The Task of Question Answering in NLP: A Comprehensive Review

Sagnik Sarkar, Pardeep Singh, Namrata Kumari,
and Poonam Kashtriya

Abstract An important task of natural language processing is a question answering (QA) (NLP). It provides an automated method of pulling the information from a given context. Thus, QA is made up of three separate modules, each of which has a core component in addition to auxiliary components. These three essential elements are answer extraction, information retrieval, and question classification. By classifying the submitted question according to its type, question classification plays a crucial role in QA systems. Information retrieval is crucial for finding answers to questions because, without the presence of the right ones in a document, no further processing can be done to come up with a solution. Last but not least, answer extraction seeks to locate the response to a query posed by the user. This paper sought to provide a comprehensive overview of the various QA methods, assessment criteria, and benchmarking tools that researchers frequently use.

Keywords Natural language processing (NLP) · QA system · Encoder · Decoder · Attention · Transformer · BERT · T5 · Knowledge graph

1 Introduction

Question answering (QA) is to provide accurate responses to questions based on a passage. In other words, QA systems enable users to ask questions and retrieve answers using natural language queries [1] and can be viewed as an advanced form

S. Sarkar · P. Singh (✉) · N. Kumari · P. Kashtriya
NIT Hamirpur, Hamirpur, HP, India
e-mail: pardeep@nith.ac.in

S. Sarkar
e-mail: 20mcs011@nith.ac.in

N. Kumari
e-mail: namrata_phd@nith.ac.in

P. Kashtriya
e-mail: poonam_phdcse@nith.ac.in

Y. Singh et al. (eds.), *Proceedings of International Conference on Recent Innovations in Computing*, Lecture Notes in Electrical Engineering 1011,
https://doi.org/10.1007/978-981-99-0601-7_46

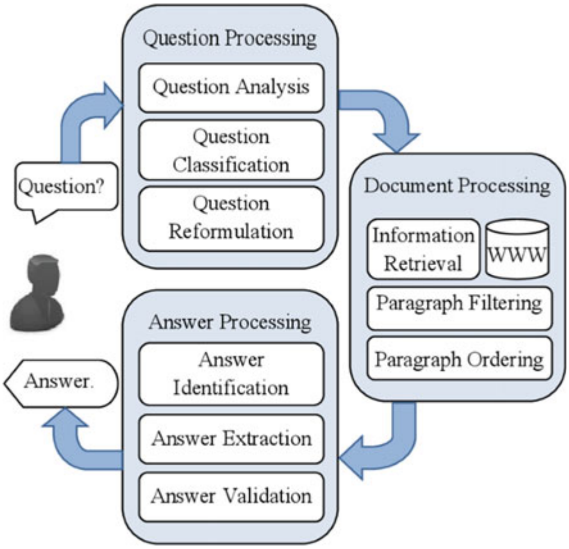
of information retrieval [2]. Additionally, the QA has been utilized to create dialogue systems and chatbots designed to simulate human conversation. There are two main procedures for processing questions. The first step is to examine the structure of the user's query. The second step is to convert the question into a meaningful question formula that is compatible with the domain of QA [3]. The majority of modern NLP problems revolve around unstructured data. This entails extracting the data from the JSON file, processing it, and then using it as needed. An implementation approach categorizes the task of extracting answers from questions into one of four types:

1. IR-QA (Information retrieval based)
2. NLP-QA (Natural language processing based)
3. KB-QA (Knowledge based)
4. Hybrid QA.

2 General Architecture

The following is the architecture of the question answering system: The user asks a question. This query is then used to extract all possible answers for the context. The appropriate architecture of a question answering system is depicted in the Fig. 1.

Fig. 1 Question answering systems [4]



2.1 Question Processing

The overall function of the question processing module, given a question as an input, is to process and analyze the input question so that the machine can understand the context of the question.

2.2 Document Processing

After giving the question as an input, the next big task is to parse the entire context passage to find the appropriate answer locations. The related results that satisfy the given queries are collected in this stage in accordance with the rules and keywords.

2.3 Answer Processing

The similarity is checked after the document processing stage to display the related answer. Once an answer key has been identified, a set of heuristics is applied to it in order to extract and display only the relevant word or phrase that answers the question.

3 Background

"Can digital computers think?" was written by Alan Turing in 1951. He asserted that a machine could be said to be thinking if it could participate in a conversation using a teleprinter and imitate a human completely, without any telltale differences. In 1952, the Hodgkin–Huxley model [5] showed how the brain creates a system that resembles an electrical network using neurons. According to Hans Peter Luhn [6], "the weight of a term that appears in a document is simply proportional to the frequency of the term". Artificial intelligence (AI), natural language processing (NLP), and their applications have all been influenced by these events. The BASEBALL program, created in 1961 by Green et al. [7] for answering questions about baseball games played in the American league over the course of a season, is the most well-known early question answering system. The LUNAR system [8], created in 1971 to aid lunar geologists in easily accessing, comparing, and evaluating the chemical composition of lunar rock and soil during the Apollo Moon mission, is the most well-known piece of work in this field. A lot of earlier models, including SYNTHEX, LIFER, and PLANES [9], attempted to answer a question. Figure 2 depicts the stages of evolution of the NLP models.

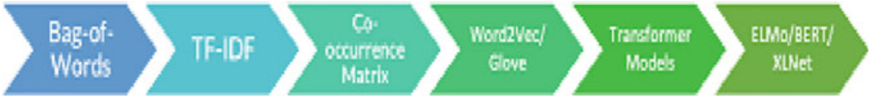


Fig. 2 Evolution of NLP models [10]

4 Benchmarks in NLP

Benchmarks are basically some set of some standard used for assessing the performance of different systems or models agreed upon by large community. To ensure that the benchmark is accepted by large community, people use multiple standard benchmarks. Some of the most renowned benchmarks that are used largely are as follows: GLUE, SuperGLUE, SQuAD1.1, and SQuAD2.0

4.1 GLUE (General Language Understanding Evaluation)

General Language Understanding Evaluation, also known as GLUE, is a sizable collection that includes a variety of tools for developing, testing, and analyzing natural language understanding systems. It was released in 2018, and NLP enthusiasts still find it to be useful today. The components are as follows:

1. A benchmark of nine sentence- or sentence-pair language understanding tasks constructed on well-established existing datasets and chosen to cover a wide range of dataset sizes, text genres, and degrees of difficulty;
2. A leaderboard to find the top overall model;
3. A diagnostic dataset to assess and analyze the model's performance in relation to a variety of linguistic issues encountered in the natural language domain.

4.2 SuperGLUE

General Language Understanding Evaluation, also known as GLUE, is a large collection of dataset that includes a variety of tools for developing, testing, and analysis. SuperGLUE is an updated version of the GLUE benchmark. SuperGLUE benchmark is designed after GLUE but with whole new set of improved and more difficult language understanding tasks, improved reasoning, and a new canvas of public leaderboard. It was introduced in 2019. Currently, Microsoft Alexander v-team with Turing NLRv5 is leading the scoreboard with URL score of 91.2.

4.3 SQuAD1.1 (Stanford Question Answering Dataset 1.1)

SQuAD or Stanford Question Answering Dataset was introduced in 2016 which consists of Reading Comprehension Datasets. These datasets are based on the Wikipedia articles. The previous version of the SQuAD dataset contains 100,000+ question answer pairs on 500+ articles.

4.4 SQuAD1.1 (Stanford Question Answering Dataset 2.0)

SQuAD2.0 or Stanford Question Answering Dataset combines all the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written so that it may look similar to answerable ones. SQuAD2.0 tests the ability of a system to not only answer questions when possible, but also determine when no answer can be found in the comprehension. Currently, the IE-NET (ensemble) by RICOH_SRCB_DML is leading the scoreboard with **EM** score of 90.93 and **F1** score of 93.21.

5 Research

In this systematic literature review (SLR), we tried to address the various steps based on the guidelines provided by the Okoli and Schabram [11], Keele [12], which emphasizing as: Purpose of the Literature Review, Searching various Literature, Practical Screen, Quality Appraisal, and Data Extraction. The amount of written digital information has increased exponentially, necessitating the use of increasingly sophisticated search tools. Pinto et al. [13], Bhoir and Potey [14]. Unstructured data is being gathered and stored at previously unheard-of rates, and its volume is growing. Bakshi et al. [15], Malik et al. [16], and Chali et al. [17], among others. The main difficulty is creating a model that can effectively extract data and knowledge for various tasks. The tendency in this situation of the question answering systems is to glean as many answers from the questions as you can. This SLR will be guided by the research questions in Table 1 in an effort to comprehend how question answering systems techniques, tools, algorithms, and systems work and perform, as well as their dependability in carrying out task.

We gathered as many journals and papers written in English in different digital libraries and reputed publications through the various keywords and tried to provide some strong evidence related to the research questions that have been tabulated earlier.

RQ_1: Fig. 3 tried to show the popularity of various models on the basis of the number of paper published in the category in every year. Here, we can observe that the BERT-based model is the most popular in this category.

Table 1 Research questions to be addressed

Question No.	Research question
RQ1	What are the popular QA techniques?
RQ2	Which domains use the question answering models?
RQ3	How it is improving the existing model?
RQ4	Contribution of other authors in the field of QA?

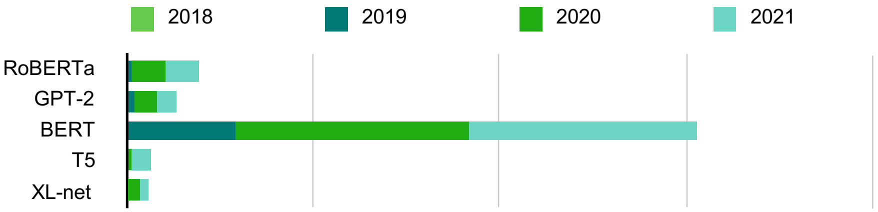


Fig. 3 A graph showing the popularity of the models

RQ_2: Fig. 4 tries to show the various question answering fields the QA models are used. We can see that general domain QA is dominantly used here.

RQ_3: The fine-tuning of different models have given rise to various improvements in the existing models. Moreover, using the different techniques over the existing model can give rise to different model which can improve the existing the model. **For Example**: The different BERT-based models like AlBERT, RoBERTa, DistilBERT with different parameters are used according to the need as shown in Table 2.

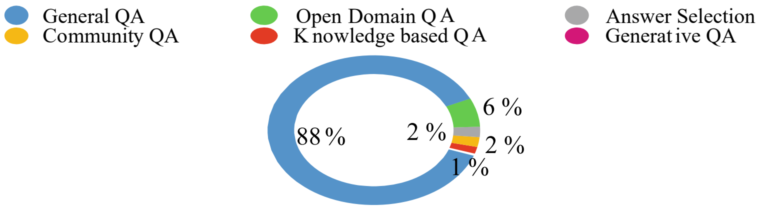


Fig. 4 Chart shows the different types of question answering area

Table 2 Different application using different models

Tasks	BERT	T5	GPT-2
Language modeling	4	3	3
Text generation	1	3	3
Question answering	7	4	3

Table 3 Table showing the area of working the models

Tasks	BERT	T5	GPT-2
Language modeling	4	3	3
Text generation	1	3	3
Question answering	7	4	3
Machine translation	2	2	1
Text classification	1		1
Text summarization	2	2	
Sentiment analysis	1	6	

RQ_4: This is the main purpose of the literature review. This question is answered in support with Table 3. Many papers have been taken into consideration for this comparison [8, 18–38]. Here, we took only three models as these are the main base models that predominate the question answering domain.

6 Conclusion

Question answering system using NLP techniques is much complicated process as compared to other type of information retrieval system. The closed domain QA systems is able to give more accurate answer than that of open domain QA system but is restricted to a single domain only. After the screening phases, we can see that the attention-based model is must preferable among the researchers. We also observed that researchers have equally turned themselves to the hybrid approaches like graph attention and applying different styles of mechanism over the base model to make their job easy. The contributions of this work are a systematic outline of different question answering systems that are able to perform better in all the different tasks. The future should try to explore the possibility of any such model that can outperform all models.

References

1. Abdi A, Idris N, Ahmad Z (2018) QAPD: an ontology-based question answering system in the physics domain. Soft Comput 22(1):213–230
2. Cao YG, Cimino JJ, Ely J, Yu H (2010) Automatically extracting information needs from complex clinical questions. J Biomed Inform 43(6):962–971
3. Joulin A, Grave E, Bojanowski P, Mikolov T (2016) Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759
4. Allam AMN, Haggag MH (2012) The question answering systems: a survey. Int J Res Rev Inf Sci (IJRRIS) 2(3)
5. Hamed SK, Ab Aziz MJ (2016) A question answering system on Holy Quran translation based on question expansion Technique and neural network classification. J Comput Sci 12(3):169–177

6. Papineni K, Roukos S, Ward T, Zhu WJ (2002) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp 311–318
7. Hyndman RJ, Koehler AB (2006) Effect of question formats on item endorsement rates in web surveys. Int J Forecast 22(4):679–688
8. Liang T, Jiang Y, Xia C, Zhao Z, Yin Y, Yu PS (2022) Multifaceted improvements for conversational open-domain question answering. arXiv preprint arXiv:2204.00266
9. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H (2020) Ernie 2.0: a continual pre-training framework for language understanding. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, no 05, pp 8968–8975
10. Hogan A, Blomqvist E, Cochez M, d'Amato C, Melo GD, Gutierrez C et al (2021) Knowledge graphs. Synthesis Lectures on Data, Semantics, and Knowledge 12(2):1–257
11. Okoli C, Schabram K (2010) A guide to conducting a systematic literature review of information systems research
12. So D, Mańke W, Liu H, Dai Z, Shazeer N, Le QV (2021) Searching for efficient transformers for language modeling. Adv Neural Inf Process Syst 34:6010–6022s
13. Turing AM (1951) Can digital computers think? The Turing test: verbal behavior as the hallmark of intelligence, pp 111–116
14. Bhoir V, Potey MA (2014) Question answering system: a heuristic approach. In: The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014). IEEE, pp 165–170
15. Bakshi K (2012) Considerations for big data: architecture and approach. In: 2012 IEEE aerospace conference. IEEE, pp 1–7
16. Malik N, Sharan A, Biswas P (2013) Domain knowledge enriched framework for restricted domain question answering system. In: 2013 IEEE international conference on computational intelligence and computing research. IEEE, pp 1–7
17. Chali Y, Hasan SA, Joty SR (2011) Improving graph-based random walks for complex question answering using syntactic, shallow semantic and extended string subsequence kernels. Inf Process Manage 47(6):843–855
18. Yao X (2014) Feature-driven question answering with natural language alignment. Doctoral dissertation, Johns Hopkins University
19. Zhang J, Zhang H, Xia C, Sun L (2020) Graph-BERT: only attention is needed for learning graph representations. arXiv preprint arXiv:2001.05140
20. Zhang X, Hao Y, Zhu XY, Li M (2008) New information distance measure and its application in question answering system. J Comput Sci Technol 23(4):557–572
21. Mozafari J, Fatemi A, Nematbakhsh MA (2019) BAS: an answer selection method using BERT language model. arXiv preprint arXiv:1911.01528
22. Sun C, Qiu X, Xu Y, Huang X (2019) How to fine-tune BERT for text classification? In: China national conference on Chinese computational linguistics. Springer, Cham, pp 194–206
23. Wang A, Cho K (2019) BERT has a mouth, and it must speak: BERT as a Markov random field language model. arXiv preprint arXiv:1902.04094
24. Wang Z, Ng P, Ma X, Nallapati R, Xiang B (2019) Multi-passage BERT: A globally normalized BERT model for open-domain question answering. arXiv preprint arXiv:1908.08167
25. 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. arXiv preprint arXiv:1902.01718
26. Kale M, Rastogi A (2020) Text-to-text pre-training for data-to-text tasks. arXiv preprint arXiv:2005.10433
27. Lin BY, Zhou W, Shen M, Zhou P, Bhagavatula C, Choi Y, Ren X (2019). CommonGen: a constrained text generation challenge for generative commonsense reasoning. arXiv preprint arXiv:1911.03705
28. Ribeiro LF, Schmitt M, Schütze H, Gurevych I (2020) Investigating pretrained language models for graph-to-text generation. arXiv preprint arXiv:2007.08426
29. Agarwal O, Kale M, Ge H, Shakeri S, Al-Rfou R (2020). Machine translation aided bilingual data-to-text generation and semantic parsing. In: Proceedings of the 3rd international workshop on natural language generation from the semantic web (WebNLG+), pp 125–130

30. Moorkens J, Toral A, Castilho S, Way A (2018) Translators' perceptions of literary post-editing using statistical and neural machine translation. Translation Spaces 7(2):240–262
31. Ethayarajh K (2019) How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. arXiv preprint arXiv:1909.00512
32. Frydenlund A, Singh G, Rudzicz F (2022) Language modelling via learning to rank. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, no 10, pp 10636–10644
33. Mager M, Astudillo RF, Naseem T, Sultan MA, Lee YS, Florian R, Roukos S (2020) GPT-too: a language-model-first approach for AMR-to-text generation. arXiv preprint arXiv:2005.09123
34. Qu Y, Liu P, Song W, Liu L, Cheng M (2020) A text generation and prediction system: pre-training on new corpora using BERT and GPT-2. In: 2020 IEEE 10th international conference on electronics information and emergency communication (ICEIEC). IEEE, pp 323–326
35. Puri R, Spring R, Patwary M, Shoeybi M, Catanzaro B (2020) Training question answering models from synthetic data. arXiv preprint arXiv:2002.09599
36. Wang A, Singh A, Michael J, Hill F, Levy O, Bowman SR (2018) GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461
37. Wang A, Pruksachatkun Y, Nangia N, Singh A, Michael J, Hill F, Levy O, Bowman S (2019) Superglue: a stickier benchmark for general-purpose language understanding systems. Advances in neural information processing systems, 32
38. Hsu HH, Huang NF (2022) Xiao-Shih: a self-enriched question answering bot with machine learning on Chinese-based MOOCs. IEEE Trans Learn Technol