Deep Learning Based Question Answering System: A Review

[1]Mohammad Aalam, [2]Prof. Suaib Ahmad, [3]Prof. Mohd Haroon

[1]M.Tech. Scholar, [2]Assistant Professor, [3]Associate Professor
Integral University

***Abstract*: Selecting the answer clause is the task of selecting sentences that contain an answer to a particular question. This is an important problem in itself and in the wider context of answering open-ended questions. In this paper, we examine the various details of the question-answering system and identify the various challenges in developing the system for answering new questions.**

Term Index: Question Answering, Deep Learning, Information Retrieval.

INTRODUCTION

Querying information from structured as well as unstructured data becomes very significant. There are lots of textual data, FAQs, newspapers, articles, documents, user cases, customer service requests, and so on. It is very difficult to take into account all this information. Who won the last football cup? What is Bitquin? When is a famous singer's birthday? To discover some information, we necessitate searching a document and spending a few minutes reading before you find the answer [1].
We classify all questions in the following types:
General questions, with yes / no answers
WH questions, start with: who, what, where, when, why, how, how much
Choice questions, where you have some options within the question Factoid questions, where the full answer can be found within the text. The answer to these
Questions consists of one or several words after another

QUERY RESPONSE APPROACH:

Naturally, a query response system consists of numerous logical building blocks:
Source of data,
Information reclamation (IR) system,
Machine reading conception representation (MR model),
One or more additional blocks, such as modules for text pre-processing, answers post processing, checking, and stabilization.
Interaction of these blocks with each other is illustrated in figure 1:

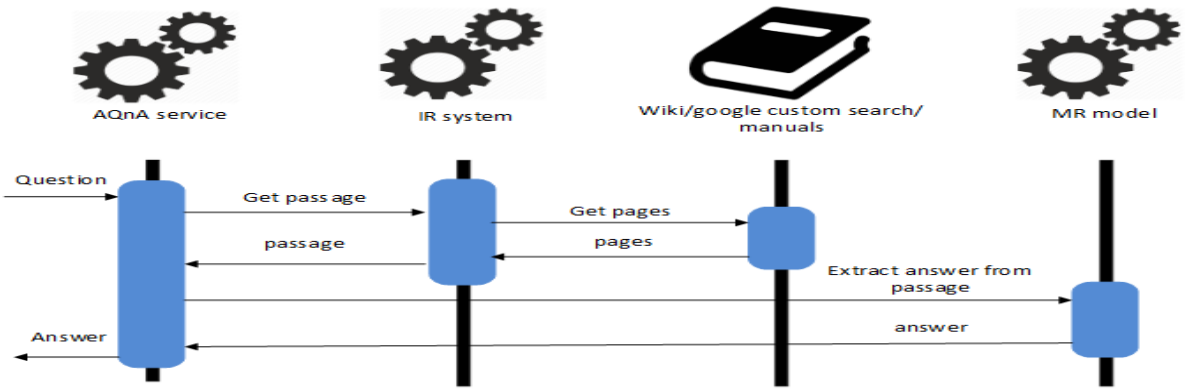

Figure 1: Interaction of blocks of question answering system

As seen in Figure 1, when the query is stirred to the system, the foremoststage to accomplish is to look for texts that can hold the response to the query. Foundation of these texts can comprise diverse web pages, Wikipedia, stored documents on a local disk etc. [2]

The next step is to select the most appropriate text from all the information found. For this purpose, the following methods can be used: TF-IDF similarity, Jaccard index, Word decorations, text deep learning similarity algorithms. All of these methods have pros and cons, different time / memory needs for pre-processing and access to the index of similarity. There is no preeminent approach to all state of affairs. The approach that is best for a concrete task depends on several factors such as domain, data set, and data quantity.

OPEN DATASETS AVAILABLE FOR QUESTION ANSWERING

The Stanford Questionnaire Data Collection (SQAD) [3] is a collection of data for reading comprehension consisting of queries posed by workers on aassortment of Wikipedia articles, where the response to each query is part of the text, or extends from the consequent reading passage, or the query possibly may not be answerable.

WikiQA dataset [4], an available set of question-and-answer pairs available to the public, collected and annotated to search for answers to open-ended questions. It was created using a more natural process and an order of magnitude greater than the previous data set. In addition, the WikiQA dataset also includes questions that have no correct sentences, enabling researchers to raise the answer, a critical element of any QA system.

The TREC-QA dataset includes queries, responses and patterns, as well as a set of documents returned by the participating teams.

The NewsQA dataset [5] is to facilitate the research commune construct algorithms capable of answering questions that require human comprehension and thinking skills. By taking advantage of CNN articles from the Deep Mind Q & A dataset, the authors prepared a dataset to understand the resource collection machine from 120,000 pairs of questions and answers.

TYPES OF QUESTION ANSWERING

There are three main modern models to answer the questions:

a) **IR-BASED FACTOID QUESTION ANSWERING:** The goal of answering IR-based real-world questions is to answer the user's question by finding short text clips on the Internet or other set of documents. In the process of processing questions, a number of information is extracted from the question. The answer type determines the type of entity that consists of the answer (person, location, time, etc.). The query determines which keywords to use for the infrared system for use in document search.

b) **ANSWERING QUESTIONS BASED ON KNOWLEDGE:** It is the idea of answering the natural language question by assigning it to a query via an organized database. Consequently, the logical form of the query is either in the form of a query or can be effortlessly transformed to one. A database can be a complete relational database, or a simpler structured database such as three-time RDF sets. The mapping systems are called from a text string to any logical form of semantic analyzers. The semantic parsers are usually assigned to answer the question either to some calculus or query language such as SQL or SPARQL.

c) Using multiple sources of information: IBM Watson [6] [7] from IBM, which won the Jeopardy honour! The 2011 Challenge is an example of a system that relies on a wide range of resources to answer questions. The first stage is to address the question. The DeepQA system analyzes, labels the named entities, and extracts the relationship to the question. Then, like text-based systems, the DeepQA system extracts the focus and type of answer (also called the LAT) and compiles the questions and the questions section. DeepQA follows the focus of the question. Ultimately, the question by type is classified as a definition question, multiple choice, puzzles, or fill in the blank. The following is the stage of creating a candidate answer by question type, where the question is combined with external documents and other sources of information to suggest many candidate answers. These candidate responses can be extracted from text documents or from structured knowledge bases.It is then passed through the candidate's scoring phase, which uses many sources of evidence to record candidates. One of the most important is the type of lexical answer. In the final merger and consolidation of points, you first combine the answers of the equal candidate. The merge and the order are actually run repeatedly. First, the candidates are sorted according to the workbook, with an approximate initial value for each candidate's answer. This value is then used to determine which variables are from a name to be identified as a combined answer, then the combined answers are rearranged.

1.4 Challenges in Question Answering

The main challenges [8] posed by the question-answer system are outlined below:

1. **LINGUISTIC GAP:**In natural language, the same meaning can be articulated in dissimilar ways. Since the question can usually be answered only if each concept is assigned, filling this gap greatly increases the proportion of questions that can be answered by the system.
2. **VAGUENESS:**It is the phenomenon of the same phrase that has different meanings; it may be structurally and synthetically (eg "flying planes") or lexicon and semantic (eg "bank"). The same string mistakenly refers to different concepts (as in a money bank vs. a riverbank) and polysemy, where the same chain refers to different but related concepts (as in a bank as a company versus a building bank).

MULTILINGUALISM: Knowledge is expressed on the Internet in different languages. While RDF resources can be described in multiple languages at the same time using language tags, there is not one language that is always used in web documents. Additionally, users have different native languages. The quality assurance system is expected to recognize language and get results on the move.

IR-BASED FACTOID QUESTION ANSWERING

The purpose of answering information-based questions is to answer the user's question by finding short text clips on the Internet or other set of documents. Figure 2 shows the three stages of the IRR-based real-time response system: question processing, section retrieval and order, and answer extraction.

RESPONSE PROCESSING

The foremostobjective of the query processing phase is to extract the keywords passed to the IR system to match the potential documents. Some systems derive additional information such as:

Response type: Entity type (person, location, time, etc.).

Focal point: The string of words in query is to be expected to be replaced byany answer string found.

Query type: Is this a question of definition, math question, and question list?

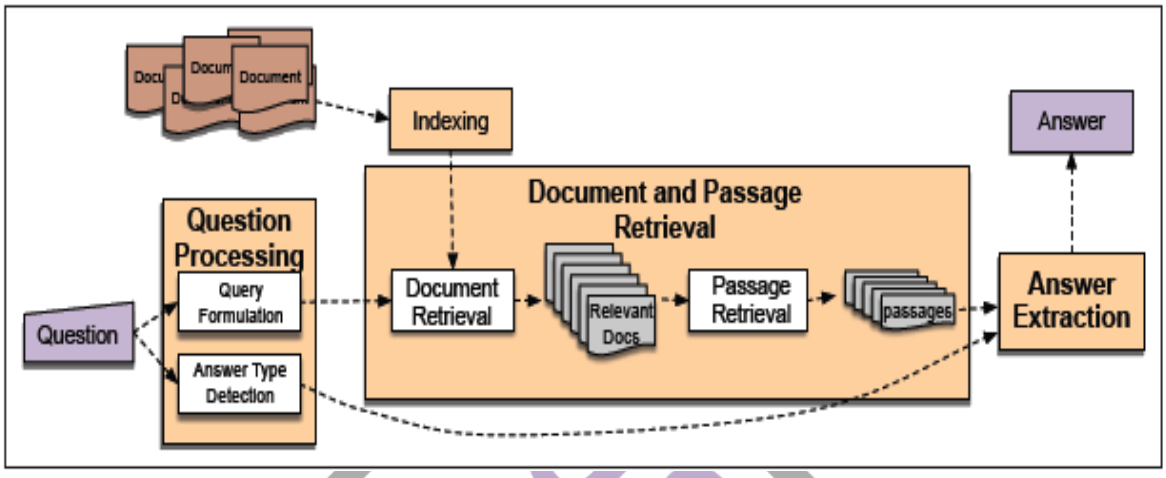

Figure 2: IR-based factoid query answering

QUERY FORMULATION

Query wording is the task of creating a query - a list of symbols - to be sent to the retrieval system to retrieve documents that may contain the answer strings. For responses to queries from the web, we can plainly pass the completequery to a web search engine, at most leaving the question word (where, when, etc.). To answer questions from smaller sets of documents such as company information pages or Wikipedia, we still use the infrared drive to index and search our documents, generally using the tf-idf matching, but we may need to do more processing.

ANSWER TYPES

Some systems use question classification, the task of finding the answer type for the answer question, and the named entity that classifies the answer. A question like "Who founded Virgin Airline?" Expected PERSON response. A question like "What is the Canadian city with the largest population?" You expect an answer of type CITY. If we know that the answer type for a question is a person, we can avoid checking each sentence in the document set, rather than focusing on sentences that remind people.

No	Number of words in questions	No of Keywords	Matching Keywords	Replying Answer
1	8	What, topic, paragraph	What, paragraph	Yes
2	10	Person, who, lost, money	Who, person	No
3	7	How, hand, many	How, hand	Yes
4	12	Nasa, budget, billion	Budget, Nasa	Yes
5	11	Boom, market	boom	No
6	10	What, sold, year	What, sold	Yes
7	9	Find, sequence, paragraph	Paragraph. sequence	Yes
8	20	Medieval, average, human, life, medicine	Medieval, life, human	Yes
9	25	Galileo, who, Roman, church, Stephen, pope	Who, Galileo, Roman	Yes
10	10	Kashmir, place, beautiful, earth	Kashmir, place, earth	Yes
11	23	Web, plant, water, air, home	Web, air, home	No

Table1: answer type and the decision for answer

DOCUMENT AND PASSAGE RETRIEVAL

The resulting IR query is sent from the question processing stage to the IR drive, resulting in a set of documents arranged by relevance to the query. Because most methods of extracting answers are designed to apply to smaller areas such as paragraph paragraphs, quality assurance systems then split upper documents into smaller sections such as sections, paragraphs, or sentences. These may already be fragmented in the source document, or we may need to run a paragraph fragmentation algorithm. The simplest form of clip retrieval is to simply pass all the bass retrievers back to the stage of extracting the answers. The most complex alternative is filtering sections by running a specific entity classification or type of answer to recovered segments. The paragraphs that do not contain the type of answer that is assigned to the question are ignored.

Supervised learning can also be used to categorize the remaining sections in full, using features such as:

• Number of named entities of the correct type in section

• Number of keywords question in section

• The longest accurate sequence of keywords for the question that occurs in the section

• The rank of the document from which the section was extracted

• Near the core words of the original query to each other [9] [10].

• Number of grams n that intersect between section and question [11].

CONCLUSION

In this paper, we have introduced the problem of question answering over linked data. We have further presented the main challenges involved in the task and presented the typical architecture or anatomy of systems addressing the task. We have further provided an overview of state-of-the-artapproaches to the problem, highlighting their features and drawbacks. Finally, we have described a selected set of systems, comprising those developed by ourselves, in more detail and concluded with a summary of open issues that need to be addressed in future research.

References

[1] Manna, Riyanka, ParthaPakray, Somnath Banerjee, Dipankar Das, and Alexander Gelbukh, CookingQA: “A question answering system based on cooking ontology”, In Mexican International Conference on Artificial Intelligence, pp. 67-78. Springer, Cham, 2016.

[2] Shobhit Srivastava, MohdHaroon, AbhishekBajaj :“Web document information extraction using class attribute approach”, 2013 4th International Conference on Computer and Communication Technology (ICCCT), Pages 17-22.Publication date2013/9/20.

[3] R Khan, M Haroon, MS Husain, “Different technique of load balancing in distributed system”: A review paper, 2015 Global Conference on Communication Technologies (GCCT), Pages 371-375, Publication date 2015/4/23

[4] Mohammad Haroon, Mohd Husain, “Interest Attentive Dynamic Load Balancing in distributed systems”, 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom), Pages 1116-1120, Publication date2015/3/11

[5] Wu, Qi, et al. Ask me anything: Free-form visual question answering based on knowledge from external sources, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[6] Xiong, Caiming, Victor Zhong, and Richard Socher. Dynamic coattention networks for question answering." arXiv preprint arXiv:1611.01604, 2016.

[7] Yang, Y., Yih, W.T. and Meek, C., Wikiqa: A challenge dataset for open-domain question answering, In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing pp. 2013–2018.

[8] Trischler, A., Wang, T., Yuan, X., Harris, J., Sordoni, A., Bachman, P. and Suleman, K., Newsqa: A machine comprehension dataset. arXiv preprint arXiv:1611.09830, 2016.

[9] Kalyanpur, A., Patwardhan, S., Boguraev, B.K., Lally, A. and Chu-Carroll, J., Fact-based question decomposition in DeepQA. IBM Journal of Research and Development, 56(3.4), pp.13–1, 2012.

[10] Watson, I. B. M. "What is watson." IBM Watson, 2015.https://ieeexplore.ieee.org/xpl/tocresult.jsp?isnumber=6177717.

[11] Höffner, K., Walter, S., Marx, E., Usbeck, R., Lehmann, J. and NgongaNgomo, A.C.,,Survey on challenges of question answering in the semantic web. Semantic Web, 8(6), pp.895–920, 2017.

[12] Pasca, M., Open-Domain Question Answering from Large Text Collections. CSLI, 2003.

[13] Monz, C., Minimal span weighting retrieval for question answering. In SIGIR Workshop on Information Retrieval for Question Answering, pp. 23–30, 2004.

[14] Brill, E., Dumais, S. T., and Banko, M., An analysis of the AskMSRquestion-answering system. In EMNLP 2002, pp. 257–264, 2002.