Integrated passage retrieval with fuzzy logic for Indonesian question answering system

Syopiansyah Jaya Putra, Muhammad Zidny Naf'an, Muhamad Nur Gunawan

Dept of Science and Technology, Syarif Hidayatullah State Islamic University, Jakarta, Indonesia
syopian@uinjkt.ac.id
Dept of *Informatics, Institute of Technology Telkom, Purwokerto, Indonesia*
zidny@ittelkom-pwt.ac.id
Dept of Science and Technology, Syarif Hidayatullah State Islamic University, Jakarta, Indonesia
nur.gunawan@uinjkt.ac.id

***Abstract*— Question Answer System (QAS) has many methods in determining candidate answer and must have the right answer for each question. Previous QAS using Fuzzy logic focused on candidate and ranking of answer. However, the QAS needs improvement in determining the relevant answer from the question and generating the correct answer in the question answer process. In this paper, we propose a new method to combine fuzzy logic and retrieval passages to obtain a collection of relevant answers in order to obtain high answer accuracy. We take relevant answers from the collection of the answer document and choose the exact answer using fuzzy logic. Methods for the QAS are preprocessing, question analyzer, passage retrieval, passage scoring, scoring for similar text, measuring keyword and candidate answers, fuzzy logic controller, rules, and extraction answer. This study produced significantly relevant answers compared to the TF-IDF method. The performance of the system has improved the accuracy of QAS by 80% which is better than the previous study.**

Keywords— passage retrieval, fuzzy logic, passage scoring, answer extraction, Indonesian language

I. Introduction

Question Answer System is a system to find answers to human questions through input interfaces using natural language [1]. QAS implementation is mostly done in the form of input and output text rather than the sound [2,3], images or videos. Answer needs can be in the form of sentences, documents, pictures, sounds, or in the form of videos. However, the challenge of QAS is how the context of the answer is the same as the question [4] Another requirement of QAS is how to process languages from diverse users. QAS in this study uses Indonesian Language Natural Language Processing (NLP). So far, many studies on NLP in Indonesian have been done [5]. The QAS response to user questions is influenced by how much the answer document is available. If the answer documents are widely available, the questions will be easily answered correctly, but if the opposite, the user has little or no right answer. QAS consists of various kinds of developing algorithms, like Fuzzy, weighted, ontology, artificial intelligence, Natural Language Processing (NLP) and Information Retrieval (IR) [6-11]. Some researchers also combine algorithms [1] to produce quality answers that are in accordance with the wishes of users.

The QAS model for the Indonesian language used in solving the accuracy of answer questions continues to be developed such as Putra et al [12] and Naf'an et al [13]. The study of QAS [12] emphasizes the use of semantic with ontology in the answer document. While the QAS study [13] emphasizes the eliminating unanswered question of user questions. In this study, combining the methods used are passage retrieval method and Fuzzy logic. Passage retrieval method is an effective method in responding answers from answer documents. While Fuzzy Logic is used when taking the right answer in the answer candidate [15]. To obtain a part of the text of the answer document, the passage retrieval process is very necessary, but this is not always easy because it can lead to a high computational process [16]. Therefore, it is necessary to use a method so that the collection of answer documents is the right set of documents to reduce the high computational process. In this study testing the passage retrieval method will be carried out in the QA Putra System et al. after Question Analysis and the output from the passage retrieval will be forwarded to the scoring, matching and Fuzzy logic processes. This paper aims to answer questions about how the retrieval passage and fuzzy can be implemented in the QAS system to get the best candidate answer for each question. In this study testing the passage retrieval method will be carried out in the QAS Putra et al. after Question Analysis and the output from the passage retrieval will be forwarded to the scoring, matching and fuzzy logic processes... This paper aims to answer questions about how the retrieval passage and fuzzy can be implemented in the QAS to get the best candidate answer for each question. The methodology is preprocessing, question analyzer, passage retrieval, passage scoring, scoring for similar text, measuring keyword and candidate answer, fuzzy logic controller, rules, and extraction answer. The result of the retrieval passage is the candidate answer text from the answer document according to the query of the questions. The results of this study are that there are 42 correct answers, 2 inexact answers, 15 incorrect, 1 unsupported, 0 unanswered. This study shows the performance of the system has improved the accuracy of QAS compared to TF and TF-IDF method.

II. Literature Review

Question Answering systems have shifted to use the machine learning method in past six years. It was in the year 2012 when the IBM Watson [6] first QA System was winning the quiz over the human contestant. It was based on IBM research where IBM proposed a question answering system based on hundreds algorithm called DeepQA. The evidence behind the Watson is not only combining two or three algorithms to get natural interaction between human and the computer but also a hundred algorithm to take decision of choosing best answer. Users will also get a direct answer from the system quickly and accurately, without sort through the documents to get a reply. The IBM Watson conducted answer the question from English language, it has impact to other researchers to study and develop the QA System in other language. For Indonesian language, the previous research of QAS used the factoid question which was received by the system. This QAS is employed to close the domain collection about the Khulafaa Al-Rashidin history which referred to [5]. The aims of this theme are to help the Muslim people in studying the history of Khulafaa Al-Rashidin. The story tells the successor to the Prophet Muhammad Peace to be Upon Him (PBUH), which consists of four companions, namely Abu Bakr Radhiyaallahu 'anhu Ra., Umar ibn alKhattab Ra., Uthman ibn Affan Ra., and Ali ibn Abi Thalib Karamallahu wajhah. The QAS that refers to M.N. Zidny, et,al [5] is called Question Answering System 2 Framework (QAS2F), which uses improvement on the scoring of the candidate answers. The scoring part just does sorting the score of Lucene, and the score of the algorithm by S. H. Wijono, et.al [10] which is called Wijono's algorithm and the distance between the keyword with the candidate answers. The right answers on QAS2F method are often not in the first rank so that the system does not select that answer. To solve this problem, the authors conducted an experiment by implementing fuzzy logic in choosing the answer assuming that the fuzzy logic is more objectively selects the solution compared to using only sorting. The other technique in question answering system on Islamic field related to ITQ are QAS used weighted vector and QAS used semantic approach. First, Term FrequencyInverse Document Frequency (TF-IDF) is used for generating weight vector for each concept in ITQ to perform the QAS for weighting the vector [12]. The candidate answer in this QAS based on the concept that has weighted vector variation from 0 and 1. The most related document to the concept has a highest weighted vector score. So, the answer to the question in QAS is based on the strong relationship between the concept and the document using the weighted vector score. Second, Semantic Approach is used in the QAS [7]. According to Putra, et.al [12] the weight vector is combined with the semantic approach for selecting the best answer for the question in input system. The method used by Putra, et. al [7] is Cosine similarity

which is used to find the semantic from the question with the candidate answer in the ITQ dataset. The previous research of QAS in ITQ and Khulafaa AlRashidin dataset has one goal, which finds the exact answer from the candidate answer [5,7]. The architecture and technique to find the exact answers are various, from very simple architecture QAS using question analyzer, passage retrieval, and Name Entity recognizer [5], and then reconstruct the architecture with QAS2F to improve the answer using scoring of the candidate [13]. However, those two studies do not use fuzzy logic in QAS. The previous research in QAS using passage retrieval also conducts the several methods.

III. METHODOLOGY

The combination method consists of passage retrieval and fuzzy logic. This QAS uses architecture from M.N. Zidny, D.E. Mahmudah, S.J. Putra, and A.F. Firmansyah [13] or (QAS2F) as the base structure. As seen in Fig.1, the development was in the extracting dataset to fuzzy logic controller. To proof of concept of combining passage retrieval and fuzzy logic, we conduct the TF, and TF-IDF as comparation of the study.

The following details of QAS Process are as follows;

A. *Preprocessing.*

At this stage, the text processing has been run and the system will build an index of the passages using Lucene 6.0.1.
Algorithm:

```
Pseudo-code Algorithm for Preprocessing:
1: D ← input documents (Indonesian texts)
STEP 1: Preprocessing
2: for all d ← D do (D: Input documents)
 3: perform tokenization of d ;( d: document)
4: normalize into letter format;
5: if language of d recognized then
6: apply Indonesian stemming and mark stop
```

B. *Question Analyzer*

At this stage, a user input a question in Indonesian language, for example: "Siapa nama lengkap Abu Bakar?" (Who is the complete name of Abu Bakar?). It will be processed to get a Boolean query, keywords, and keyword entity. The Boolean query is the input to Passage Retrieval. The keyword is the keyword of the question, while keyword of body is the keyword object in question. Determination word of an object is based on a dictionary that has been made by Zidny, et. al [5]. Type of entity which handles this QAS consists of three categories: Person, Time, and Location.

It is used to determine the object using string matching of library LingPipe (http://alias-i.com/lingpipe/).

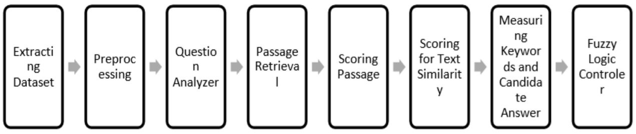


FIG . 1 PROCESS OF QAS COMBINED PASSAGE RETRIEVAL AND FUZZY LOGIC

ALGORITHM:

```
1: Q <- INPUT A QUESTION.
2: GET BOOLEAN QUERY, KEYWORD, AND KEYWORD ENTITY.
3: ANALYZE THE QUERY
```

C. *Passage Retrieval*

Boolean query results sent to the Question Analyzer module on Passage Retrieval. It is a search engine of the QAS to do search corpus relevant to the query using the Lucene library version 6.0.1.

D. *Scoring Passage*

This stage will run Wijono's algorithm [10] to score all passages retrieved from Passage Retrieval

E. *Scoring Passage*

Adding a scoring process for reading. This calculates the level of similarity between text on the Passage and query of the question. This similarity measurement algorithm is using Jaro-Winkler Distance from LingPipe library.

F. *Measuring Keyword and Candidate Answer*

At this stage, the system has been measuring the distance between keyword and ccandidate answer. The result was the table of distance.

G. *Fuzzy Logic Controller*

The fuzzy rule-based system has three phases; fuzzification, inference, and defuzzification. Fuzzification converts the crisp input into the fuzzy contribution, while Inference contains fuzzy rules and reasoning process takes into account all the rules stored in the knowledge base. The final stage is the defuzzification which will convert the output fuzzy inference result into a crisp production. JfuzzyLogic, one of a fuzzy logic library, has a step by step of the fuzzy rule-based system [15,17].

H. *Fuzzy Logic Controller*

The input of this phase is the crisp value of each variable fuzzy. Furthermore, based on the membership function and domain of each fuzzy set, membership value will be generated for each fuzzy set of 4 variables

I. Inference

The second phase of fuzzy logic is inference. This step will run the rules stored in the knowledge base. Fuzzy rules model used is Mamdani models which is defined as follows : [14].

IF x1 is A1 AND … AND Qn is An THEN y is B

Where: A1,.., An and B is fuzzy set and "x1 is A1" means the value of the variable x1 is a member of the fuzzy set A1. There are 54 possibilities to set rules for fuzzy inference.

J. *Defuzzification*

The last stage of the fuzzy rule-based system is defuzzification. It will convert values into a single result fuzzification crisp output. The method used in the current research is the Center of Gravity (COG) defined as:

$$y^* = \frac{\sum y\mu R(y)}{\sum \mu R(y)} \qquad (1)$$

With y^* is a crisp production value, y is input and $\mu\ R(y)$ membership value for the set R from y [14].

K. *Rules of Previous Work*

Based on the architecture by Zidny, et/al [13], all of the questions were successfully answered, then the rules used in the study will be retained in the present experiment.

L. *Answer Extraction*

It is the last phase in which the system will extract the correct answer and be displayed to the user.

M. *Performance Evaluation*

The testing of this QAS is to input the questions written in natural language. 60 items are consisting of 20 questions for the type of person (PERSON), 20 questions for the kind of location (LOCATION), and 20 questions for the kind of time (TIME). The list of items used refers to what is defined by Zidny, et al [5], then the results of the test are the value of accuracy of the QAS as stated by Zidny [5]. There are five types of results for the system response which are Correct, Unsupported, Inexact, Incorrect, and Unanswered.

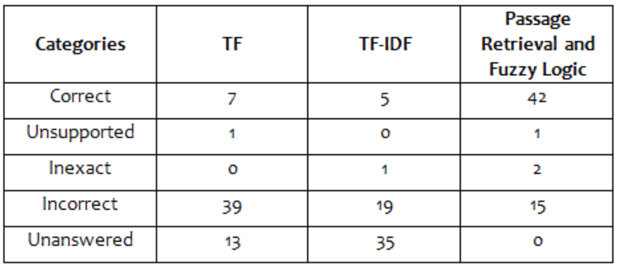
IV. RESULTS AND DISCUSSION

Based on the architecture and testing methodology mentioned previously. It indicates that the correct answer scoring is 42 out of 60 questions, or 70% with the right answer. For a scoring results of Unsupported, Inexact, incorrect and unanswered are 2%, 3%, 15%, and 0% respectively.

Comparison results of current research with TF and TFIDF can be seen in Table 1.

The results obtained in the current study have increased. It increases the number of correct answers, however, it has reduced the number of wrong answers.

TABLE I COMPARISON OF TEST RESULTS



Categories	TF	TF-IDF	Passage Retrieval and Fuzzy Logic
Correct	7	5	42
Unsupported	1	0	1
Inexact	0	1	2
Incorrect	39	19	15
Unanswered	13	35	0

The answer to question number 8 generated from Naf'an et al [13] is "Hasan," however the correct answer is "Ruqayyah." It was happening because the candidate answer "Ruqayyah" is placed on the 5th ranking so that the system does not choose this answer. The current research, as a result of ranks that uses fuzzy logic, string "Ruqayyah" is the 1st ranked, so the system chooses the candidate answers in response. The candidate answer is occupying the 1st ranking due to having the highest defuzzification crisp output results.

This research is necessary to improve the number of correct answers for question answering system using the Indonesian language. Future research directions define the variable fuzzy similarity using other features such as n-gram, language models, etc. and redefining fuzzy rules to obtain optimal levels of accuracy.

V. CONCLUSIONS

This research succeeded in implementing combination of passage retrieval and the fuzzy logic on passage the document and the scoring process of Question Answering System (QAS) and succeeded in increasing the accuracy level of QAS. From the experiment result, implementing the combination method for answers ranks on Indonesian language QAS will increase the number of correct answers.

REFERENCES

[1] Kolomiyets, O., & Moens, M. F. (2011). A survey on question answering technology from an information retrieval perspective. Information Sciences, 181(24), 5412-5434.

[2] Gunawan, A. A., Mulyono, P. R., & Budiharto, W. (2018). Indonesian Question Answering System for Solving Arithmetic Word Problems on Intelligent Humanoid Robot. Procedia Computer Science, 135, 719-726.

[3] Bouziane, A., Bouchiha, D., Doumi, N., & Malki, M. (2015). Question answering systems: survey and trends. Procedia Computer Science, 73, 366-375.

[4] Putra, S. J., & Khalil, I. (2017, August). Context for the intelligent search of information. In Cyber and IT Service Management (CITSM), 2017 5th International Conference on (pp. 1-4). IEEE.

[5] M.N. Zidny, D.E. Mahmudah, S.J. Putra, and A.F. Firmansyah, "Eliminating Unanswered Questions from Question Answering System for Khulafaa Al-Rashidin History". In 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M) (pp. 140–143). Jakarta. 2016.

[6] Ferrucci, D. A. (2012). Introduction to "this is watson". IBM Journal of Research and Development, 56(3.4), 1-1.

[7] S.J. Putra, R. H. Gusmita, K. Hulliyah, and H.T. Sukmana, "A semantic-based Question Answering System for Indonesian Translation of Quran." In 18th International Conference on Information Integration and Web-based Applications and Services (pp. 504–507). Singapore. 2016.

[8] C. Monz, "From Document Retrieval to Question Answering". Disertation. Amsterdam: Institute for Logic, Language, and Computation, Universiteti van Amsterdam. 2005.

[9] Tomek Strzalkowski, Sanda Harabagiu, "Advances in Open Domain Question Answering." PO Box 17,3300, AA Dordrecht, The Netherlands: Springer.

[10] S. H. Wijono, I. Budi, L. Fitria, and M. Adriani, "Finding answers to Indonesian questions from English documents." CEUR Workshop Proceedings, vol.1172, pp.1–4. 2006.

[11] Y. Rochmawati, and R. Kusumaningrum, "Studi Perbandingan Algoritma Pencarian String dalam Metode Approximate String Matching untuk Identifikasi Kesalahan Pengetikan Teks." Jurnal Buana Informatika, 7(2), 125–134. 2015.

[12] S.J. Putra, K. Hulliyah, N. Hakiem, R.P. Iswara, and A.F. Firmansyah, "Generating Weighted Vector for Concepts in Indonesian Translation of Quran." In 18th International Conference on Information Integration and Web-based Applications and Services, pp. 293–297). 2016

[13] M.N. Zidny, R. H. Gusmita, and M.T. Rosyadi, "Developing an Indonesian Question Answering System about Khulafaur Rasyidin History." In International Conference on Cyber & IT Service Management (CITSM) (pp. A-41). Bandung. 2012.

[14] Suyanto., "Artificial Intelligence (2nd ed.)." Bandung: Penerbit Informatika. 2014

[15] P. Cingolani, and J. Alcalá-Fdez, "JFuzzyLogic: A robust and flexible Fuzzy-Logic inference system language implementation." IEEE International Conference on Fuzzy Systems. 2012.

[16] Melucci, M. (1998). Passage retrieval: A probabilistic technique. Information Processing & Management, 34(1), 43-68.

[17] P. Cingolani, and J. Alcalá-Fdez, "JFuzzyLogic: a Java Library to Design Fuzzy Logic Controllers According to the Standard for Fuzzy Control Programming." International Journal of Computational Intelligence Systems, 6(sup1), pp. 61–75. 2013.