A BRIEF SURVEY OF QUESTION ANSWERING SYSTEMS

Michael Caballero

University of California, Berkeley, CA, USA

ABSTRACT

Question Answering (QA) is a subfield of Natural Language Processing (NLP) and computer science focused on building systems that automatically answer questions from humans in natural language. This survey summarizes the history and current state of the field and is intended as an introductory overview of QA systems. After discussing QA history, this paper summarizes the different approaches to the architecture of QA systems -- whether they are closed or open-domain and whether they are text-based, knowledge-based, or hybrid systems. Lastly, some common datasets in this field are introduced and different evaluation metrics are discussed.

KEYWORDS

Question Answering, Natural Language Processing,Information Extraction, Artificial Intelligence, QA Systems

1. INTRODUCTION

Question Answering (QA) is a subfield of Natural Language Processing (NLP) that aims to build systems that automatically answer questions from humans in natural language. QA has a rich history and is now a very popular topic in computer science and NLP. Current research has shifted from simple questions, like "Who is the President of the US?", to complex queries that require explanation.

This survey aims to provide a succinct overview of the history and current state of question answering systems. First, it details the origin and history of these systems, mentioning the most popular early systems and significant accomplishments. Next, QA architecture is discussed; the two main differentiating factors in QA systems arewhether the system is open or closed-domain and whether the system is text-based, knowledge-based, or hybrid. The fourth section discusses the key processes of QA systems, detailing the differences between text-based and knowledge-based systems. After key processes, this survey explores the most popular datasets used for QA systems and evaluation metrics to compare systems.

2. ORIGIN & HISTORY

The development of QA systems has been a staple in NLP research for well over the last half-century. Early systems were built on quite basic computers and focused on simply retrieving information in a small closed-domain setting. As the first survey on QA systems remarked, "the domains of these programs were quite restricted, however, so their designers were able to ignore many of the issues in understanding a question and answering it adequately" [1].

DOI: 10.5121/ijaia.2021.12501

The first acknowledged QA system arrived in 1961; Green et al.'s BASEBALL answered simple questions relating to the month, place, team, or score of American League baseball games for one year of play [2]. One interesting aspect of BASEBALL which reveals how rudimentary the computers were at the time is that "the program reads the question from punched cards" [2]. The main emphasis of this research was to show a very basic proof-of-concept for QA systems. Improvements to build much better syntactic and semantic parsers led to the next generation of QA systems that allowed more freedom of expression in the inputted question; the most famous of these was LUNAR[3]. LUNAR was a QA system for lunar geologists on the Apollo moon mission that enabled them to access and evaluate the chemical analysis composition data on lunar rocks and soil.

QA systems have advanced quite significantly since these early stages. About a decade ago they surpassed what many would consider a great baseline of question answering ability. IBM Research built Watson, a deep learning QA system, that competed and won facing the best contestants on "Jeopardy" [4]. This was a proud moment when the field surpassed human-level intelligence at one key test of question-answering ability. proceedings.

3. QA ARCHITECTURE

The first significant differentiating factor among QA systems is whether the system is open-domain or closed-domain. This factor relates to the type of question that can be inputted into the system. "Closed domain question answering deals with questions under a specific domain (music, weather, forecasting, etc.)" [5]. This factor is primarily based on when the QA system was developed as the field's early focus was mainly on closed-domain systems like LUNAR and BASEBALL.

Current research is almost solely focused on open-domain QA to not restrict the question inputted. This is likely due to two reasons: technological advancement and applicability. This field of research has advanced substantially with exponentially more powerful computers, allowing for complex systems that can rise to the challenge of open-domain QA. Additionally, an open-domain system more resembles the archetypal AI that we can imagine in society, with Zhu et al. writing that "building an OpenQA system that is capable of answering any input questions is deemed as the ultimate goal of QA research" [6]. Since the general established goal in current research is an open-domain QA system, the rest of the paper will focus on these systems.

With open-domain QA systems, there are three main architectures: text-based, knowledge-based, and hybrid. This is the main decision in creating QA systems as the architecture depends on the underlying source of information. Text-based QA systems rely on unstructured documents like textual excerpts from Wikipedia or similar sites and answer questions by finding the most similar answer of all possible answers. Knowledge-based QA sources answers from structured data in knowledge bases that include relations, facts, and entities usually in RDF graphs or SQL databases. Until recently, knowledge-based QA struggled to correctly formulate answers to complex questions which require combing and reasoning over multiple facts. But, Yu et al. achieved state of the art accuracy by implementing a hierarchical recurrent neural network that was able to detect relations between data points in a knowledge base [7].

While most research chooses between structured and unstructured data, the future of the field is likely in semi-structured data, employing hybrid QA systems that read both types of data to maximize efficiency and accuracy. As Dimitrakis et al. writes:

"A limitation of the previous approaches [text-based and knowledge-based QA systems] is that a single type of source can be insufficient to provide the ideal answer. The latter [hybrid systems] could be the synthesis of information extracted from various resources with different types of information representation" [8].

And, the impact of hybrid systems are now evident in research as state of the art results in open QA now come from models like Retrieval-Augmented Generation (RAG). Mao et al. show that "fusing their results consistently yields better retrieval accuracy" [9].

4. KEY PROCESSES

Because the data QA systems query varies so significantly from structured to unstructured to semi-structured, the key processes of systems change dramatically. As expected, hybrid systems simply share key processes from both text-based and knowledge-based systems. While this survey does not have enough space to deeply examine different methods for each key process, it will identify and explain popular methods. The first key process in both types of systems is Question Analysis. This step analyzes the input, and when performed well, greatly reduces the size of the search space which has cascading effects on the rest of the steps. The research on Question Analysis contains many popular methods including morphological, syntactical, and semantic analysis. One other traditional method is focus recognition where the purpose defined by Moldovan et al. is to locate the word or sequence of words that pertain to the information requested [10]. Ferret et al. refined this definition to identifying both the head of the focus and a list of modifiers [11].

4.1. Text-Based Methods

After performing Question Analysis, the two key steps in unstructured QA systems are Passage Retrieval and Answer Extraction. Passage Retrieval is the core component of QA and is comprised of three key steps: document retrieval, searching the data for relevant documents; passage extraction, finding small excerpts that relate to the query; and passage ranking, which ranks the excerpts by their potential to contain the answer. IBM Watson popularized the use of Information Retrieval systems to optimize the first task of Document Retrieval [12].

Answer Extraction is the last step and involves generating the final answer and validating it with a confidence score. While historic QA systems typically presented five answers, QA systems have been required to present just one answer since 2002 [13].

4.2. Knowledge-Based Methods

With structured data, there are two main methods post Question Analysis [14]. The first and more traditional method is Information Retrieval which focuses on sorting over candidate answers [15]. While very effective in the past, these systems are not interpretable and do not perform well with complex questions. The second method, Semantic Parsing, focuses on converting sentences to their semantic representation which then can be used for executable queries. Berant and Liang have greatly progressed research in this space by converting sentences into what they name logical forms[16, 17]. Semantic Parsing based on neural networks, or Neural Semantic Parsing, is the state-of-the-art method of choice for knowledge-based QA systems, greatly enhancing parsing capability and scalability.

5. POPULAR DATASETS

While research into open-domain QA utilizes many different datasets, there are many common datasets in research. These popular datasets are utilized in multiple papers, allowing for comparisons between two different models. The more sophisticated datasets utilize complex questions to test how QA systems can reason. As the QA systems are either text-based or knowledge-based, evaluation datasets normally cater to one type of QA system and not both. Below, several of the major datasets for each separate method are introduced.

5.1. Unstructured Data

One of the first common unstructured datasets for QA was popularized by Wang et al. in their attempt to use quasi-synchronous grammar for QA [18]. This dataset, TREC-QA, was gathered from the Text REtrevial Conference (TREC) 8-13 QA. TREC-QA includes editor-generated questions with potential sentences for answer extraction selected by matching similar context words in the question. Wang et al. established a popular training, evaluation, and test set split with 1325 questions used to train a QA system, 82 for validation, and 100 for testing the system.
Another popular unstructured dataset, WikiQA was proposed by Yang et al. in response to weaknesses they found in the TREC-QA dataset [19]. WikiQA is substantially larger than TREC-QA, and its creators claim that WikiQA was constructed more naturally than TREC-QA. Furthermore, "the WikiQA dataset includes questions for which there are no correct sentences, enabling researchers to work on answer triggering," a challenge that tests a QA system's ability to detect if there is a valid answer in a set of candidate sentences [19]. WikiQA contains a much more robust set of questions with 3047 in total, more than twice as many as TREC-QA's 1507.

5.2. Structured Data

The first and older structured dataset that became popular in the research of knowledge-based QA systems is WEBQUESTIONS. WEBQUESTIONS was constructed by Berant et al. to train a semantic parser on question-answer pairs instead of annotated logical forms [20]. They used the Google Suggest API to obtain realistic questions that start with a wh-word and only contained one entity. In order to annotate the data, they submitted 100k of their million questions to Amazon Mechanical Turk. As this was one of the earlier common structured datasets, one of its major flaws is that 84% of the questions are simple with only a few requiring complex reasoning.
WEBQUESTIONS was iterated upon and this major flaw was ameliorated by Talmor and Berant with COMPLEXWEBQUESTIONS, where they attempted to demonstrate the capability of question decomposition [21]. This dataset contains the semantic phenomena from questions like those in WEBQUESTIONS which contain one or multiple entities, but also adds "four compositionality types by generating compound questions (45% of the times), conjunctions (45%), superlatives (5%), and comparatives (5%)" [21].

6. EVALUATION

While using a common dataset for two different methods is a step towards comparing the two, the most important factors in comparison are the evaluation metrics. There is no "best" evaluation metric used for QA systems, but rather a number of different metrics that help researchers understand the performance of different parts of a system. One roadblock in the research is that articles use different evaluation metrics which makes comparing the success of different architectures difficult. In the future, there will hopefully be a few common evaluation metrics that are always reported.

Of the four main metrics for comparing the performance of QA systems, two of the most common are the standard classification metrics, Accuracy and F-score. Accuracy(1) is used for evaluating datasets whose questions have just one answer and is simply the fraction of all answered questions that were answered correctly.

$$\text{Accuracy} = \frac{\#\text{ of Correctly Answered Qs}}{\#\text{ of Answered Qs}} \tag{1}$$

When evaluating a question where there is not just one answer and rather several correct answers are possible, F-score is the main metric to compare performance. F-score(4) is broken down into Precision and Recall, two separate metrics that are combined together in F-score. The Precision(2) is the fraction of answers outputted that are correct answers, while Recall(3) is the fraction of correct answers that the system outputted. F-score combines these two metrics together into one number, representing the harmonic mean of Precision and Recall.

$$\text{Precision} = \frac{\#\text{ of Correct Answers Found}}{\#\text{ of Answers Found}} \tag{2}$$

$$\text{Recall} = \frac{\#\text{ of Correct Answers Found}}{\#\text{ of Correct Answers}} \tag{3}$$

$$\text{F-score} = 2\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The next evaluation metric, the Mean Reciprocal Rank (MRR), is one of the best for measuring the performance of QA systems. MRR (6) indicates a system's ability to answer a question and is just the mean of the Reciprocal Rank (RR) which is calculated for each question. The RR (5) is zero if no correct answer is provided or otherwise it is the inverse of the rank of the first correct answer.

$$\text{RR} = \frac{1}{\text{rank}} \tag{5}$$

$$\text{MRR} = \frac{1}{\text{Q}}\sum_{\text{i}=1}^{\text{Q}} \text{RR}_\text{i} \text{ , where Q} = \text{Number of Questions} \tag{6}$$

The last evaluation metric commonly used is the Mean of Average Precision (MAP). The Average Precision (AP) (7) for a question is the average value of the Precision as a function of Recall which is obtained by computing the Precision and Recall at every position in the ranked list of answers. MAP (8) simply averages the AP for every question in the dataset.

$$\text{AP} = \sum_{\text{k}=1}^{\text{K}} \frac{\text{p(k)}}{\text{k}} \tag{7}$$

$$\text{p(k)} = \frac{\#\text{ of Correct Answers in First k Results}}{\text{k}}$$

$$\text{MAP} = \frac{1}{Q}\sum_{i=1}^{Q} \text{AP}_i \text{ , where Q} = \text{Number of Questions} \quad (8)$$

7. CONCLUSION

This survey attempts to provide a briefexamination of QA systems from their early origin to the current systems in 2021. It provides an overview of the architecture and key processes in both text and knowledge-based QA systems. This paper focuses on introductory aspects of the field like popular datasets and evaluation metrics which enable comparison between methods. Future research for QA systems will likely be focused on more complex hybrid systems that enable combining the strengths of text-based and knowledge-based systems to maximize efficiency and accuracy. These systems will draw from semi-structured data enabling the future of the field to potentially become quintessential AI systems with fantastic question answering abilities.

REFERENCES

[1] C. Paris,(1985)"Towards More Graceful Interaction: A Survey of Question-Answering Programs", Columbia University Computer Science Technical Reports. https://doi.org/10.7916/D8765PBX.

[2] B. Green, A. Wolf, C. Chomsky, and K. Laughery, (1961)"BASEBALL: An automatic question answerer",Proceedings of Western Joint IRE-AIEE-ACM Computing Conference, Los Angeles, CA. https://doi.org/10.1145/1460690.1460714.

[3] W. Woods, (1973)"Progress in Natural Language Understanding: An Application to Lunar Geology",Proceedings of the National Conference of the American Federation of Information Processing Societies. https://doi.org/10.1145/1499586.1499695.

[4] D.Ferrucci, E. Brown, J. Chu-Carroll, J. Fan, D. Gondek, A. Kalyanpur, A. Lally, J. Murdock, E. Nyberg, J. Prager, N.Schlaefer, and C. Welty,(2010) "Building Watson: An Overview of the DeepQA Project",AI Magazine, 31(3), 59-79. https://doi.org/10.1609/aimag.v31i3.2303.

[5] A. Allam and M. Haggag, (2012) "Question Answering Systems: A Survey", International Journal of Research and Reviews in Information Sciences (IJRRIS), 2,(3). https://doi.org/10.1016/j.procs.2015.12.005.

[6] F. Zhu, W. Lei, C. Wang, J. Zheng, S.Poria, and T. Chua, (2021)"Retrieving and Reading: A Comprehensive Survey on Open-Domain Question Answering".arXiv: 2101.00774v2. Version 2.

[7] M. Yu, W. Yin, K. Hasan, C. Santos, B. Xiang, and B. Zhou, (2017) "Improved Neural Relation Detection for Knowledge Base Question Answering", Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),Association for Computational Linguistics, pages 571-581. http://dx.doi.org/10.18653/v1/P17-1053.

[8] E.Dimitrakis, K.Sgontzos, and Y.Tzitzikas, (2020)"A survey on question answering systems over linked data and documents", Journal of Intelligent Information Systems, 55, 233–259. https://doi.org/10.1007/s10844-019-00584-7.

[9] Y. Mao, P. He, X. Liu, Y. Shen, J. Gao, J. Han, and W. Chen, (2020) "Generation-Augmented Retrieval for Open-domain Question Answering". arXiv: 2009.08553v2.

[10] D. Moldovan, S.Harabagiu, M.Pasca, R.Mihalcea, R. Goodrum, R.Girju, and V. Rus, (1999)"Lasso: A Tool for Surfing the Answer Net",Proceedings of the Eighth Text Retrieval Conference (TREC-8).

[11] O. Ferret, B. Grau, M.Hurault-Plantet, G. Illouz, L. Monceaux, I.Robba, and A. Vilnat,(2001)"Finding an answer based on the recognition of the question focus",TREC.

[12] J. Chu-Carroll, J. Fan, N.Schlaefer, and W.Zadrozny, (2012) "Textual resource acquisition and engineering",IBM Journal of Research and Development, vol. 56, no. 3.4, pp. 4:1-4:1. https://doi.org/10.1147/JRD.2012.2185901.

[13] E. Voorhees, (2002)"Overview of the TREC 2002 Question Answering Track", Proceedings of the Text Retrieval Conference (TREC 2002).

[14] X. Yao, J.Berant, and B.Durme, (2014) "Information extraction or semantic parsing",Proceedings of ACL. http://dx.doi.org/10.3115/v1/W14-2416.

[15] X. Yao and B. Durme, (2014)“Information extraction over structured data: Question answering with freebase”, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers),Association for Computational Linguistics, pages 956–966. https://doi.org/10.3115/v1/P14-1090.

[16] J.Berant and P. Liang, (2014) “Semantic parsing via paraphrasing”, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, pages 1415–1425. http://dx.doi.org/10.3115/v1/P14-1133.

[17] J.Berant and P. Liang, (2015) “Imitation learning of agenda-based semantic parsers”,Transactions of the Association for Computational Linguistics, 3, pages 545–558. http://dx.doi.org/10.1162/tacl_a_00157.

[18] M. Wang,(2006) “A survey of answer extraction techniques in factoid question answering”,Computational Linguistics, 1(1).

[19] Y. Yang, W.Yih, and C. Meek, (2015) “WikiQA: A Challenge Dataset for Open-Domain Question Answering”, Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D15-1237.

[20] J.Berant, A. Chou, R.Frostig, and P. Liang, (2013) “Semantic Parsing on Freebase from Question-Answer Pairs”, Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pages 1533-1544.

[21] A. Talmor and J.Berant, (2018) “The Web as a Knowledge-base for Answering Complex Questions”, Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), NAACL, pages 641-651. http://dx.doi.org/10.18653/v1/N18-1059.

[22] A. Bouziane, D.Bouchiha, N.Doumi, and M.Malki, (2015) “Question Answering Systems: Survey and Trends”, Procedia Computer Science, 73, pages 366-375. https://doi.org/10.1016/j.procs.2015.12.005.

[23] A. Mishra and S. Jain, (2016)“A survey on question answering systems with classification”, Journal of King Saud University - Computer and Information Sciences, Volume 28, Issue 3, pages 345-361. https://doi.org/10.1016/j.jksuci.2014.10.007.

[24] B. Fu, Y.Qiu, C. Tang, Y. Li, H. Yu, and J. Sun, (2020)“A Survey on Complex Question Answering over Knowledge Base: Recent Advances and Challenges”. arXiv: 2007.13069.

[25] E. Choi, H. He, M. Iyyer, M.Yatskar, W.Yih, Y. Choi, P. Liang, and L.Zettlemoyer, (2018) “QuAC: Question Answering in Context”, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pages 2174-2184. http://dx.doi.org/10.18653/v1/D18-1241.

[26] M. Wang, N. Smith, and T.Mitamura, (2007) “What is the Jeopardy Model? A Quasi-Synchronous Grammar for QA”, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing (EMNLP) and Computational Natural Language Learning, Association for Computational Linguistics, pages 22-32.

[27] M.Iyyer, J. Boyd-Graber, L.Claudino, R.Socher, H. Daume´ III, (2014)“A neural network for factoid question answering over paragraphs”, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, pages 633–644. http://dx.doi.org/10.3115/v1/D14-1070.

[28] Z.AbbasianTaeb and S.Momtazi, (2020) “Text-based Question Answering from Information Retrieval and Deep Neural Networks Perspective: A Survey”.arXiv: 2002.06612. Version 2.

AUTHOR

Michael Caballero recently graduated from the Honors Data Science program at the University of California, Berkeley with a Computer Science minor.