UNCC QA: biomedical question answering system, pp 66–71. Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/w18-5308
49. Lee K, Chang MW, Toutanova K (2019) Latent retrieval for weakly supervised open domain question answering. In: ACL 2019—57th annual meeting of the association for computational linguistics, proceedings of the conference, pp 6086–6096. https://doi.org/10.18653/v1/p19-1612
50. Su L, Guo J, Fan Y, Lan Y, Cheng X (2019) Controlling risk of web question answering. In: SIGIR 2019—proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 115–124. https://doi.org/10.1145/3331184.3331261
51. Kanagarajan K, Arumugam S (2019) Intelligent sentence retrieval using semantic word based answer generation algorithm with cuckoo search optimization. Clust Comput 22(s3):7003–7013. https://doi.org/10.1007/s10586-018-2054-x
52. Parikh S, Vohra Q, Tiwari M (2020) Automated utterance generation. In: AAAI 2020—34th AAAI conference on artificial intelligence, pp 13344–13349. https://doi.org/10.1609/aaai.v34i08.7047
53. Karpukhin V, Oguz B, Min S, Lewis P, Wu L, Edunov S, Chen D, Yih W-t (2020) Dense passage retrieval for open-domain question answering. In: Proceeding of the EMNLP, pp 6769–6781. https://doi.org/10.18653/v1/2020.emnlp-main.550
54. Lee S, Sedoc J (2020) Using the poly-encoder for a COVID-19 question answering system
55. Bakari W, Neji M (2020) A novel semantic and logical-based approach integrating RTE technique in the Arabic question-answering. Int J Speech Technol 56:1–17

56. Zhang Y, Zhang X, Hu Y, Wang G, Yan R (2021) WULAI-QA: web understanding and learning with AI towards document-based question answering against COVID-19. In: WSDM 2021—Proceedings of the 14th ACM international conference on web search and data mining, pp 898–901. https://doi.org/10.1145/3437963.3441707
57. Wu Y, Zhao S, Guo R (2021) A novel community answer matching approach based on phrase fusion heterogeneous information network. Inf Process Manag 58(1):102408. https://doi.org/10.1016/j.ipm.2020.102408
58. Lin D, Tang J, Li X, Pang K, Li S, Wang T (2022) BERT-SMAP: paying attention to essential terms in passage ranking beyond BERT. Inf Process Manag 59(2):102788. https://doi.org/10.1016/j.ipm.2021.102788
59. Christmann P, Saha Roy R, Weikum G (2022) Beyond NED: fast and effective search space reduction for complex question answering over knowledge bases. In: Proceedings of the fifteenth ACM international conference on web search and data mining, pp 172–180. Association for Computing Machinery. https://doi.org/10.1145/3488560.3498488
60. Yan R, Liao W, Cui J, Zhang H, Hu Y, Zhao D (2021) Multilingual COVID-QA: learning towards global information sharing via web question answering in multiple languages. In: The web conference 2021—proceedings of the world wide web conference, WWW 2021, pp 2590–2600. https://doi.org/10.1145/3442381.3449991
61. McElvain G, Sanchez G, Matthews S, Teo D, Pompili F, Custis T (2019) WestSearch plus: a non-factoid question-answering system for the legal domain. In: SIGIR 2019—proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 1361–1364. https://doi.org/10.1145/3331184.3331397
62. Zhang S, Lau JH, Zhang X, Chan J, Paris C (2019) Discovering relevant reviews for answering product-related queries. In: Proceedings of the IEEE international conference on data mining, ICDM, pp 1468–1473. https://doi.org/10.1109/ICDM.2019.00192
63. Rowsell J (2013) VQA: visual question answering Stanislaw. In: Proceedings of the IEEE international conference on computer vision, pp 1–182. https://doi.org/10.4324/9780203071953
64. Kacupaj E, Zafar H, Lehmann J, Maleshkova M (2020) VQuAnDa: Verbalization QUestion ANswering DAtaset. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 12123 LNCS, pp 531–547. https://doi.org/10.1007/978-3-030-49461-2_31
65. Fan A, Jernite Y, Perez E, Grangier D, Weston J, Auli M (2020) ELi5: long form question answering. In: ACL 2019—57th annual meeting of the association for computational linguistics, proceedings of the conference, pp 3558–3567. https://doi.org/10.18653/v1/p19-1346
66. Berant J, Chou A, Frostig R, Liang P (2013) Semantic parsing on freebase from question-answer pairs. In: EMNLP 2013—2013 conference on empirical methods in natural language processing, proceedings of the conference, pp 1533–1544
67. Banchs RE (2012) Movie-DiC: a movie dialogue corpus for research and development. In: 50th annual meeting of the association for computational linguistics, ACL 2012—proceedings of the conference, vol 2, pp 203–207
68. Rajpurkar P, Zhang J, Lopyrev K, Liang P (2016) SQuad: 100,000+ questions for machine comprehension of text. In: EMNLP 2016—conference on empirical methods in natural language processing, proceedings, pp 2383–2392. https://doi.org/10.18653/v1/d16-1264
69. Cesario E, Folino F, Manco G, Pontieri L (2005) An incremental clustering scheme for duplicate detection in large databases. In: 9th international database engineering & application symposium (IDEAS'05), pp 89–95. https://doi.org/10.1109/IDEAS.2005.10
70. Chinchor N, Sundheim M (1992) MUC-5 evaluation metrics. In: Fifth message understanding conference (MUC-5): proceedings of a conference held in Baltimore, Maryland, pp 69–78
71. Cormack GV, Lynam TR (2006) Statistical precision of information retrieval evaluation. In: Proceedings of the twenty-ninth annual international ACM SIGIR conference on research and development in information retrieval, pp 533–540. https://doi.org/10.1145/1148170.1148262
72. Wu Y, Mukunoki M, Funatomi T, Minoh M, Lao S (2011) Optimizing mean reciprocal rank for person re-identification. In: 2011 8th IEEE international conference on advanced video and signal based surveillance, AVSS 2011. IEEE, pp 408–413. https://doi.org/10.1109/AVSS.2011.6027363
73. Yu CT, Salton G (1977) Effective information retrieval using term accuracy. Commun ACM 20(3):135–142. https://doi.org/10.1145/359436.359441
74. Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. http://acl.ldc.upenn.edu/W/W05/W05-09.pdf#page=75

75. Mikolov T, Deoras A, Kombrink S, Burget L, Černocký JH (2011) Empirical evaluation and combination of advanced language modeling techniques. In: Proceedings of the annual conference of the international speech communication association, INTERSPEECH, pp 605–608
76. Papineni K, Roukos S, Ward T, Zhu W-J (1922) BLEU: a method for automatic evaluation of machine translation. In: 40th annual meeting of the association for computational linguistics (ACL), pp 437–461. https://doi.org/10.1002/andp.19223712302
77. Rodrigo A, Peñas A (2017) A study about the future evaluation of Question-Answering systems. Knowl Based Syst 137:83–93. https://doi.org/10.1016/j.knosys.2017.09.015
78. Azmi AM, Al-Qabbany AO, Hussain A (2019) Computational and natural language processing based studies of hadith literature: a survey. Artif Intell Rev 52(2):1369–1414. https://doi.org/10.1007/s10462-019-09692-w
79. Abdi A, Hasan S, Arshi M, Shamsuddin SM, Idris N (2020) A question answering system in hadith using linguistic knowledge. Comput Speech Lang 60:101023. https://doi.org/10.1016/j.csl.2019.101023

Faiza Qamar is a Ph.D. student at National University of Sciences and Technology (NUST), Pakistan. She completed her Bachelor's and master's in computer science from University of Gujrat (UOG) and University of Engineering and Technology (UET), respectively. Her research interests are Natural Language Processing (NLP), Machine learning, Deep learning, and Inter-disciplinary research.

Seemab Latif is an associate professor and a researcher at National University of Sciences and Technology (NUST), Pakistan. She received her Ph.D. from the University of Manchester, UK. Her research interest includes artificial intelligence, machine learning, data mining, and NLP. Her professional services include Industry Consultations, Conference Chair, Technical Program Committee Member, and reviewer for several international journals and conferences. In the last 3 years, she has established research collaborations with national and international universities and institutes. She has also secured grants from Asian Development Bank, HEDP, World Bank, Higher Education Commission Pakistan, National ICT, HEC Technology Development Fund, and UK ILM Ideas. She has received the School Best Teacher award in 2016 and the University Best Innovator Award in 2020. She is also the founder of NUST spin-off company, Aawaz AI Tech.

Asad Shah received the B.Sc. degree in computer science from the COMSATS Institute of Information Technology, Islamabad, Pakistan, in 2006, the M.Sc. degree in advanced computer science from The University of Manchester, Manchester, U.K., in 2008, and the D.Phil. degree in computer science from the University of Malaya, Kuala Lumpur, Malaysia, in 2017. He is currently an Assistant Professor with the School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad. His research interests include information retrieval, information processing, web credibility, and question answering systems. He has produced several articles in the area.