A Review on Question Analysis, Document Retrieval and Answer Extraction Method in Question Answering System

1st Irfandy Thalib
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
irfandythalib@mail.ugm.ac.id

2nd Widyawan
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
widyawan@ugm.ac.id

3rd Indah Soesanti
Department of Electrical Engineering and Information Technology
Universitas Gadjah Mada
Yogyakarta, Indonesia
indahsoesanti@ugm.ac.id

2020 International Conference on Smart Technology and Applications (ICoSTA) 978-1-7281-3083-5/20/$31.00 ©2020 IEEE 10.1109/ICoSTA48221.2020.1570614175

***Abstract* — Question Answering System (QAS) is a technique to extract answers in the available source like electronic documents. It has been used in many domains, such as Education, Medical, and Religion. Furthermore, QAS can extract answers in the open-domain. In this research, we analyze the state-of-the-art method used by QAS in question analysis, document retrieval and answer extraction in recent years. Furthermore, we present two classifications in the source type and baseline method. We also give the view in terms of the strengths and weaknesses of each method as well as the issue to solve in the future QAS.**

Keywords — question answering system, natural language processing, machine learning, big data, information retrieval

I. INTRODUCTION

Question answering system (QAS) is a technique to extract answers from available sources. QAS has become one of the most popular information retrieval applications [1]. A QAS allows the user to give a query in a natural language, and the answer to the query will be automatically returned by the system.

QAS application has been studied by many researchers to solve problems in many domain areas, such as medical domain [7][13][21], education domain [3][4], religion domain [15][19], and electronic commerce domain [27].

Furthermore, the general domain of QAS can be divided into two types, which are open-domain and close/restricted domain. QAS can be categorized as an open-domain when the QAS can answer an independent question. Some popular examples of this domain are Apple's Siri and IBM's Watson, where the QAS can answer domain-independent questions given by the user. On the other hand, QAS can be categorized as a restricted domain if it can answer only domain-specific questions [5].

The architecture of a QAS consists of three major stages, which are question processing, relevant passage retrieval, and answer processing [2][6]. The whole architecture can be seen in Fig. 1.

In this work, we will give an analysis of the state-of-the-art method used by the QAS in recent years as well as the strengths and weaknesses.

The rest of the paper organized as follows: in section two, we review the method used in question analysis, document retrieval and answer extraction and classify them based on the same characteristic, then in section three, we give a classification based on data source type, baseline method, and comparison of method and result by each method. Lastly, in section 4 we draw the conclusion as well as the suggestion for the future development of QAS.

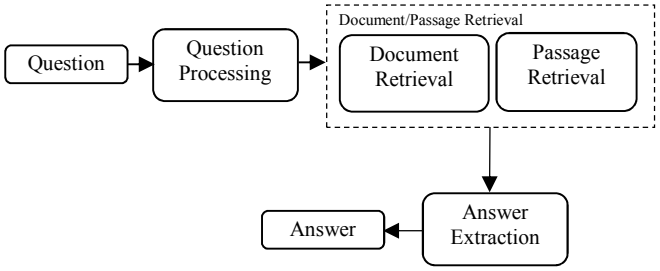

Fig. 1. The architecture of QAS [6]

II. METHOD CLASSIFICATION

Three major stages used in QAS are question analysis, document/passage retrieval, and answer extraction. In this section, we will focus on reviewing the method used in these stages mentioned above.

A. Method Used in Question Analysis

Question classification is considered as one of the most significant phases of a typical Question Answering (QA) system [7]. It has a vital role [8] to help the system find or construct an accurate answer, which results in an improvement of the quality of question answering systems [9][10][25].

The output of this step is the expected answer type of the question query as well as the list of important keywords that will be used in document or passage retrieval and the answer retrieval process.

The question classification process can significantly help a QAS in predicting the type of entities needed to be presented in the relevant passage candidates by classifying the question into various correct answer types taxonomy [11] such as location, person, time, etc. Therefore, predicting the correct answer type is an essential factor in the success of a QAS in producing the correct answer for a question. An example of this is the question "Who invented the camera?", then the expected answer is "person". Another example is "When was the camera invented?" then the answer needed is time.

In general, there are two main approaches for question classification, which are rule-based approach and learning-based approach [12]. However, in this review, we categorize them into Natural Language Processing (NLP) approach (lexical, semantic and pattern-based), machine learning

approach, and hybrid approach (the combination of NLP and machine learning).

1. Natural Language Processing Approach

Dodiya et. al. in [13] proposed a rule-based method that classifies the question into coarse-fine categories that were introduced by Li and Roth in [12]. In question analysis stages, the system extracts a list of keywords and then labels them into a primary keyword, secondary keyword, and question object. Next, the question classification module matches the keyword with the pre-defined pattern. For example, pattern type for "what" type questions: "What is the procedure for kidney stone removal?", as the word "what" found in the sentence, the question then labeled as "description." The result shows the module got the highest accuracy in the "why" type question with 66%. However, the accuracy of "how" type questions was 44.54%, which is quite low compared to other results. This was affected by the fact of "how" type questions that have two types of categories: numeric and description. As a result, the classifier was not able to identify the matched pattern of the question and assigned it to the correct category. Furthermore, the drawback of rule-based techniques is it is not performing well on diverse datasets as it requires many rules to classify the specified questions [7].

On the other hand, Hamed et al. [15] used a semantic approach to create a question expansion in English Quran QAS. They used available WordNet synonyms and Islamic synonyms to create several questions that will be used in verses retrieval.

2. Machine Learning Approach

Some researchers also have involved machine learning in classifying the question. Puteh et al. [31] made a question classification model for Malay QAS. They adopted coarse-fine abbreviation taxonomy but using only the coarse abbreviation as the label for the question. The dataset used was translated questions from Islamic websites. The questions were labeled manually. After going through pre-processing (removing stopwords, unigram bag-of-words, and TFIDF for the questions vector representations), the labeled question is used to build a machine learning classifier model. Machine learning algorithms used were Support Vector Machine (SVM), Naïve Bayes and Random Forest. After 10-fold cross-validation, the result revealed that the classifier based on SVM has the best score (86.8% mean accuracy).

SVM is a more complex algorithm but can model non-linear decision boundaries. Furthermore, it is also quite robust against overfitting, especially in high-dimensional space, like in the question and text classification problems [31].

3. Hybrid Approach

In the hybrid approach, the combination of NLP and machine learning have been used. Mohasseb et al. in [17] used three features of NLP (grammatical, domain-specific, and patterns). The grammatical features consist of the seven-word classes in English (verb, noun, adjective, preposition, determiner, adverb, and conjunction) and six categories (what, where, when, who, which, and how). In domain-specific features, a label is added to identify the type of question; for example, Products is labeled as "P". The grammatical pattern of the question is then used to classify the question type. These features define two layers of class for each word in question and the expected answer type. The example of a pre-processed question: "What is the smallest country in Asia?" has the following grammatical pattern $QW_{What} + LV + D + Adj + CN_{OS} + P + PN_{G}$ and the label is LOC. The pre-processed question and label are then employed as the feature in the machine learning classifier. The classification result of Decision Tree: ABBR, DESC, ENTITY, HUM, NUM, LOC had 100%, 99%, 92.8%, 99%, 95.9%, 88.1% respectively. On the other hand, the result of SVM: ABBR, DESC, ENTITY, HUM, NUM, LOC had 100%, 99.8%, 91.8%, 99.8%, 95.8%, 88.4% respectively. The results validated that combining grammatical and domain-specific grammatical features enhances the classification accuracy. Moreover, it helps the machine learning algorithms to better differentiate between different class types.

In research [14], Xu et al. used 6000 questions (5500 questions for training while 500 for testing). The features of the questions are extracted, such as their part of speech, their synonyms, named entities, their semantic types, interrogative words, top words, and dependency relation. After that, the features are used in the machine learning classifier using the SVM algorithm. The experiment resulted in a top accuracy in the Coarse category classification with 92.2%, and top accuracy in the Fine category with 87.2%.

The issue that should be marked in question analysis stage is the flexibility and accuracy in defining the expected answer type of the question. Therefore, an improvement in the method is still needed.

B. Method Used in Document/Passage Retrieval

The document retrieval stage is the process for choosing the candidate document or passage that further will be used as a source in the answer extraction stage, and commonly this is linked by the result of the question analysis process.

A search engine website is one example of document retrieval systems. However, the existing search engine systems still face problems like retrieving many irrelevant documents and word mismatch, especially when the query given by the user is not specific enough [18][19]. The failure to find the correct document will lead to an incorrect answer. QAS is expected to be able to find the documents that contain the answer in the document collection [20].

After conducting an analysis, we categorize the method used by researches to find the correct documents or passages that contain the answer into two types, namely document similarity score approach and machine learning approach.

1. Document Similarity Score

Sarrouti et al. in [21] employed Stanford CoreNLP sentence boundary, stemmed words and Unified Medical Language System (UMLS) concepts as features for the BM25 model to retrieve relevant passages in biomedical QAS. The query is inputted into the PubMed search engine to get relevant documents based on UMLS similarity between concepts of the query and each title of the returned documents. Subsequently, the system employed stemmed words and UMLS concepts as features for the BM25 model to rerank the candidate passages and keep the N top-ranked ones. The proposed method achieved a 6.84% improvement in terms of

MAP (Mean Average Precision) compared to the current state-of-the-art methods.

2. Machine Learning

Hamed et. al. in [15] employed a Neural Network classifier to labeling the Quran verses in Al-Baqarah Surah into two (Pilgrimage and Fasting). In the document retrieval stage, the purpose is to extract relevant verses, not to find actual verses to the question, before sending them to the answer extraction stage. The system then classified the question label (either pilgrimage and fasting) and chose the verses which have the same label.

Artificial Neural Network (ANN) classifier can identify the complex patterns existing in the data. It has been successfully applied in many fields and has given a promising performance in document classification tasks [22][23].

C. Method Used in Answer Extraction

The last stage in typical QAS is the answer extraction process. In this stage, the answer to the question is extracted from the chosen document or passages in the document or passages retrieval process. This stage is typically applied to the top-ranked passages rather than to all available documents [21].

Hamed et. al. in [15] used a word matching score technique. The candidate answers were the selected verses in the document retrieval passage. This research employed the N-gram technique (bigram and unigram) to extract a list of words from the question. The verses that contain similar words with the list of extracted words will be retrieved. Subsequently, the system performs words matching scoring for each of these retrieved verses to calculate the number of similar words between the expanded questions and these verses. Lastly, the system ranks all these scored verses where the top verses are returned as the answer.

N-grams have been successfully used in many language processing applications, which includes measuring and identifying text reuse in journalism [30]. Furthermore, the selection of the size of the n-gram is essential since the small size would cause many matchings, while the large size would generate very few matches [15].

III. DISCUSSION

In this work, we present two classifications for QAS, which are document/passage source type categorization and baseline technique type.

A. Document/Passage Source Type

Different type of source requires a different retrieval approach in QAS. We classify the data source type used by QAS as answer source into three categories, namely question-answer pair (QA-pair), structured database, free documents.

1. QA-Pair

The characteristic of this source is that the question already has the answer. The system only needs to find a similar question in the available source and returns the answer to the question. However, to retrieve the correct QA-pair, a deeper understanding of QA texts is needed [28]. Examples of sources that fall in this type are Frequently Asked Question (FAQ) list and question-answer in forum website. Sakata et. al. [28] used localgovFAQ (a Japanese administrative municipality data) and data available on the StackExchange website. In this research, the authors used the similarity measurement technique. They proposed two similarity algorithms, which were TSUBAKI and BERT to retrieve the most similar QA-pair with the query. They claimed that the method got a high-performance retrieval.

The advantage of QAS using this source type is the processing time needed to find the answer is shorter than in other source types due to available pre-defined question-answer. However, the limitation is that the system needs a large number of pre-defined question-answer pairs. When there is no available similar question in the source, the system will not return the correct answer.

2. Structured Database

In a structured database, question in QAS is transformed into a database query (Natural Language Interface to Database) using tools like Named Entity Recognition (NER) to detect all entities that available in the question. For example, the question "what is the price of Samsung J5" will be converted into "select price in product-table where brand=Samsung and type=J5". Wudaru et. al. [29] transformed the question into a Structured Query Language (SQL) query. The question is first preprocessed using NLP tools such as CoreNLP, POS tagger, dependency parser, and NLTK Lemmatizer. Furthermore, they used pre-defined SQL templates and a list of rules to convert the question into an SQL query.

3. Free Document

The free document has longer content than other source types. Examples of this type are blog articles, news articles, books, etc. Since the text is longer, the system needs more effort to extract the answer. The research [21] retrieved medical documents and performed similarity measurement using BM25 ranking algorithm to get the relevant documents. The query must perform a longer scanning time to find relevant documents as well as the answer.

B. Baseline Technique (NLP and Machine Learning Approach)

1. NLP Baseline

In QAS that using NLP [13][21], a popular tool that we can see is NER. Mainly, NER is used to extract entities that exist in question. The entity will be used to find the document or passage that contains the entity to be chosen as a candidate document. However, the drawback of this baseline is the technique is not independent if they are used in different answer sources and different languages.

2. Machine Learning Baseline

Many prior studies have used machine learning algorithms for question classification tasks [14]-[17]. SVM is one of the most used algorithms [10][24][25]. Other studies like [9] and [26] have performed other machine learning algorithms such as Nearest Neighbors, Decision Tree, and Naive Bayes.

On the other hand, some researchers have employed deep learning techniques such as research by Sharma et. al. [16]. The dataset used is an open dataset from Facebook, namely Babi Dataset. They implemented Long Short Term Memory (LSTM) baseline and two modified LSTM called Memory

Networks and Dynamic Memory Networks (DMN). They found that DMN reached the highest mean accuracy by 91%. The authors also found that LSTM-based models are only helpful for a generic and simple task that does not need much processing. Giving the memory to the network serves as a significant enhancement. However, the model has its weakness since it can have only a single pass over the story for answering questions. The network still struggles with tasks that require too much knowledge from the story. DMN, which allowing multiple passes over story tackle this problem and therefore solves most of the Babi tasks.

We found that combining different features in NLP task like lexical and semantic attributes with machine learning algorithms such as SVM can improve the classification accuracy. However, the machine learning baseline is highly dependent on the quality and the domain of the dataset. The application of machine learning in cross-domain still needs to be explored.

C. Result Comparison

After doing a deep analysis of each work, we finally can list the strengths and weaknesses of each method, as is presented in Table I below.

TABLE I. METHOD AND RESULT COMPARISON

Paper, Method and Result Level	Strength	Weakness
Paper: [13] **Result level**: Question classification **Method**: NLP (rule-based syntactical pattern matching) **Output**: coarse categories English	Faster to find categories since the pattern has been defined. Fit for an open-domain question. The answer returned is more specific since the EAT has been defined.	Accuracy in the "how" category is still low. It requires more patterns for a particular question type. It still needs to be explored in other restricted domains. Not performing well on diverse datasets as it requires numerous rules to classify the specified questions.
Paper: [31] **Result level**: Question classification **Method**: Machine learning, N-gram, and TFIDF as the feature of the classifier.	Can classify the category of unstructured terms in question.	Still get low accuracy in some categories due to the limitation of the dataset.
Paper: [17] **Result level**: Question classification **Method**: Hybrid (NLP's grammatical pattern as a feature and Machine learning as a classifier). **Output**: coarse-fine categories English	Can get EAT of unstructured terms in question.	Since the classifier using grammatical features, it is highly dependent on the language (English) and still needs to be explored in other languages.
Paper: [14] **Result level**: Question classification **Method**: Hybrid (NLP's top words + dependency relation as a feature and machine-learning as a classifier) **Output**: coarse-fine categories English	Can get the semantic meaning of the question.	Highly dependent on the language (English) and still need to be explored in other languages. After adding the dependency relationship, the feature of the space dimension is higher, which leads to the overfitting of the classifier.
Paper: [21] **Result level**: Document/passage retrieval **Method**: Similarity retrieval (using UMLS concept and BM25 algorithm)	Can retrieve documents semantically.	The first retrieval step is highly dependent on the PubMed search engine, where the result can not be evaluated.
Paper: [15] **Result level**: Answer extraction **Method**: NLP (semantic question expansion)	Can get users query meaning. Therefore, the result in answer retrieval is more accurate.	The question does not have EAT. Therefore, the system will return a longer sentence as an answer. Time processing is longer as the number of expanded questions are not limited and following as many words in the ontology that has the same meaning.
Paper: [16] **Result level**: Answer Extraction **Method**: Deep Learning (Word2vecmodel as a feature in LSTM)	Can handle tasks that need transitive reasoning like a QAS.	Highly dependent on the quality and the domain of the dataset.

IV. CONCLUSION

Finally, the authors conclude that each stage in a QAS has different approaches. In question analysis, NLP and machine learning have been proposed to handle the question classification. In document retrieval stages, similarity ranking algorithm and machine learning have been used to choose the relevant documents. In the answer extraction, the word matching score technique has been applied. Furthermore, we classify the type of answer source into three types, which are QA-pair, structured database, and free documents. In the baseline method, we categorize into NLP baseline and machine learning baseline.

We also have presented the strengths and weaknesses of each method. Therefore, it can give a brief view of the future development of QAS.

Lastly, even though QAS has been developed in various domains and techniques, the available QAS mainly built for English. Therefore, it still needs to be evaluated in other languages.

REFERENCES

[1] A. Mohasseb, M. Bader-El-Den, and M. Cocea, "Question categorization and classification using grammar-based approach," *Inf. Process. Manag.*, vol. 54, no. 6, pp. 1228–1243, 2018.

[2] L. Hirschman and R. Gaizauskas, "Natural language question answering: The view from here," *Nat. Lang. Eng.,* vol. 7, no. 4, pp. 275–300, 2001.

[3] A. Samadi, E. F. Hanaa, M. Qbadou, M. Youssfi, and F. Akef, "A syntactic and semantic multi-agent based question answering system for collaborative e-learning," *2018 4th International Conference on Optimization and Applications (ICOA), Mohammedia,* 2018, pp. 1-4.

[4] S. Sinha, S. Basak, Y. Dey and A. Mondal, "An Educational Chatbot for Answering Queries", in Emerging Technology in Modelling and Graphics., vol. 937., *Springer*, Singapore 2019, pp 55-60.

[5] A. Mishra and S. K. Jain, "A survey on question answering systems with classification," *J. King Saud Univ. - Comput. Inf.* Sci., vol. 28, no. 3, pp. 345–361, 2016.

[6] D. J. Jurafsky and H. Martin. *Speech and Language Processing (2nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2009.

[7] M. Wasim, M. N. Asim, M. U. Ghani Khan, and W. Mahmood, "Multi-label biomedical question classification for lexical answer type prediction," *J. Biomed. Inform*., vol. 93, no. March, p. 103143, 2019.

[8] P. Jacquemart and P. Zweigenbaum, "Towards a medical question-answering system: A feasibility study," *Stud. Health Technol. Inform*., vol. 95, pp. 463–468, 2003.

[9] Zhang, Dell, Lee, S. Wee(s). "Question Classification Using Support Vector Machines." *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 26-32.

[10] N. Van-Tu and L. Anh-Cuong, "Improving question classification by feature extraction and selection," *Indian J. Sci. Technol*., vol. 9, no. 17, 2016.

[11] S. K. Ray, S. Singh, and B. P. Joshi, "A semantic approach for question classification using WordNet and Wikipedia," *Pattern Recognit. Lett*., vol. 31, no. 13, pp. 1935–1943, 2010.

[12] X. Li and D. Roth, "Learning question classifiers: The role of semantic information," Nat. Lang. Eng., vol. 12, no. 3, pp. 229–249, 2006.

[13] T. Dodiya and S. Jain, "Question classification for medical domain Question Answering system," WIECON-ECE 2016 - 2016 IEEE Int. WIE Conf. Electr. Comput. Eng., no. December, pp. 204–207, 2016.

[14] S. Xu, G. Cheng, and F. Kong, "Research on question classification for Automatic Question Answering," in *Proceedings of the 2016 International Conference on Asian Language Processing, IALP 2016*, 2016, pp. 218–221.

[15] S. K. Hamed and M. J. A. Aziz, "A question answering system on Holy Quran translation based on question expansion technique and Neural Network classification," *J. Comput. Sci*., vol. 12, no. 3, pp. 169–177, 2016.

[16] Y. Sharma and S. Gupta, "Deep Learning Approaches for Question Answering System," *Procedia Comput. Sci.*, vol. 132, pp. 785–794, 2018.

[17] A. Mohasseb, M. Bader-El-Den, and M. Cocea, "Classification of factoid questions intent using grammatical features," *ICT Express*, vol. 4, no. 4, pp. 239–242, 2018.

[18] H. Imran and A. Sharan, "Thesaurus and query expansion," *J. Comput. Sci.*, vol. 1, no. 2, pp. 89–97, 2009.

[19] H. Ishkewy and H. Harb, "ISWSE: Islamic Semantic Web Search Engine," *Int. J. Comput. Appl.*, vol. 112, no. 5, pp. 975–8887, 2015.

[20] Bilotti, Matthew and Nyberg, Eric. "Improving Text Retrieval Precision and Answer Accuracy in Question Answering Systems", *Proceedings of the 2nd workshop on Information Retrieval for Question Answering*, 2008, pp. 1-8.

[21] M. Sarrouti and S. Ouatik El Alaoui, "A passage retrieval method based on probabilistic information retrieval and UMLS concepts in biomedical question answering," *J. Biomed. Inform.*, vol. 68, pp. 96–103, 2017.

[22] Ramlall, I., "Artificial Intelligence: Neural Networks Simplified", *International Research Journal of Finance and Economics*, vol. 39, 2009.

[23] A. Patra and D. Singh, "A Survey Report on Text Classification with Different Term Weighting Methods and Comparison between Classification Algorithms," *Int. J. Comput. Appl.*, vol. 75, no. 7, pp. 14–18, 2013.

[24] Hao T., Xie W., Xu F., "A WordNet Expansion-Based Approach for Question Targets Identification and Classification." *In Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, 2015,2015 pp. 333-344.

[25] D. Metzler and W. B. Croft, "Analysis of statistical question classification for fact-based questions," Information Retrieval., vol. 8, no. 3, pp. 481–504, 2005.

[26] M. Mishra, V. K. Mishra, and S. H.R., "Question Classification using Semantic, Syntactic and Lexical features," *Int. J. Web Semant. Technol.*, vol. 4, no. 3, pp. 39–47, 2013.

[27] X. Pan and T. Zhang, "Research on E-Commerce Automatic Question Answering System Model Based on Data Mining," *J. Phys. Conf. Ser.*, vol. 1069, no. 1, 2018.

[28] W. Sakata, R. Tanaka, T. Shibata, and S. Kurohashi, "FAQ retrieval using query-question similarity and BERT-based query-answer relevance*," in SIGIR 2019 - Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 1113–1116.

[29] V. Wudaru, N. Koditala, A. Reddy, and R. Mamidi, "Question Answering on Structured Data using NLIDB Approach," *5th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2019*, pp. 1–4, 2019.

[30] R. M. A. Nawab, M. Stevenson, and P. Clough, "Detecting text reuse with modified and weighted N-grams," *in *SEM 2012 - 1st Joint Conference on Lexical and Computational Semantics*, 2012, vol. 1, pp. 54–58.

[31] N. Puteh, M. Zabidin, Husin, H. M. Tahir and A. Hussain, "Building a Question Classification Model for a Malay Question Answering System", *Int. J. of Innovative Technology and Exploring Engineering*, vol. 8, no. 5, 2019