Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 59 (2015) 305 – 311

Procedia Computer Science

International Conference on Computer Science and Computational Intelligence (ICCSCI 2015)

Using Vector Space Model in Question Answering System

Jovita, Linda, Andrei Hartawan, Derwin Suhartono*

Bina Nusantara University, Computer Science Department, Kemanggisan, Jakarta, Indonesia

Abstract

Question answering system is an information retrieval system in which the expected response givesdirectly the answer as requested rather than set of references which have possibilities as the answer. The objective of this research is to represent knowledge and retrieve the answer for a given question by utilizing Vector Space Model. While some mechanisms have been developed in question answering system such as N-gram, template-driven response, reversible transformation, etc., an attempt to use Vector Space Model is conducted. The query will be compared to the knowledge based by measuring their similarity. Data sample to test the model comes from 2 Ministers of Indonesia; they are The Minister of Education and Culture and The Minister of Tourism andCreative Economy Culture.In the experiment, 150 questions are given to the system. Each question words are given 25 questions. The experiment gives 0.662 of recall, 0.548 of precision, and 0.580 of F-measure. But unfortunately it needs around 29 seconds in average to give the answer to the users.

*Keywords:*Question Answering System; Vector Space Model; Information Retrieval

1. Background

Generally, information can be found through social media and internet. However, too many available resources make people difficult to understand the information they have received. Natural Language Processing (NLP) is a computer field and technique which is developed from language study and computational linguistic in artificial intelligence[1]. Some applied researches can be produced by involving NLP technique. One of them is

* Corresponding author. Tel.: +62215345830 ext 2188; fax: +62215300244.
*E-mail address:*dsuhartono@binus.edu

doi:10.1016/j.procs.2015.07.570

question answering system. The ideal question answering system (QAS) is that the system should be able to give the answer in a short description as close as possible with the question.User should be satisfied with the given answers. QAS has been developed since 1960 and still continues till now. On that period, there were also some another question answering researches including AskMSR which uses rule-based[2], web site as the data source[3], template-driven responses[4], START application[5], and NLQA as a new architecture of question answering system[6].

Vector space model is a natural approach which is based on vector space from each word. Document and query are parts of the space[7]. Document is assumed as a vector which has magnitude (distance) and direction. In vector space model, term is represented by using dimension from vector space. Relevance from a document to query is based on the similarity between document vector and query vector[8]. The related research which uses Vector Space Model is conducted for information retrieval[9].

By looking at the result of the previous researches especially the last one, it is indicated that Vector Space Model can likely give a better result than another approaches. To check further about this hypothesis, an attempt to utilize Vector Space Model in the question answering system is conducted.

2. Previous Works

Question answering system is an information retrieval system in which the expected response gives directly the answer as requested rather than set of references which have possibilities as the answer[10]. Question answering system uses basic technique in Natural Language Processing. The aim to build a question answering system is to get answer from the given question. The answer is not in the form of one whole document or the exact match from the statement as the other information retrieval (IR) system did, but the answer is in the form of some words or a sentence which directly gives the answer. The challenge of building a question answering system is how the system can get the answer as accurate as it can.

In Rule-based Question Answering System[2], for each 5W (what, who, where, when, and why) will have different type of answer. For example, the answer to “who” questions will more likely be name of a person, or “when” and “where” questions will return dateline as an answer. A “why” question chooses the sentence that appears latest in the story, and all other types of question choose the sentence that appears earliest in the story. However, if there is no sentence receives a positive score, then “when” and “where” questions return dateline as a default, “why” questions return the last sentence in the story, and all other questions return the first sentence in the story. This method was tested and it showed that Quarc (the name of the system) is best performed on “when” questions since it achieved 55% accuracy and performed the worst on “what” and “why” questions, reaching only 28% accuracy. The “what” questions are the most difficult because they may have variety of answers, whereas the only general pattern/rule was that they often look for a description of an event/object.

Microsoft Research made a product of question answering system named AskMSR[3]. AskMSR uses web site as the big resource of the data. In this research, the system architecture can evaluate the accuracy from different system components. A system that gives the wrong answers to the users is worse than a system that does not give the answer from the user’s questions. Therefore, they explore new strategy to predict in answering the question which is indicated will give the wrong answers. Starting by rewriting the question which is submitted by the user, the system will then send it to the search engine. After that, system will collect the summaries and doing N-gram calculation until the best answer is obtained. The N-gram calculation consists of mining, filtering, and tilling. AskMSR changes the query *“Who is Bill Gates married to?”* becomes *“Bill Gates is married to”*. From the query analysis result, AskMSR extracts some substrings, filters them based on the question words and collects the answers from the substring that occursfrequently. The system presented here is a step toward the final objective in using web as self-update and comprehensive knowledge repositories that can automatically answer any questions with less effort than it should be. Accuracy is measured by calculating Mean Reciprocal Rank (MRR). Its value is 0.507. For the evaluation, the system is still not able to give optimal answers to question words *how*, *why*, and *what if.* The system also depends on the information repetition in web site, and does not have any methods to check the validity of the answers.

Another system is a template-driven response which runs like a psychotherapist gives response like a human responds to the question from users[4]. The system works by using simple parsing from the user’s input which is combined to the simple substitution. It will be shown by phrase and template to respond to the user. However, the system ability is limited to a conversation based on the domain knowledge as it is template-driven response. Not only the conversation limitation, but also the user interface of this system is still not quite attractive as well.

There is also a question answering system research which is implemented in an application named START (Syntactic Analysis using Reversible Transformations)[5]. START is established by Boris Katz and his team. It uses sentences and annotation phrase to describe the content which is closely related with the snippets of information. The excess of the system is the usage of Log Answer System which shows some relevant answers. In Log Answer System, the question will be first inputted by user using web based user interface. Later on, the Log Answer gives list of answers in corresponds with the available context based on the relevant textual sources. The answer is passed on from extensive knowledge base which is translated from German Wikipedia. At the beginning of this system, there is an initialization process. It uses questions to identify the relevant knowledge base fragment. The question is analyzed by linguistic method which is then translated into MultiNet and FOL representation. Furthermore, it will be processed through some steps such as filtering, ML-based proof ranking, ML-based filtering and so on. The result of this research is 0.35 as the highest value of recall and 0.089 as the precision value.

A new architecture of Natural Language Question Answering (NLQA) was proposed[6]. It consists of 3 (three) phases; they are question processing, document processing and answer extraction. Question processing is a process to analyze and classify question and formulate user's request. Document processing is used to choose some relevant documents set and it will be extracted into one paragraph to answer user's questions. Answer extraction is responsible to choose response based on relevant fragment from the document.

A quite different with the question answering system researches, Amin conducted an information retrieval system as his research[8]. By using vector space model, the achieved precision was 0.54 while the recall is 0.19. The average computation time is 1.5 seconds. Thus, it is indicated that by using Vector Space Model we can improve the performance of question answering system.

3. Proposed Methodology

We propose a method which uses Vector Space Model to represent the knowledge while the similarity is measured using the comparison of document vector and query vector. Figure 1 depicts the proposed methodology.

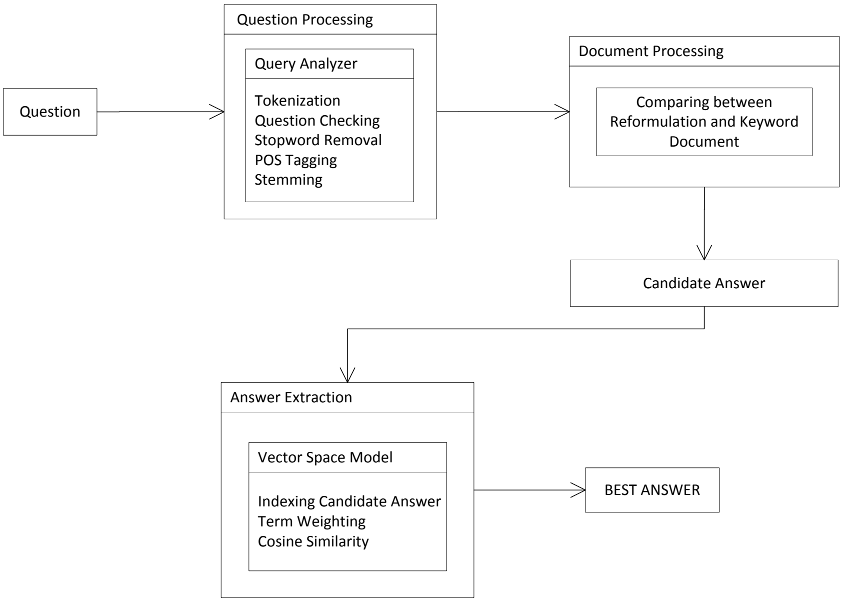

Fig. 1. Proposed methodology using Vector Space Model

In the document processing, the similar activities will be done. The documents are converted into documents representation by using tokenization and stemming process. Subsequently, comparison is conducted between question reformulation to the database. This process results on list of candidate answers. If the keyword is existed in the documents, it will be marked with 1 (one) value while it is not, it will be marked with 0 (zero) value. The indexing result becomes the *df* (document frequency) value, and *idf* (inverse document frequency) will then be calculated by using the formula:

$$idf = \log\frac{\mathrm{N}}{df} \tag{1}$$

The result will be inputted to the matrix by multiplying *tf* and *idf* values. They will be transferred to one new document. The document is indexed according to the existing candidate answer. Every candidate answer is also attached with the weight value. The weight value comes from multiplication between term vector which occurs in the query (W_{iq}) and document vector (W_{ij}). Distance between documents is calculated by using the formula:

$$|d_j| = \sqrt{\sum_{i=1}^{t}(W_{ij})^2} \tag{2}$$

Calculation of distance between queries is also done in this stage by using the formula:

$$|q| = \sqrt{\sum_{j=1}^{t}(W_{i,q})^2} \tag{3}$$

The candidate answer will then allocated to the vector space model. Afterwards, the similarity between the query and documents is measured. It is calculated by multiplying query vector with candidate answer vector and divided by the absolute value between query distance and document distance by using the following formula:

$$Sim(q, d_j) = \frac{\mathrm{q}.d_j}{|q|*|d_j|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{t}} W_{iq}.W_{ij}}{\sqrt{\sum_{j=1}^{t}(W_{iq})^2 * \sum_{i=1}^{t}(W_{ij})2}} \tag{4}$$

|q| : query distance
|dj| : document distance
Wiq : query weight which is calculated from indexing result and term weighting from query
Wij : document weight which is calculated from indexing result and term weighting from candidate answers

From the calculation, the similarity value between query and document will be obtained. The highest value of similarity will be claimed as the best answer got by the system.

4. Results and Discussion

Every question inputted to the system will go through many processes; they are question processing (tokenization, question checking, stopword removal, POS tagging, reformulation), document processing (tokenization, stemming, comparing keyword between reformulation and document), and answer extraction by using vector space model (candidate answer allocation, indexing, term weighting, similarity measurement). At the end, the final score which comes from vector space calculation will be ordered from the highest to the lowest so that we can get the recall and precision values. The one that has the highest score will be taken as the best answer. It will be given by the system to the user.

50 users have tried to input the question. The users are taken randomly. From those inputs, 25 questions are taken from each question word. Speed and accuracy are measured by using those 25 questions. The question words consist of 6 items; they are what, who, when, why, where, and how. Therefore, we have 150 answers. The data comes from 2 ministers from Indonesia; they are Minister of Education & Culture and Minister of Tourism & Creative Economy Culture. All of the questions are limited only for the Indonesian cultures and foods. From the

time measurement, the average time to do the retrieval process is 29.41 seconds.Some examples of the question given to the system are as follows:

Question : What is Tari Pendet? (Literally means what is Pendet dance)
Answer : Tari Pendet is performing arts of Bali.
Execution time : 30 seconds

Question : Where is Danau Toba? (Literally means where is Toba lake)
Answer : Located in the middle of the northern part of the Indonesian island of Sumatra with a surface elevation of about 900 meters (2,953 feet).
Execution time : 11 seconds

Question : When was Monas opened for public?
Answer : Monas was opened to the public in 1975.
Execution time : 36 seconds

Question : Who built Prambanan temple?
Answer : A Prambanan temple was first built at the site around 850 CE by Rakai Pikatan.
Execution time : 19 seconds

Question : Why was Jakarta city so crowded?
Answer : Jakarta has 10.187 million people in 2013 with the 661.52 km² area, Jakarta is most crowded and populous city in Indonesia
Execution time : 30 seconds

Question : How many hours to go to Seven Mount lake?
Answer : Mountain with an elevation of 1.653 meters above sea level it is a mountain sacred to the Hindu Religion
Execution time : 29 seconds

As it can be seen from the samples above, the answers for the question words “why” and “how” are not really relevant with the question while another question words give a good result. The question words “what”, “where”, “when”, and “who” show relevant answers for the given questions. However, the average value of relevant and not relevant answers is measured by using precision and recall. We also involve F-measurement as it is closely related to give prior summarization from precision and recall values.

$$Recall = \frac{\text{Number of relevant items retrieved}}{\text{Total number of relevant items in collection}} \tag{5}$$

$$Precision = \frac{\text{Number of relevant items retrieved}}{\text{Total number of items retrieved}} \tag{6}$$

$$Fmeasure = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{7}$$

The result from all the conducted experiments are presented by using their average value. The average value from each question words are presented in table 1.

Table 1.Recall, Precision, and F-measure Value from the Experiments

Question Word	Precision	Recall	F-Measure
What	0.54520464	0.59429	0.56869
Where	0.473581	0.747534	0.579828
When	0.936232	0.694034	0.797142
Who	0.649472	0.655185	0.652316
Why	0.254378	0.580569	0.353757
How	0.423771	0.703157	0.528832
Average	**0.547106**	**0.662462**	**0.580094**

While the performance that has been achieved by the system is compared to the previous researches. The result is summarizedand presented in table 2.

Table 2. Performance Comparison to Previous Researches

Previous Researches	Precision	Recall	F-Measure	MRR	Time (seconds)
Quarc	-	-	-	0.46	-
QAS AskMSR (2002)	-	-	-	0.507	-
TREC QAS (2005)	0.279	0.262	0.27	-	-
START (2011)	0.089	0.35	-	-	-
IRS (2012)	0,54	0,19	0.28	-	1.5
NLQA (2013)	-	0.82	-	-	-
Proposed Methodology	**0.547106**	**0.662462**	**0.580094**	**-**	**29.40667**

Precision from our proposed methodology is 0.547190. If it is rounded up in 2 decimals position, it will be 0.55. This precision value is higher than the available precision values of TREC QAS (0.279), START (0.089), and IRS (0.54). Recall from our system reaches 0.662462, it is higher than the available recall values of TREC QAS (0.262), START (0.35), and IRS (0.19). F-measure from our system is 0.580094. It is higher than TREC QAS (0.27) and IRS (0.28).

5. Conclusion

Based on the evaluation, it is indicated thatvector space model can be considered in question answering system. This is because the precision and recall have successfully outperformed the result of previous researches. As it can be seen table 2, the recall is 0.547106, precision is 0.662462, and F-measure is 0.580094. Unfortunately, it needs around 29 seconds in average to give answer to the users. The time needed is quite long. Thus, we suggest for future improvement to use topic modelling or some other approaches like Latent Semantic Analysis to construct the model of documents representation. By using data training, the complexity to retrieve the information should be better.

References

1. James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning*. Beijing: O'Reilly. 2012.
2. Ellen Riloff and Michael Thelen. *A Rule-based Question Answering System for Reading Comprehension Tests*. Proceeding ANLP/NAACL-ReadingComp '00 Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems, Volume 6, Pages 13-19. 2000.
3. Eric Brill, Susan Dumais and Michele Banko. *An Analysis of the AskMSR Question-Answering System*. Proceedings of 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2002.
4. Bayan Abu Shawar and Eric Atwell. *Using Dialogue Corpora to Train a Chatbot*. In: Archer, D., Rayson, P., Wilson, A., McEnery, T. (eds.) Proceedings of the Corpus Linguistics 2003 Conference, pp. 681-690. Lancaster University. 2003.
5. Tiansi Dong, Ingo Glockner, Ulrich Furbach, and Bjorn Pelzer. *A Natural Language Question Answering System as a Participant in Human Q&A Portals*. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, pp. 2430-2435. 2011.

6. Athira P. M., Sreeja M. and P.C. Reghuraj. *Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System*. International Journal of Web & Semantic Technology (IJWesT), Vol. 4, No. 4, October 2013.
7. Stephen Clark. *Vector Space Models of Lexical Meaning*. A draft chapter for the Wiley-Blackwell Handbook of Contemporary Semantics, second edition, University of Cambridge Computer Laboratory. 2014.
8. Fatkhul Amin. *Sistem Temu Kembali Informasi dengan Metode Vector Space Model*. Jurnal Sistem Informasi Bisnis, Vol 2, No 2, 2012.
9. Richard Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. New York: Addison Wesley. 1999.
10. R. Mervin. *An Overview of Question Answering System*. International Journal of Research In Advance Technology in Engineering (IJRATE). Vol. 1, Special Issue, October 2013.