Paper Review of "Reading Wikipedia to Answer Open-Domain Questions"

Alexandra Ioana Neagu
TU Delft
neagu.alexandra0206@gmail.com

Abstract

In this review, I evaluated a paper on machine reading at scale (MRS), ***Reading Wikipedia to Answer Open-Domain Questions*** (Chen et al., 2017), that utilizes Wikipedia as the primary knowledge source for open-domain question answering (QA). The paper proposes a comprehensive system called DrQA, consisting of a Document Retriever and a Document Reader, which together tackle the challenge of MRS. Key strengths of the paper include its innovative approach, thorough experimental evaluation, and clear presentation of results. However, there are areas for improvement, such as the need for further exploration into end-to-end training and aggregation of information across documents. Despite this, the work presents a significant contribution to the field of natural language processing and demonstrates promising results for open-domain QA tasks.

1 What is this paper about, and what contributions to NLP does it make?

The paper introduces a method for open-domain question answering (QA) using Wikipedia as the sole knowledge source. This approach is named "machine reading at scale" (MRS), as it combines the challenges of document retrieval with machine comprehension of text. The proposed system, DrQA, includes two main components: the **document retriever**, which utilizes bigram hashing and TF-IDF matching to efficiently retrieve relevant articles from Wikipedia based on a given question, and the **document reader**, a multi-layer recurrent neural network (RNN) trained to identify answer spans within the retrieved documents. The system is evaluated using multiple existing QA datasets, demonstrating that both components are highly competitive with existing counterparts and that multitask learning and distant supervision improve overall performance.

In my opinion, the main contribution of the paper is DrQA, an integrated QA system that combines document retrieval with machine comprehension to handle the vast and unstructured information in Wikipedia. This integration addresses the dual challenge of locating relevant documents and accurately extracting answers, which is critical for open-domain QA tasks. The **document retriever** component is notable for its efficiency, which is crucial for scaling QA systems to handle large knowledge bases like Wikipedia. The **document reader** leverages an RNN to achieve state-of-the-art results in machine comprehension tasks, specifically on benchmarks like SQuAD. This demonstrates the model's ability to deeply understand and process natural language text to extract precise answers. By using Wikipedia as the only knowledge base, the paper emphasizes the potential of leveraging large, unstructured datasets for QA. This approach contrasts with other systems that rely on multiple sources, highlighting the importance of precise information retrieval and comprehension from a single, extensive dataset. Lastly, the paper also contributes to the methodology of training QA systems. It shows that multitask learning and distant supervision can enhance system performance across various datasets, suggesting a more generalized and robust approach to training QA models.

1.1 Paper summary

The *related work* section outlines the evolution of open-domain QA. It originated from the need to find answers in collections of unstructured documents, as seen in the TREC competitions. With the advent of knowledge bases (KBs) like Freebase, QA has shifted towards using structured data, but limitations in KBs, such as incompleteness, have led researchers back to using raw text for QA. Machine comprehension of text has advanced significantly due to deep learning models and new datasets like SQuAD and CNN/Daily Mail. Previous work on using Wikipedia for QA involved combining it with other resources or using it for answer validation, while this paper focuses solely on text comprehension from Wikipedia to emphasize machine reading at scale. Notable QA systems like AskMSR, IBM's DeepQA, and YodaQA utilize both structured and unstructured data, whereas this paper's approach uses Wikipedia exclusively. The paper also explores multitask learning to improve performance across different QA datasets, building on similar efforts in NLP and machine learning.

Section 3 describes the DrQA's approach. It is a system for MRS that comprises two main components: the document Retriever and the Document Reader. The Document Retriever is a non-machine learning-based module that efficiently narrows down the search space by retrieving relevant Wikipedia articles using TF-IDF weighted bag-of-words vectors and bigram hashing. This module processes a question and returns five relevant articles for further analysis. The Document Reader is a neural network model inspired by recent advancements in machine comprehension. It encodes paragraphs using multi-layer bidirectional LSTMs, incorporating various features such as pre-trained Glove word embeddings, exact match indicators, token properties (POS, NER, TF), and aligned question embeddings. The model then predicts the answer span by processing the paragraph and question encodings and using bilinear classifiers to determine the start and end positions of the answer within the text. This two-component system effectively combines document retrieval with advanced machine comprehension to tackle open-domain QA using Wikipedia as the sole knowledge source.

Section 4 talks about the types and sources of data used in the paper's research. The data for the DrQA system includes three main types: Wikipedia as the knowledge source, the SQuAD dataset for training the Document Reader, and three additional QA datasets (CuratedTREC, WebQuestions, and WikiMovies) for testing open-domain QA and evaluating multitask learning and distant supervision. The 2016-12-21 Wikipedia dump is used, consisting of over 5 million articles with all structured data removed. The SQuAD dataset, with 87k training examples, is used to train the Document Reader on machine comprehension tasks. The open-domain QA evaluation uses SQuAD along with CuratedTREC, WebQuestions, and WikiMovies, each offering diverse question types and answer formats. Distant supervision is employed to associate paragraphs with questions in datasets lacking explicit document references, enhancing the training set.

Section 5 details evaluations of the Document Retriever and Document Reader modules individually and their combined performance in DrQA for open-domain QA using Wikipedia. The Document Retriever was tested on various QA datasets, outperforming the Wikipedia Search Engine, especially when using bigram hashing. Okapi BM25 and cosine distance in the word embeddings space performed worse. The Document Reader, evaluated using SQuAD, employs 3-layer bidirectional LSTMs and various features such as word embeddings and exact matches. It achieved 70.0% exact match and 79.0% F1 on the SQuAD test set, outperforming published results. Ablation analysis showed that aligned question embeddings and exact match features are complementary and crucial for performance. Finally, DrQA was tested on open-domain QA using four datasets. Three versions of DrQA were evaluated: trained on SQuAD only, fine-tuned with distant supervision, and multitask trained with all datasets. The multitask model with distant supervision performed best. Despite the challenging task, DrQA showed reasonable performance across all datasets. The performance drop from SQuAD to the full Wikipedia setting highlighted the complexity added by needing to find the correct document.

The paper concludes by suggesting that future improvements could involve training the Document Reader to aggregate information from multiple paragraphs and documents and conducting end-to-end training for both the Document Retriever and Document Reader rather than treating them as separate systems.

2 What strengths does this paper have?

This paper demonstrates several strengths. It effectively integrates various components, including document retrieval, machine comprehension, distant supervision, and multitask learning, to develop a comprehensive system for open-domain QA. This approach allows for a more holistic solution to the task, addressing both search and comprehension aspects. It thoroughly evaluates the performance of each component and the full system across multiple benchmarks, providing detailed results and analysis. This comprehensive evaluation demonstrates the efficacy of the proposed approach and allows for a clear understanding of its strengths and weaknesses. Furthermore, it introduces innovative methodologies such as using Wikipedia as the unique knowledge source and incorporating distant supervision and multitask learning for training. These approaches contribute to the novelty and originality of the work, pushing the boundaries of current research in machine reading at scale.

Lastly, the research has practical implications for real-world applications such as question answering systems. By demonstrating the feasibility and effectiveness of the proposed approach, the paper lays the groundwork for the development of more advanced and accurate QA systems capable of handling large-scale knowledge sources like Wikipedia.

3 What weaknesses does this paper have?

While the paper demonstrates several strengths, it also has some weaknesses. While the paper evaluates the proposed system across multiple benchmarks, it primarily focuses on performance metrics such as exact match and F1 score. However, these metrics may not fully capture the real-world usability and robustness of the system. Additional evaluation on factors such as user satisfaction, response time, and handling of ambiguous queries could provide a more comprehensive understanding of the system's effectiveness. The evaluation is also primarily focused on synthetic benchmarks such as SQuAD, CuratedTREC, WebQuestions, and WikiMovies. While these benchmarks are commonly used in the field, they may not fully capture the real-world complexity and diversity of open-domain QA tasks. Including evaluation on more diverse and challenging datasets could provide a more robust assessment of the system's performance.

Furthermore, while the paper discusses the performance of the proposed system, it does not provide a detailed comparison with existing state-of-the-art approaches in the field. A comparative analysis with other QA systems, especially those utilizing different knowledge sources or methodologies, would help to highlight the strengths and weaknesses of the proposed approach in relation to its counterparts.

The proposed system involves multiple components and sophisticated methodologies such as distant supervision and multitask learning. While these techniques contribute to the effectiveness of the system, they also introduce complexity in implementation and training. Replicating the system and adapting it to new domains or datasets may require significant computational resources and expertise, limiting its accessibility to researchers with limited resources. On top of this, the paper heavily relies on Wikipedia as the sole knowledge source for open-domain QA. While Wikipedia is a vast and comprehensive resource, it may not cover all domains or provide the most up-to-date information. Depending solely on Wikipedia for knowledge may limit the system's ability to handle queries outside the scope of Wikipedia or those requiring real-time information.

Lastly, the paper briefly mentions some limitations, such as the ambiguity of SQuAD questions and the need for additional resources for further improvement. However, a more thorough discussion of the limitations and potential challenges of the proposed approach would provide readers with a better understanding of its constraints and areas for future research and improvement.

4 What are the research questions that are answered in the experimental evaluation?

In the experimental evaluation, the paper addresses several research questions:

1. **RQ1: Performance of Document Retriever**: The experimental evaluation addresses the effectiveness of the Document Retriever module in finding relevant articles given a question. The research question aims to determine whether the proposed approach outperforms existing methods, particularly the Wikipedia Search Engine. Key results include the comparison of retrieval performance metrics, such as the ratio of questions for which the answer appears in the top 5 retrieved pages. The results demonstrate that the proposed approach, especially with bigram hashing, surpasses the performance of the Wikipedia Search Engine.

2. **RQ2: Performance of Document Reader on SQuAD**: Another research question focuses on evaluating the performance of the Document Reader component on the SQuAD dataset for machine comprehension. The evaluation aims to assess the effectiveness of the proposed model in extracting accurate answers from paragraphs. Key results include exact match (EM) and F1 scores on the SQuAD development and test sets. The results show that the proposed model achieves

competitive performance, surpassing existing state-of-the-art systems and achieving high accuracy in answering questions based on the provided paragraphs.

3. **RQ3: Full System Performance in Open-domain QA**: The experimental evaluation investigates the overall performance of the full system, DrQA, for open-domain question answering using multiple datasets. The research question aims to assess how well the integrated system performs across various benchmarks and whether the incorporation of search, distant supervision, and multitask learning improves overall performance. Key results include F1 scores and performance comparisons between different versions of DrQA. The results indicate that the DrQA system provides reasonable performance across different datasets, with the multitask learning approach yielding the best overall results.

5 Can you think of how you would improve the work presented in this paper?

Following the authors' suggestions for future research avenues, I likewise believe that the following improvements are worth being investigated:

1. **RQ1: How would end-to-end training across the Document Retriever and Document Reader modules impact the performance of the DrQA system for open-domain question answering?** Currently, the two components are trained independently. End-to-end training could potentially enhance the integration between the retrieval and comprehension stages, leading to improved overall performance by allowing the system to learn more directly from the interaction between these components.

2. **RQ2: How would incorporating multi-document understanding directly into the training process affect the performance of the Document Reader component?** Currently, the Document Reader aggregates over multiple paragraphs and documents during inference, but it trains on paragraphs independently. Training the model to understand and extract information from multiple documents simultaneously could enhance its ability to handle complex queries that require information synthesis from multiple sources, thereby improving the system's performance in open-domain question answering tasks.

By addressing these research questions, we could potentially enhance the versatility and performance of DrQA across a wider range of data sources, making it more applicable to real-world applications in various fields.

6 How reproducible is the work?

I would rate the reproducibility of this work as a **2 out of 4**. While researchers and practitioners could possibly reproduce the results described here with some difficulty, there are challenges due to the absence of code provided in the paper (Raff, 2019). The paper offers detailed descriptions of the datasets used, the methods employed in the experiments, and the evaluation metrics. However, without access to the specific implementation details or parameter settings, replicating the results may require additional effort and experimentation.

7 How confident are you in your assessment of this paper?

I would rate my confidence in the assessment of this paper as a **4 out of 5**. I am quite sure about the key points I've evaluated, and I've made efforts to check important aspects of the paper carefully. However, there is a possibility that I might have missed some details or didn't fully understand some central points, particularly regarding the novelty of the work. Overall, I feel confident in my evaluation, but I acknowledge the potential for overlooked nuances.

References

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading wikipedia to answer open-domain questions. *arXiv preprint arXiv:1704.00051*.

Edward Raff. 2019. A step toward quantifying independently reproducible machine learning research. *Advances in Neural Information Processing Systems*, 32.