A Rule-Based Question Answering System on Relevant Documents of Indonesian Quran Translation

Ria Hari Gusmita[1], Yusuf Durachman[2], Salman Harun[3], Asep Fajar Firmansyah[4], Husni Teja Sukmana[5], Adam Suhaimi[6]
Department of Informatics Engineering[1,5], Department of Information System[2,4,6], Department of Islamic Education[3]
State Islamic University of Syarif Hidayatullah[12345], International Islamic University Malaysia[6]

***Abstract*- This paper presents work in development of a question answering (QA) system by using a combination of two different architectures i.e. the one used relevant documents and another used rule-based method, which those two contribute for answer extraction. Base on previous researches testing result, it could be inferred that each of the methods could be a complement for another method in order to increase system performance. This QA was purposed to gather information from Indonesian Quran Translation. The new architecture was designed to gather relevant documents toward the keywords and be used subsequentially to gather answer candidates by using rule-based method. The initial results indicate that system still restricted with retrieved relevant documents, and caused delivering only 60% correct answers. This achievement is not better than the previous one that used rule-based method only.**

I. INTRODUCTION

Question answering (QA) exist as one of efforts to solve weaknesses of search engines. Both of them are an implementation of information retrieval (IR), where IR focuses on retrieving relevant information from particular corpus base on keywords entered by user. Instead delivering several relevant documents just like search engines, question answering systems provide the exact answer related to user question. Another nice-facility of QA is in which user can use natural language to define their question just like as they want it to be.

Some QA researchs on Indonesian documents are done with employ several different architectures or methods. Standard architecture was applied in [5][11][13][18], which there are four components i.e. question analysis, document retrieval, document analysis, and answer selection. Another architecture also be used by other researchs as described in [4][8][14]. Those were practice rule-based method to gather answer from all of documents. First architecture still lacks at document retrieval and answer extraction, as not all retrieved documents were relevant and answer extraction's algorithm still trap on lexical sense. Second architecture is inefficient since rule-base method on answer extraction applied to all of documents.

Weaknesses of two architectures as mentioned above led this research to build new architecture by combine them. All components of first architecture are utilized except for document analysis that is replaced by implementation of rule-based method on relevant documents. This will subsequentially followed by answer extraction process that is a calculation adopted from second architecture. This new QA architecture is implemented on Indonesian Quran translation, which for the first version of it processed the first chapter in Quran namely Al-Baqarah. There are three kinds of question type allowed to be used in the system i.e. who, when, and where. In order to yield good performance, this new architecture equipped with several useful components defined in [18] that function at increase time execution and question processing reliability. Those components comprising usage knowledge and question structure analysis.

II. Question Answering System

A. *Standard Architecture*

Question answering system (QAS) is one field in information retrieval that process question in natural language form and return system's correct answer. In a dissertation's report entitled "From Information Retrieval to Question-Answering", [10] mentioned that "Information retrieval sistems that allow for users to pose natural language questions are known as question answering systems". A question answering system aims at processing a natural language question and find a location in a document that contains answer[2]. Architecture of question answering system depicted in the following figure:

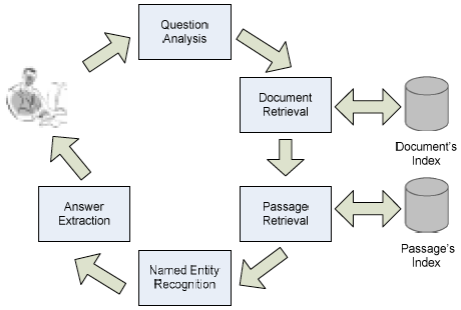

Fig. 1. Standard Architecture of Question Answering System

Figure 1 shows that QAS start with analyze user's question to get question type, keywords, and keywords entity. Keyword is used by document retrieval to gather relevant documents from related corpus. Retrieved documents are splitted into passages that is will benefit system in finding answer. Furthermore, named entity recognition will put tag on every word within each passage that has entity's name. Hereafter, every passage will scored base on number of words that has similar entity's

name of answer type. Question type gathered in the first step Answer type is useful to get answer type. Finally, system will implement an algorithm to find the answer within the high-score passages that have enriched with name entity's tag.

B. Rule-based Method

Quarc is a reading comprehension test developed by [12] that utilize rule-based method to find the correct answer for each question. Rule-based methods uses lexical and semantic heuristics to look for evidence that a sentence contains the answer to a question[12]. Here is an example of a rule used on question type "who":

1. Score(S) += WordMatch(Q,S)
2. If ¬ contains(Q,NAME) and contains(S,NAME)
   Then Score(S) += **confident**
3. If ¬ contains(Q,NAME) and contains(S,*name*)
   Then Score(S) += **good_clue**
4. If contains(S,{NAME,HUMAN})
   Then Score(S) += **good_clue**

Fig. 2. Rule for Question Type "Who" in English

C. New Architecture of Question Answering System

As explained in chapter I, new architecture of question answering system developed by combining two different question answering system's architectures in order to enhance system's performance. This following figure shows the new architecture:

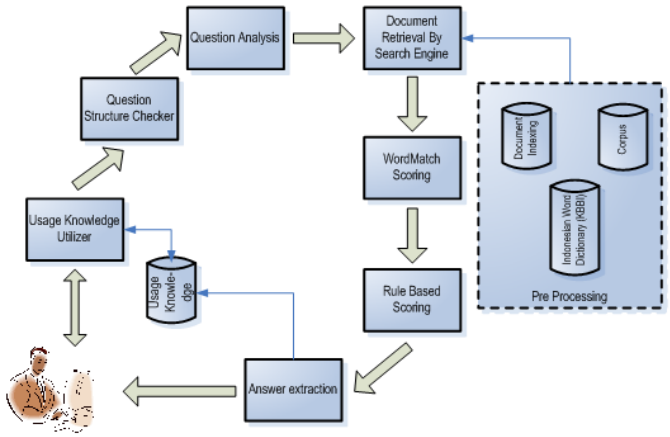

Fig. 3. New Architecture of Question Answering System

New process in this architecture could be seen after system got relevant documents. Word match scoring function is applied to each of those documents (not all documents definitely) to count number of similar words between question and document. Relevant score is set to each document and it will be an initial score for each of them. Furthermore, relevant documents are getting processed by rule-based scoring component to get final score. This is done by implement a suitable rule of question type. System will rank all scored documents, and find the correct answer within the highest scored document.

III. IMPLEMENTATION

This part describes what rule implemented for all of question type namely who, what, and where in the new architecture. Moreover, testing result also explained in this part along with related analysis.

A. Rule of Who Question Type

a) Score (S) + = WordMatch (Q,S)

b) If ¬contain(Q, NAMA) and contains (S, NAMA)

Then Score (S) += confident

c) If ¬contains (Q, NAMA) and contains (S, *nama*)

Then Score (S) += good_clue

d) If contains (S,{NAMA, ORANG})

Then Score(S)+= good_clue

Fig. 4. Rule of Who Question Type on Indonesian Quran Translation

B. Rule of When Question Type

a) If contains (S, WAKTU)

Then Score (S) += good_clue

Score (S) += WordMatch (Q,S)

b) If contains (Q, {*akhirnya, terakhir*}) and contains (S, {*saat, sesaat, ketika, kala, semenjak, sejak, waktu, setelah, sebelum*})

Then Score (S) == slam_dunk

c) If contains (Q, {*mulai, memulai, pertama*}) and contains (S, {*saat, ketika, kala, semenjak, sejak, waktu, setelah, sebelum*})

Then Score (S) += slam_dunk

Fig. 5. Rule of When Question Type on Indonesian Quran Translation

C. *Rule of Where Question Type*

a) Score (S) += WordMatch (Q,S)

b) If contains (S, KATA_DEPAN)

Then Score (S) += good_clue

c) If contains (S, TEMPAT)

Then Score (S) += confident

Fig. 6. Rule of Where Question Type on Indonesian Quran Translation

D. *Testing Result*

Testing is done by using ten question for each of question type. In order to compare system's performance, all questions come from previous research that used rule-based method only[4]. Base on the testing result, system still not able to be better than the previous one as explained in this following table:

TABLE I
COMPARISON OF SYSTEM'S RESULT AND PREVIOUS SYSTEM[12]

Question Type	Percentage of number of correct answer in current research	Percentage of number of correct answer in previous research
Who	60%	97,5%
When	60%	90%
Where	40%	68,03%

Table I shows that system's performance was still not as good as previous system's performance. This is caused of several reasons:

1. System was not always get relevant documents from search engine. This implied answer extraction's performance that system delivered incorrect answers or even none answers. Figure below explained the evidence:

```
QUESTION: Siapa yang termasuk orang-orang merugi
QUERY: orang orang merugi
BOOLEAN QUERY: orang AND orang AND merugi
CATEGORY: PERSON
KEYWORD ENTITY: []
KEYWORD : [yang, termasuk, orang, orang, merugi]

# PASSAGES #
Found 0 documents in 16ms that matched query = orang AND
orang AND merugi

# SCORING PASSAGE #
java.lang.NullPointerException
   at
org.qa.evaluation.EvaluationQuestionAnswering.main(Evalua
tionQuestionAnswering.java:310)
```

Fig. 7. Screen Shoot of System's Failure in Answering Question

2. Incorrect answers produced when system found an irrelevant document with highest score. This can be seen at following figure:

```
Siapa yang memiliki langit dan bumi?
# PASSAGES #
Found 9 documents in 20ms that matched query = langit
AND bumi

# SCORING PASSAGE #
2 - 33 - Allah berfirman: "Hai Adam, beritahukanlah
kepada mereka nama-nama benda ini". Maka setelah
diberitahukannya kepada mereka nama-nama benda itu,
Allah berfirman: "Bukankah sudah Ku-katakan kepadamu,
bahwa sesungguhnya Aku mengetahui rahasia langit dan
bumi dan mengetahui apa yang kamu lahirkan dan apa yang
kamu sembunyikan?"
```

Fig. 8. System found Answer from The Highest Scored Document

Relevant document in above figure is the highest scored document as it hadhighest number of person's semantic class's words (question type was who) as figure out below:

```
Allah berfirman: "Hai Adam, beritahukanlah kepada mereka
nama-nama benda ini". Maka setelah diberitahukannya
kepada mereka nama-nama benda itu, Allah berfirman:
"Bukankah sudah Ku-katakan kepadamu, bahwa sesungguhnya
Aku mengetahui rahasia langit dan bumi dan mengetahui apa
yang kamu lahirkan dan apa yang kamu sembunyikan?"
```

Fig. 9. Existence of Many Person's Semantic Class's Words within Irrelevant Document

Number of person's semantic class's words within above document is 8 (eight). This number is bigger than number from others document that actually relevant as described below:

```
First relevant document, contains 3 words from semantic
class NAMA and ORANG :
Ayat 107 - Tiadakah kamu mengetahui bahwa kerajaan langit
dan bumi adalah kepunyaan Allah? Dan tiada bagimu selain
Allah seorang pelindung maupun seorang penolong.

Second relevant document, contains 5 words from semantic
class NAMA and ORANG:
Ayat 116 - Mereka (orang-orang kafir) berkata: "Allah
mempunyai anak". Maha Suci Allah, bahkan apa yang ada di
langit dan di bumi adalah kepunyaan Allah; semua tunduk
kepada-Nya.

Third relevant document, contains 7 words from semantic
class NAMA and ORANG:
Ayat 284 - Kepunyaan Allah-lah segala apa yang ada di
langit dan apa yang ada di bumi. Dan jika kamu melahirkan
apa yang ada di dalam hatimu atau kamu menyembunyikan,
niscaya Allah akan membuat perhitungan dengan kamu
tentang perbuatanmu itu. Maka Allah mengampuni siapa yang
dikehendaki-Nya dan menyiksa siapa yang dikehendaki-Nya;
dan Allah Maha Kuasa atas segala sesuatu
```

Fig. 10. List of Relevant Documents along with Number of Person's Semantic Class's Words

IV. CONCLUSION

We have presented a development of new architecture of question answering system that combine two existing architectures for gathering information from Indonesian Quran translation. Due to limitation of lucene as search engine's library in processing Indonesian documents and finally imply document retrieval function, our system still not able to increase performance in delivering correct answers.

We believe that refinement of lucene toward Indonesian documents along with further analysis of Indonesian question type's rule will benefit this research in the future, and so development of question answering system for whole chapter in Quran can be established.

REFERENCES

[1] Alwi, H, Soenjono Dardjowidjojo, Hans Lapoliwa, Anton M. Moeliono, *Tata Bahasa Baku Bahasa Indonesia*, PT Balai Pustaka, Departemen Pendidikan dan Kebudayaan Republik Indonesia, Jakarta, 1998.

[2] Bilotti, Matthew. W. dan Eric Nyberg., "Improving Text Retrieval Precision and Answer Accuracy in Question Answering Sistems," 22nd

International Conference on Computational Linguistics, Manchester, UK: Association for Computational Linguistics (ACL), p. 1-8, 2008.

[3] Bütcher, S., dkk., *Information Retrieval; Implementing and Evaluating Search Engines*, London: MIT Press, 2010.

[4] Dwi Anggraeni Meynar, *Implementasi Question Answering Sistem dengan Metode Rule-Based pada Terjemahan Al-Qur'an Surat Al-Baqarah*, Skripsi Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Pertanian Bogor, 2007.

[5] Gunawan dan Gita Lovina, "Question Answering Sistem dan Penerapannya pada Al-Kitab," *Jurnal Informatika*, Universitas Kristen Petra, 2006.

[6] Hovy Eduard, Gerber Laurie, Hermjakob Ulf, Junk Michael, Lin Chin-Yew, "Question Answering in Webclopedia," *The Ninth Text REtrieval Conference*, November 13 – 16, 2000.

[7] Jurafsky Daniel, Martin H. James, *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall, 2000.

[8] Lutfi Citra Rosiana, *Question Answering Sistem pada Terjemah Juz Amma Menggunakan Metode Rule Based*, Tugas Akhir Fakultas Sains dan Teknologi, UIN Maulana Malik Ibrahim, 2012.

[9] Manning, Cristopher D., et al., *Introduction to Information Retrieval*. New York: Cambridge University Press, 2008.

[10] Monz, Christof, *From Document Retrieval to Question Answering*, Amsterdam: Universiteit van Amsterdam, 2003.

[11] Naf'an Zidny, Hari Gusmita Ria, *Sistem Tanya Jawab Berbahasa Indonesia tentang Sejarah Khulafaur Rasyidin*, Skripsi Fakultas Sains dan Teknologi, UIN Syarif Hidayatullah Jakarta, 2012.

[12] Riloff Ellen, Thelen Michael, "A Rule-Based Question Answering Sistem for Reading Comprehension Tests," *ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Sistems*, 2000.

[13] Wijono, Sri Hartati, dkk. *Finding Answers to Indonesian Questions from English Documents*. Laporan Kerja Wokrshop Cross Language Evaluation Forum, 2006.

[14] Ikhsani N, *Implementasi Question Answering Sistem dengan Metode Rule-Based untuk Temu Kembali Informasi Berbahasa Indonesia, Bogor*, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor, 2006.

[15] Panitia Pengembangan Bahasa Indonesia, *Pedoman Umum Ejaan Bahasa Indonesia yang Disempurnakan*, Pusat Bahasa Departemen Pendidikan Nasional, Indonesia, 2000.

[16] Josef Essberger, *English Prepositions List*, EnglishClub.com, England, United Kingdom, 2012.

[17] Croft W. Bruce, Metzler Donald, Strohman Trevor, *Search Engine Information Retrieval in Practice*, 1st ed., Pearson Education, 2009

[18] Lestari, Antania Hanjani dan Gusmita, Ria Hari, "Designing of Improvement of Question Answering Sistem about Khulafaur Rasyidin's History by Implementing Usage Knowledge, Question Grammatical Checker, Question Structure Analyzer, and Indonesian Big Dictionary on Stemming Process," in *Proceedings of The 2nd International Conference on Information Technology for Cyber & IT Service Management (CITSM) 2013 in Conjunction With The ITIL 2011, Prince2, And COBIT5 Workshop, Training, and Certification*, Jakarta, 2013