Improving The Scoring Process of Question Answering System in Indonesian Language using Fuzzy Logic

Syopiansyah Jaya Putra[1], Muhammad Zidny Naf'an[2], Muhamad Nur Gunawan[3]
[1,3]Faculty of Science and Technology, Syarif Hidayatullah State Islamic University, Jakarta, Indonesia
[2]Department of Informatics, Institute of Technology Telkom, Purwokerto, Indonesia
syopian@uinjkt.ac.id[1], zidny@ittelkom-pwt.ac.id[2], nur.gunawan@uinjkt.ac.id[3]

***Abstract*— Basically, Question Answering System (QAS) show the answer from sources with several candidates. The key success factor is how system knows that the candidate answer is exactly the correct answer. To determine the correct answer, this paper aims at applying the fuzzy logic to improve the score of the candidate answers for the QAS using Indonesian language. 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. There are four fuzzy variables applied namely Lucerne's score, Wijono's algorithm score, a similarity between passage and query text, the distance between keywords and text entity. This research focuses on applying the fuzzy logic on the scoring process of QAS and succeeded in increasing the accuracy level of QAS.**

Keywords—fuzzy logic, scoring process, answer extraction, question answering system, Indonesian language

I. INTRODUCTION

Question Answering System (QAS) is a system that allows users to declare the information needed in the form of a natural language question, and returns short text excerpts or even phrases in response. The availability of vast and varied sources of information, and the rapid development of Natural Language Processing (NLP), Information Extraction (IE), and Information Retrieval (IR) techniques have greatly affected the development of the Question Answering System, which was previously only able to answer the questions which were limited to a particular field (domain) based on structured information sources such as databases, now it can answer various types of questions by source from information from an unstructured text collection.

The research of QAS in ITQ and Khulafaa Al-Rashidin dataset has one goal, which finds the exact answer from the candidate answer [1-4]. 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 [2]. However, those two studies do not use fuzzy logic in QAS.

Therefore, this paper aims at answering the questions : How far fuzzy logic can be implemented in QAS system to get the best candidate answer for each question. In the following, the researcher will describe the structure and the working of fuzzy logic system which consist of : preprocessing, question analyzer, passage retrieval, scoring passage, scoring for text similarity, measuring key word and candidate answer, fuzzy logic controller, fuzzyfication, inference, defuzzyfication, rules of previous work, answer extraction, and performance evaluation.

II. LITERATURE REVIEW

Question Answering systems have transformed much in past four decades at par with the whole of natural language processing. It was in the year 1978 when the first classic QA book was published. It was based on Lehnert's thesis where she proposed a question answering system based on semantics and reasoning. We must not be mistaken, question answering systems were at place much before the publication of the book, the story dating back to sixties [6]. Question Answering System (QAS) is one of the fields in information retrieval that allow the users to input a question in the form of natural language [5] to get natural interaction between human and the computer. Users will also get a direct answer from the system quickly and accurately, without sort through the documents to get a reply.

The previous research of QAS for the Indonesian language 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 [1]. 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 al-Khattab Ra., Uthman ibn Affan Ra., and Ali ibn Abi Thalib Karamallahu wajhah.

The QAS that refers to M.N. Zidny, et,al [2] 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 [7] 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 Frequency-Inverse Document Frequency (TF-IDF) is used for generating weight vector for each concept in ITQ to perform the QAS for weighting the vector [3]. 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 [4]. According to Putra, et.al [3] 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 [4] is Cosine similarity which is used to find the semantic from the question with the candidate answer in the ITQ dataset.

III. METHODOLOGY

Fuzzy logic consists of three main phases: fuzzification, inference, and defuzzification. The researcher uses four fuzzy variables in this QAS: lucerne's score, Wijono's algorithm score, a similarity between passage and query text, and the distance between keywords and text entities. The inferencing phase uses the Mamdani model, as well as the Center of Gravity method used in the defuzzification stage.

This QAS uses architecture from M.N. Zidny, D.E. Mahmudah, S.J. Putra, and A.F. Firmansyah [2] or (QAS2F) as the base structure. As seen in Fig.1, the development was in the scoring process, by implementing fuzzy logic in that section.

The following details of QAS Architecture 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.

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 [1]. 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/). The result of the Question Analyzer to the question can be seen in the following table.

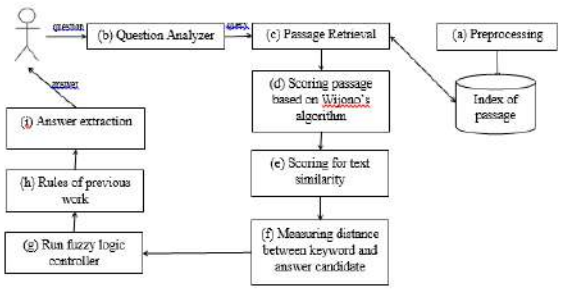

Fig. 1. QAS Architecture

TABLE I. RESULT OF QUESTION ANALYZER

Question	**:**	***Siapakah nama lengkap abu bakar?***
Boolean Query	:	nama AND lengkap AND abu AND bakar
Keyword Entity	:	[abu bakar]
Keyword	:	[nama, lengkap]

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 [7] to score all passages retrieved from Passage Retrieval.

E. Scoring for Text similarity

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 [8] from LingPipe library.

F. Measuring Keyword and Candidate Answer

At this stage, the system has been measuring the distance between keyword and candidate 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 [9]. JfuzzyLogic, one of a fuzzy logic library, has a step by step of the fuzzy rule-based system [10, 11].

H. Fuzzification

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. Membership functions and fields for all four variables are as seen in Table 2.

TABLE II. FUZZY SET

Input/ Output	Variable	Fuzzy Set	Membership Function	Domain
Input	Lucene Score	low	Trapezoidal	[0, 5]
		medium	Triangular	[2, 12]
		high	Trapezoidal	[6, ∞]
	Wijono Algorithm Score	low	Trapezoidal	[0, 2.75]
		medium	Triangular	[2, 3.25]
		high	Trapezoidal	[2.75, ∞]
	Similarity	similar	Trapezoidal	[0, 0.3]
		not similar	Trapezoidal	[0.175, 1.0]
	Distance	close	Trapezoidal	[0, 10]
		quite close	Trapezoidal	[7, 30]
		Far	Trapezoidal	[27, ∞]
Output	Worthiness	low	Trapezoidal	[0, 80]
		high	Trapezoidal	[50, 100]

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 : [9]

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 [9].

K. Rules of Previous Work

Based on the architecture by Zidny, et/al [2], 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 [1], then the results of the test are the value of accuracy of the QAS as stated by Zidny [1]. 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, the result is presented in Table 3. 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 Naf'an et al [1] and Naf'an et al [2] can be seen in Table 4.

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. Two wrong answers on the Naf'an et al [2] become true in the current research. Those are the question number 8 ("*Siapakah putri Rasulullah yang menikah dengan Utsman?*") and number 16 ("*Siapakah sahabat pilihan Rasulullah SAW?*").

TABLE III. TEST RESULT

Categories	Result	Percentage
Correct	42	70%
Unsupported	1	2%
Inexact	2	3%
Incorrect	15	25%
Unanswered	0	0%

TABLE IV. COMPARISON OF TEST RESULTS

Categories	(Naf'an et al., 2012)	(Naf'an et al., 2016)	Current Research
Correct	37	40	42
Unsupported	1	1	1
Inexact	1	2	2
Incorrect	8	17	15
Unanswered	13	0	0

The answer to question number 8 generated from Naf'an et al [2] 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. The current research, Crisp input value for the candidate answers "Ruqayyah" can be seen in Table 5, whereas the rule for the value of variable worthiness on Table 3 is high.

TABLE V. CRISP INPUT AND FUZZIFICATION RESULT FOR THE 16TH QUESTION

Variable Fuzzy	**Lucene score**	**Wijono algorithm score**	**Similarity**	**Distance**
Crisp Input	12.89	7	0.47	2
Fuzzification	μ high = 0.766	μ high = 1.0	μ not similar = 0.662	μ close = 1.0

Based on the results of Table 5, the fuzzy rule that meets the state is the 4th and 22nd rule. Both of these rules incorporate defuzzification results in a set of high worthiness with the score 0.948.

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 the fuzzy logic on the scoring process of Question Answering System and succeeded in increasing the accuracy level of Question Answering System. From the experiment result, implementing the fuzzy rules for answers ranks on Indonesian language QAS will increase the number of correct answers. It succeeded to convert the two incorrect answer become the right answer since fuzzy rules are made to place candidate correct answer in the first rank successfully.

REFERENCES

[1] 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.

[2] 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.

[3] 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

[4] 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.

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

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

[7] 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.

[8] 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.

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

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

[11] 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.