Towards Explainable Question Answering (XQA)

Saeedeh Shekarpour,[1] **Faisal Alshargi,**[2] **Mohammadjafar Shekarpour**

[1] University of Dayton, Dayton, United States
[2] University of Leipzig, Leipzig, Germany
sshekarpour1@udayton.org, alshargi@informatik.uni-leipzig.de, mj.shekarpour@gmail.com

Abstract

The increasing rate of information pollution on the Web requires novel solutions to tackle that. Question Answering (QA) interfaces are simplified and user-friendly interfaces to access information on the Web. However, similar to other AI applications, they are black boxes which do not manifest the details of the learning or reasoning steps for augmenting an answer. The Explainable Question Answering (XQA) system can alleviate the pain of information pollution where it provides transparency to the underlying computational model and exposes an interface enabling the end-user to access and validate provenance, validity, context, circulation, interpretation, and feedbacks of information. This position paper sheds light on the core concepts, expectations, and challenges in favor of the following questions (i) What is an XQA system?, (ii) Why do we need XQA?, (iii) When do we need XQA? (iv) How to represent the explanations? (iv) How to evaluate XQA systems?

Introduction

The increasing rate of information pollution [1–4] on the Web requires novel solutions to tackle. In fact there major deficiencies in the area of computation, information, and Web science as follows: (i) Information disorder on the Web: content is shared and spread on the Web without any accountability (e.g., bots [6–9] or manipulative politicians [10] posts fake news). The misinformation is easily spread on social networks [11]. Although tech companies try to identify misinformation using AI techniques, it is not sufficient [12–14]. In fact, the root of this problem lies in the fact that the Web infrastructure might need newer standards and protocols for sharing, organizing and managing content (ii) The incompetence of the Information Retrieval (IR) and Question Answering (QA) models and interfaces: the IR systems are limited to the bag-of-the-words semantics and QA systems mostly deal with factoid questions. In fact, they fail to take into account the other aspects of the content such as provenance, context, temporal and locative dimensions and feedbacks from the crowd during the spread of content. In addition, they fail to 1) provide transparency about their exploitation and ranking mechanisms, 2) discriminate trustworthy content and sources from untrustworthy ones, 3) identify manipulative or misleading context, and 4) reveal provenance.

Question Answering (QA) applications are a subcategory of Artificial Intelligence (AI) applications where for a given question, an adequate answer(s) is provided to the end-user regardless of concerns related to the structure and semantics of the underlying data. The spectrum of QA implementations varies from statistical approaches (Shekarpour, Ngomo, and Auer 2013; Shekarpour et al. 2015), deep learning models (Xiong, Merity, and Socher 2016; Shekarpour, Ngomo, and Auer 2013) to simple rule-based (i.e., template-based) approaches (Unger et al. 2012; Shekarpour et al. 2011). Also, the underlying data sets in which the answer is exploited might range from Knowledge Graphs (KG) holding a solid semantics as well as structure to unstructured corpora (free text) or consolidation of both. Apart from the implementation details and the background data, roughly speaking, the research community introduced the following categories of QA systems:

- **Ad-hoc QA:** advocates simple and short questions and typically relies on one single KG or Corpus.
- **Hybrid QA:** requires federating knowledge from heterogeneous sources (Bast et al. 2007).
- **Complex QA:** deals with complex questions which are long, and ambiguous. Typically, to answer such questions, it is required to exploit answers from a hybrid of KGs and textual content (Asadifar, Kahani, and Shekarpour 2018).
- **Visualized QA:** answers texual questions from images (Li et al.).
- **Pipeline-based QA:** provides automatic integration of the state-of-the-art QA implementations (Singh et al. 2018b,a).

A missing point in all types of QA systems is that in case of either success or failure, they are silent to

AAAI Fall 2020 Symposium on AI for Social Good.

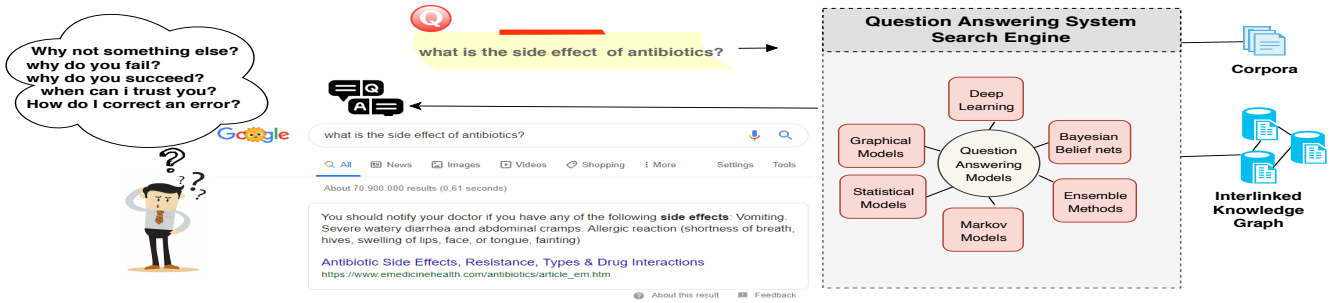

Figure 1: The existing QA systems are a black box which do not provide any explanation for their inference.

the question of why? Why have been a particular answer chosen? Why were the rest of the candidates disregarded? Why did the QA system fail to answer? whether it is the fault of the model, quality of data, or lack of data? The truth is that the existing QA systems similar to other AI applications are a black box (see Figure 1) meaning they do not provide any supporting fact (explanation) about the represented answer with respect to the trustworthiness rate to the source of information, the confidence/reliability rate to the chosen answer, and the chain of reasoning or learning steps led to predict the final answer. For example, Figure 1 shows that the user sends the question `'what is the side effect of antibiotics?'` to the QA system. If the answer is represented in a way similar to the interface of Google, then the end-user might have a mixed feeling as to whether s/he can rely on this answer or how and why such an answer is chosen among numerous candidates?

The rising challenges regarding the credibility, reliability, and validity of the state-of-the-art QA systems are of high importance, especially on critical domains such as life-science involved with human life. The **Explainable Question Answering (XQA) systems** are an emerging area which tries to address the shortcomings of the existing QA systems. The recent article (Yang et al. 2018) published a data set containing pairs of question/answer along with the supporting facts of the corpus where an inference mechanism over them led to the answer. Figure 2 is an example taken from the original article (Yang et al. 2018). The assumption behind this data set is that the questions require multi-hops to conclude the answer, which is not the case all the time. Besides, this kind of representations might not be an ideal form for XQA; for example, whether representing solely the supporting facts is sufficient? how reliable are the supporting facts? Who published them? And how credible is the publisher? And furthermore, regarding the interface, is not the end-user overwhelmed if s/he wants to go through all the supporting facts? Is not there a more user-friendly approach for representation?

The XQA similar to all applications of Explainable AI (XAI) is expected to be transparent, accountable and fair (Sample). If QA is biased (bad QA), it will come

Paragraph A, Return to Olympus:
[1] *Return to Olympus is the only album by the alternative rock band Malfunkshun.* [2] *It was released after the band had broken up and after lead singer Andrew Wood (later of Mother Love Bone) had died of a drug overdose in 1990.* [3] Stone Gossard, of Pearl Jam, had compiled the songs and released the album on his label, Loosegroove Records.

Paragraph B, Mother Love Bone:
[4] *Mother Love Bone was an American rock band that formed in Seattle, Washington in 1987.* [5] The band was active from 1987 to 1990. [6] *Frontman Andrew Wood's personality and compositions helped to catapult the group to the top of the burgeoning late 1980s/early 1990s Seattle music scene.* [7] *Wood died only days before the scheduled release of the band's debut album, "Apple", thus ending the group's hopes of success.* [8] The album was finally released a few months later.

Q: What was the former band of the member of Mother Love Bone who died just before the release of "Apple"?
A: Malfunkshun
Supporting facts: 1, 2, 4, 6, 7

Figure 2: An example from (Yang et al. 2018) where the supporting facts necessary to answer the given question **Q** are listed.

up with discriminating information which is biased based on race, gender, age, ethnicity, religion, social or political rank of publisher and targeted user (Buranyi 2017). (Gunning 2017) raises six fundamental competency questions regarding XAI as follows:

1. Why did the AI system do that?
2. Why did not the AI system do something else?
3. When did the AI system succeed?
4. When did the AI system fail?
5. When does the AI system give enough confidence in the decision that you can trust?
6. How can the AI system correct an error?

In the area of XQA, we adopt these questions; however, we apply sufficient modifications as follows:

1. Why did the QA system choose this answer?
2. Why did not the QA system answer something else?
3. When did the QA system succeed?

4. When did the QA system fail?
5. When does the QA system give enough confidence in the answer that you can trust?
6. How can the QA system correct an error?

This visionary paper introduces the core concepts, expectations and challenges in favor of the questions (i) What is an Explainable Question Answering (XQA) system?, (ii) Why do we need XQA?, (iii) When do we need XQA? (iv) How to represent the explanations? (iv) How to evaluate XQA systems? In the following sections, we address each question respectively.

What is XQA?

To answer the question of **What is XQA?**, we feature two layers i.e., **model** and **interface** for XQA similar to XAI (Gunning 2017). Figure 3 shows our envisioned plan for XQA where at the end, the end user confidently conclude that he can/cannot trust to the answer. In the following, we present a formal definition of XQA.

Definition 1 (Explainable Question Answering) *XQA is a system relying on an **explainable computational model** for exploiting the answer and then utilizes an **explainable interface** to represent the answer(s) along with the explanation(s) to the end-user.*

This definition highlights two major components of XQA as (i) **explainable computational model** and (ii) **explainable interface**. In the following we discuss these two components in more details:

Explainable Computational Model. Whatever computational model employed in XQA system, (e.g., learning-based model, schema-driven approach, reasoning approach, heuristic approach, rule-based approach, or a mixture of various models) it has to explain all intermediate and final choices meaning the rationale behind the decisions should be **transparent**, **fair**, and **accountable** (Sample). The responsible QA system distinguishes misinformation, disinformation, mal-information, and true facts (Wardle and Derakhshan 2017). Furthermore, it cares about the untrustworthiness and trustworthiness of data publisher, information representation, updated or outdated information, accurate or inaccurate information, and also the interpretations that the answer might raise. Whereas, the fair QA system is not biased based on the certain characteristics of the data publisher, or the targeted end user (e.g., region, race, social or political rank). Finally, the transparency of QA systems refers to the availability and accessibility to the reasons behind the decisions of the QA system in each step upon the request of involving individuals (e.g., end user, developer, data publisher, policymakers).

Explainable Interface. The explainable interface introduced in (Gunning 2017) contains two layers (i) a cognitive layer and (ii) an explanation layer. The cognitive layer represents the implications learned from the computational model in an explainable form (abstractive or summarized representation), and then the explanation layer is responsible for delivering them to the end user in an interactive mode. We introduce several fundamental features which the future generation of XQA have to launch. We extensively elaborate on our view about the interface in Section 5.

Why do we need XQA?

We showcase the importance of having XQA using the two following arguments.

Information Disorder Era. The growth rate of mis-, dis-, mal- information on the Web is getting dramatically worsened (Wardle 2018). Still, the existing search engines fail to identify misinformation even where it is highly crucial (Kata 2010). It is expected from the information retrieval systems (either keyword-based search engines or QA systems) to identify mis-, dis-, mal- information from reliable and trustworthy information.

Human Subject Area. Having XQA for areas being subjected to lives particularly human subject is highly important. For example, bio-medical and life-science domains require to discriminate between the hypothetical facts, resulting facts, methodological facts, or goal-oriented facts. Thus XQA has to infer the answer of informational question based on the context of the question as to whether it is asking about resulting facts, or hypothetical facts, etc.

When do we need XQA?

Typically in the domains that the user wants to make a decision upon the given answer, XQA matters since it enables the end user to make a decision with trust. There are domains that traditional QA does not hurt. For example, if the end user is looking for the `'nearby Italian restaurant'`, QA systems suffice. On the contrary, in the domain of health, having the explanations is demanding otherwise the health care providers can not entirely rely on the answers disposed by the system.

How to represent explanations?

We illustrate the life cycle of information on the Web in Figure 4 which can be published as a stack of the metadata. Each piece of information has a publishing source. Further, genuine information might be framed or manipulated in a context. Then, the information might be spread on social media. Concerning its circulation on social media or the Web, it might be annotated or commented on by the crowd.

We feature the explainable QA interface with respect to its life cycle as it should enable the end-user to 1) access context, 2) find the provenance of information, 3)

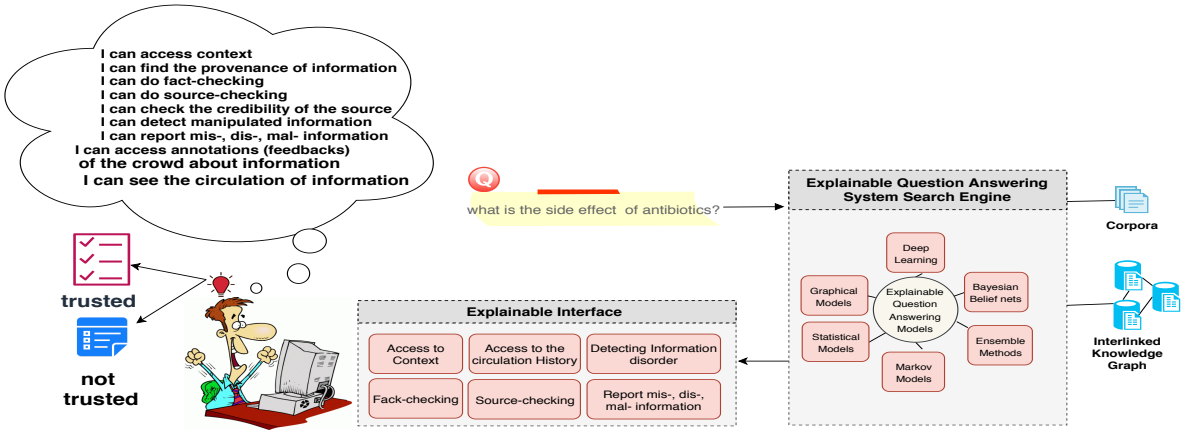

Figure 3: The explainable question answering exposes explainable models and explainable interface; then the user can make a decision as to whether to trust or not.

Figure 4: The life cycle of information on the Web.

do fact-checking, 4) do source-checking, 5) check the credibility of the source, 6) detect manipulated information, 7) report mis-, dis-, mal- information, 8) access annotations (feedbacks) of the crowd, 9) reveal the circulation history.

How to evaluate XQA systems?

The evaluation of an XQA system has to check its performance from both qualitative and quantitative perspectives. The Human-Computer Interaction (HCI) community already targeted various aspects of the human-centered design and evaluation challenges of black-box systems. However, the QA systems received the least attention comparing to other AI applications such as recommender systems. Regarding XQA, the qualitative measures can be (i) **adequate justification:** thus the end user feels that she is aware of the reasoning steps of the computational model, (ii) **confidence:** the user can trust the system, and place the willing for the continuation of interactions, (iii) **understandability:** educates the user as how the system infers or what are the causes of failures and unexpected answers, and (iv) **user involvement:** encourages the user to engage in the process of QA such as question rewriting. On the other hand, the quantitative measures are concerned with the questions such as *"How effective is the approach for generating explanations?"*. For example, it measures the effectiveness in terms of the preciseness of the explanations. However, this area is still an open research area that requires the research community introduce metrics, criteria, and benchmarks for evaluating various features of XQA systems.

Conclusion

In this paper, we discussed the concepts, expectations, and challenges of XQA. The expectation is that the future generation of QA systems (or search engines) rely on computational explainable models and interact with the end-user via the explainable user interface. The explainable computational models are transparent, fair and accountable. Also, the explainable interfaces enable the end-user to interact with features for source-checking, fact-checking and also accessing to context and circulation history. In addition, the explainable interfaces allow the end-user to report mis-, dis-, mal- information.

We are at the beginning of a long-term agenda to mature this vision and furthermore provide standards and solutions. The phenomena of information pollution is a dark side of the Web which will endanger our society, democracy, justice service and health care. We hope that the XQA will be the attention of the research community in the next couple of years.

References

Asadifar, S.; Kahani, M.; and Shekarpour, S. 2018. HCqa: Hybrid and Complex Question Answering on Textual Corpus and Knowledge Graph. *CoRR* abs/1811.10986.

Bast, H.; Chitea, A.; Suchanek, F.; and Weber, I. 2007. Ester: efficient search on text, entities, and relations. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

Buranyi, S. 2017. Rise of the racist robots – how AI is learning all our worst impulses. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses.

Gunning, D. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* .

Kata, A. 2010. A postmodern Pandora's box: anti-vaccination misinformation on the Internet. *Vaccine* 28(7): 1709–1716.

Li, Q.; Fu, J.; Yu, D.; Mei, T.; and Luo, J. ???? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018*.

Sample, I. ???? Computer says no: why making AIs fair, accountable and transparent is crucial. https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial. Accessed: 2017-11-05.

Shekarpour, S.; Auer, S.; Ngomo, A. N.; Gerber, D.; Hellmann, S.; and Stadler, C. 2011. Keyword-Driven SPARQL Query Generation Leveraging Background Knowledge. In *Proceedings of the 2011 IEEE/WIC/ACM International Conference on Web Intelligence, WI 2011, Campus Scientifique de la Doua, Lyon, France, August 22-27, 2011*.

Shekarpour, S.; Marx, E.; Ngomo, A. N.; and Auer, S. 2015. SINA: Semantic interpretation of user queries for question answering on interlinked data. *J. Web Semant.* .

Shekarpour, S.; Ngomo, A. N.; and Auer, S. 2013. Question answering on interlinked data. In *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*.

Singh, K.; Both, A.; Radhakrishna, A. S.; and Shekarpour, S. 2018a. Frankenstein: A Platform Enabling Reuse of Question Answering Components. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece,Proceedings*.

Singh, K.; Radhakrishna, A. S.; Both, A.; Shekarpour, S.; Lytra, I.; Usbeck, R.; Vyas, A.; Khikmatullaev, A.; Punjani, D.; Lange, C.; Vidal, M.; Lehmann, J.; and Auer, S. 2018b. Why Reinvent the Wheel: Let's Build Question Answering Systems Together. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France*.

Unger, C.; Bühmann, L.; Lehmann, J.; Ngomo, A. N.; Gerber, D.; and Cimiano, P. 2012. Template-based question answering over RDF data. In *Proceedings of the 21st World Wide Web Conference 2012, WWW 2012, Lyon, France, April 16-20, 2012*.

Wardle, C. 2018. DISINFORMATION GETS WORSE. https://www.niemanlab.org/2017/12/disinformation-gets-worse/.

Wardle, C.; and Derakhshan, H. 2017. Information Disorder: Toward an interdisciplinary framework for research and policymaking. https://shorensteincenter.org/information-disorder-framework-for-research-and-policymaking/.

Xiong, C.; Merity, S.; and Socher, R. 2016. Dynamic memory networks for visual and textual question answering. In *International conference on machine learning*.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W. W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*.