A Survey on Dataset Development Techniques for QA Systems*

Aicha. Aggoune[1,2,*,†]

[1]Computer science department, University 8th May 1945, Guelma, Algeria
[2]LabSTIC Laboratory, University 8th May 1945, Guelma, Algeria

Abstract

Question-answering (QA) systems are pivotal in natural language processing, driving advancements in conversational AI, virtual assistants, and automated knowledge retrieval. The quality and structure of datasets play a critical role in the performance, reliability, and adaptability of these systems. This paper presents a comprehensive review of dataset development techniques for QA systems. We classify these techniques into three categories: manual techniques, which are based on expert domain and crowdsourcing, and automatic techniques, which are divided into two classes: knowledge-based methods and machine learning, and innovative techniques by using data augmentation methods. We introduce a comparison of some important datasets for QA systems according to different criteria with a special focus is given to evaluation metrics used to assess dataset quality. The study can guide practitioners in developing robust, high-quality datasets for future QA systems.

Keywords

QA systems, Dataset development, Metrics, Techniques

1. Introduction

Natural language processing (NLP) has seen remarkable advancements in recent years, with question-answering (QA) systems emerging as one of the most impactful applications. QA systems, designed to retrieve precise answers from vast textual information, are now integral to technologies such as search engines, virtual assistants, and knowledge-based systems. The performance of these systems hinges not only on sophisticated algorithms and model architectures but also on the quality and relevance of the datasets used to train them. High-quality datasets provide the essential foundation for these models to understand complex language structures, reason over context, and accurately respond to user queries [1].

Developing robust datasets for QA is a complex and resource-intensive process. Key challenges in dataset development include ensuring data diversity and balancing language complexity. Various techniques have emerged to address these challenges, ranging from traditional manual annotation to innovative method by using data augmentation methods.

This paper aims to provide a comprehensive review of the techniques used in developing datasets for QA systems, focusing on their strengths, limitations, and areas of application. By systematically examining these methods, we seek to illuminate best practices and emerging trends in QA dataset development. Furthermore, this review addresses the importance of dataset validation and quality metrics, highlighting how they contribute to the reliability and effectiveness of QA systems. Ultimately, our goal is to guide researchers and practitioners in creating datasets that better serve the needs of future QA models, fostering continued innovation and performance improvements in the field.

The remainder of this paper is organized as follows: In Section 2, we introduce the theoretical foundations. Section 3 reviews the techniques for dataset development. In Section 4, we present a

Proceedings of the International IAM'24: International Conference on Informatics and Applied Mathematics, December 04–05, 2024, Guelma, Algeria

[*]Corresponding author.

[†]These authors contributed equally.

aggoune.aicha@univ-guelma.dz (Aicha. Aggoune)
0000-0003-2422-2591 (Aicha. Aggoune)

comparison between dataset structures. Section 5. describe the important metrics for Assessing Datasets. Conclusions are drawn in the last section.

2. Theoretical foundations

2.1. Question-Answering systems

Question-answering (QA) systems offer an intuitive interface for querying vast stores of information across diverse data formats, including both structured and unstructured data in natural languages. These systems play a crucial role in transforming raw data into usable knowledge, enabling users to retrieve specific answers to questions rather than sifting through large documents or databases [2]. QA systems are increasingly employed in applications ranging from customer support and virtual assistants to research and education, where they can quickly extract insights from sources such as documents, databases, and even multimedia content.

To operate effectively, QA systems need to handle the variability and complexity of natural language, requiring them to interpret nuanced questions and extract relevant answers accurately. This involves the integration of techniques from fields such as natural language processing (NLP), information retrieval (IR), and machine learning (ML). Additionally, QA systems must accommodate the inherent diversity in question formulations and adapt to different data types, including text documents, tables, knowledge graphs, and multimodal data.

2.2. Closed-domain Question-Answering systems

Closed-domain Question-answering systems (CQA) are specialized to respond to queries within defined subject areas, such as sports, healthcare, education, or entertainment [3]. These systems leverage domain-specific knowledge, often structured in detailed ontologies or databases, to streamline information retrieval and enhance accuracy in answering questions. The focus on a particular domain simplifies the task for natural language processing (NLP) models, as the system can utilize a well-defined vocabulary, set of concepts, and relationships unique to that domain. For example, in a medical QA system, structured knowledge about diseases, symptoms, and treatments can help the system precisely interpret and respond to health-related inquiries.

Unlike closed-domain systems, open-domain QA systems rely on vast, unstructured sources of information, such as large text corpora, encyclopedic databases (like Wikipedia), or even the internet itself, rather than predefined, domain-specific knowledge structures. This allows them to provide answers on diverse subjects, from historical events and scientific concepts to general trivia and current events.

Closed-domain QA systems are specifically tailored to operate in contexts where general-purpose, open-domain solutions may lack the required depth, precision, or contextual understanding [4]. The development of high-quality datasets specifically tailored for QA systems is essential to training models that are reliable, accurate, and generalizable across domains. These datasets need to account for linguistic diversity, context sensitivity, and a wide range of question types, from simple fact-based queries to complex, reasoning-based questions.

3. Techniques of dataset development for CQA systems

A variety of techniques have been developed to construct datasets for question-answering (QA) systems, each designed to address particular challenges in generating comprehensive and high-quality data for training and evaluation purposes. In this survey, we categorize these techniques into three main types: manual methods, automated methods, and innovative approaches.

3.1. Manual methods

Manual Methods refer to dataset creation techniques that rely on human effort for data collection, question generation, and answer annotation [5]. These methods are highly valuable for ensuring data quality, relevance, and contextual accuracy, as they allow human annotators to apply their expertise and judgment in curating the dataset. However, manual methods are often labor-intensive, time-consuming, and costly, especially for large-scale datasets. Human annotators create question-answer pairs based on a given text or knowledge source. Annotators carefully read through documents, extract meaningful information, and formulate questions that can be answered directly from the content [6]. Another method is based on crowdsourcing, which involves outsourcing the task of question and answer generation to a large pool of workers on platforms like Amazon Mechanical Turk or Figure Eight [7]. This approach allows for rapid data collection from a diverse group of contributors.

In specialized fields, such as medicine, law, or finance, domain experts are employed to create or validate question-answer pairs. Their expertise ensures that the information is accurate, contextually relevant, and adheres to domain-specific standards.

3.2. Automated methods

These methods significantly reduce the time and cost required to produce vast amounts of question-answer pairs, making it possible to construct datasets for training and evaluating models on a large scale. Automatic techniques for creating question-answering (QA) datasets can be broadly divided into two main classes: knowledge-based methods and machine learning-based methods.

Knowledge-based methods rely on structured information sources, such as ontologies, knowledge graphs, and databases, to automatically generate question-answer pairs [8]. These methods use predefined rules, templates, and structured data to produce questions and identify corresponding answers.

Machine learning-based methods, especially those using natural language processing (NLP) and deep learning, have transformed QA dataset creation by automating the generation of complex, context-rich question-answer pairs [9]. These methods use trained models to generate or extract questions and answers from unstructured text, offering greater flexibility and adaptability [10].

More advanced automated approaches involve using machine learning models, particularly large pre-trained language models (e.g., GPT-3, BERT, T5), to generate question-answer pairs synthetically [11, 12]. These models are trained on extensive text corpora, enabling them to produce realistic and contextually varied questions based on input content.

3.3. Innovative approaches

In recent years, data augmentation techniques have gained traction as a way to enhance and diversify QA datasets without the need for entirely new data sources. These techniques manipulate existing question-answer pairs to create new, varied versions, expanding the dataset and exposing models to a wider range of language patterns, contexts, and question types [13]. Data augmentation approaches are particularly useful for improving model generalization and robustness, helping QA systems perform better in real-world scenarios [14].

Data augmentation techniques like synonym substitution, paraphrasing, and entity replacement are used to increase dataset size and diversity automatically [15]. By modifying existing question-answer pairs, these methods create variations that expose models to different phrasings and vocabulary without needing new data sources.

4. Comparison between datasets structures

When evaluating QA datasets, it is crucial to consider the structure of the dataset and the type of question-answer (Q&A) pairs it contains. Different datasets follow various organizational structures based on their intended use.

The most existing QA datasets typically consist of pairs of questions and corresponding answers. For example, SQuAD (Stanford Question Answering Dataset): Questions are based on a paragraph, and answers are specific spans of text from the paragraph [16]. TriviaQA: Similar to SQuAD, the dataset contains questions with answers that are directly extracted from documents or web pages [17]. Natural Questions (NQ): Contains questions where answers are extracted from long documents.

Another innovative approach involves query generation from natural language questions. This structure focuses on generating queries that can be used to retrieve answers from a database, knowledge graph, or other structured data sources [18]. This type of dataset emphasizes the process of converting a natural language question into a structured query that can be executed on a structured database or system, such as SQL. WikiSQL [2] is a large-scale dataset for natural language to SQL query generation. It contains questions based on data tables from Wikipedia and includes SQL queries that extract answers from these tables.

More recent work focuses on the generation of Mongo queries from natural questions with the application of three data augmentation techniques: paraphrasing, back translation, and named entity substitution [19]. An extended work aims to generate more complex queries with auto-validation of the augmented data [20].

Query generation-based datasets are a valuable tool for developing information retrieval systems that bridge the gap between natural language and structured data. By converting natural language questions into executable queries (e.g., SQL, SPARQL, MQL), these datasets enable systems to access and retrieve information from sources.

Table 1 outlining key criteria used to assess various datasets for Question-Answering (QA) systems.

Table 1
Review of some popular datasets

Ref	Dataset	Source	Field	Methodology	Data size
[16]	SQuAD	Wikipedia	Diverse	Selection of Articles, Question Generation, Answer Annotation	+100K
[21]	DBPal	Synthetic	Diverse	Generator, data augmentation, Lemmatizer	3 million
[18]	NarratiQA	books	Movies	Data collection, question generation	46,765
[22]	BabiMovie	Wikipedia	Movies	data collection, data structuring, dialog generation, question formulation	10.000
[19]	M2Q2	Mflix	Movies	Creating templates, data augmentation, data revision	88,100
[20]	M2Q2+	Mflix	Movies	Creating templates, data augmentation, auto-validation	100k

5. Metrics for Assessing Datasets

For datasets designed for generative QA, where the model must generate queries in natural language, different metrics are used to evaluate the quality of the generated queries.

Automatic evaluation using BLEU and ROUGE scores: BLEU is a widely recognized metric in the field of machine translation, while ROUGE is commonly used for evaluating text summarization and other natural language generation tasks. A higher score of these metrics indicates greater similarity and thus a more accurate translation.

BLEU is a widely recognized metric in the field of machine translation [23], while ROUGE is commonly used for evaluating text summarization and other natural language generation tasks [23]. A higher score of these metrics indicates greater similarity and thus a more accurate translation.

$$BLEU = BP \times exp \sum_{n=1}^{N} (w_n log P_n) \,. \tag{1}$$

Where:

- N is the maximum n-gram size (usually up to 4).
- Pn is the precision for n-grams.
- Wn is the weight assigned to the precision, usually set to 1/N
- BP (Brevity Penalty) adjusts the score for the shorter translations.

ROUGE evaluates the n-gram overlap between the output summary and one or more reference summaries [24]. The following formula of ROUGE measure:

$$ROUGE = \frac{ROUGE\_N}{m} + \frac{ROUGE\_L}{m} + \frac{ROUGE\_S}{m}. \tag{2}$$

Where:

$$ROUGE\_N = \frac{Total\ number\ of\ unigrams}{Number\ of\ overlapping\ unigrams}. \tag{3}$$

$$ROUGE\_L = \frac{\sum_{ref\ summaries}(longest\_common\_sequence)}{\sum_{ref\ summaries}(summary\ length)}. \tag{4}$$

$$ROUGE\_S = \frac{\sum_{ref\ summaries}\sum_{skip\ bigram}(count\ match(skip)}{\sum_{ref\ summaries}\sum_{skip\ bigram}(count(skip\ bigram))}. \tag{5}$$

METEOR (Metric for Evaluation of Translation with Explicit ORdering) [25]: Evaluates text generation based on synonyms, stemming, and word order. It is more flexible than BLEU and rewards synonyms and paraphrased text. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

The METEOR score is calculated as follows:

$$METEOR = F_{\text{mean}} \times (1 - \text{Penalty}). \tag{6}$$

where:

Harmonic Mean of Precision and Recall:

$$F_{\text{mean}} = \frac{10 \cdot \text{Precision} \cdot \text{Recall}}{9 \cdot \text{Precision} + \text{Recall}}, \tag{7}$$

$$\text{Penalty} = \gamma \cdot \left(\frac{\text{chunks}}{\text{matches}}\right)^{\beta}, \tag{8}$$

matches: Total number of matched unigrams,

chunks: Groups of matches in the same order,

γ and β: Tunable parameters to control the penalty's impact (default values are usually $\gamma = 0.5$ and $\beta = 3.0$).

Finally, a key metric is how well a model performs on the dataset: Training Loss/Accuracy: These metrics reflect how well the model learns from the dataset during training. A lower loss and higher accuracy indicate a model that fits the data well.

A low training loss and high accuracy on tasks like extractive QA or question answering from a knowledge base suggest that the dataset is well-constructed and provides enough relevant information.
A low training loss and high accuracy on tasks like extractive QA or question answering from a knowledge base suggest that the dataset is well-constructed and provides enough relevant information.

6. Conclusion

Various techniques for dataset creation and validation in the field of question-answering (QA) systems. These techniques are essential for advancing the effectiveness of QA systems across multiple domains and ensuring that they can handle a diverse set of questions and answer types. this survey offers valuable insights into the diversity of datasets available for training and evaluating QA systems. The datasets reviewed here span a wide range of domains, question types, and answer formats, each designed to address specific challenges in QA. While progress has been made in creating large-scale, diverse, and specialized datasets, challenges related to scalability, dataset quality, and domain generalization remain. As QA systems continue to evolve, the development of new datasets and evaluation metrics will play a crucial role in advancing the capabilities of these systems, allowing them to handle increasingly complex tasks in real-world applications.

Declaration on Generative AI

During the preparation of this work, the author used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] P. Wang, T. Shi, C. K. Reddy, Text-to-sql generation for question answering on electronic medical records, in: Proceedings of The Web Conference 2020, 2020, pp. 350–361.

[2] V. Zhong, C. Xiong, R. Socher, Seq2sql: Generating structured queries from natural language using reinforcement learning, arXiv preprint arXiv:1709.00103 (2017).

[3] J. Qi, J. Tang, Z. He, X. Wan, Y. Cheng, C. Zhou, X. Wang, Q. Zhang, Z. Lin, Rasat: Integrating relational structures into pretrained seq2seq model for text-to-sql, arXiv preprint arXiv:2205.06983 (2022).

[4] J. A. Alzubi, R. Jain, A. Singh, P. Parwekar, M. Gupta, Cobert: Covid-19 question answering system using bert, Arabian journal for science and engineering 48 (2023) 11003–11013.

[5] G. Paulin, M. Ivasic-Kos, Review and analysis of synthetic dataset generation methods and techniques for application in computer vision, Artificial intelligence review 56 (2023) 9221–9265.

[6] S. Jin, Y. Wang, S. Liu, Y. Zhang, W. Gu, Optimizing dataset creation: A general purpose data filtering system for training large language models (2024).

[7] S. Lipping, P. Sudarsanam, K. Drossos, T. Virtanen, Clotho-aqa: A crowdsourced dataset for audio question answering, in: 2022 30th European Signal Processing Conference (EUSIPCO), IEEE, 2022, pp. 1140–1144.

[8] A. Pereira, A. Trifan, R. P. Lopes, J. L. Oliveira, Systematic review of question answering over knowledge bases, IET Software 16 (2022) 1–13.

[9] S. U. Amin, A. Hussain, B. Kim, S. Seo, Deep learning based active learning technique for data annotation and improve the overall performance of classification models, Expert Systems with Applications 228 (2023) 120391.

[10] K. Nassiri, M. Akhloufi, Transformer models used for text-based question answering systems, Applied Intelligence 53 (2023) 10602–10635.

[11] K. M. Hossen, M. N. Uddin, M. Arefin, M. A. Uddin, Bert model-based natural language to nosql query conversion using deep learning approach, International Journal of Advanced Computer Science and Applications 14 (2023).

[12] A. Tola, Towards User-Friendly NoSQL: A Synthetic Dataset Approach and Large Language Models for Natural Language Query Translation, Ph.D. thesis, Politecnico di Torino, 2024.

[13] J. Chen, D. Tam, C. Raffel, M. Bansal, D. Yang, An empirical survey of data augmentation for

limited data learning in nlp, Transactions of the Association for Computational Linguistics 11 (2023) 191–211.

[14] X. Chen, Y. Zhang, J. Deng, J.-Y. Jiang, W. Wang, Gotta: generative few-shot question answering by prompt-based cloze data augmentation, in: Proceedings of the 2023 SIAM International Conference on Data Mining (SDM), SIAM, 2023, pp. 909–917.

[15] L. F. A. O. Pellicer, T. M. Ferreira, A. H. R. Costa, Data augmentation techniques in natural language processing, Applied Soft Computing 132 (2023) 109803.

[16] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for squad, arXiv preprint arXiv:1806.03822 (2018).

[17] M. Joshi, E. Choi, D. S. Weld, L. Zettlemoyer, Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, arXiv preprint arXiv:1705.03551 (2017).

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

[19] A. Aggoune, Z. Mihoubi, M2q2: A text-to-mql dataset for movie qa systems, in: Proceedings of the 6th Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI), 2024, pp. 10–18.

[20] A. Aggoune, Z. Mihoubi, Towards efficient dataset development: A case study of m2q2+ in movie qa systems, in: Proceedings of the the 6th Edition of the International Conference on Advanced Aspects of Software Engineering (ICAASE), 2024, pp. 15–22.

[21] N. Weir, P. Utama, A. Galakatos, A. Crotty, A. Ilkhechi, S. Ramaswamy, R. Bhushan, N. Geisler, B. Hättasch, S. Eger, et al., Dbpal: A fully pluggable nl2sql training pipeline, in: Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data, 2020, pp. 2347–2361.

[22] S. Blank, F. Wilhelm, H.-P. Zorn, A. Rettinger, Querying nosql with deep learning to answer natural language questions, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 2019, pp. 9416–9421.

[23] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[24] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004, pp. 74–81.

[25] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.