Original Research Article

A Hybrid Question Answering System

ABSTRACT

In this study, we propose a hybrid Question Answering (QA) system for Arabic language. The system combines textual and structured knowledge-Base(KB) data for question answering. It make use of other relevant text data, outside the KB, which could enrich the available information. The system consists of four modules. 1) a KB, 2) an online module, and 3) A Text- to-KB transformer to construct our own knowledge base from web texts. Using these modules, we can query two types of information sources: knowledge bases, and web text. Text-to-KB uses web search results to identify question topic entities, map question words to KB predicates, and enhance the features of the candidates obtained from the KB. The system scored f-measure of .495 when using KB. The system performed better with f-measure of .573 when using both KB and Text-to-KB module. The system demonstrates higher performance by combining knowledge base and text from external resources.

Keywords: Question Answering, Information Retrieval, Information Extraction, Knowledge Base

1. INTRODUCTION

Whenever a user needs information about a specific topic, it simply supplies a query to any search engine, e.g. Google. Traditional search engines returns a list of links to documents which may contain the answer. The user has to browse these links and tries to locate the answer. QA systems retrieves specific answers in response to user questions , rather than a lists of links to documents. Two approaches for Question Answering (QA) have evolved: text-centric, and knowledge base-centric. Text-Centric QA systems use collection of text documents to return passages relevant to a user's question and extract candidate answers [1]. The KB-Centric QA systems, which are emerged from the database community, depends on large scale knowledge bases, such as Freebase [2], DBpedia [3], WikiData [4] which store a massive amount of knowledge about various kinds of entities. Knowledge Base Question Answering(KBQA) systems have been classified into two major approaches: semantic parsing, and Information Extraction (IE) [5]. The semantic parsing focuses on understanding the question, and tries to parse sentences into their logical forms (semantic representations)[6, 7, 8]. IE approaches [9, 10, 11] are based on detecting topic entities in the question, and employing predefined templates for mapping the question to predicates, exploring these entities' neighborhood in a KB. Various QA systems based on various information sources have been proposed: QA systems based on KB approach[12][13], non-web based systems [14][15], web-based systems[16][17], machine learning-based systems[17]. QA systems are developing from systems based on Information Retrieval (IR) to ones based on KBs, KBs and IRs. KBQA systems provides very high precision, but requires curated KBs; However, these KBs cannot include all the information that web text can communicate. To overcome this limitation, other information sources besides curated KBs are required. In this paper, we present a hybrid QA system that utilizes multiple information sources: a curated KB and web text. To the best of our knowledge, this work will be the first on Arabic QA that combines both KB and the web text as sources of information.

2. METHODOLOGY

The following figure shows the architecture of the system which consists of Knowledge base, Online module and Text-to-KB(Fig. 1)

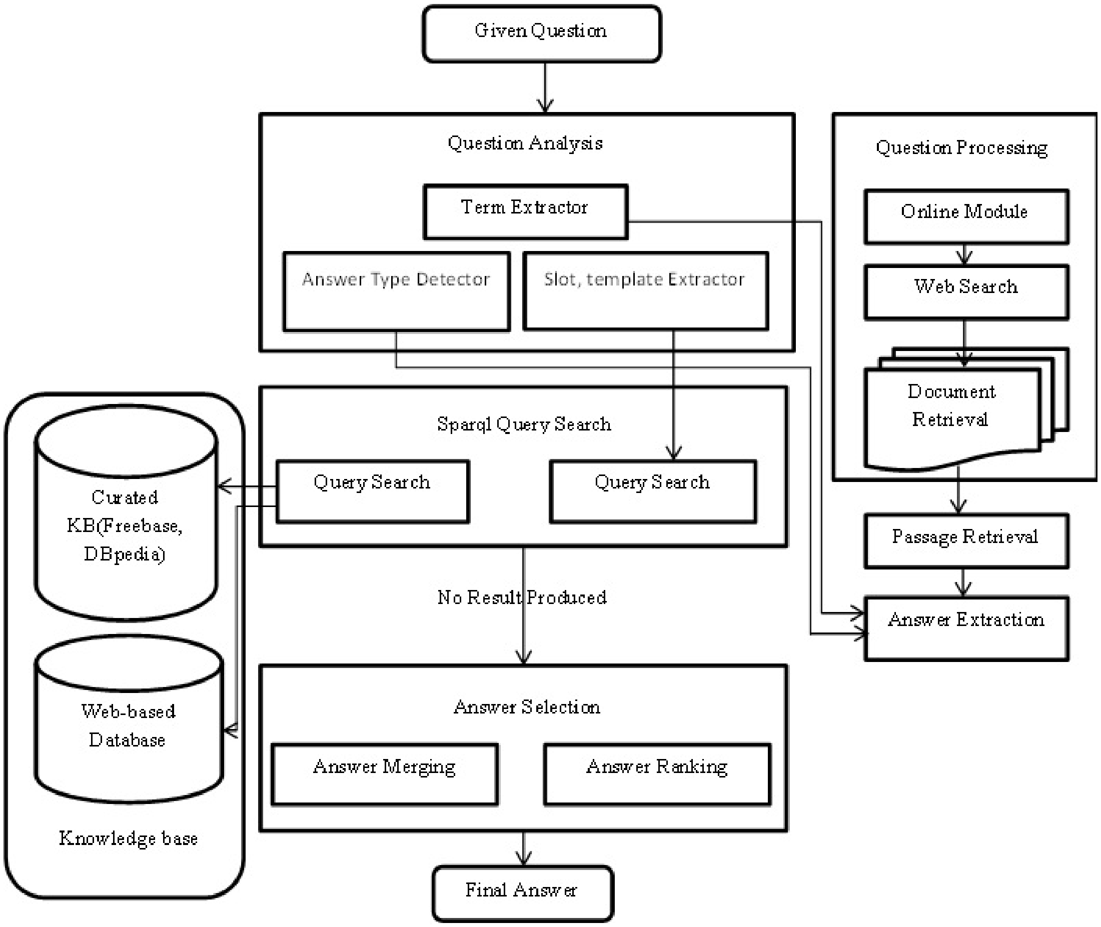

Fig. 1. QA Architecture

2.1 Knowledge Base

A Question Answering system based on KB takes a natural language (NL) question as its input and uses structured KBs like DBpedia to retrieve the answer. A KB-based QA system employs structured information sources, so it generates very specific answers. First the NL question is segmented/tokenized into individual words/tokens(for this we use Stanford Word Segmenter for Arabic[18], the segmenter will split the question into individual tokens, e.g. Given the question: “ما هو أطول نهر بالعالم؟”, it will be segmented into individual tokens, “ما”,”هو”,”أطول”,”نهر”,"في","العالم”); then string based methods are employed and NL phrases(extracted using a small set of hand-crafted rules) along with KB node mapping dictionary are automatically generated to match KB vocabulary to the tokens. We generate query candidates by using a limited set of hand-crafted grammar rules(manually annotated grammar rules based on Arabic syntax) to combine tokens into a single unified representation of meaning.

The major phrases available in Arabic are Noun Phrase(NP), Verb Phrase(VP), Prepositional Phrase(PP). An NP starts with a noun or a pronoun which expresses the entity of person, place, or animal about which the phrase is referring. The nominal sentence consists of “starting” “المبتدأ” which is followed by information” “الخبر” which is the complement of the starting. A VP is one which starts with a verb in any of the three forms (present , past and order verb). The VP consists of verb “الفعل” which is followed by “subject” “الفاعل”. This means that the verb requires no more than the subject to complete the meaning. A PP in Arabic language is used in the same manner of English. It comes in the form of a preposition followed by a noun or a noun phrase. There exist 20 particles "حرف جر" in the Arabic language, they come in the form of one-letter, two-letter and three-letter word groups. The grammar rules to be considered are the rules that are used to identify NP, VP, PP chunk boundaries. Based on the typical grammatical structure of Arabic for NP, the rules to build a noun phrase which are correct grammatically. The following are 7 general rules to get NPs:

NP:{ (<SN>│ <SPN >)* < POSS_PRON >?<ADJ >*}
NP: { <SN> < POSS_PRON >? <ADJ>*<NOUN_PROP >*}
NP :{ (<SPN>│<SN>)* <ADJ >* <SN>?<CD>?< NOUN_PROP >? <ADJ >?}
NP: { < SN >* <DEM_PRON> (< SN >│< SPN>)? <ADJ>*}
NP: {<ADJ> (<SPN>│<SN>) * (<POSS_PRON>│<ADJ >*)? }
NP: {<SN> < POSS_PRON>? <CD>?(<SN>│<SPN>│<NOUN_PROP>)? <ADJ>*}
NP: { < X : POS(X) NP components> (<CC>< Y: POS(Y)= POS(X) >)* }

Two general rules for building a grammatically correct VP are derived.
VP:{ (X: POS (X) {<PAS>, <PRV > , <IV >}) <PPRON> }
VP :{ (W: POS (X) {<PAS>, < PASSV>,<PRV > , <IV >})(Y : POS(Y) NP components>and Y is the last word)}

The third type of chunks is PP and they are defined as a combination of a preposition and a word or phrase, in our case.
PP :{ <PREP > <PPRON>}
PP: {<PREP> <Y: POS (Y) is an NP and Y is the last word >}

These rules are applied to derive the NP, VP, and PP(when available) to construct the query candidates from the question.

In the LSP approach, regular expression patterns that express the POS(for getting the POS tags we use Stanford POS Tagger for Arabic[19], e.g. Given the question:”ما هي عاصمة ألمانيا؟”, the output of the POS tagger will be a tag assigned to each word in the question, ما/WP,هي/PRP, عاصمة/NN, ألمانيا/DTNN) ,lexical or chunk type patterns of a NL question and a SPARQL query template are generated. If a match is found, slots in the SPARQL[20] query template are occupied with the word-matched chunks from NL question. However, there is no context information for KB-based QA modules , and therefore it cannot score/rank its answer candidates; instead KB-based module forwards its answer candidate to an answer merging task in the online module and this module rank the answer candidates.

2.2 Online Module

The online module searches text to find answers. The online module performs four tasks (Fig 1): first is question classification. and; the second is the passage retriever. In question classification, it analyzes the question semantically and identifies the answer type (Table 1) where the answer type is a label generated based on the semantic classification of the question. E.g. the question “Who invented the television?” is classified as “Human:individual”, this means, the answer type that the question is looking for is a name of human(individual).

Table 1. Answer Types: Two Level Taxonomy

Main class/Main answer type	**Sub-class/Sub-answer type**
ABBREV	Abbreviation, explanation(explanation for

	abbreviation)
ENTITY	product, religion, sport, substance, symbol, technique, other, term, vehicle, word , animal, body, color, currency, event, food, instrument, language, letter, plant,
DESCRIPTION	Definition, description, manner, reason
HUMAN	Group, individual, title, description
NUMERIC	Code, count, date, distance, money, order, period, percent, speed, temp, size, weight, other
LOCATION	City, country, mountain, state, other
ORGANIZATION	Organization or institute, group or committee

The passage retriever retrieves relevant passages by segmenting the documents that are related to the user question; the third task is the answer extractor. It extracts answer candidates; the fourth task merges answer candidates from the online module and KB, it then ranks the answer candidates and returns the final list of answers. Context information are used to score answer candidates which are the output of the SPARQL not only from online module answer extraction task. Lexical, syntactic and semantic analysis are employed for question processing, which includes extracting terms by a Support Vector Machines (SVM) [21]. Lucene [22] is utilized for indexing web pages dump and for searching and processing relevant documents and passages which contain the answer. After the analysis of passages is performed, sentences in the passages are scored. Named Entities(NEs) which have the same or similar answer types as answer candidates from top-n sentences in passages are extracted. Finally, our system ranks answer candidates from answer extraction task using semantic similarity between question and sentences that include answer candidates and the final answer list is delivered to the user.

2.3 Text-to-KB

The limitation of the KB is that it can only store small amount of information as compared to its original unstructured text. To overcome this problem, we use Text-to-KB component which converts unstructured text into triples to be fed in the knowledge base. In order to extract triples from unstructured text, we use the semantic role labels of a sentence and the dependency tree. Extraction templates are constructed that specify, for each dependency tree structure pattern, how triples should be extracted. A full document is retrieved to detect sentences that include word tokens that occur in arguments and relation words of each seed triple. Then a dependency tree of the sentence for each seed triple is constructed, sentence pair, and a linear path that contains arguments and relation words is identified. This path with location of arguments and relation words can generate an extraction template. Semantic rule labeling provide similar results that can be converted to triple format. Predicates of the results are considered as relation phrases and each argument and argument modifier are considered as each argument of triples. A small set of rules is also used to convert semantic rule labeling results to triples.

3. RESULTS AND DISCUSSION

Classical Text-QA systems depend on search results to return relevant documents, and then from those relevant documents answers to users' questions are extracted. Text-to-KB process the output of the online module. Text-to-KB detects KB triples in both snippets and documents and then store them in the KB. We incorporated Text-to-KB module in our system, that goes beyond the basic KBQA model by adding external textual sources during the QA process. A main challenge in KBQA is that questions given in natural language are not easily mapped to entities and predicates in a KB. An applicable approach for handling this task is supervised machine learning, which employs examples of questions with their labels and the corresponding answers (for this questions) to learn this mapping. We use a dataset consisting of a collection of labeled question-answer pairs (1000 question-answer pairs) to calculate the associations between question keywords and predicates to extend system's lexicon where the domain of the dataset is the Arabic Wikipedia. The results of using knowledge base approach alone and Text-to-KB along with the knowledge base approach are provided in Table 2. The result reported for our QA system is computed using precision, Recall and F1-measure. As we can see, Text-to-KB significantly improves over the baseline system.

Table 2. System performance using KB only & Using both KB and Text-to-KB module

System	Precision	Recall	F1-measure
Knowledge base	.635	.406	.495
Knowledge base+ Text-to-KB (Web Search)	.642	.519	.573

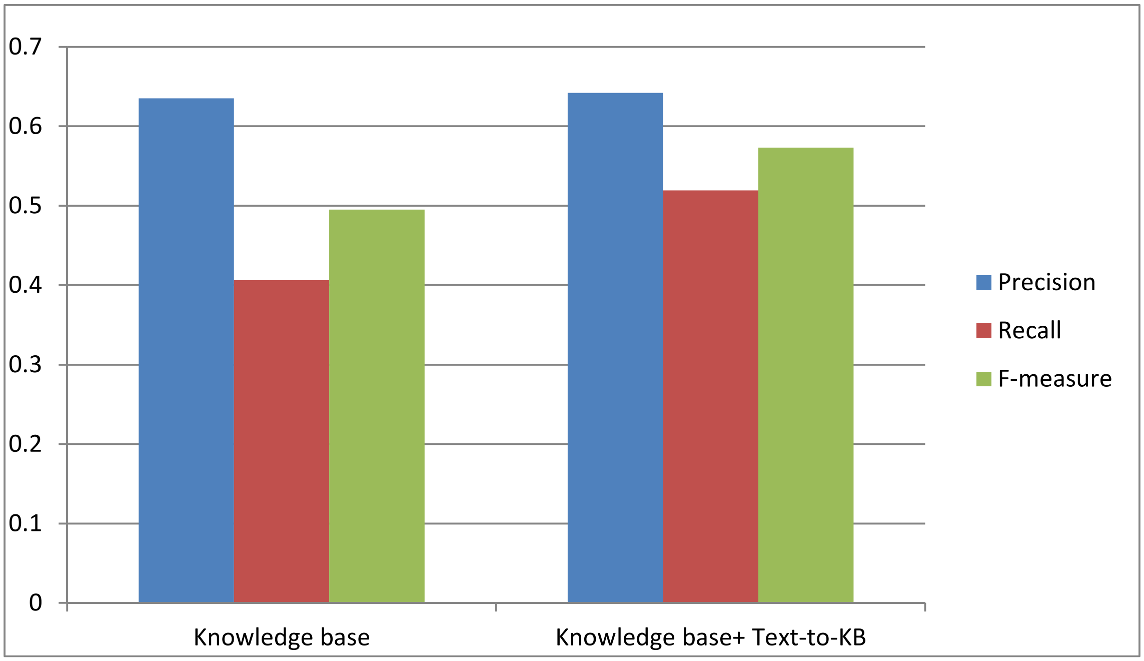

Fig. 2. System performance using KB & KB+ Text-to-KB

We demonstrated that by coupling evidence from knowledge base and text from external resources the system performance can be boosted. The system scored .495 for the f-measure using only KB. The performance of the system the system is proved to be better using both KB and text-to-KB. The computed f-measure using both methods is .573. Comparison of some existing systems on English are presented in Table 3.

Table 3. Performance comparison of our system with existing systems

System	Precision	Recall	F-measure
Jacana [23]	.458	.517	.486
Kitt AI [24]	.526	.526	.535
STAGG [25]	.607	.528	.565
Our System	.642	.519	.573

4. CONCLUSION

In this paper, we show that unstructured text resources can be used for knowledge base question answering to enhance query understanding, generation of candidate answer and ranking. The proposed system uses semantic relatedness among question and sentences to rank answer candidates from KB and from online module and provide the final answer list to user.

REFERENCES

1. Dang HT, Kelly D, and Lin JJ. Overview of the TREC 2007 question answering track. Proceedings of TREC, 2007.

2. Bollacker K, Evans C, Paritosh P, Sturge T, and Taylor J. Freebase: A collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (SIGMOD '08), 2008; 1247-1250. DOI:10.1145/1376616.1376746.

3. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, and Ives Z. Dbpedia: A nucleus for a web of open data. The Semantic Web. Lecture Notes in Computer Science. Springer, 2007; 4825: 722-735. DOI: 10.1007/978-3-540-76298-0_52.

4. Vrande D and Krotzsch M. Wikidata: A free collaborative knowledgebase. Communications of ACM, Sept. 2014;57(10): 78-85. DOI: 10.1145/2629489.

5. Yao X, Berant J, and Durme BV. Freebase qa: Information extraction or semantic parsing?. In Proceedings of the ACL 2014 Workshop on Semantic Parsing, ACL, 2014. DOI: 10.3115/v1/W14-2416.

6. Berant J, Chou A, Frostig R, and Liang P. Semantic parsing on freebase from question-answer pairs. In Proceedings of EMNLP, 2013.

7. Berant J and Liang P. Semantic parsing via paraphrasing. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics , ACL, 2014; 1: 1415–1425 . DOI: 10.3115/v1/P14-1133.

8. Berant J and Liang P. Imitation learning of agenda-based semantic parsers. Transactions of the Association for Computational Linguistics, ACL, 2015; 3: 545–558.

9. Bast H and Haussmann E. More accurate question answering on freebase. In Proceedings of the 24th ACM International on Conference on Information and Knowledge Management(CIKM '15), ACM, 2015;1431-1440.DOI:10.1145/2806416.2806472.

10. Yih W-T, Chang M-W, He X, and Gao J. Semantic parsing via staged query graph generation: Question answering with knowledge base. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, ACL, 2015; 1321–1331. DOI: 10.3115/v1/P15-1128.

11. Yao X and Van Durme BV. Information extraction over structured data: Question answering with freebase. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL, 2014; 1. DOI: 10.3115/v1/P14-1090.

12. Mohammed F, Nasser K, Harb H .A knowledge-based Arabic Question Answering System (AQAS). In Proceedings of ACM SIGART Bulletin, 1993; 21-33.

13. Al-Shawakfa E. A Rule-Based Approach to Understand Questions in Arabic Question Answering. Jordanian Journal of Computers and Information Technology (JJCIT), 2016; 2(3):210-231.

14. Akour M, Abufardeh S, Magel K and Al-Radaideh Q. QArabPro: A Rule Based Question Answering System for Reading Comprehension Tests in Arabic. American Journal of Applied Sciences. 2011; 8(6):652-661.

15. Bekhti S, Rahman A, Al-Harbi M and Saba T. AQUASYS: An Arabic Question-Answering System Based on Extensive Question Analysis and Answer Relevance Scoring. International Journal of academic Research, 2011;3:45-54.

16. Kurdi H, Alkhaider S, and Alfaifi N. Development and Evaluation of a Web Based Question Answering System for Arabic Language. Fourth International conference on Computer Science & Information Technology(CS&IT). 2014; 187-202.

17. Ahmed W, and Anto B. Web-based Arabic Question Answering System using Machine Learning Approach. International Journal of Advanced Research in Computer Science. 2017;8(1):40-45.

18. Stanford Word Segmentor. Accessed 22 February 2017. Available: http://NLP.st.edu/software/segmenter.shtml.

19. Stanford POS Tagger. Accessed 22 February 2017. Available: http://nlp.stanford.edu/software/tagger.shtml.

20. SPARQL. Accessed 27 January 2017. Available: https://www.w3.org/TR/sparql11-query/.

21. Schlaefer N, Ko J, Betteridge J, Pathak MA, Nyberg E and Sautter G . Semantic Extensions of the Ephyra QA System for TREC 2007. In Proceedings of The Sixteenth Text REtrieval Conference(TREC 2007), 2007.

22. Lucene. Accessed 03 February 2017. Available: http://lucene.apache.org/core.

23. Yao X and Durme BV. Information extraction over structured data: Question answering with freebase. Proceedings of ACL, 2014.

24. Yao X. Lean question answering over freebase from scratch. Proceedings of NAACL Demo, 2015.

25. Yih W-T, Chang M-W, He X, and Gao J. Semantic parsing via staged query graph generation: Question answering with knowledge base. Proceedings of ACL, 2015.