Question Answering Survey: Directions, Challenges, Datasets, Evaluation Matrices

Hariom A. Pandya[a], Dr. Brijesh S. Bhatt[a]

[a]*Computer Engineering Department, Dharmsinh Desai University, Nadiad, Gujarat, India*

Abstract

The usage and amount of information available on the internet increase over the past decade. This digitization leads to the need for automated answering system to extract fruitful information from redundant and transitional knowledge sources. Such systems are designed to cater the most prominent answer from this giant knowledge source to the user's query using natural language understanding (NLU) and thus eminently depends on the Question-answering(QA) field.

Question answering involves but not limited to the steps like mapping of user's question to pertinent query, retrieval of relevant information, finding the best suitable answer from the retrieved information etc. The current improvement of deep learning models evince compelling performance improvement in all these tasks.

In this review work, the research directions of QA field are analyzed based on the type of question, answer type, source of evidence-answer, and modeling approach. This detailing followed by open challenges of the field like automatic question generation, similarity detection and, low resource availability for a language. In the end, a survey of available datasets and evaluation measures is presented.

Keywords: Question Answering, Machine Reading Comprehension, Knowldge based Question Answering, Video based Question answering, Question answering datasets, Question answering evaluation measures

1. Introduction

Information retrieval(IR) has been an active area of research for past many decades as manual retrieval of fruitful information from enormous data is a tedious and time-consuming task. Handling this issue and retrieval of updated and important information brings the attention of many researchers [102, 120, 20, 58]. The notable performance improvement in the area can be observed by digital personal assistants (such as Amazon Alexa, Apple Siri, Google Home, etc.), robots-communication, clinical uses of conversation agents for mental health, Chatbot, etc.

Generally, these applications are tuned to have users input in the form of text, video, or speech [118, 103, 9, 115, 74, 22, 25, 83, 64]. Based on the question or input, the system will perform the required action. The pivotal requirement here is to understand and process the output produced by the information retrieval(IR) module [109, 40], and is the role of the question answering system.

Question answering is a field of information retrieval[58] and natural language processing (NLP), concerned with building systems that automatically answer questions posed by humans in a natural language.[1]

As many syntax, semantic, and discourse level challenges involved in this field, it is one of the prominent research directions of recent decades. This can be witnessed by figure 1 which indicates statistics of total submitted and accepted papers of QA field in Association for Computational Linguistics (ACL). It clearly indicates the surge in the popularity of the field.

Pragmatically helpful resources for answering the query include Wikipedia, quora, Reddit, tweeter, stackoverflow, etc.[85, 84, 110, 57, 38, 30, 49, 24, 114]. Many researchers are working in the applications which entertain users query by identifying desired information from such colossal datasets and cater it in accordance to the query.

Finding the best possible answer from one of these datasets is a challenging task [19, 98, 86, 101] due to features selection, its mapping with model and finding relevancy ranking of all candidates answers etc. common challenges. Apart from this, there are some

---

[1]The definition is borrowed from the Wikipedia. https://en.wikipedia.org/wiki/Question_answering

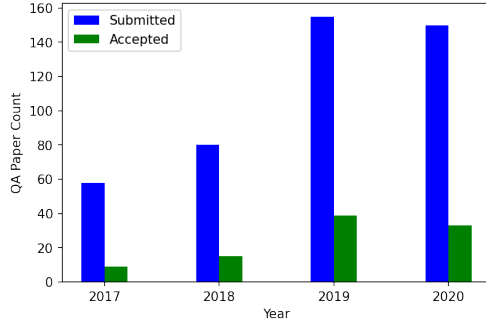*Corresponding Author : hariompandya.ce@ddu.ac.in (Hariom A. Pandya)*

Figure 1: Count of submitted and accepted QA paper in Association for Computational Linguistics(ACL)
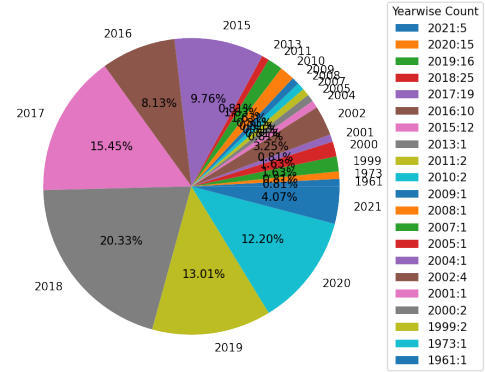

Figure 2: Year wise Surveyed Papers

application-specific key challenges involved in question answering tasks, some of them are highlighted in this paper.

Some of the early notable work of the IR tasks are the BASEBALL and the LUNAR system. BASE-BALL is a program developed by Green et al. [32], which is considered as the foundation of QA field, by many researchers. The system was aimed to answer questions about baseball game played in the American League. The idea there is to fetch the answer, stored in a punch card, with the help of a dictionary. The next breakthrough of the QA field is in 1971, LUNAR System(Lunar Sciences Natural Language Information System) [104], a system designed for chemical analysis of Lunar rocks. It contains database files provided by NASA with 13,000 entry tables of chemical, isotope, and age analyses of the Apollo 11 samples.

Starting from such Information Retrieval (IR) tasks to current encoder-decoder based question-answering, the field has witnessed tremendous progress especially after the involvement of machine learning and deep learning techniques in the field [87, 11, 102, 120, 20, 58, 42, 71]. The availability of high computing machines is another key factor for this exponential advancement.

This survey covers advancement of the QA field with predominant focus on the deep learning based QA approaches. The year wise distribution of the surveyed papers is mentioned in Figure 2.

The remainder of this paper is organized as follows: In the next section, we explore the QA categorization based on the parameters like types of question, answer type, source of evidence or answer, and modeling approaches of the field. Next, we explore the machine reading comprehension (MRC) datasets and Knowledge graph(KB) based datasets followed by evaluation matrices of QA field. In the end observations with statistics are catered.

2. Categorization of QA system

Research in the field of Question Answering is diversified due to the reasons like type of question, expected answer type, source of answer evidence, answer retrieval modeling approach etc. In this study we have covered these categories as mentioned in the Figure 3. In further subsections survey of each is catered. The table 1 indicates summary of individual category used in the selected papers between years 2017-20.

Paper	Year	QT	AT	ES	AS	MA
Reading Wikipedia to Answer Open-Domain Questions [11]	2017	RC	Def	RT	RT	DL
Leveraging video descriptions to learn video question answering [116]	2017	Visual	Fact	Hybrid	Hybrid	DL
End to end long short term memory networks for non-factoid question answering [19]	2017	RC	Def	RT	RT	DL
Improving Deep Learning for Multiple Choice Question Answering with Candidate Contexts [76]	2018	MCQ	Def	RT	RT	DL
A Better Way to Attend: Attention with Trees for Video Question Answering [108]	2018	Visual	Fact	Hybrid	Hybrid	DL
Knowledge base question answering via encoding of complex query graphs [70]	2018	MCQ	Hybrid	KB	KB	ML
Bert with history answer embedding for conversational question answering [82]	2019	Conv	Hybrid	RT	RT	DL
Framework for QA in Sanskrit through Automated Construction of Knowledge Graphs [96]	2019	MCQ	Fact	KB	KB	RB
Careful Selection of Knowledge to solve Open Book Question Answering [6]	2020	MCQ	Fact	RT	RT	DL
Open-Retrieval Conversational Question Answering [81]	2020	Conv	RT	Hybrid	RT	DL
Bert representations for video question answering [111]	2020	Visual	Fact	Hybrid	Hybrid	DL
Counterfactual Samples Synthesizing for Robust Visual Question Answering [13]	2020	Visual	Fact	Hybrid	Hybrid	DL
Decomposing visual question answering deep network via tensor decomposition and regression [4]	2020	Visual	Hybrid	Hybrid	Hybrid	DL
Graph-Based Reasoning over Heterogeneous External Knowledge for Commonsense QA [71]	2020	MCQ	Fact	RT,KB	KB	DL
Research on Medical Question Answering System Based on Knowledge Graph[47]	2021	MCQ	Fact	KB	KB	RB

Table 1: Study of question type, answer type, evidence and answer sources, modeling approach for selected QA papers between the years 2017-2021. Full forms for the used short forms are given in [2]

---

[2]QT: Question Type, AT: Answer Type, ES: Evidence Source, AS: Answer Source, MA: Modeling Apporach, RC:Reading Comprehension, Conv:Conversational QA, Fact:Factoid QA, Def:Definition Based QA, RT:Raw Text, RB:Rule Based
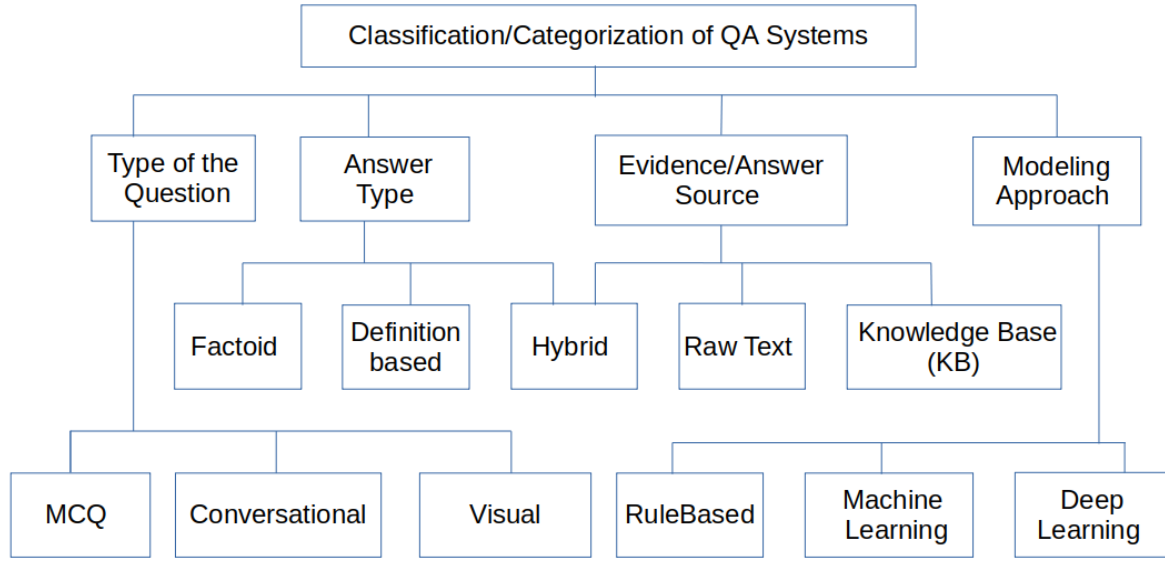
Figure 3: Question Answering : Categories

2.1. Categorization based on type of question

In this section the we are presenting major three directions of QA based on the type of questions, which includes multiple choice questions (MCQ) with the key challenge of eliminating the misleading options, conversational QA in which predominant dependency is on the conversation history and visual QA where evidence reference is derived from the farrago of text and visual features.

2.1.1. MCQ Questions Answering

Along with linguistic and relevant evidence retrieval challenges, answering the question with multiple options involves the challenge of equivocal reference[76] where mapping candidate answers with evidence using standard IR approach can produce the misleading solution due to distraction of distorted options.

The technique of abduction is used by authors [6] by serving question hypothesis and domain knowledge to logical solver with the responsibility to abandon the possible misleading answers. Copynet Seq2Seq [33] is adopted by [6] as a logic solver.

2.1.2. Conversational QA

Discourse plays important role in conversational question answering. Here the direct answering of question is not possible by looking at the context only as along with context the question depends on the conversational history [81, 9, 82]. The conversation history

plays vital role in answer generation. In the conversational question answering, the expected answer can be factoid, a definition or complex question either. The conversational question can be as shown in Figure 4.

Question 1: What space station supported three manned mission in 1973-74?
Answer1   : Skylab
Rationla1 : Skylab, a space station that supported three manned mission in 1973-74.

Question2: It comes under which program?
Answer2   : Apollo Application Program
Rational2 : Apollo uses Saturn family rocket as launch vehicles. Apollo/Saturn vehicles are also used for Apollo Application Program which consisted of Skylab, a space station that supported three manned mission in 1973-74.

Figure 4: SQuAD example with the need of conversation history

By combining conversational response generation and reading comprehension problem, batter performance is observed by [87] in the conversational questions with free from text answers.

To handle conversational response generation problem, BiLSTM is used with encoder decoded in which at the encoder end passage, question-answer history is feeded as "p q_{i-n} a_{i-n} ,,,,,q_{i-1} a_{i-1} q_i" and decoder is responsible to generate the answer. As the answer words

are likely to appear in the original passage, copy mechanism in the decoder is used. This model is referred to as the Pointer-Generator network, PGNet. Document Reader proposed by [11] is used to handle the reading comprehension problem by predicting the rationale for given question.

2.1.3. Visual Question Answering

Answering a question from visual content is referred to as Visual Question Answering (VQA) which embroil natural language processing and computer vision techniques [111, 90]. Here, the visual content can be either image or video and the expected answer can be of factoid, non factoid, yes/no or any other types. The most recent approaches of VQA domain[13] involve the use of a convolution neural network and a recurrent neural network to map the image and question features to common feature space. As per authors [106], the object identification, counting, or appearance, interactions, social relationships, or inferences about why or how something is occurring, are some example features which can be retrieved from visual content[36]. Figure 5 indicates examples of visual question answering derived from the image dataset COCOQA[89] and TVQA[64], a dataset for video based question answering.

The model proposed by [3] uses LSTM to encode the question text and VGGNet[94] to encode the reference image. Features derived are combined after normalization and passed to a fully connected network. A softmax layer at the end obtains a distribution over answers.

- **Video based Question Answering**

Focus of [64] is on visual question answering, in which to generate the feature set for network, visual features are used along with semantic features. To extract video features Faster R-CNN and ResNet101 is used. Bi-directional LSTM (BiLSTM) is used by author to encode both textual and visual sequences. A subtitle is flattened into a long sequence of words and GloVe is used to embed the words. Along with visual features, character name and subtitles are also specified to revamp the performance[4].

Authors [108] uses the datasets YouTube-QA, annotated video clip data [113] and TGIF-QA, annotated GIF data [67] for the video question answering system. The hierarchical parse tree based representation of question with the attention model[115] of video frames is core idea proposed by them. The model is trained on TreeLSTM which is diminished to a classification problem by adding softmax layer. The most promising answer from the candidate set is selected by the model.

The attention input is designed from the input video frames using BiLSTM. In the parse tree, the leaf nodes represent the words of the question. The state of the leaf nodes comes from the word embedding[29] and attention input to the important words.

An approach to automatically harvest videos with description, candidate question answering pair generation and train the model based on modified existing approaches, is proposed by [116]. The focus of [95] is to design end to end memory network with the data as, statements followed by question with single word answer. This architecture is modified by [116] by replacing the statements (sequence of words) by video (sequence of frames). The bi-directional LSTM is used to encode the sequence of frame representation. Visual Question Answering model proposed by [3] is for the question answering for a single image. Here, in the modified approach [116] sequence of frames and the questions are encoded by two separate LSTM. To join both result element wise multiplication is performed.

In the soft attention model [112], the dynamic soft attention[107] is applied to frames. In the modified version [116], the question is also encoded to while paying the attention to the frame.

The encoder decoder architecture of [99] learns to encode the video and decode the sentence. video and question are encoded followed by decoding of answer in the modified version of [116].

2.2. Categorization based on Answer Type

One way to classify current research of question answering filed is, based on the types of expected answer [109, 59, 93, 57, 11]which is generally further divided into factoid, definition, or complex question answering.

2.2.1. Factoid Questions Answering

Factoid questions are the ones that ask about a simple fact and can be answered in a few words. A question *"Which is the Capital of India?"* can be answered by single word *"Delhi"*, is an example of factoid question.

The task of answering knowledge based factoid question rely on inference, requires the aggregation of information from multiple sources. Automatic aggregation often fails because it combines semantically unrelated facts leading to bad inferences. The works by[59, 93] address this inference drift problem using modified page ranking and develops unsupervised and supervised mechanisms to control random walks on Open Information Extraction(OIE) knowledge graphs.

Figure 5: a) Example imageframe b) questionanswer pair based on the visual content c) dataset from which example is derived

2.2.2. Definition based Question Answering

In the definition based question answering [57, 11] the aim is to derive the short passage from the knowledge base, which satisfies the question constrain. A question - answer pair *"Q: What is Machine Learning?", "A: Machine learning is a field which provides the ability to learn the knowledge automatically, to computer systems."* is an example of definition based question answering.

In the research work authors [57] uses Wikipedia pages to generate training model of short and long answers. Generation of dataset involves, annotator with a question, Wikipedia page pair. The annotator returns a (long answer, short answer) pair. Wikipedia articles are used to answer the question by authors [11]. Bigram hashing with TF-IDF matching is used to train multi layer recurrent neural network. The trained model can detect the answer from Wikipedia paragraph.

The authors [19] approach the task using deep learning methods without the need of feature extraction. The end to end training is achieved with a Bidirectional Long Short Term Memory (BLSTM) network and embedding layer is also involved in the training finetune the word representation as shown in the Figure 6 drawn with reference to [19].

2.2.3. Hybrid Question Answering

Answering the hybrid question like: *Which is the second largest planet in the solar system?* involves many natural language challenges [109]. Answer to such question can be factoid or non factoid, but the major challenge here is the selecting statements from the knowledge base/graph to satisfy the question statements. we need to use several semantic clues like : the answer is a planet, the answer is contained by solar system, the answer ranks second, the shorting order is descending order ,etc.

Multiple predicates are required to constrain the answer set for such questions complex questions, if the knowledge graph is used [42, 66, 8]. To tackle this problem Luo et al. [70], first generate multiple query graphs for a given question. Next semantic similarities between the question and each query graph using deep neural networks is calculated and final result from each candidate graph is combined for predicting the correct answer.

Approach of [98] divides the model in two parts: question interpretation, and answer inference. In the question interpretation phase the sets of entities and predicates that are relevant for answering the input question is identified with confidence score. In the second phase these confidence scores are propagated and aggregated over the structure of the knowledge Graph, to provide a confidence distribution over the set of possible answers.

2.3. Categorization based on evidence or answer source

Information can be stored as raw text or can also be represented as a knowledge graph. Based on the storage where we are looking for the answer we can divide this
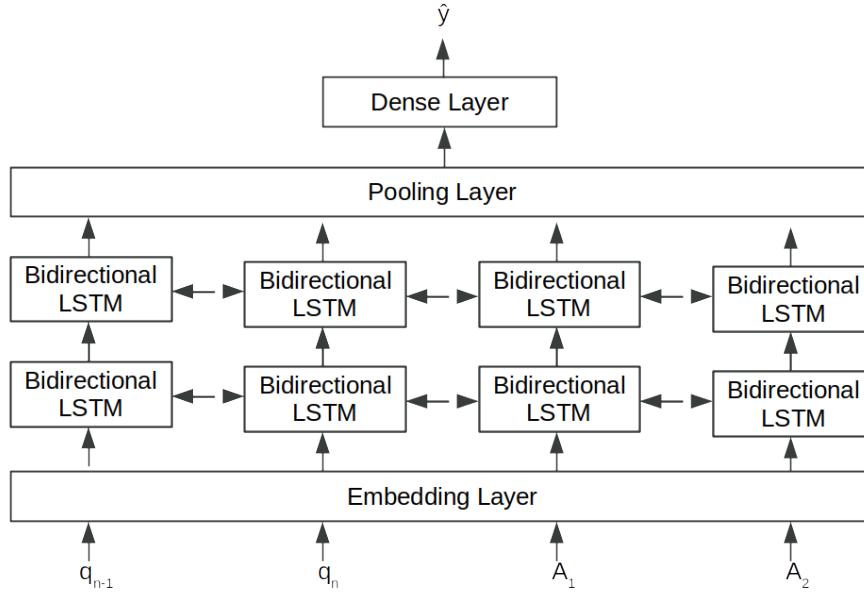
Figure 6: Bidirectional LSTM with n length question and answer text

type further as raw text based QA and Knowledge based QA(KBQA).

2.3.1. Raw text based Question Answering

The reasoning disparity between humans and machines is at a peak when it comes to answering a question from a passage. For a given question, humans can detect the answer paragraph or sentences from a given passage at ease while it is the challenge for the machine [45, 52]. An example of context, question and answer pair of SQuAD is mentioned in the Figure 7 to represent this context.

Context:

Apollo ran from 1961 to 1972, and was supported by the two-man Gemini program which ran concurrently with it from 1962 to 1966. Gemini missions developed some of the space travel techniques that were necessary for the success of the Apollo missions. Apollo used Saturn family rockets as launch vehicles. Apollo/Saturn vehicles were also used for an Apollo Applications Program, which consisted of Skylab, a space station that supported three manned missions in 1973–74, and the Apollo–Soyuz Test Project, a joint Earth orbit mission with the Soviet Union in 1975.

Question: What space station supported three manned missions in 1973–1974?

Answer : Skylab

Figure 7: SQuAD:Context,Question-Answer example

Much research in this direction is going on and still not able to beat the human performance. Datasets like SQuAD [85, 84], WikiQA [110], MS Marco [75] ,TriviaQA [49] and many more are designed to tackle this challenge and to improve the machine comprehensive performance by using deep neural networks techniques involves but not limited to Bi-LSTM, attention based learning, distinct supervise learning, BERT, GPT models etc.

Authors [15] approach this issue on [85, 49] datasets. The method followed by them includes pre-trained Glove word embedding combined with character embedding where the character embedding is generated using CNN followed by max-pooling. Bi-directional GRU is used to map the above embedding to context aware embedding. Lateral question and context are merged using attention mechanism. Further using bi-directional GRU start and end token of probable answers are identified.

The model of [11] is used to find the answer for given question using Wikipedia. The model is of two part 1) Document retriever : to find relevant articles 2) Machine comprehension model : document reader for extracting answer from small set of documents. In the Document retriever part, for a given question 5 Wikipedia articles

6

are retrieved using TF-IDF and bigram matching methods. In the Document Reader part, for all paragraphs, Bi-LSTM is trained using features like 1) Glove word embedding, 2) exact matching of paragraph word pi with one question word, 3) POS,Named Entity Recognition and Term frequency of each paragraph words and 4) attention score of similarity between each question word qj and current paragraph word pi. Result of all hidden layers are concatenated to obtain final features of pi. For a given question, on the top of word embedding of qi RNN layer is added. Result of all hidden units are concatenated in single vector {q1,q2,,,ql}. Once we have vectors for paragraphs and question then for each pair the goal is to predict span of token that is most likely the correct answer. Paragraph vector and question vectors are taken to predict two ends of span using two separate classifiers.

2.3.2. Knowledge based Question Answering

The ontology based information represented is commonly refer as a knowledge graph. In the knowledge graph the information can be represented as [subject, relation, object] triplet, where *relation* indicates the connection via which *subject* and *object* are connected. One example of such triplet is shown in the Figure 8. Here, the relation *'is a university in'* connects subject *'DDU'* and object *'INDIA'*.

Figure 8: Knowledge Graph : Triplet Example

Main challenge here is to map question words onto the graph and finding appropriate connections between nodes to locate the answer [71, 21]. Starting with simple mapping of question-answer to dividing the question into multiple subset and processing each with deep learning based model [118, 42], this direction has attracted many researchers.

2.4. Categorization based on Modeling approach

Starting with word matching model to recent transformer based models [10, 23] general classification of model is in rule based, machine learning based and deep learning based models.

2.4.1. Rule based Models

In the majority of approaches of rule based models, based on WH question type different rules are designed[92]. Language processing tasks like, semantic class tagging, and entity recognition plays a major role in such systems. A rule can be a logical combination of any of the above tasks. Each rule awards some points to a all sentences of the input [50]. After all of the rules have been applied, the sentence that obtains the highest score is returned as the answer. In the current era mainly models for low resource languages are based on such rules [96] while others are diverted to machine learning or deep learning based system where significant performance improvement is observed.

2.4.2. Machine Learning based models

Majorly parsing result of question is given to machine learning models like support vector machine(SVM), decision tree(DT), naive bayes(NB) for the purpose of classification[48, 21, 60]. Other prediction problem using machine learning can be a prediction whether the given community question [73] will entertained by some answer or not using predefined feature sets [72, 69].

2.4.3. Deep Learning based models

With the availability of computing power at ease and an introduction of recurrent neural network(RNN) based models in the text processing field, the research progress of QA is diverted from pure machine learning based models to deep learning based models [11, 98, 87, 64].

For a given question answering task, finetuning of pretrained transformer models like Google's BERT [23] or OpenAI's GPT [10] is the current state of art.

3. Neoteric challenges of QA field

Current advancement in the technology and eminent progress in deep learning leads question answering field beyond simple text based question and answer [17] and thus have opened a door for new directions. The directions where notable progress is expressed in contemporary era includes semantic or discourse related challenges, automatic question generation, similarity detection between questions and challenges for low resource languages etc.

3.1. Semantic related challenges

Semantics, deals with the meaning of word or statement, is playing pivotal role in the understanding the

knowledge structure, finding relationship among entities and objects and retrieval of evidence from ambiguous knowledge[54]. Majority of QA systems are divided in two components named as evidence generation and answer retrieval [11, 76]. To accomplish the evidence generation part in knowledge graph based QA systems generation of the sub-graph with relevant nodes is required while the evidence passage retrieval method is needed in the reading comprehension to achieve it. Semantic nets,first order predicate logic,conceptual graphs, conceptual dependency, are some techniques used for finding the proper meaning of word.

3.2. Discourse related challenges

Discourse plays a vital role in the conversational question answering system [14] in which the current question depends not only on the evidence but also on the history of conversation.

Anaphora Resolution- the method of deriving the correct meaning by detecting anaphora-antecedent pairs[100, 80] and Discourse Structure Recognition - the arrangement of text structure [43], are two major directions which comes under discourse level challenges.

3.3. Automatic question generation

The task of automatically creating questions that can be answered by a certain span of text within a given passage, is explored by [119]. This involve the challenge of natural language generation (NLG) [65]. Long text has posed challenges for sequence to sequence neural models in question generation – worse performances were reported if using the whole paragraph (with multiple sentences) as the input [26]. The maxout pointer mechanism with gated self-attention encoder is proposed to address the challenges of processing long text inputs for question generation. In the model design at the encoder side recurrent neural network (RNN)[44] is used and in the decoding stage, the decoder is another RNN that generates words sequentially conditioned on the encoded input representation and the previously decoded words.

Entity Relationship triplet from the knowledge graph is extracted by authors [88]. These keywords are given to recurrent neural network to generate question from that sequence. Answer to that question will be one of that keyword.

3.4. Question Similarity Detection

The direction of finding question similarities involves challenge of few words when it come to detect the similarity. For example, words used for framing the questions might different but still have semantic, like 'Is it a

watchable movie?' and 'Shall we plan for that movie?' should identified as similar question. In opposition to that 'Is it a watchable news channel?' should identified as different question even it shares most of the words and language model of that words.

Especially after removing the stop words, the vector of each question is of very limited words. Finding the similarity between them involves advance level processing instead of applying only the methods like TF-IDF similarity, Resnik similarity, n-gram matching and other similarity measures and thus it attracts many researcher[37, 79]. Supervised clustering of questions into users intent categories, by using semantic classifiers is the focus direction of [37].

Approach used by [79] is like, for a given question POS tagging followed by stop words removal is applied to identify keywords and that are used in tf-idf based ranking of all questions of dataset. Next top n results are compared with the user query for finding similarity scores.

3.5. Challenges related to low resource languages

Different natural languages behold different levels of progress in this field of question answering. For example, the aim of [96] is to automatic generation of the knowledge graph in the form of a triplet which indicates human relationships (like father, mother, sister, etc.) from Mahabharata. Input question consists of Sanskrit text(in digital Unicode format) and type of human relationship intended. The output is a set of triplets in the form [subject, predicate, object] where the predicate is of the relationship type intended and subject and object are entities. If [a, R, b] is an output triplet, then it implies that object b is relation R of a subject a.

As the focused natural language is Sanskrit, many pre-processing tasks are available to some limited extends only. Pre-processing and knowledge base generation from raw data are major challenges that need to handle by a researcher working with less popular or less explored natural languages [2, 51, 1, 34, 50]. Many Indian languages are testaments for this argument. Authors [7, 46, 31, 12] developed morphological analyzer using paradigm based approach. For resolving ambiguities among the rules, dictionary based methods and statistical methods are used [53]. Major challenge of this work is hand crafted rule building for a language that is highly inflectional with limited availability of the resources. Cross-lingual learning techniques are gaining popularity to tackle the resource constrain issue of low resource languages [62, 41, 35, 105, 77].

4. Question Answering Popular Datasets

In the field of question answering a gigantic sequence of questions and answers produced by a human is necessary to train a network. This can be accomplished using largely available datasets which are in the form of row text documents or in the form of graph based structure referred as the knowledge graph. Mostly all the datasets are generated either by crowd-sourcing or by manual annotation. In this survey, some of the ferociously used datasets for machine reading comprehensive for single answer and multiple choice question answering(MCQ), are highlighted. Figure 9 indicates Google Scholar[3] citation count of referred dataset papers till the date 17-Feb-2021.

4.1. Machine Comprehensive Datasets

- **Stanford Question Answering Dataset (SQuAD)**
  Stanford Question Answering Dataset (SQuAD) is a reading comprehension dataset. It is available in two versions[85, 84] SQuAD1.1 and SQuAD2.0. Here, questions and answering pairs are created through crowd-sourcing by considering Wikipedia as a base. SQuAD 1.1, contains 107,785 question-answer pairs on 536 articles. SQuAD2.0 combines question-answering pairs of SQuAD1.1 with 53,775 unanswerable questions from same paragraphs.

- **WikiQA**
  The WikiQA dataset [110] contains 3,047 questions by extracting queries from Bing query logs, using simple heuristics like starts with WH question and end with '?'. To identify the answer topic, the summary paragraph of the Wikipedia page related to a user query is retrieved. As the summary section of the Wikipedia page usually provides important information, sentences of it are assumed as candidates for the answer. A question with extracted paragraph is crowd-sourced to identify answer sentences.

- **Natural Questions**
  In [57], questions consist of queries submitted to Google search engine. Along with a question, the Wikipedia page from the top 5 search results is used to find candidate answer. A long answer (typically a paragraph) and a short answer (one or more

entities) if present in the Wikipedia page is supplied. The public release consists of 307,373 training examples with single annotations, 7,830 examples with 5-way annotations data for training and 7,842 for testing.

- **MS Marco**
  The dataset [75] contains 10,10,916 questions sampled from Bing's search query logs with a human-generated answer and 1,82,669 human rewritten generated answers. It also contains 88,41,823 passages, extracted from 35,63,535 web documents. The dataset contains Questions, types of question, answers, well formed answer and document from which answer passage is extracted.

- **NewsQA**
  NewsQA [97] contains 119,633 natural language questions posed by crowd-workers on 12,744 news articles from CNN. Text span of article, which contains answer is highlighted by crowd-workers. For 1000 randomly selected examples of NewsQA and SQuAD, to identify the reasoning category, manual labeling is applied in one of the five categories as 1)Word Matching, 2) Paraphrasing, 3) Inference, 4)Synthesis, 5)Ambiguous/Insufficient.

- **CNN/DAILY MAIL**
  In paper [38] almost 93k articles from the CNN and 220k articles from the Daily Mail, are scraped with corresponding questions. Questions are constructed synthetically by deleting a single entity from summary of each article. To identify the correct answer involves recognition of textual entailment between the article and the question. The named entities within an article are identified in a pre-processing step.

- **CoQA**
  Dataset [87] contains 127k question-answering pair, obtained from 8,000 conversations of text passages. Crowd-workers are grouped in two: questioners and answerers. For a given question, the questioner asks questions for which answerer submit a reply. Both can see the ongoing conversation during this process. The passages are from seven includes children's stories from MCTest [91], literature from Project Gutenberg, 5 middle and high school English exams from RACE [61], news articles from CNN [38], articles from Wikipedia, Reddit articles from the Writing Prompts dataset [30], and science articles from AI2 Science Questions. For each in-domain dataset, 100 passages are kept
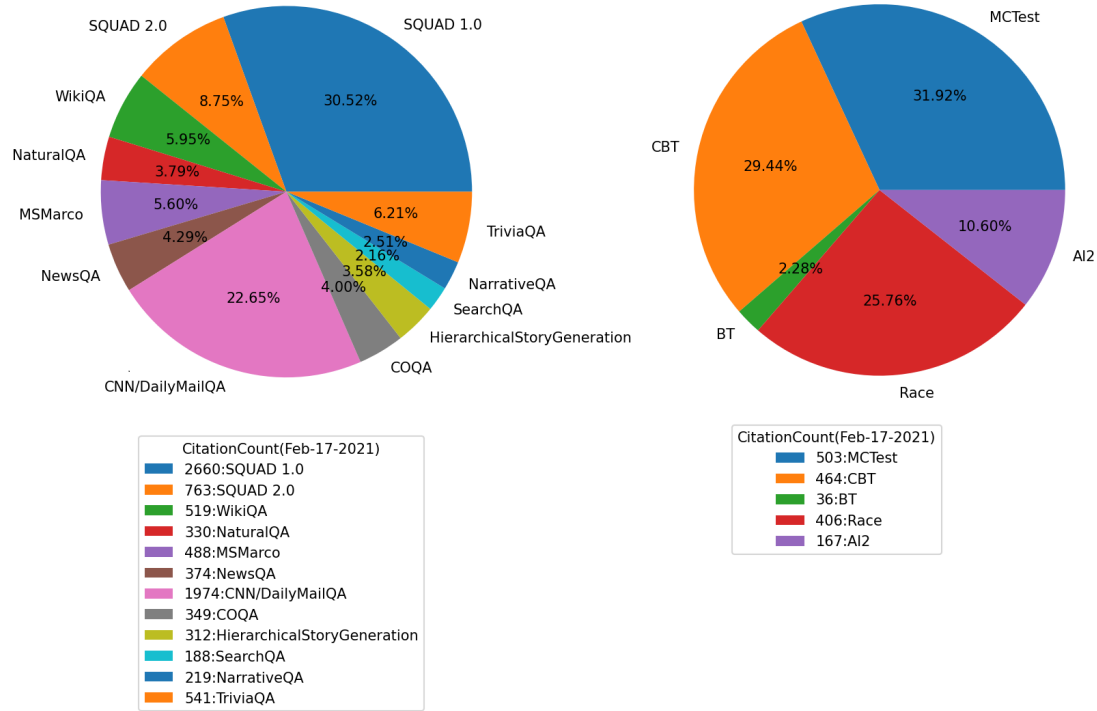

Figure 9: Dataset Citation Count

for development and 100 for the test set, remaining passages are supplied as a training set.

- **SearchQA**
  SearchQA, [27] consists of more than 140k question-answer pairs with each pair having 49.6 snippets on average. Each question-answer-context tuple of the SearchQA comes with additional meta-data such as the snippet's URL which can be valuable resources for future research as per the authors. Set of question-answer pairs are crawled from J! Archive, which has archived all the question-answer pairs from the quiz competition television show Jeopardy!.

- **NarrativeQA**
  In NarrativeQA [55] dataset total 1,567 stories are the collection from 1) books: fetched from Project Gutenberg and 2) movie scripts: scraped from the web. Full data is partitioned into non-overlapping training, validation, and test portions, along with stories/summaries. The dataset contains 46,765 question-answer pairs written by human annotators. Most of the questions are form WH with approx 9.8 tokens.

- **TriviaQA**

TriviaQA [49] is the collection of total 650k question, answer and evidence triple. Most of the question often requires multi-sentence reasoning and retrieved from 14 trivia and quiz-league websites. For retrieval of the evidence two approaches are used:1) Wikipedia articles,2) Using top 50 web search result using Bing web search API.

4.2. MCQ Datasets

- **MCTEST**
  To collect the dataset [91], crowd-sourced workers were instructed to write a short children story and four multiple-choice questions-answers for that story. Workers were supposed to generate option set such that all option words are from the story. At least two questions are such that it requires multiple sentences from the story to answer. The MCTest corpus contains two sets of stories, named MC160 of 160 stories and MC500 of 500 stories. Grammar errors done by a worker are manually corrected in MC160 while in MC500 writing task is assigned after taking grammar test.

- **CHILDREN'S BOOK TEST**
  The Children's book test (CBT) [39] is built from

10

freely available 108 books. Total 6,69,334 training 8,000 validation and 10,000 question are developed. To design question-answer pairs, initial 20 sentences from the context, and expected answer word (which must be one of four POS class Named Entities, Noun, Verb or Preposition), is removed from the 21 st sentence, which becomes the query. Models must identify the answer word among a selection of 10 candidate answers appearing in the context sentences and the -query.

- **BOOK TEST**
  Approach to collect the dataset [5] is similar to CBT [39]. Authors identifies whether each sentence contains either a named entity(NE) or a common noun(CN) that already appeared in one of the preceding twenty sentences. If so, it is then replaced by a blank question. The preceding 20 sentences are used as the context document. Total 3555 copyright-free books to extract CN examples and 10507 books for named entity NE examples. The dataset contains 14, 140, 825 training examples and 7, 917, 523, 807 tokens. The validation and test set consists of 10, 000 NE and 10, 000 CN questions each. The training data consist of tuples (q, d, a, A), where q is a question, d is a document that contains the answer to question q, A is a set of possible answers and a ∈ A is the ground truth answer. Both q and d are sequences of words from the vocabulary.

- **RACE**
  RACE [61], consists of 27,933 passages and 97,687 questions of English exams for middle-school and high-school Chinese students. Each MCQ question is associated with four candidate options, one out of them is correct. Questions and candidate answers are free from text spans of the passage.

- **AI2 Reasoning Challenge (ARC)**
  The ARC dataset [18] consists of 7787 multiple choice non-diagram science questions. The dataset is divided into two parts, a Challenge Set of 2590 "hard" questions and an Easy Set of 5197 questions. The question vocabulary uses 6329 distinct words. The questions of Challenge Set are answered incorrectly by a retrieval-based algorithm and a word co-occurrence algorithm both.

5. Evaluation Metrics

There are many evaluation methods available for the field of question answering. Depends on the problem addressed, different authors compare their performance with benchmark using different evaluation matrix [28]. Starting with Precision, Recall and F1 score upto some benchmark evaluation like BLEU,ROUGE are covered in this section.

- **Precision, Recall,F1 Score**
  For the MCQ based Question Answering, standard way to measure performance is to calculate precision, recall or F1 score as given by next equations.

$$Precision = \frac{TruePositive}{PredictedPositive} \quad (1)$$

$$Recall = \frac{TruePositive}{ActualPositive} \quad (2)$$

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (3)$$

- **BLEU[78]**
  It is the evaluation approach based on modified n-gram precision where candidates are scored based on count of n-grams that are also present in any reference. Score would be aggregated with Geometric Mean. In the calculation of BLUE-1, for each Candidate word of machine generated output find:

  1) W_{max}: Maximum number of times a word occurs in any single reference translation .

  2) $Count_{clip}$: Clipped word count is a min from (W_{max}, W_{count}). Where W_{count}, indicates count of candidate word in candidate output.

  3) Add $Count_{clip}$ for all words and divide it by total words of candidates output.

- **METEOR[63]**
  METEOR (Metric for Evaluation of Translation with Explicit ORdering) is a metric originally designed for the evaluation of machine translation output with the aim of avoiding the limitations of BLEU.

  A term mapping can be thought of as a line between a unigram in candidate statement, and a unigram in reference statement with the constraints as every unigram in the candidate translation must map to zero or one unigram in the reference. The

alignment is a set of mappings between unigrams and incrementally produced with a series of stages.

Each stage consists of two phases: 1) Using external modules ((i) Exact Match module ii) Porter steam module iii) WN synonyms module.), list all possible unigram mapping. 2) In the second phase, a large subset of this module is selected. Each stage maps unigrams that have not been mapped to any unigram in any of the preceding stages. By default, the first stage uses the "exact" mapping module, the second the "porter stem" module and the third the "WN synonymy" module and it is feasible to change this sequence to produce a different result. For large segment common in both segment, n-gram matches are used to compute a penalty p for the alignment.

- **ROUGE[68]**

  ROUGE is a set of metrics for evaluating automatic summarization of texts as well as machine translations and question answering. ROUGE-1 considers unigram, ROUGE-2 bigrams and so on, ROUGE-N measure ngrams. To calculate ROUGE score precision, recall and F Score are measured for the overlapping ngrams. Recall refers to how much of the reference summary the system summary is recovering or capturing. Precision indicates how much of the system summary was, in fact, relevant. Variants of ROUGE are ROUGE-N, ROUGE-S, and ROUGE-L.

- **Word Mover's Distance(WMD) [56]** To find the distance between two documents the word mover's distance is hyper-parameter free, interpretable and high retrieval accurate method. The WMD distance measures the dissimilarity between two text documents as the minimum amount of distance that the embedded words of one document need to "travel" to reach the embedded words of another document.

- **Sentence Mover's Similarity (SMS)[16]**

  This techniques is modified WMD for evaluating multi-sentence texts by basing the score on sentence embeddings. Computation is same to WMD except each document is represented as a bag of sentence embeddings rather than a bag of word embeddings.

- **BERT Score [117]**

  Computes a score by leveraging contextualized word representations, allowing it to go beyond exact match and capture paraphrases better.

An alignment is computed between candidate and reference words by computing pairwise cosine similarity. This alignment is then aggregated in to precision and recall scores before being aggregated into a (modified) F1 score that is weighted using inverse-document-frequency values.

- **Mean Reciprocal Rank (MRR)[101]**

  In the document retrieval field, MRR is used to identify the ability of system in finding the correct answer. The Reciprocal Rank (RR) is the inverse of rank of first correct answer for individual question. i.e. for a given question i, it's rank ($rank_i$) is k, from all n answers generated by the system if kth answer is correct. Average of RR score of each question is considered as MRR.

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{rank_i} \qquad (4)$$

6. Observations

In the recent era, human performance has been surpassed by many state-of-the-art models for pure factoid-based QA and thus, specific branches of QA involving conversational QA, visual QA, multi-hop QA, open-domain QA are gaining the attention of researchers. With the availability of sizeable datasets like SQUAD in addition to the support from hardware advancements in terms of high computing capability, deep Learning based techniques like transformers, RNN, attention mechanism are giving promising results over rule-based and machine learning-based approaches and hence, are part of the major state of the art models. The presence of pre-trained models like BERT, GPT, XLnet, Distil-BERT, Text-to-Text Transfer Transformer(T5), UniLM, and Reformer, has changed the model designing approach to the fine-tuning of pretrained models over the training from scratch.

Table 2 indicates the result obtained by top-performing state-of-the-art models on popular QA datasets. The evaluation approach used by a dataset depends on the problem at hand and the representation of information in it thus, we have mentioned evaluation parameter with result. Additionally F1 score is also mentioned, if present otherwise the field related to that is kept blank in the table.

The majority of the above-mentioned progress is witnessed by resource-rich languages like English or French. For low resource languages, cross-lingual or multilingual approaches are better alternatives over

Dataset	Model	Result	F1 Score
SQuAD 2.0	Human performance	86.831(EM)	89.452
	SA-Net on Albert (ensemble)	90.724(EM)	93.011
SQuAD 1.1	Human performance	82.304(EM)	91.221
	LUKE (single model)	90.202(EM)	95.379
TVQA	Human performance	93.44(Val)	—
	ZGF (single model)	70.29(Val)	—
TVQA	Human performance	89.61(val)	—
	STAGE (span) (single model)	70.5(Val)	—
NaturalQA	PoolingFormer(Long answer)	0.62448(R@P=90)	0.79823
	ReflectionNet(Short answer)	0.35046(R@P=90)	0.64114
MS Marco	DML	0.47(MRR@100Dev)	—
COQA	Human performance	89.4 (InDomain)	—
	RoBERTa + AT + KD (ensemble)	91.4(InDomain)	—
TriviaQA		82.99(Full EM)	87.18
AI2 (ARC)	UnifiedQA + ARC MC/DA + IR	81.4(Accuracy)	—
MCTest	Parallel-Hierarchical on Sparse	71% (MC500 Test)	—

Table 2: Performance of state-of-the-art models on QA datasets

language-specific models due to the limited availability of datasets to design the deep learning-based model.

7. Conclusion

In this paper, we have presented a survey of recent state-of-the-art QA systems. Based on the type of questions different research directions are analyzed followed by open challenges like reading comprehension, conversational question answering, visual question answering, question generation, challenges for low resource language. Some approaches used by researchers are presented with detailing. Finally, a survey of question answering datasets and evaluation measures is presented. We have also provided the observation that the deep learning techniques are gaining popularity for resource-rich languages; it is far off for low-resource languages where still rule-based or machine learning approaches are prevailing.

References

[1] Adams, O., Makarucha, A., Neubig, G., Bird, S., and Cohn, T. (2017). Cross-lingual word embeddings for low-resource language modeling. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 937–947.

[2] Al-Sallab, A., Baly, R., Hajj, H., Shaban, K. B., El-Hajj, W., and Badaro, G. (2017). Aroma: A recursive deep learning model for opinion mining in arabic as a low resource language. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 16(4):1–20.

[3] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: visual question answering. *CoRR*, abs/1505.00468.

[4] Bai, Z., Li, Y., Woźniak, M., Zhou, M., and Li, D. (2021). Decomvqanet: Decomposing visual question answering deep network via tensor decomposition and regression. *Pattern Recognition*, 110:107538.

[5] Bajgar, O., Kadlec, R., and Kleindienst, J. (2016). Embracing data abundance: Booktest dataset for reading comprehension. *CoRR*, abs/1610.00956.

[6] Banerjee, P., Pal, K. K., Mitra, A., and Baral, C. (2019). Careful selection of knowledge to solve open book question answering.

[7] Bapat, M., Gune, H., and Bhattacharyya, P. (2010). A paradigm-based finite state morphological analyzer for marathi. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing*, pages 26–34.

[8] Bauer, L., Wang, Y., and Bansal, M. (2018). Commonsense for generative multi-hop question answering tasks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4220–4230, Brussels, Belgium. Association for Computational Linguistics.

[9] Braun, D., Mendez, A. H., Matthes, F., and Langen, M. (2017). Evaluating natural language understanding services for conversational question answering systems. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 174–185.

[10] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

[11] Chen, D., Fisch, A., Weston, J., and Bordes, A. (2017). Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

[12] Chen, E., Park, H. H., and Schwartz, L. (2020a). Improved finite-state morphological analysis for st. lawrence island yupik using paradigm function morphology. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2676–2684.

[13] Chen, L., Yan, X., Xiao, J., Zhang, H., Pu, S., and Zhuang, Y. (2020b). Counterfactual samples synthesizing for robust visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[14] Chen, Z., Yang, R., Zhao, Z., Cai, D., and He, X. (2018). Dialogue act recognition via crf-attentive structured network. In *The 41st international acm sigir conference on research & development in information retrieval*, pages 225–234.

[15] Clark, C. and Gardner, M. (2018). Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 845–855, Melbourne, Australia. Association for Computational Linguistics.

[16] Clark, E., Celikyilmaz, A., and Smith, N. A. (2019). Sentence mover's similarity: Automatic evaluation for multi-sentence texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2748–2760, Florence, Italy. Association for Computational Linguistics.

[17] Clark, J. H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., and Palomaki, J. (2020). Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

[18] Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the AI2 reasoning challenge. *CoRR*, abs/1803.05457.

[19] Cohen, D. and Croft, W. B. (2016). End to end long short term memory networks for non-factoid question answering. In *Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval*, ICTIR '16, page 143–146, New York, NY, USA. Association for Computing Machinery.

[20] Cui, H., Sun, R., Li, K., Kan, M.-Y., and Chua, T.-S. (2005). Question answering passage retrieval using dependency relations. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 400–407.

[21] Cuteri, B., Reale, K., and Ricca, F. (2019). A logic-based ques-

tion answering system for cultural heritage. In Calimeri, F., Leone, N., and Manna, M., editors, *Logics in Artificial Intelligence*, pages 526–541, Cham. Springer International Publishing.

[22] Dave, P. R. and Pandya, H. A. (2020). Satellite image classification with data augmentation and convolutional neural network. In *Advances in Electrical and Computer Technologies*, pages 83–92. Springer.

[23] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[24] Dhingra, B., Mazaitis, K., and Cohen, W. W. (2017). Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*.

[25] Diefenbach, D., Lopez, V., Singh, K., and Maret, P. (2018). Core techniques of question answering systems over knowledge bases: a survey. *Knowledge and Information systems*, 55(3):529–569.

[26] Du, X., Shao, J., and Cardie, C. (2017). Learning to ask: Neural question generation for reading comprehension. *arXiv preprint arXiv:1705.00106*.

[27] Dunn, M., Sagun, L., Higgins, M., Güney, V. U., Cirik, V., and Cho, K. (2017). Searchqa: A new q&a dataset augmented with context from a search engine. *CoRR*, abs/1704.05179.

[28] Durmus, E., He, H., and Diab, M. (2020). Feqa: A question answering evaluation framework for faithfulness assessment in abstractive summarization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

[29] Esposito, M., Damiano, E., Minutolo, A., De Pietro, G., and Fujita, H. (2020). Hybrid query expansion using lexical resources and word embeddings for sentence retrieval in question answering. *Information Sciences*, 514:88 – 105.

[30] Fan, A., Lewis, M., and Dauphin, Y. (2018). Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

[31] Goyal, V. and Lehal, G. S. (2008). Hindi morphological analyzer and generator. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 1156–1159. IEEE.

[32] Green, B. F., Wolf, A. K., Chomsky, C., and Laughery, K. (1961). Baseball: An automatic question-answerer. In *Papers Presented at the May 9-11, 1961, Western Joint IRE-AIEE-ACM Computer Conference*, IRE-AIEE-ACM '61 (Western), page 219–224, New York, NY, USA. Association for Computing Machinery.

[33] Gu, J., Lu, Z., Li, H., and Li, V. O. (2016). Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

[34] Gu, J., Wang, Y., Chen, Y., Cho, K., and Li, V. O. (2018). Meta-learning for low-resource neural machine translation. *arXiv preprint arXiv:1808.08437*.

[35] Guan, Z., Liu, X., Wu, L., Wu, J., Xu, R., Zhang, J., and Li, Y. (2020). Cross-lingual multi-keyword rank search with semantic extension over encrypted data. *Information Sciences*, 514:523–540.

[36] Gupta, A. K. (2017). Survey of visual question answering: Datasets and techniques. *CoRR*, abs/1705.03865.

[37] Haponchyk, I., Uva, A., Yu, S., Uryupina, O., and Moschitti, A. (2018). Supervised clustering of questions into intents for dialog system applications. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2310–2321, Brussels, Belgium. Association for Computational Linguistics.

[38] Hermann, K. M., Kočiský, T., Grefenstette, E., Espeholt, L.,

Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1693–1701, Cambridge, MA, USA. MIT Press.

[39] Hill, F., Bordes, A., Chopra, S., and Weston, J. (2016). The goldilocks principle: Reading children's books with explicit memory representations. In Bengio, Y. and LeCun, Y., editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

[40] Horvitz, E. J. (2009). System and methods for inferring informational goals and preferred level of detail of results in response to questions posed to an automated information-retrieval or question-answering service. US Patent 7,519,529.

[41] Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., and Johnson, M. (2020). Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

[42] Huang, X., Zhang, J., Li, D., and Li, P. (2019). Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 105–113.

[43] Huber, P. and Carenini, G. (2019). Predicting discourse structure using distant supervision from sentiment. *arXiv preprint arXiv:1910.14176*.

[44] Indurthi, S. R., Raghu, D., Khapra, M. M., and Joshi, S. (2017). Generating natural language question-answer pairs from a knowledge graph using a rnn based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385.

[45] Izacard, G. and Grave, E. (2020). Leveraging passage retrieval with generative models for open domain question answering.

[46] Jena, I., Chaudhury, S., Chaudhry, H., and Sharma, D. M. (2011). Developing oriya morphological analyzer using lt-toolbox. In *International Conference on Information Systems for Indian Languages*, pages 124–129. Springer.

[47] Jiang, Z., Chi, C., and Zhan, Y. (2021). Research on medical question answering system based on knowledge graph. *IEEE Access*, 9:21094–21101.

[48] Joachims, T. (2002). *Learning to classify text using support vector machines*, volume 668. Springer Science & Business Media.

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

[50] Kahaduwa, H., Pathirana, D., Arachchi, P. L., Dias, V., Ranathunga, S., and Kohomban, U. (2017). Question answering system for the travel domain. In *2017 Moratuwa Engineering Research Conference (MERCon)*, pages 449–454.

[51] Karakanta, A., Dehdari, J., and van Genabith, J. (2018). Neural machine translation for low-resource languages without parallel corpora. *Machine Translation*, 32(1):167–189.

[52] Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih, W. (2020). Dense passage retrieval for open-domain question answering.

[53] Khalifa, S., Hassan, S., and Habash, N. (2017). A morphological analyzer for gulf arabic verbs. In *Proceedings of the Third Arabic Natural Language Processing Workshop*, pages 35–45.

[54] Khurana, D., Koli, A., Khatter, K., and Singh, S. (2017). Natural language processing: State of the art, current trends and chal-

lenges. *arXiv preprint arXiv:1708.05148*.

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

[56] Kusner, M. J., Sun, Y., Kolkin, N. I., and Weinberger, K. Q. (2015). From word embeddings to document distances. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, page 957–966. JMLR.org.

[57] Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

[58] Kwok, C. C., Etzioni, O., and Weld, D. S. (2001). Scaling question answering to the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 150–161.

[59] Kwon, H., Trivedi, H., Jansen, P., Surdeanu, M., and Balasubramanian, N. (2018). Controlling information aggregation for complex question answering. In Pasi, G., Piwowarski, B., Azzopardi, L., and Hanbury, A., editors, *Advances in Information Retrieval*, pages 750–757, Cham. Springer International Publishing.

[60] Lahbari, I., Alaoui, S. O. E., and Zidani, K. A. (2018). Toward a new arabic question answering system. *Int. Arab J. Inf. Technol.*, 15:610–619.

[61] Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. (2017). RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

[62] Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining. *arXiv preprint arXiv:1901.07291*.

[63] Lavie, A. and Agarwal, A. (2007). Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, page 228–231, USA. Association for Computational Linguistics.

[64] Lei, J., Yu, L., Bansal, M., and Berg, T. (2018). TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

[65] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

[66] Li, Y., Li, W., and Nie, L. (2021). Incremental knowledge based question answering. *arXiv preprint arXiv:2101.06938*.

[67] Li, Y., Song, Y., Cao, L., Tetreault, J. R., Goldberg, L., Jaimes, A., and Luo, J. (2016). TGIF: A new dataset and benchmark on animated GIF description. *CoRR*, abs/1604.02748.

[68] Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

[69] Liu, Q., Agichtein, E., Dror, G., Gabrilovich, E., Maarek, Y., Pelleg, D., and Szpektor, I. (2011). Predicting web searcher satisfaction with existing community-based answers. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 415–424.

[70] Luo, K., Lin, F., Luo, X., and Zhu, K. (2018). Knowledge base question answering via encoding of complex query graphs. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

[71] Lv, S., Guo, D., Xu, J., Tang, D., Duan, N., Gong, M., Shou, L., Jiang, D., Cao, G., and Hu, S. (2020). Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8449–8456.

[72] Molino, P., Aiello, L. M., and Lops, P. (2016). Social question answering: Textual, user, and network features for best answer prediction. *ACM Transactions on Information Systems (TOIS)*, 35(1):1–40.

[73] Morris, M. R., Teevan, J., and Panovich, K. (2010). What do people ask their social networks, and why? a survey study of status message q&a behavior. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1739–1748.

[74] Moses, D. A., Leonard, M. K., Makin, J. G., and Chang, E. F. (2019). Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature communications*, 10(1):1–14.

[75] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A human generated machine reading comprehension dataset. *CoRR*, abs/1611.09268.

[76] Nicula, B., Ruseti, S., and Rebedea, T. (2018). Improving deep learning for multiple choice question answering with candidate contexts. In *European Conference on Information Retrieval*, pages 678–683. Springer.

[77] Ocquaye, E. N., Mao, Q., Xue, Y., and Song, H. (2021). Cross lingual speech emotion recognition via triple attentive asymmetric convolutional neural network. *International Journal of Intelligent Systems*, 36(1):53–71.

[78] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

[79] Philip, B. A., Jog, M., and Upasani, A. M. (2018). Dr. tux: A question answering system for ubuntu users. *CoRR*, abs/1808.08357.

[80] Poesio, M., Grishina, Y., Kolhatkar, V., Moosavi, N. S., Roesiger, I., Roussel, A., Simonjetz, F., Uma, A., Uryupina, O., Yu, J., et al. (2018). Anaphora resolution with the arrau corpus. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 11–22.

[81] Qu, C., Yang, L., Chen, C., Qiu, M., Croft, W. B., and Iyyer, M. (2020). Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 539–548.

[82] Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., and Iyyer, M. (2019). Bert with history answer embedding for conversational question answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1133–1136.

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

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

[85] Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016). SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas.

Association for Computational Linguistics.

[86] Ravichandran, D. and Hovy, E. (2002). Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual meeting of the association for Computational Linguistics*, pages 41–47.

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

[88] Reddy, S., Raghu, D., Khapra, M. M., and Joshi, S. (2017). Generating natural language question-answer pairs from a knowledge graph using a RNN based question generation model. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 376–385, Valencia, Spain. Association for Computational Linguistics.

[89] Ren, M., Kiros, R., and Zemel, R. S. (2015a). Exploring models and data for image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2953–2961, Cambridge, MA, USA. MIT Press.

[90] Ren, M., Kiros, R., and Zemel, R. S. (2015b). Image question answering: A visual semantic embedding model and a new dataset. *CoRR*, abs/1505.02074.

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

[92] Riloff, E. and Thelen, M. (2000). A rule-based question answering system for reading comprehension tests. In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*.

[93] Shwartz, V., West, P., Bras, R. L., Bhagavatula, C., and Choi, Y. (2020). Unsupervised commonsense question answering with self-talk.

[94] Simonyan, K. and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

[95] Sukhbaatar, S., Szlam, A., Weston, J., and Fergus, R. (2015). End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015-January:2440–2448. 29th Annual Conference on Neural Information Processing Systems, NIPS 2015 ; Conference date: 07-12-2015 Through 12-12-2015.

[96] Terdalkar, H. and Bhattacharya, A. (2019). Framework for question-answering in Sanskrit through automated construction of knowledge graphs. In *Proceedings of the 6th International Sanskrit Computational Linguistics Symposium*, pages 97–116, IIT Kharagpur, India. Association for Computational Linguistics.

[97] Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P., and Suleman, K. (2017). NewsQA: A machine comprehension dataset. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 191–200, Vancouver, Canada. Association for Computational Linguistics.

[98] Vakulenko, S., Fernandez Garcia, J. D., Polleres, A., de Rijke, M., and Cochez, M. (2019). Message passing for complex question answering over knowledge graphs. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 1431–1440, New York, NY, USA. Association for Computing Machinery.

[99] Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., and Saenko, K. (2015). Sequence to sequence – video to text. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4534–4542.

[100] Voita, E., Serdyukov, P., Sennrich, R., and Titov, I. (2018). Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.

[101] Voorhees, E. M. et al. (1999). The trec-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

[102] Wang, D. and Nyberg, E. (2015). A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 707–712.

[103] Wang, W., Yang, N., Wei, F., Chang, B., and 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)*, pages 189–198.

[104] Woods, W. A. (1973). Progress in natural language understanding: An application to lunar geology. In *Proceedings of the June 4-8, 1973, National Computer Conference and Exposition*, AFIPS '73, page 441–450, New York, NY, USA. Association for Computing Machinery.

[105] Wu, H., Wang, Z., Qing, F., and Li, S. (2021). Reinforced transformer with cross-lingual distillation for cross-lingual aspect sentiment classification. *Electronics*, 10(3):270.

[106] Wu, Q., Teney, D., Wang, P., Shen, C., Dick, A. R., and van den Hengel, A. (2016). Visual question answering: A survey of methods and datasets. *CoRR*, abs/1607.05910.

[107] Xiong, C., Zhong, V., and Socher, R. (2016). Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

[108] Xue, H., Chu, W., Zhao, Z., and Cai, D. (2018). A better way to attend: Attention with trees for video question answering. *IEEE Transactions on Image Processing*, 27(11):5563–5574.

[109] Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., Li, M., and Lin, J. (2019). End-to-end open-domain question answering with bertserini. *arXiv preprint arXiv:1902.01718*.

[110] Yang, Y., Yih, W.-t., and Meek, C. (2015). WikiQA: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2013–2018, Lisbon, Portugal. Association for Computational Linguistics.

[111] Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., and Takemura, H. (2020). Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1556–1565.

[112] Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., and Courville, A. (2015). Describing videos by exploiting temporal structure.

[113] Yu, H., Wang, J., Huang, Z., Yang, Y., and Xu, W. (2015). Video paragraph captioning using hierarchical recurrent neural networks. *CoRR*, abs/1510.07712.

[114] Yu, J., Qiu, M., Jiang, J., Huang, J., Song, S., Chu, W., and Chen, H. (2018). Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 682–690.

[115] Yu, Z., Yu, J., Cui, Y., Tao, D., and Tian, Q. (2019). Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

[116] Zeng, K.-H., Chen, T.-H., Chuang, C.-Y., Liao, Y.-H., Niebles, J. C., and Sun, M. (2017). Leveraging video descriptions to learn video question answering. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4334–4340.

16

AAAI Press.

[117] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2019). Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

[118] Zhang, Y., Dai, H., Kozareva, Z., Smola, A., and Song, L. (2018). Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

[119] Zhao, Y., Ni, X., Ding, Y., and Ke, Q. (2018). Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3901–3910, Brussels, Belgium. Association for Computational Linguistics.

[120] Zheng, Z. (2002). Answerbus question answering system. In *Human Language Technology Conference (HLT 2002)*, volume 27. Citeseer.