EVE: AN AUTOMATED QUESTION ANSWERING SYSTEM FOR EVENTS INFORMATION

Ivan Christanno[1], Priscilla[2], Jody Johansyah Maulana[3],
Derwin Suhartono[4], and Rini Wongso[5]

[1,2,3,4,5]Computer Science Department, School of Computer Science, Bina Nusantara University
Jln. K.H. Syahdan No 9, Jakarta Barat, DKI Jakarta, 11480, Indonesia
[1]ivan_christanno@binusian.org; [2]priscilla@binusian.org; [3]nikel_squarehead@binusian.org;
[4]dsuhartono@binus.edu; [5]rwongso@binus.edu

Received: 11th November 2016/ **Revised:** 19th January 2017/ **Accepted:** 26th January 2017

***Abstract* -** The objective of this research was to create a closed-domain of automated question answering system specifically for events called Eve. Automated Question Answering System (QAS) is a system that accepts question input in the form of natural language. The question will be processed through modules to finally return the most appropriate answer to the corresponding question instead of returning a full document as an output. The scope of the events was those which were organized by Students Association of Computer Science (HIMTI) in Bina Nusantara University. It consisted of 3 main modules namely query processing, information retrieval, and information extraction. Meanwhile, the approaches used in this system included question classification, document indexing, named entity recognition and others. For the results, the system can answer 63 questions for word matching technique, and 32 questions for word similarity technique out of 94 questions correctly.

Keywords: closed domain, question answering system, event information

I. INTRODUCTION

At the present time, information has been a very important part of everyday life. One of the main sources of information retrieval is from the Internet. However, the Internet is always growing which means there is more and more information added every moment possible. Because of this rapid growth of Internet, the information may be available most of the time, but it is not always accurate, especially with the growing redundancy of available data and information.

On the other hand, information may not even be available on the Internet. For example, information about an event may be rather exclusive or not published on the Internet. People will have to ask the contact person of the event to find specific information. In another case, information of an event may be available on the web, but many people are not fond of reading thoroughly. They may prefer obtaining the exact information they are looking for to reading all the articles. Therefore, asking the contact person is easier. However, sometimes there are short comings when asking the contact person for information. First, people may be reluctant to ask an individual. Second, the contact person may not be available at all times. Third, the contact person may not provide an exact or accurate answer to the person who enquires a question.

To overcome this problem, a question answering system for the event information is developed. A Question Answering (QA) system is an application that allows a user to ask a question in natural language, and an unstructured document collection to look for the correct answer (Buscaldi, Rosso, Gómez-Soriano, & Sanchis, 2010). The goal of a QA system is to retrieve answers to questions rather than full documents or best-matching passages (Kaur & Rimpi, 2013). This research is conducted to combine techniques to improve Question Answering System (QAS), and also to find which techniques with better performance. In addition, the system uses 2 scenarios which are rule-based with Word Matching (WM) technique and rule based on Word Similarity (WS) checking. The creation of this QAS is to find whether the users prefer to use QAS rather than phoning the contact person.

Most QA systems consist of three main modules: Query Processing, Information Retrieval, and Information Extraction. Many different systems develop different techniques in each of these modules to achieve better results. Query Processing (QP) is the task to convert a natural language question into an unambiguous form that a computer is capable of understanding (Bhatia, Madaan, Sharma, & Dixit, 2015). Next, Information Retrieval (IR) is the task of finding documents that are relevant to a user's need for information (Russell & Norvig, 2010). Then, Information Extraction (IE) is the task to find the sentences that are most likely the answer to the question from the documents found in IR (Russell & Norvig, 2010). In this research, many previous researches are included to be used as comparison to Eve QAS such as Exact Phases in IR for QA (Stoyanchev, Song, & Lahti, 2008), Named Entity Recognition (NER) in QA (Mollá, Van Zaanen, & Cassidy, 2007), Monolingual Indonesian QA (Zulen & Purwarianti, 2011), and AskIndo (Jovita, Linda, Hartawan, & Suhartono, 2015).

Besides the mentioned researches, many other related researches had also been noted. Rinaldi, Dowdall, Kaljurand, Hess, and Mollá (2003) created ExtrAns that concentrated in paraphrase problems. It is challenging for the system to understand sentences with different structures and words, but they have the same meaning. ExtrAns transforms documents and queries into a semantic representation called Minimal Logic Form (MLF), and the answers will be derived from logical proof from the document (Molla, 2009). The other technique is to measure the relatedness of words as done by Salahli (2009) who measured the relatedness of words by using the combination of the relationship existing between words, depending on the context or importance. According to Islam, Milios, and Keselj (2012), word relatedness as the degree of how much one word had to do with another word. Commonly, existing researches that use word relatedness can be broadly categorized into three major groups of corpus-based, knowledge-based, and hybrid methods.

The objective of this research is the system can retrieve relevant documents and extract an answer from the question related to the scope of research. With this QAS, the users can find more information about the events. Moreover, the event organizers can also make use of the system to share the information about an event to the developers so the developers can include it in the system.

II. METHODS

Eve is a QA system in which the system can answer the factoid questions from the user about an event. There are 3 main modules proposed for the method in this research. The modules consist of Query Processing, Information Retrieval, and Information Extraction. Before the system can work properly, a document preprocessing aims to store named entities from the text corpora into gazetteer to be run in an independent process at the beginning (NE Preprocessing). The text in corpora is pre-annotated with LOCATION, PERSON, ORGANIZATION, and EVENT tags. The CURRENCY and TIME tags are not pre-annotated in the documents. Instead, during information extraction module, the system to annotate the text with these tags. Eve system is built using Phyton 2.7.9 with NLTK 3.0 and BeautifulSoup 4.4.1 libraries. The overall method proposed is described in Figure 1.

Query processing module performs 5 processes including tokenization, validating question word, case folding, question classification, and stopwords removal. The system will split the whole question into words and symbols. Then, the system validates whether the question is valid or not by checking if the question contains "what"/"where"/"when"/"who" or not. In the case of folding, each letter in question is lower-cased. Next, the system must identify the question, so the system will know what the Expected Answer Type (EAT) for that question is in finding the answer. Table 1 shows the EAT of each question type.

"What" question type may have several EAT. There is a need for the additional term, called identifier. The identifier is a list of words that can be paired with "what" question to form a different "what" question classification. To classify "what" question, the system calculates the distance between "what" and the identifier. The maximum distance between "what" and the identifier is 3 words. If the distance is within the maximum distance, the new classification of "what" can be classified. If the distance is larger than the maximum distance, the identifier cannot be paired with 'what' question, and the EAT will be definition or description instead. The classification of "what" question can be seen in Table 2.

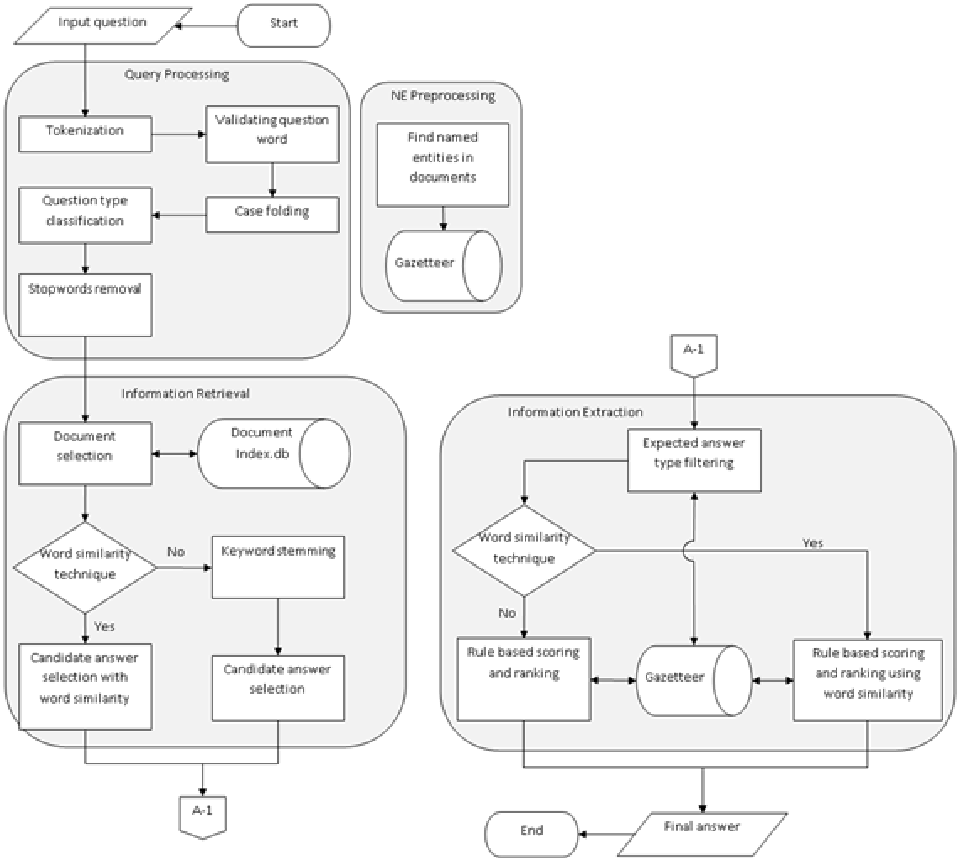
Input question
Start
Query Processing
Tokenization
Validating question word
Question type classification
Case folding
Stopwords removal
NE Preprocessing
Find named entities in documents
Gazetteer
Information Retrieval
Document selection
Document Index.db
Word similarity technique
No
Keyword stemming
Yes
Candidate answer selection with word similarity
Candidate answer selection
A-1
A-1
Information Extraction
Expected answer type filtering
Word similarity technique
Yes
No
Rule based scoring and ranking
Gazetteer
Rule based scoring and ranking using word similarity
End
Final answer

Figure 1 Flow of the Automated Question Answering System

Table 1 List of Expected Answer Type by Question Word

Question Classification	Expected Answer Type	Example
What	DEFINITION / DESCRIPTION	What is Euphoria?
Who	PERSON	Who is the chairman of Hexion?
Where	LOCATION	Where can I register to the Hexion?
When	TIME	When Hiesta takes place?

Table 2 "What" Question Further Classification

Question Classification	Expected Answer Type	Example
"What" & "time", "date", "day", "month", "year"	TIME	What time does the workshop of Hiesta take place?
"What" & "location", "venue", "place"	LOCATION	- What is the venue for EUPHORIA? - What is the location of the second seminar of Hiesta?
"What" & "name"	PERSON	What is the name of the chairman of Hexion?
"What" & "organization", "company", "association", "network", "firm", "party"	ORGANIZATION	What company is the speaker of Hiesta workshop from?
"What" & "event"	EVENT	What events are available in Hexion competition?
"What" & "price", "fee", "cost"	CURRENCY	What is the fee for Hiesta seminar registration?

In the case of "what" and "name", there are some additional conditions. If the system finds there is "what" and "name" combination, the system will seek another identifier that follows the "name". The maximum distance between the identifier and "name" is 2 words. Like before, if the distance is larger than the maximum distance, the identifier cannot be paired with "what" and "name" question and the EAT will be PERSON. In preventing the mismatch in the next process, stopwords are removed.

Information Retrieval module selects appropriate documents by checking each keyword in the query and find whether the keywords match with the index in the database or not. Stemming is also done in this module using Snowball Stemmer provided by NLTK. The candidate answers are selected by matching keywords between the question and the text in the sentences from the retrieved documents. If the sentence contains the keyword from the query, it will be retrieved and added to the list of candidate sentences. If no sentence is retrieved from the documents, the system will state that it cannot find the answer to the question. Then, the candidate answers are filtered based on corresponding EAT. During this process, the gazetteer will be loaded based on EAT. A special case is given to TIME and CURRENCY EAT. Using regular expression (regex), the system will read the sentence and tag the text as it matches the pattern that represents the EAT.

The patterns of TIME are as following. First, the pattern 1 checks a number (e.g. 1, 12, 24 or one, twelve, twenty-four) followed by the day(s), week(s), month(s), year(s), and special expression (before, after, earlier, later, ago). Second, the pattern 2 analyzes the word "this", "next", "last" followed by day(s), week(s), month(s), year(s), and day of the week (Sunday, Monday, Wednesday) or month (January, February, March). Third, pattern 3 sees the relative identifier of time like today, yesterday, tomorrow, tonight, and tonight. Fourth, pattern 4 follows the checking of date time format DD/MM/YYYY, and HH:MM:SS.SS. It does not matter how many digits are used or in what order the date format is. Fifth, pattern 5 checks the phrase "in", followed by a 4-digit year. Last, pattern 6 checks a different date pattern from pattern 4, which is a number with optional "st", "nd", "rd", and "th", followed by month, and optional year. The date number can also be placed after the month alternatively. Meanwhile, CURRENCY has simpler pattern recognition. It needs "Rp" followed by a space, digits with optional thousand separators, and optional cent digits.

Moreover, the candidate answers will be scored using a rule based on scoring and ranking from Rule Based QA. Because the scope and domain of Eve are different from Rule Based QA, the scoring rules are altered. However, the base is preserved. Like the Rule Based technique, there are four values of scoring. There are clue (+3), good_clue (+4), confident (+6), and slam_dunk (+20). The evaluation is based on the EAT, not the question word. All the rules share a common scoring method, which is WordMatch. WordMatch matches the stemmed keywords from the query with the text, similar to how the system retrieves sentences in the previous process. Rules for each EAT are described in Table 3.

The sentence with the highest score will be chosen as the final answer. If there are multiple candidate sentences that share the same high score, the first two sentences will be chosen as the final answers together. The sequence of the sentences is based on which comes first in the original document.

Furthermore, Eve has an alternate system. This alternate system matches keywords by checking the word similarity between the keywords in the query and the text in the documents. The checking process is done by comparing synsets of keywords from question and sentence by using Wordnet. The synset word similarity comparison is done using Leacock and Chodorow (lch) relatedness measure algorithm. If the value of the measure exceeds the threshold, which is 2, 15, these words will be considered similar. The threshold is taken from Cause of Why (https://github.com/bwbaugh/causeofwhy). Due to the checking the word similarity, the alternate system does not stem the keywords. It is also because it will lose its meaning when checking the word similarity.

Table 3 Rules for EAT

Rules	Details
LOCATION	Score(S) += WordMatch(Q, S) The sentence contains a preposition: Score(S) += good_clue The S contains LOCATION EAT: Score(S) += slam_dunk
PERSON	Score(S) += WordMatch(Q, S) Q does not contain PERSON EAT: S contains PERSON EAT: Score(S) += *slam_dunk* S contains “name”: Score(S) += *good_clue*
TIME	S contains TIME EAT: Score(S) += *slam_dunk* Score(S) += WordMatch(Q, S) Q contains “last” and S contains “first”, “last”, “since”, or “ago”: Score(S) += good_clue Q contains “start” or “begin” and S contains “start”, “begin”, “since”, or “year”: Score(S) += good_clue
CURRENCY	Score(S) += WordMatch(Q, S) S contains CURRENCY EAT: Score(S) += slam_dunk
ORGANIZATION and EVENT	Score(S) += WordMatch(Q, S) S contains corresponding EAT: Score(S) += slam_dunk
Definition / Description	Score(S) += WordMatch(Q, S)

III. RESULTS AND DISCUSSIONS

Testing process measures precision, recall, Mean Reciprocal Rank (MRR), accuracy, and the system response time in answering the question. Because recall and precision are antagonistic to one another, an additional measurement called F-measure is necessary to balance recall and precision (Jurafsky & Martin, 2008). Meanwhile, the Mean Reciprocal Rank (MRR) is to measure the probability of correctness in the list of possible responses (Craswell, 2009). Accuracy is the number of questions that are answered correctly divided by all questions.

Testing is performed by 30 random users from Bina Nusantara University in computer science major, where each user is asked about 4 questions with 4 different question types (“what”, “who”, “when”, and “where”) to the system. From the questions that are asked to the system, the measures will be examined. All question asked are limited to HIMTI (Himpunan Mahasiswa Teknik Informatika) events only.

The system has been tested using 2 scenarios: one is using Word Matching (WM) technique, and the other is using Word Similarity (WS) technique. Both techniques are used in retrieving candidate answers and in Rule Based ranking algorithm. During the test, 120 questions were asked, but there are only 94 unique questions consisting of 24 “what” questions, 21 “who” questions, 22 “when” questions, and 27 “where” questions. This is because some users ask similar questions, so there is an overlap in some questions. Some examples of the question given to the system are as described in Table 4.

Moreover, the average precision, recall, MRR, F-measure, response time, and accuracy of each question type using Eve WM are presented in Table 5. Meanwhile, the same factors using Eve WS are in Table 6.

According to the test, the total number of correct answers is 63 by using WM and 32 by using WS out of 94 questions. The accuracy of the systems which are measured by the total number of the correct answers is divided by a total number of questions. It is 0,67 for WM and 0,34 for WS. Then, the overall average of precision is 0,376 for WM and 0,197 for WS. The overall average of recall is 0,676 for WM and 0,746 for WS. Meanwhile, the overall average of response time is 0,774 for WM and 27,456 for WS. Next, the overall average of MRR is 0,657 for WM and 0,388 for WS, while the average of F-measure is 0,477 for WM and 0,304 for WS.

It can be seen that WM is better than WS in precision, response time, MRR, and accuracy. WM has higher precision because WM extracts fewer candidate answers (less divisor) than WS. WS extracts the sentences with a word that is assumed similar with the keywords in the query, even it is not relevant to the question. Thus, it has more candidate answers and reduces the precision. Regarding recall, WS is higher than WM as WS extracts the sentences with words that are similar to the keywords meaning the possibility of WS in obtaining relevant sentences is higher than WM. This results in a higher recall. Meanwhile, in relation to time, WS takes more time in extracting answers because of the complexity in checking similarity between two words as well as high loading time of the library due to its large content.

For the F-measure, WM is higher than WS because high precision usually drops recall and vice versa. F-measure is a good measurement to balance them. WM is better in precision, and WS is better in the recall, but the F-measure concludes that WM has more favorable results. The higher accuracy of WM than WS means that WM can answer questions correctly more often than WS.

Eve WM and Eve WS are compared to other previous researches. Eve WM and WS are higher in recall (0,6775 and 0,7455 respectively) than StoQA (0,53) and Ask Indo (0,662). Meanwhile, the average response time of Eve WM is 0,774 seconds. This is higher than the other researches mentioned previously. The average MRR of Eve WM is 0,64675. This is higher than the previous researches in 2008 and 2011. The details of the comparison can be seen in Table 7.

Table 4 Testing with Eve WM & Eve WS

No.	Question	Answer (Eve WM)	Answer (Eve WS)
1.	What is the theme of HTTP?	HTTP 2015 arises a theme "Strengthening Harmony and Inspiring New Experiences (SHINE)" to give the unity, courage and also the inspiration to the freshman.	HTTP 2015 arises a theme "Strengthening Harmony and Inspiring New Experiences (SHINE)" to give the unity, courage and also the inspiration to the freshman.
2.	Who is the speaker of seminar 1 of Hiesta?	The speaker of seminar 1 is Su Rahman, CEO from Simple C.	The speaker of seminar 1 is Su Rahman, CEO from Simple C.
3.	Where is the location of seminar 2 of Hiesta?	Seminar 2 will be located at ASA 902, Alam Sutera Main Campus Binus University	The participant can register for seminar 1 at HIMTI Stand and on the website. The participant can register for seminar 2 at HIMTI Stand and on the website.
4.	When is the registration for DotA 2 Hexion?	The period of the registration for DotA 2 competition is from 1st October 2014 to 11th December 2014 at 09.00 WIB - 17.00 WIB	The time of the game competitions are from Monday, 15th December 2014 to Saturday, 20th December 2014 at 09.00-17.00 WIB. The period of the registration for DotA 2 competition is from 1st October 2014 to 11th December 2014 at 09.00 WIB - 17.00 WIB.

Table 5 Measurement of Each Question Type Using Eve WM

	Precision	Recall	Response Time (s)	MRR	F-measure	Accuracy
What	0,366	0,847	1	0,756	0,51	0,75
Who	0,048	0,095	0,296	0,071	0,063	0,095
Where	0,39	0,296	0,91	0,79	0,524	0,814
When	0,7	0,071	0,89	0,97	0,81	0,955
Average	**0,376**	**0,676**	**0,774**	**0,647**	**0,477**	**0,67**

Table 6 Measurement of Each Question Type Using Eve WS

	Precision	Recall	Response Time (s)	MRR	F-measure	Accuracy
What	0,218	0,917	29,84	0,451	0,35	0,375
Who	0,048	0,095	24,75	0,071	0,063	0,095
Where	0,16	1	28,35	0,43	0,276	0,22
When	0,26	0,97	26,89	0,6	0,525	0,681
Average	**0,197**	**0,746**	**27,46**	**0,388**	**0,304**	**0,34**

Table 7 The Comparison with Previous Researches

System	Precision	Recall	MRR	F-measure	Response time	Accuracy	Year
Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question	-	-	0,6191	-	-	-	2011
Exact Phrases in Information Retrieval for Question Answering (StoQA)	0,73	0,53	0,631	-	-	-	2008
Ask Indo	0,548	0,662	-	0,58	29,407	-	2015
Eve with word matching technique (WM)	0,376	0,6755	0,64675	0,48	0,774	0,67	2016
Eve with word similarity technique (WS)	0,1965	0,7455	0,388	0,304	27,48	0,34	2016

IV. CONCLUSIONS

There are two conclusions in this research. First, Eve WM can give the number of relevant answers to user's factoid question in natural language more than Eve WS. Eve WM can answer most of the question accurately, especially in "when" question type. There are 21 out of 22 "when" question type that can be answered correctly. The overall accuracy of Eve WM is 63 out of 94 questions. Meanwhile, the Eve WS has 32 correct answers out of 94 questions.

Second, Eve WM and Eve WS can surpass the recall of other previous researches. The average value of Eve WM recall is 0,6755. Meanwhile, the Eve WS has the highest recall which is 0,7455. It is because the word similarity checking can extract the sentences with the word that is assumed similar with keywords. Both Eve WM and Eve WS perform badly in "who" question type. "Who" questions cannot be answered most of the time because the system can only retrieve an answer with a named entity of the expected answer type, whereas most "who" questions do not always expect a named entity as the answer. Eve WM has the fastest response time about 0,774 seconds while Eve WS has the longest response time of 27,46 seconds.

REFERENCES

Bhatia, P., Madaan, R., Sharma, A. K., & Dixit, A. (2015). A comparison study of question answering systems. *Journal of Network Communications and Emerging Technologies, 5*(2), 192-198

Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., & Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems, 34*(2), 113-134.

Craswell, N. (2009). *Encyclopedia of database systems: Mean Reciprocal Rank.* US: Springer.

Hartawan, A., & Suhartono, D. (2015). Using Vector Space Model in Question Answering System. *Procedia Computer Science*, *59,* 305-311.

Islam, A., Milios, E., & Kešelj, V. (2012). Comparing word relatedness measures based on Google n-grams. *Proceedings of COLING 2012: Posters,* 495–506.

Jurafsky, D., & Martin, J. H. (2008). *Speech and language processing.* Englewood Cliffs: Prentice Hall

Kaur, H., & Rimpi. (2013). A review on novel scoring system for identify accurate answers for factoid questions. *International Journal of Science and Research (IJSR), 2*(9), 154-157.

Mollá, D. (2009). From minimal logical forms for answer extraction to logical graphs for question answering. In *Searching answers: Festschrift in Honour of Michael Hess on the Occasion of His 60th Birthday* (pp. 101-108).

Mollá, D., Van Zaanen, M., & Cassidy, S. (2007). Named entity recognition in question answering of speech data. In *Proceedings of the Australasian Language Technology Workshop* (pp. 57-65).

Rinaldi, F., Dowdall, J., Kaljurand, K., Hess, M., & Mollá, D. (2003, July). Exploiting paraphrases in a Question Answering System. In *Proceedings of the second international workshop on Paraphrasing-Volume 16* (pp. 25-32). Association for Computational Linguistics.

Russell, S., & Norvig, P. (2010). *Artificial intelligence: A modern approach* (3rd ed.). New Jersey: Prentice Hall.

Salahli, M. A. (2009). An approach for measuring semantic relatedness between words via related terms. *Mathematical and Computational Applications, 14*(1), 55-64.

Stoyanchev, S., Song, Y. C., & Lahti, W. (2008, August). Exact phrases in information retrieval for question answering. In *Coling 2008*: *Proceedings of the 2nd workshop on Information Retrieval for Question Answering* (pp. 9-16). Association for Computational Linguistics.

Zulen, A. A., & Purwarianti, A. (2011). Study and Implementation of Monolingual Approach on Indonesian Question Answering for Factoid and Non-Factoid Question. In *25th Pacific Asia Conference on Language, Information and Computation* (pp. 622-631).