Information Extraction from Web as Knowledge Resources for Indonesian Question Answering System

Abdiansah ABDIANSAH[1*] and Alvi Syahrini UTAMI[2]

[1,2]AIRLab, Comp. Sci. Dept., Universitas Sriwijaya, Indonesia
**Corresponding author: abdiansah@unsri.ac.id*

Abstract
Research in the field of Open Domain Question Answering System (OD-QAS) generally involves external knowledge which are dynamic and require high-level representation. Strong external knowledge is one of the key success of QAS. Therefore, intensive research is needed in this area. Web is one of the big source of information that can be used as external knowledge by QAS. However, the main problem is the Web contains a lot of unstructured data. Hence, a model is needed to extract information from the web. The model developed in this research based on pipeline architecture and consists three main processes: pre-processing, information extraction processing, and text processing. The input model is factoid questions, and the output are snippets or set of sentences that contains target answers. There are three search engines assist to finding relevant information from the Web, i.e, Yahoo!, Bing, and Ask. The result of average precision and deviation value for the each search engines are slightly different. The highest total number of snippets (true positive) generated by Yahoo! is 65 snippets, while the best average precision obtained by Bing is 25.33%.

Keywords: *information extraction, Indonesian Question Answering System*

Introduction

Many studies in Question Answering System (QAS) has been conducted, especially for English target, which is QAS that processes questions and produces answers in English. Nowadays, QAS still actively researched by scientis [1]. However, research of QAS with Indonesian target or Indonesian Question Answering System (IQAS) is still rarely explored. In fact, Indonesian is the official language used by more than 230 million people [2]. Several scientists have been studied about IQAS [3, 4]. Their research focus on "how to find the correct answer" from a question based on set of documents. In addition, the IQAS that has been they developed is still specific domain and used static knowledge. One of the studies in the field of QAS related to knowledge domain is searching for answers from unlimited sources known as the Open Domain Question Answering System (OD-QAS).

Research in the field of OD-QAS is still being investigated [11, 12, 13, 14, 15, 16]. It usually involves dynamic external knowledge with high-level abstraction and representation. Many researchers focus on external knowledge because it determines the success of QAS. There are two kinds of external knowledge, i.e, structured knowledge like open-data (ex. DBPedia) and unstructured knowledge (ex. Web). The use of open-data for IQAS is not yet maximal because the available data is still limited for the Indonesian language. Therefore, the only way is to extract information directly from the Web, but the obstacles that will be faced may quite a lot, such as HTML to text conversion, eliminating text noise, proper sentence separation based on context, and so on.

Web is one of information source that contains large and growing data [5]. Data from the Web is more unstructured and contains noise. In addition, the absence of a general framework for handling unstructured data has led scientists to develop techniques for specific domains [6]. One important issue relating to data retrieval from the Web is "how to transform unstructured data into structured data". Data from the Web can be processed by a system if the data has well structured.

Based on the previous information, there are two problems discussed in this paper, as follows:
How to extract information directly from the Web?
How can the information be used as knowledge for IQAS?
The above problems are tried to be solved by building and testing model of information extraction from the Web. The model aims to extract information from the web based on a question. The output of model is set of text (snippets) that expected contains "potential answers" to the question. The results of model are measured using precision by observing how many snippets are true (true positive) and not true (false positive) to the actual value.

RELATED WORKS

IQAS research is becoming active in 2005, which was pioneered by Adriani and Rinawati [7] while attending the Cross-Language Evaluation Forum (CLEF): Question Answering Track. The latest IQAS research publication is conducted by Gunawan et al. [4], who created a question-answering system for solving arithmetic in a robot. Furthermore, the following sub-chapter will explain the latest research on IQAS in last five years.

Suwarningsih et al. [8] developed IQAS that focused on Question Generation, which is a component that aims to create questions automatically from various inputs such as texts, databases, and semantic representations. The IQAS is made especially for the medical field, called Indonesian Medical Question Answering (IMQA). Furthermore, Putra et al. [9] developed IQAS based on semantic for Quran Terjemah Indonesia (QTI) or Indonesian Translated Quran. They made specific ontology in QTI field based on three factoid questions (who, where, and when).

The research by Suwarningsih et al. [10] continued the earlier research, which focused on Question Classification module, especially for medical question. Furthermore, Saelan et al. [3] developed IQAS, especially for comparison purpose called Comparative Question Answering System (CQAS). The research focused on Question Classification module by examining Comparative Question (CQ), which is a question comparing two or more entities. Finally, Gunawan et al. [4] create IQAS for the field of Intelligent Humanoid Robot (IHR). The research focused on solving the Arithmetic Word Problem, which is problem-related to mathematical representation in the form of words or sentences.

Previous researches shows that there are no studies using information from the Web as source for IQAS knowledge. Generally, the IQAS that was build is still to specific domain and focused on finding the right answer. Besides, the research related to Open Domain Question Answering System [15, 16] has been conducted, especially for English target, but for Indonesian target as our knowledge there are no scientists focused on this area.

RESEARCH METHOD

Figure 1 shows the model of information extraction from the Web. Its architecture based on pipeline processing, so that the sequence of processes run sequentially. The model consists of three main processes:

Pre-processing aims to turning the questions into the queries of search engine.

Information extraction processing aims to extracting and processing text from the Web. Text processing aims to turning text into set of sentences that contains potential answers

The model in Figure 1 is implemented using Python programming language and using third-party libraries such as NLTK, Requests, BeautifulSoup, and others. Next sub-chapter will explain the three processes above and discuss measuring result of the model.

Pre-processing

In Question Answering System, a question is main input for the system. Until now, there is no standard how to modeled the question. typically, question is formed by using question words

(5W1H) in the begin of sentence and question mark in the end of sentence (?). In this research, we are only using "who" question type and seven questions as follows:

"siapa nama presiden indonesia pertama?"

(who is the name of the first indonesian president?)
"siapa nama presiden indonesia kedua?"
(who is the name of the second indonesian president?) "siapa nama presiden indonesia ketiga?"

(who is the name of the third indonesian president?)
"siapa nama presiden indonesia keempat?"
(who is the name of the fourth indonesian president?)
"siapa nama presiden indonesia kelima?"
(who is the name of the fifth indonesian president?)
"siapa nama presiden indonesia keenam?"
(who is the name of the sixth indonesian president?)
"siapa nama presiden indonesia ketujuh?"
(who is the name of the seventh indonesian president?)

The seven questions above seem to be similar and only differ in the last word of the sentence. Although, all of these questions are similar, the answers to these questions are different. This research focuses on how to find potential answers as many as posibble based on information available on the Web. Therefore, we are not given much attention to the form and the number of questions.

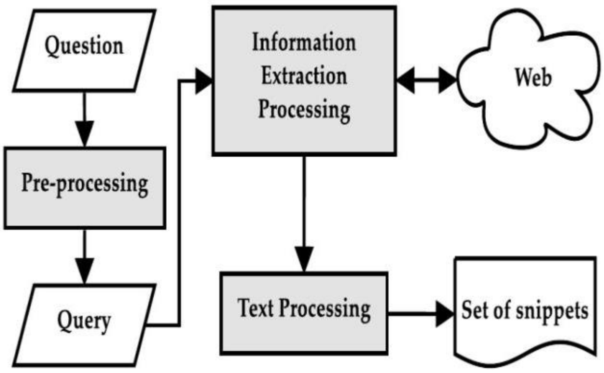
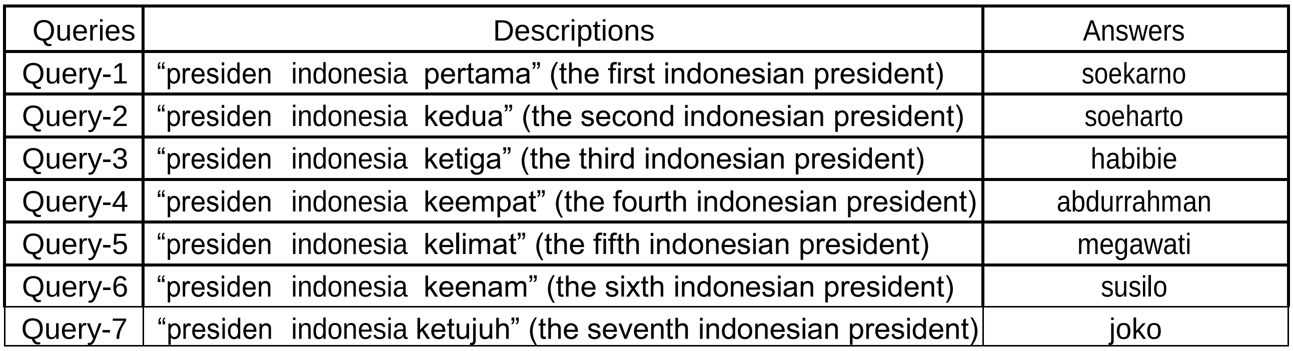


Figure 1. Model of extraction information from Web. The model based on pipeline processing model. There are three main process i.e pre-processing, information extraction processing, and text processing. The model 's output is set of snippets that have potential answers.

Next, all questions will be converted into queries to be processed by the search engine. Table 1 contains list of queries that correspond to the expected answers. The table consist of three column, i.e. queries, descriptions, and answer. In the description column, there are manual removal of question word (who) and question mark (?). This removal aims to reduce the queries of search engine. The answer column contains the correct answer from the given question, for example, "soekarno" is the answer for question "siapa nama presiden indonesia pertama?" (who is the name of first Indonesian president?).

We have limited for a question with multiple potential answers in order to focuses on process of information extraction from the Web. For example, the question "siapa nama presiden indonesia keempat?" (who is the name of the fourth Indonesian president?) has two potential answers: "abdurrahman wahid" (formal) and "gusdur " (informal). In this study we only takes just one answer and the first word is chosen if the answer has more than one word. For the previous example, we choose "abdurrahman" as answer for the given question.

Table 1. List of queries that correspond to expected answers.



Queries	Descriptions	Answers
Query-1	"presiden indonesia pertama" (the first indonesian president)	soekarno
Query-2	"presiden indonesia kedua" (the second indonesian president)	soeharto
Query-3	"presiden indonesia ketiga" (the third indonesian president)	habibie
Query-4	"presiden indonesia keempat" (the fourth indonesian president)	abdurrahman
Query-5	"presiden indonesia kelimat" (the fifth indonesian president)	megawati
Query-6	"presiden indonesia keenam" (the sixth indonesian president)	susilo
Query-7	"presiden indonesia ketujuh" (the seventh indonesian president)	joko

Information Extraction Processing

Queries obtained in the previous process then sent to the search engine. We are used three search engine in this study i.e. Yahoo!, Bing, and Ask. They are chosen because more accessible and available than Google (popular search engine)1 . The search engine gives HTML pages contains URLs that are relevant to the query. The HTML page then processed to remove HTML tags and retrieve the URLs on pages. Generally, there are 10 URLs on the first page of the search engine results. The URLs then processed online to retrieve the HTML files. After that, all HTML files converted into text file (HTML to text) using BeutifulSoup2 and the results combined into one text file.
The text which results from BeautifulSoup still have a lot of noise. One kind of the noise is two or more words was merged as effect from process separating text and HTML tags (ex. "seharusnyakita", "presidenpertama" etc). Therefore, we build noise removal process which separate keywords and non-keywords. For example, the word "presidenpertama" will be "presiden" and "pertama" after being processed by noise removal process. We do not process words as non-eywords which merge together, because they are not related to the process of finding answer (ex. "seharusnyakita").

Text Processing

After the text is assumed cleaning from the noise, the next step is doing tokenization process. It aims to separate sentences from the text. We uses Punkt algorithm to perform tokenization which is available in the NLTKs library. Punkt algorithm is build using supervised technique, so that we need data training in Indonesia language3. After set of sentences obtained, the next step is finding answers in sentences using Regex (Regular Expression) technique. If a sentence contains answers or called snippet, the sentence is considered to have potential answer, even though it is not correct. The final result of text processing is set of snippets.

Measurement

The result of model is measured by precision because we just considered both the number of snippets contains appropriate (true positive) and inappropriate (false positive) that correspond to the actual answers. The precision formula can be seen in eq. (1) with TP as true positive, and FP as false positive. We also compare the results of three search engines (Yahoo!, Bing, and Ask). The final result is total average of precision for each search engine. The higher the average, then the better the result.

$$\text{Precision} = TP / (TP + FP) \;\ldots. \qquad (1)$$

RESULTS

Figure 2 gives the results of the first query (Query-1) and deviation standard for each search engine. The above chart shows the sample of the first query, which provides three information, i.e, the length of the sentences, the average of sentence length, and the value of deviation standard. It can be seen Bing gives the highest number of sentences compared to the others, which is 209 sentences. The difference between the result of Bing and Yahoo! is not too significant, that is 18 sentences, while the difference with Ask is quite significant at 191 sentences. However if the number of the sentences resulting from all queries are combined, Yahoo! Provides the highest result, that is 980 sentences, followed by Bing that is 832 sentences and finally Ask that is 120 sentences. Next, the average of sentence length for Yahoo! and Bing is quite normal, around 60 and 30 words for each sentence, while for Ask, the resulting sentences are too long. Finally, Bings deviation standard is smaller than the others at 154. These results provide information that Yahoo! gives the most number of sentences compared to the others, while Bing gives the smaller deviation standard value than the others.

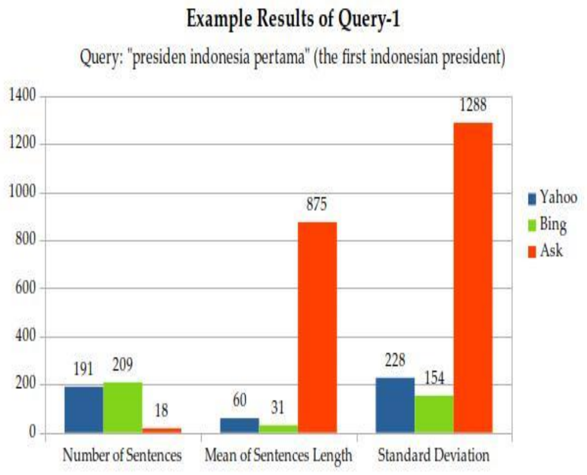


Figure 2. Show some statistic values (mean and standard deviation) for Query-1. It is yielded based on number of sentences that produces automatically by experiment tool using three search engines (Yahoo, Bing, Ask).

Figure 3 shows the result of deviation standard of each search engine for each query. It can be seen from the figure, that the difference of deviation standard between Yahoo! And Bing is not too large. Even for Query-3 and Query-4, the deviation standard of Yahoo! is smaller than Bing. If the deviation standard values of all queries are combined then Yahoo! gives smallest result, which is 2.405, followed by Bing (3.191), and Ask (7.143). The smaller deviation standard, the better result for sentence length composition. Based on this information, Yahoo! used as the main search engine for the model because it gave better results than others.

Figure 4 shows the number of potential answers from each document generated by Yahoo!. The ideal expected outcome is the number of answers from query and target answer is higher than non-arget answers. For example, in the Doc-Q1 (Document of Query-1), the number of answers “soekarno” (Q1-Soekarno) should be higher compared to the others, because the document is generated from query that matches to the target answer “soekarno”. Furthermore, in the figure, there are only four by seven that match the expectations, i.e, Doc-Q1, Doc-Q2, Doc-Q3, and Doc- 7 or about 57,14% from all documents.

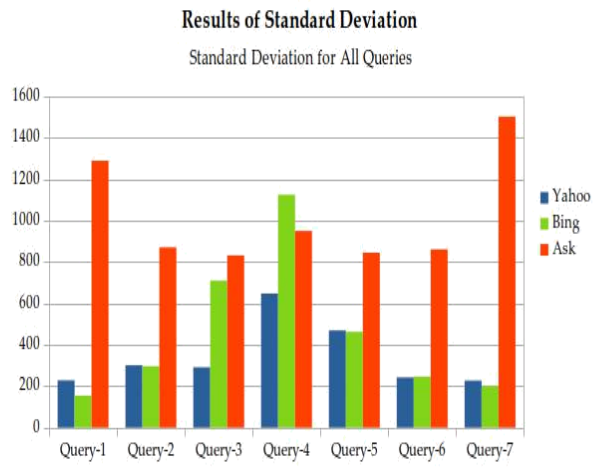


Figure 3. standard deviation values for all queries.

Next, the result of precision for each search engine. In Figure 4, the potential answers are found based on the number of answers in a document. This results obtained automatically by system using regex technique. Manual calculation is performed to find the right precision, because even if the sentence contains answers, it is not yet sure to be valid snippets. In Table 2, we can see the number of snippets generated by each search engine based on queries and answers. Yahoo! produces more sentences contains answers than the others, which is 265 sentences. Interestingly, the result of Ask gives greater number of sentences than Bing, even though the overall Bing gives the larger number of sentences that extracted from web.

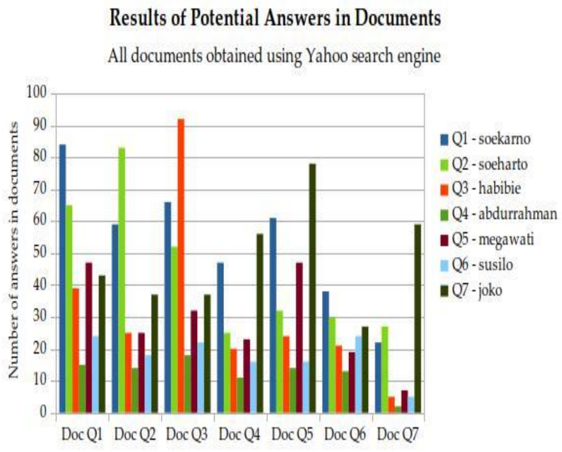


Figure 4. The chart show potensial answers in documents. Yahoo is used as search engine to obtained all documents because it outperforms than Bing or Ask. In the chart, each document have number of answers.

Table 2. Number of snippets generated by model (* is the best value).

Search Engines	QD-1	QD-2	QD-3	QD-4	QD-5	QD-6	QD-7	Total
Yahoo!	66	55	48	10	29	21	36	265*
Bing	45	33	41	8	33	21	36	217
Ask	54	42	31	16	25	14	35	227

Table 3 contains the results of precision values for each search engines. The documents that processed are documents based on the expected queries and target answers, for example, the QD-1 column is a document obtained from Query-1 (“presiden indonesia pertama”) that contains snippets with the answer “soekarno”. The experimental results show that overall precision is only slightly different from the average and deviation standard of (24,67% and 7,47), (25,33% and 8,32), and (24,77% and 6,30) respectively for Yahoo!, Bing, and Ask. Interestingly, Ask gives the highest precision value and the lowest deviation standard. However, the number of true positive for overall documents obtained by Yahoo! is 65 snippets, followed by Ask (56 snippets), and Bing (54 snippets). These results indicate that the search engine does not significantly influence to the model. The best average of precision is only 25.33% so that we need more improvement.

Table 3. Precision values obtained from manually checked (* is the best value).

Search Engines	QD-1	QD-2	QD-3	QD-4	QD-5	QD-6	QD-7
Yahoo!	27,27%	21,82%*	16,67%	30,00%	34,48%*	28,57%	13,89%
Bing	26,67%	21,21%	21,95%	37,50%*	30,30%	28,57%	11,11%
Ask	29,69%*	19,05%	22,58%*	31,25%	28,00%	28,57%	14,29%*

DISCUSSION

Based on the experimental results above, there are several things need to be discussed. It should be noted that the model developed in this study as initial model for extracting information from Web, especially in the field of Indonesian Open Domain Question Answering System. Therefore,several deficiencies need to be corrected. Three important factors need more attention. First, pre-processing that focuses on obtained good quality and unambiguous query. In this study, the process to change a question to the query is done manually, whereas in the real cases it should be done automatically. Also, the types of the question need to be expanded, not just limited to the factoid questions. Information obtained from the Web is very depends on the query provided. Query generation is needed to get more information so that more than one type of query is processed by search engine.

Second, text processing that focuses on changing raw text into collection of sentences containing candidate answers (snippets). Specific text processing tools for Indonesian language are still limited, so this is the main obstacle in this study. The model is build based on pipeline architecture, so that the output of a process will be input for the next process. This architecture causes very high level dependency, for example, the noise as result from conversion HTML to text might affect the sentences tokenization and subsequent processes. This research still uses standard and limited text processing tools, so it is possible to be improved and expanded. The outcome of text processing gives many linguistic phenomena that discussed separately. Here are some cases of linguistic phenomena that occured in this study: co-references ("soekarno-hatta dilantik menjadi presiden dan wakil presiden pertama indonesia"), contextual ("megawati soekarnoputri selaku presiden perempuan pertama di indonesia" in QD-1), incomplete-text ("soekarno presiden soekarno presiden pertama ri "), numerical ("presiden ke 2 indonesia soeharto"), and others. The model has not been made specifically for these linguistic phenomena.

Third, answer identification focuses on marking words or phrases that considered as answers. The model is made to identify the answer manually by determining the direct answer, for example, the question "siapa nama presiden indonesia pertama?" (who is the name of the first indonesia president?) will be given the direct answer "soekarno". The manual identification conducted because this study focuses on build a model to extract information from the Web and observes how many sentences (have correct answers) generated by model. In real cases, the model should identify the answer automatically, for example, by using question classification and name-entity recognition. Besides, it is possible to give weight for each entity considered as answer candidate. In the experimental result of the model, there are two things can be discussed, i.e, the number of sentences, and precision value generated based on search engine. This model has not determined the limit of the maximum sentence length. If a sentence is too long, then it is not optimal for snippets. The sentence length generated by Yahoo! and Bing is quite realistic, while the results from Ask are still too long (see Figure 2). The process of generating sentence is greatly influenced by the technique of sentence tokenization, and the previous text processing such as noise-removal and others. In Figure 3, we have 57.14% according to the expectations. Even in Doc-4, the result is opposite to the expectations but the all expected results are highest than un-expected results. One of the main factors is because the relationship between target answer and information availability, for example, in Doc-4 ("presiden indonesia keempat"), which uses formal answer "abdurrahman", it could be that "gusdur" is more popular in the Web than "abdurrahman wahid". Furthermore, the precision value is still low, with the best average is 25,33%. But these results are still open to be improved, considering the model developed is very simple.

CONCLUSION

This article describes the model of information extraction from the Web for Indonesian Question Answering System. It contains three main processes i.e. pre- processing, information extraction processing, and text processing. The output of model is set of sentences that potentially contains answers (snippets) from given queries (questions). There are three search engines used to search relevant information from the Web, i.e. Yahoo!, Bing, and Ask. The experimental results shows that Yahoo! gives better result than other, both from the number of sentences and the average deviation standard for all queries. Therefore, Yahoo! is used as the main search engine for the model.

The experimental obtained 57,14% for expectations results that used to see how many potential answers from a document using Yahoo!. This result obtained by observing how many answers are contained in a document using regex function. There many factors influences the result, such as noise removal technique and text processing tool where both not yet optimal for Indonesian language. Next, the precision involved to measure the number of snippets contains appropriate and inappropriate that correspond to the actual answers and how many matched and unmatched answers relating to the context of sentences. A sentence may contains answer that does not fit to the context of supporting sentence. The context of sentence examined manually. The final results show that the average and deviation standard of precision for the three search engines are slightly different. While the total number of snippets for the largest number produced by Yahoo!, which is 65 snippets from 265 snippets. The experimental results show that the search engine does not significantly influence the model. The best average of precision yield is 25.33%, so it needs to be improved.

FUTURE WORKS

The model of information extraction from the Web that developed in this research is highly dependent on many factors, such as the relevance information generated by search engine, target language, text processing tool (e.g. noise removal, parser, sentence tokenization, NER, syntactic, similarity function, etc), and system infrastructure (network connection, processing unit, etc).

Nevertheless, the model can be used as baseline for this field. The next research will focus on pre-rocessing, especially in cleaning text so that we have good quality snippets.

REFERENCES

[1] Mathur, A., & Haider, Question answering system: A survey. In 2015 International Conference on Smart Technologies and Management for Computing, Communication, Controls, Energy and Materials ICSTM (2015), pp. 47-57. IEEE.

[2] BPS. 2010. Badan Pusat Statistik. Information on https://sp2010.bps.go.id/index.php.

[3] Saelan, A., Purwarianti, A., & Widyantoro. Question analysis for Indonesian comparative question. In Journal of Physics: Conference Series (2017), Vol. 801, No. 1. IOP Publishing.

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

[5] Gangemi, A., Recupero, D. R., Mongiov, M., Nuzzolese, A. G., & Presutti, Identifying motifs for evaluating open knowledge extraction on the Web (2016), 108, 33-41.

[6] Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner. Web data extraction, applications and techniques: A survey (2014), 70, 301-323.

[7] Adriani, M. & Rinawati, University of indonesia participation at query answering-CLEF 2005. In Working Notes CLEF (2005) Workshop, Vienna, Austria.

[8] Suwarningsih, W., Supriana, I., & Purwarianti, A, Discovery Indonesian Medical Question-Answering Pairs Pattern With Question Generation. International Journal of Applied Engineering Research (2015), pp. 34217-34223.

[9] Putra, Gusmita, R. H., Hulliyah, K., & Sukmana, H. T. A semantic-based question answering system for indonesian translation of Quran. In Proceedings of the 18th International Conference on Information Integration and Web-based Applications and Services (2016), pp. 504-507.

[10] Suwarningsih, W., Purwarianti, A., & Supriana. Indonesian medical question classification with pattern matching. International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (2016), pp. 106-109. IEEE.

[11] Das, R., Dhuliawala, S., Zaheer, M., & McCallum. Multi-step retriever-reader interaction for scalable open-domain question answering, (2019), arXiv preprint arXiv:1905.05733.

[12] Lee, K., Chang, M. W., & Toutanova, Latent Retrieval for Weakly Supervised Open Domain Question Answering, (2019), arXiv preprint arXiv:1906.00300.

[13] Liu, J., Lin, Y., Liu, Z., & Sun, M: A Cross-lingual Open-domain Question Answering Dataset. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (2019), pp. 2358-2368.

[14] Seo, M., Lee, J., Kwiatkowski, T., Parikh, A. P., Farhadi, A., & Hajishirzi, Real-Time Open-Domain Question Answering with Dense-Sparse Phrase Index (2019) arXiv preprint arXiv:1906.05807.

[15] Sun, H., Bedrax-Weiss, T., & Cohen, PullNet: Open Domain Question Answering with Iterative Retrieval on Knowledge Bases and Text, (2019) arXiv preprint arXiv:1904.09537.

[16] Yang, W., Xie, Y., Lin, A., Li, X., Tan, L., Xiong, K., ... & Lin, End-to-end open-domain question answering with bertserini (2019), arXiv preprint arXiv:1902.01718