ICSITECH

by Viny M

Question Answering System with Hidden Markov Model Speech Recognition

Viny Christanti Mawardi, Hobert Ho, Agus Budi Dharmawan
Faculty of Information Technology
Tarumanagara University
Jl. Letjen S. Parman No 1, Jakarta, Indonesia
viny@untar.ac.id, 535130027@fti.untar.ac.id, agusd@fti.untar.ac.id

***Abstract*—Question answering system is a system that can give an answer from the user. In general, question answering can generate answer to text questions. This paper reports the result of question answering system that can receive input questions from speech and text. Hidden Markov Model (HMM) used to recognize the voice provided by the user. The HMM speech recognition used the feature value obtained from Mel Frequency Cepstrum Coefficients method (MFCC). The question answering system used Vector Space Model from Lucene search engine to retrieve relevant documents. The result shows that HMM speech recognition system's success rate in recognizing words is 83.31% which obtained from 13 tested questions. The result also shows that question answering system can answer 4 out of 6 questions that correctly identified by speech recognition system.**

***Keywords*—Hidden markov model, MFCC, question answering system, speech recognition, vector space model**

I. Introduction

Question answering is one of the topics in natural language processing. To understand the question from a user, the system needs the ability to process natural language. Natural languages that usually used as input are text and speech input. In order to process speech input, the system must recognize the text form from given speech, which can be done by using a Speech Recognition System.

There are a lot of researches about speech recognition and the question answering systems. One of the examples for speech recognition is "Automatic Speech and Speaker Recognition by MFCC, HMM, and Vector Quantization" by Desmukh and Bachute [1] and "Using Vector Space Model in Question Answering System" by Hartawan et al. [2]. Although those research yield good results, they were studied independently, separated from each other. This paper tries to combine those systems into one big system and analyze the results to see the impact caused by connecting those systems.

The objective of speech recognition is to determine what word does the speaker says. Several techniques have been proposed for reducing the mismatch between the testing and training environments. First, human speech is converted to a digital signal to produce digital data representing signal each time step. Second, The digitized speech then processed using Mel Frequency Cepstrum Coefficients method (MFCC) to extract its features and stored in a vector. Third, the feature vectors that extracted by MFCC are quantized using vector quantization to produce discrete feature vectors. The last step is to classify these feature vectors using Hidden Markov Model (HMM) and store the produced HMM parameter to a database. Viterbi algorithm will utilize those parameters when the system receive new speech to be recognized.

After obtaining the question in text form, the system will search the answer by searching and retrieving 10 Top documents related to the question using vector space model provided by Lucene. These documents then split into many passages. One passage contains five sentences, with the last sentence of the passage is repeated on the first sentence in next passage [3]. All of these passages will be ranked based on the scoring table, and top 5 passage will be retrieved for answer extraction. Answer candidates are obtained and ranked using the distance between question keyword and answer candidate position in the passages. Only top-ranked answer will be retrieved as an answer.

This paper reports the findings of speech recognition study using MFCC and HMM and question answering study using Lucene vector space model. Assessing answer is done by checking the related document and determine whether the answer is correct or not.

II. Speech Recognition System

The first thing to do before classifying speech using HMM is to extract its feature and store it in a vector using MFCC. The MFCC process as follows [4]:

1) Frame the signal into short frames (20-40 ms)
2) For each frame calculate the Discrete Fourier Transform
3) Compute the mel-spaced filter-bank
4) Take the logarithm of all filter-bank energies
5) Take the Discrete Cosine Transform (DCT) of the log filter-bank energies.
6) Keep 13 DCT coefficients and discard the rest.

The next step after obtaining feature vectors is quantized the feature vectors using vector quantization. Vector quantization quantizes the continuous feature vectors to discrete feature vectors that can be processed by discrete HMM.

A. *Hidden Markov Model*

The HMM is a very powerful statistical method of characterizing the observed data samples of a discrete-time series [4]. This model describes all the possible paths through the state space and assigns a probability to each one. To understand the concepts of HMM, the following elements are defined as: [1]

1) N : No of states in HMM
2) M :Total different symbol per state
3) π : Initial state distribution ($\pi = \pi_i$)
4) A : State transition probability distribution $A = [a_{ij}]$
5) B : Observation symbol probability distribution

π, A and B from [1] are called HMM model and notated by $\lambda(\lambda = (A, B, \pi))$. In order to apply HMM, there are 3 things that need to be done: [5]

1) Calculate Parameter
The main purpose of this step is to compute probability of observation sequence $O = \{O1, O2, , O_T\}$ given the model . The algorithm that used in this step is forward and backward algorithm.
2) Find Optimal State Sequence
The most common solution to find optimal state sequence is viterbi algorithm. Viterbi algorithm is a dynamic programming algorithm that calculates transition state path given observation sequence of symbols [6].
3) Estimating the Model parameters
Estimation of model parameters is needed in order to adjust the model parameter (A, B, π), according to a certain optimally criteria. Baum-Welch algorithm is one of the techniques that used to solve this problem. It is an iterative method to estimate the new values for the model parameters.

B. Voice Activity Detection

Another important issue in speech recognition system is to determine active speech periods and silent periods within a given speech signal [7]. Speech can be characterized as discontinue signal because information is present only if someone is speaking. The regions where information presents called active region and the pauses between talking are called inactive or silence region. An algorithm employed to detect the presence or absence of speech is referred to as a voice activity detector (VAD) [7].

In general, VAD takes the feature from an input signal, spit those signals into frames (5-40ms), and then compare those value with threshold taken from the region that only contains noise. A sound is present (VAD = 1) if the value exceeds the threshold and not present or silent if the value is lower the threshold (VAD = 0). The success of VAD algorithm in splitting the speech depends on the threshold value.

VAD algorithm that used in this system is an Energy-based VAD. A threshold value is calculated by averaging first 40 frames (10ms per frame) with the assumption that user will not speak in first 0.4 seconds after record button is clicked.

Assume that x(i) is i_{th} speech sample. If frame length is N, then the j_{th} frame can be written like (1).

$$f_j = \{x((j-1)N+1) \ldots x(j.N))\} \quad (1)$$

Energy value can be calculated by using (2)

$$E_j = \frac{1}{N} \sum_{i=(j-1)N+1}^{j.N} x^2(i) \quad (2)$$

where E_j is energy at j^{th} frame and x_i is speech sample at j^{th} frame.

III. Question Answering System

Question answering system is a system that produces an answer from a user's questions. Three types of question usually asked is Factoid, List and Definition question. This system limits the question type to Factoid question. Factoid question is a question where the answer is a simple fact and usually short.

Question answering system can also be categorized into two categories based on its scope. There are open domain question answering (the system can answer any question given by the user) and closed domain question answering (system only respond to the question in a limited topic). This system limits the scope to closed domain question answering with Indonesia's history as its topic.

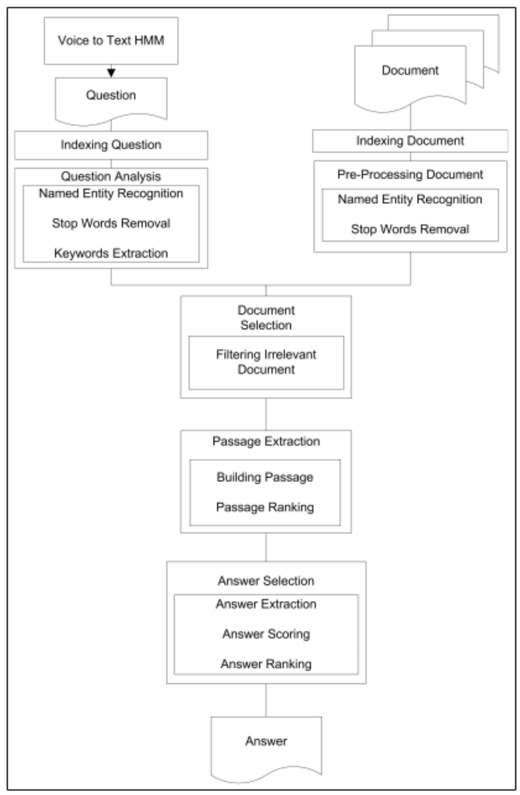

Fig. 1. Question answering system scheme

A. Preprocessing

Preprocessing in text mining consists of 4 steps: tokenization, stopword removal, stemming, and indexing [8]. Tokenization removes symbols, lowercase the text and tokenizes it. Stopword removal removes words that not necessary like conjunctions. Stemming converts the words into its base form without affix.

This system uses lucene search engine function to remove Indonesian's stopwords and stem Indonesian's words. The algorithm that lucene used to stem Indonesian words is Porter algorithm based on Fadillah Z Tala's "A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia." The Porter stemmer was chosen based on the consideration that its basic idea seems appropriate for the morphological structure of words in Bahasa Indonesia [9].

B. Question Analysis

To obtain the correct answer, user's questions must be analyzed properly. The main purpose is to determine what kind of answer that the question wants. Table I shows the relation between question mark and the answer type that question wants.

TABLE I. Question Mark and Answer Type Relations

Answer Type	Question Mark
Name	*Siapa...? Siapakah...? Apa...?*
Location	*Dimana..? Darimana..? Kemana..?*
Time	*Kapan..? Berapa..?*

C. Named Entity Recognition

Named entity recognition (NER) is a subproblem of information extraction and involves processing structured and unstructured documents and identifying expressions that refer to peoples, places, organizations and companies [10]. This task can be done by using many tools, one of them that applied in this system is Stanford named entity recognizer, developed by The Stanford Natural Language Processing Group at Stanford University. One of the main advantages from this NER is that this tool has good documentation and run on JAVA, which can run in many OS platform.

Stanford NER classifies the document by using conditional random field (CRF) model to build a classifier. CRF model is chosen because it represents state of the art in sequence modeling, allowing both discriminative training and bi-directional flow of probabilistic information across the sequence [11]. CRFClassifier builds probabilistic models to segment and labels sequential data. CRFClassifier needs to be trained before classifies the document to build classifier file. A classifier is a machine learning tool that will take data items and place them into one of k classes.

D. Vector Space Model

Vector space model is a way of representing documents through the words that they contain [12]. This model is created based on thought that meaning in document formed by its term and vector is the suitable form to represent that meaning. Since this method represents meaning in a document using a vector, the similarity between document and query can be obtained by calculating the degree that formed between document vector and query vector. This method is called cosine similarity and formulated as (3): [13]

$$SC\left(Q,D_i\right)=\frac{\sum_{j=1}^{t}W_{qj}d_{ij}}{\sqrt{\sum_{j=1}^{t}\left(d_{ij}\right)^2\ \sum_{j=1}^{t}\left(W_{qj}\right)^2}} \tag{3}$$

where t = total term; W_{qj} = weight of j^{th} term in query; d_{ij} = weight of j^{th} term in document.

Weight of term in document can be calculated from its frequency. This method called Term Frequency-Inverse Document Frequency(TF-IDF). IDF can be calculated by (4) and (5)

$$d_{ij}=\ tf_{ij}\times idf_j \tag{4}$$

$$idf_j=\ \log_2\frac{d}{df_j} \tag{5}$$

where t = total term in document; tf_{ij} = frequency of term t_j in d_i document (term frequency); df_j = total document that contains term t_j; idf_j = inverse from document frequency; d = total document.

The weakness of this TF-IDF method is that if the length of each document isn't equal, then the weight of document with less length will be more smaller than longer document. In order to solve this, TF-IDF weight score need to be normalized. The equation used to normalize the weight is

$$w\left(word_i\right)=\frac{w\left(word_i\right)}{\sqrt{\sum_{i=1}^{n}w^2\left(word_i\right)}} \tag{6}$$

where $w(word_i)$ = weight from i^{th} term; n = total term each document.

E. Passage Retrieval

A passage is the smallest part of the document. Passage retrieval works with the assumption that in the small parts of each relevant document contains information that relevant to the query. To obtain more accurate result, the document is split into passages. This short passage will be treated as a single document. In this system, one passage will contain five sentences with the last sentence of the passage will be repeated at first sentence of next passage [3].

F. Passage Ranking

The next step after forming passage is to rank the passage. In this step, NER will be executed again to determine which passage that contains the answer. The passage which had the same entities with the question will be used to the next step. The chosen passage then would be ranked based on following features [14]:

1) Total of relevant entity in passage.
2) Total of keyword in passage.

3) Total of the longest word from question keyword in document.
4) The document rank which the passage come from.

G. Answer Extraction

The last step in question answering system is answer extraction. After ranking the passage and take top N passage (which in this system is five passage), those passages will be processed to get the answer. The process is as follows [14]:

1) Count the word distance between answer candidate and each keyword from question.
2) The answer candidate that has more keyword from question will be prioritized.
3) If there are some answer candidate that have same distances, then the answer candidate that has higher document rank will be prioritized.
4) If there are some answer candidate that have same document rank, then the answer candidate that has more frequency in answer candidate list will be prioritized.
5) If there are answer candidate that have same frequencies, then answer candidate that processed first will be prioritized.

IV. Experiment

All the experiments were done in Indonesia. The configuration that used by speech recognition system in this testing is two states, 1024 codebook, 400 samples per frame with 160 samples overlap. Speech data that trained into the system was taken from one source (Hobert's speech). Total speech data that trained into the system is 900 data, with the composition of 30 words and each word repeated 30 times. Speech data recorded and stored in waveform audio format (WAV) data format.

The classifier Stanford NER that used in named entity recognition was trained using 20 documents that randomly chosen from document collections with total words are 24489 words. Each word was tagged by reading the content of document manually and determines the correct tag each word based on the context in the document.

The configuration that used by question answering system takes 10 top document from lucene search results, 5 top passage after ranking the passage, and 1 top answer from all answer candidates. The documents that used as knowledge are 402 Indonesia's history which taken from 11th and 12th-grade high school material. The document obtained from 5 websites as follows:

1) http://www.materisma.com (151 documents)
2) http://www.gurusejarah.com (127 documents)
3) http://pengertiansejarah.com (84 documents)
4) http://sejarah-interaktif.blogspot.com (35 documents)
5) https://fatihsaputro.wordpress.com (5 documents)

V. Results and Discussion

Table II shows the speech recognition experiment results for question no.5 to 10. The complete list of speech recognition results can be seen in Table III. In each experiment, the speaker (Hobert) speaks the question in one try, then the speech data is split with VAD algorithm into speech token, and each speech token is recognized using speech recognition system. The success rate of Speech Recognition System is 82.31%, obtained by calculating each question's success rate and then average it with other question's success rate. These mistakes occur because of some factors like speech tone variations that still not trained to the system, errors in splitting speech (VAD mistakes) or tone of speech of one word is nearly same with other words.

The example of the tone of speech of one word is almost same with other words can be seen at experiment no. 9 in Table II. At that experiment, the system fails to recognize *"kapan"* and *"saragosa"* speech and recognize those words as *"taman"* and *"sriwijaya."* The intonation *"kapan"* is quite similar with intonation *"taman"* in Indonesian since the suffix of the words contains "an." This tone of spelling makes viterbi algorithm think that *"kapan"* speech as *"taman"* word. For more details look at Fig. 2.

TABLE II. Speech Recognition results (Question no.5 to 10)

Experiment No.	Spoken Word	Recognized Word	Correct
5	*kapan*	*kapan*	true
	budi	*budi*	true
	utomo	*utomo*	true
	didirikan	*didirikan*	true
6	*kapan*	*kapan*	true
	diponegoro	*utomo*	false
	wafat	*wafat*	true
7	*kapan*	*kapan*	true
	hayam	*hayam*	true
	wuruk	*wuruk*	true
	wafat	*wafat*	true
8	*kapan*	*taman*	false
	konferensi	*konferensi*	true
	asia	*sriwijaya*	false
	afrika	*afrika*	true
	diselenggarakan	*kerajaan*	false
9	*kapan*	*taman*	false
	perjanjian	*perjanjian*	true
	saragosa	*sriwijaya*	false
	terjadi	*terjadi*	true
10	*kapan*	*kapan*	true
	taman	*taman*	true
	siswa	*siswa*	true
	didirikan	*didirikan*	true

Fig. 2 shows the distribution of word's probability from Viterbi algorithm result. More close the word to center (0,0) means that word has more probability as a correct representation of the speech. From the figure can be seen that word *"kapan"* probability is lower than word *"taman"* since the word *"taman"* is more close to center than word *"kapan"*. Viterbi algorithm decide that word *"taman"* is more suitable as word representation for speech *"kapan"*. n this case, because the intonation when word *"kapan"* is spelled is more similar

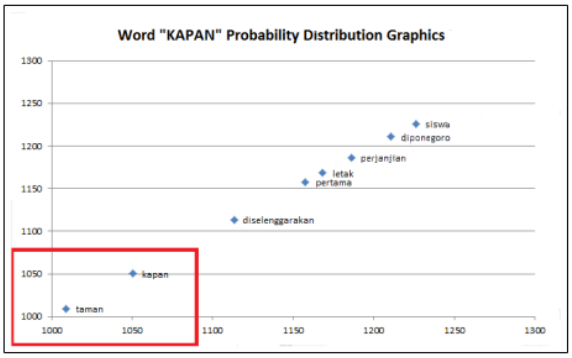

Fig. 2. Word *"kapan"* probability Distribution Graphics

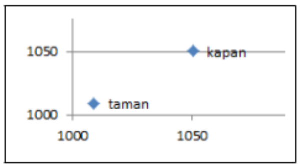

Fig. 3. Word *"kapan"* probability Distribution Graphics (Zoomed)

TABLE III. Experiment Results

No	Questions	Speech Recognition Results	Question Answering Results
1	*dimana budi utomo didirikan*	*dimana budi utomo didirikan*	*batavia*
2	*dimana konferensi asia afrika diselenggarakan*	*dimana konferensi asia afrika masa*	*jenewa*
3	*dimana letak kerajaan sriwijaya*	*dimana letak kerajaan sriwijaya*	*palembang*
4	*dimana taman siswa didirikan*	*saragosa taman siapa didirikan*	*mahasiswa*
5	*kapan budi utomo didirikan*	*kapan budi utomo didirikan*	*tahun 1911*
6	*kapan diponegoro wafat*	*kapan utomo wafat*	*tahun 1920*
7	*kapan hayam wuruk wafat*	*kapan hayam wuruk wafat*	*1350 1389*
8	*kapan konferensi asia afrika diselenggarakan*	*taman konferensi sriwijaya afrika kerajaan*	NO ANSWER
9	*kapan perjanjian saragosa terjadi*	*taman perjanjian sriwijaya terjadi*	NO ANSWER
10	*kapan taman siswa didirikan*	*kapan taman siswa didirikan*	*1922*
11	*siapa pendiri taman siswa*	*siapa pendiri taman siswa*	*pawiyatan wanito*
12	*siapa presiden pada masa orde baru*	*siapa presiden diselenggarakan masa orde baru*	NO ANSWER
13	*siapa presiden pertama indonesia*	*siapa presiden pertama indonesia*	*orang indonesia orisinil*

to the intonation of word *"taman"* trained data than the intonation of word *"kapan"* trained data.

Table III shows that when speech recognition result is incorrect, the answer from question answering system also wrong. The failure of speech recognition makes the relevant keyword from the question is lost, and question answering system becomes unable to look for the correct answer. The example of this failure can be seen at question no.6 *"kapan diponegoro wafat"*. In this case, speech recognition failed to recognize speech *"diponegoro"* and recognize it as *"utomo"*. The important keyword from that question is *"diponegoro"* which is person's name. Because of this failure, question answering system is not looking at document about *"diponegoro"* but at *"utomo"* related document, which is certainly incorrect, and produce the wrong answer.

Speech recognition also fails to recognize the question mark words like shown at question 8 and 9 in Table III. This failure makes question answering system becomes unable to determine which type of answer that question asks and resulting in "NO ANSWER" verdict.

In case that speech recognition can recognize the question correctly but still produce the wrong answer, like in question no. 5, 11, and 13, the failure is caused by question answering system. The following are the reasons why question answering systems are unable to produce a correct result:

1) Classifier Stanford NER fail to recognize answer's name entity, either the correct answer are not tagged, or the answer only half tagged. This failure happened because the training data for creating the classifier is too few.
2) The method used to look for the answers in the document only looks syntactically. This kind of method only seeks the answer based on its position in a document, not based on its meaning. That is why the system cannot look the answer if the answer using substitute words like pronoun or synonym.
3) The scoring mechanism in passage retrieval is not strong enough to rank the passage. The scoring method that used in this system makes the passage that doesn't contain the answer can have the same score or higher than the passage that contains the answers.

VI. Conclusion

Speech recognition system with HMM and vector quantization can recognize tested speech with success rate 82.31%. Even so, the system still fails to detect some word like *"kapan"* and *"taman"* because the intonation between those two words are quite similar. The question answering system also can answer 4 out of 6 question that correctly recognized by speech recognition system. The remaining 7 question cannot produce correct answer because speech recognition cannot recognize the question correctly and make the question's keyword changed.

To improve the accuracy of speech recognition system, we will use better sound classification method like dynamic time warping (DTW) or artificial neural network (ANN). We will also increase the amount of speech training data. Other than that, we also need to find other way to determine threshold in voice activity detection algorithm. For the question answering system, we will increase the amount of knowledge document and redesign the scoring mechanism for passage retrieval. We will also apply syntactic and semantic method for looking the answer in document.

REFERENCES

[1] S. Deshmukh and M. Bachute, "Automatic speech and speaker recognition by mfcc, hmm and vector quantization," *IJEIT*, vol. 3, no. 1, 2013.

[2] A. Hartawan, D. Suhartono *et al.*, "Using vector space model in question answering system," *Procedia Computer Science*, vol. 59, pp. 305–311, 2015.

[3] S. H. Wijono, I. Budi, L. Fitria, and M. Adriani, "Finding answers to indonesian questions from english documents." in *CLEF (Working Notes)*, 2006.

[4] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.

[5] G. Tiwari, M. pandey, and M. Shrestha, "Text-prompted remote speaker authentication," Master's thesis, Tribhuvan University, Nepal, 2011.

[6] U. R. Tatavarty, "Implementation of numerically stable hidden markov model," 2011.

[7] K. Sakhnov, E. Verteletskaya, and B. Simak, "Dynamical energy-based speech/silence detector for speech enhancement applications," in *Proceedings of the World Congress on Engineering*, vol. 1. Citeseer, 2009, p. 2.

[8] S. Vijayarani, M. J. Ilamathi, and M. Nithya, "Preprocessing techniques for text mining-an overview," *International Journal of Computer Science & Communication Networks*, vol. 5, no. 1, pp. 7–16, 2015.

[9] F. Z. Tala, "A study of stemming effects on information retrieval in bahasa indonesia," *Institute for Logic, Language and Computation, Universiteit van Amsterdam, The Netherlands*, 2003.

[10] A. Mansouri, L. S. Affendey, and A. Mamat, "Named entity recognition approaches," *International Journal of Computer Science and Network Security*, vol. 8, no. 2, pp. 339–344, 2008.

[11] J. R. Finkel, T. Grenager, and C. Manning, "Incorporating non-local information into information extraction systems by gibbs sampling," in *Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2005, pp. 363–370.

[12] V. K. Singh and V. K. Singh, "Vector space model: an information retrieval system," *Int. J. Adv. Engg. Res. Studies/IV/II/Jan.-March*, vol. 141, p. 143, 2015.

[13] D. A. Grossman and O. Frieder, *Information retrieval: Algorithms and heuristics*. Springer Science & Business Media, 2012, vol. 15.

[14] D. Jurafsky and J. H. Martin, "Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition."

ICSITECH

ORIGINALITY REPORT

14%	7%	11%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Pongsakorn Saipech, Pusadee Seresangtakul. "Automatic Thai Subjective Examination using Cosine Similarity", 2018 5th International Conference on Advanced Informatics: Concept Theory and Applications (ICAICTA), 2018 Publication	1%
2	www.iaeng.org Internet Source	1%
3	Submitted to Indian Institute of Technology, Kharagpure Student Paper	1%
4	Anzar S.M., Amala K., Remya Rajendran, Ashwin Mohan, Ajeesh P.S., Mohammed Sabeeh K., Febin Aziz. "Efficient online and offline template update mechanisms for speaker recognition", Computers & Electrical Engineering, 2016 Publication	1%
5	www.thesai.org Internet Source	1%

6	Submitted to TechKnowledge Student Paper	1%
7	stop-words-list-bahasa-indonesia.blogspot.com Internet Source	1%
8	nlp.stanford.edu Internet Source	1%
9	dias.library.tuc.gr Internet Source	1%
10	www.mo-data.com Internet Source	1%
11	www.ijarcsse.com Internet Source	1%
12	Lecture Notes in Electrical Engineering, 2016. Publication	<1%
13	Minh-Son Nguyen, Tu-Lanh Vo. "Vietnamese Voice Recognition for Home Automation using MFCC and DTW Techniques", 2015 International Conference on Advanced Computing and Applications (ACOMP), 2015 Publication	<1%
14	N. Thakoor, J. Gao. "Shape classifier based on generalized probabilistic descent method with hidden Markov descriptor", Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005	<1%

Publication

15 Erez Manor, Shlomo Greenberg. "Voice trigger system using fuzzy logic", 2017 International Conference on Circuits, System and Simulation (ICCSS), 2017 <1%

Publication

16 Cheng Hua Li, Soon Cheol Park. "Combination of modified BPNN algorithms and an efficient feature selection method for text categorization", Information Processing & Management, 2009 <1%

Publication

17 Y. Nakamura. "Probabilistic model of whole-body motion imitation from partial observations", ICAR 05 Proceedings 12th International Conference on Advanced Robotics 2005, 2005 <1%

Publication

18 Nithin, C., and Jini Cheriyan. "A novel approach to automatic singer identification in duet recordings with background accompaniments", 2014 Annual International Conference on Emerging Research Areas Magnetics Machines and Drives (AICERA/iCMMD), 2014. <1%

Publication

19 Jovita, Linda, Andrei Hartawan, Derwin Suhartono. "Using Vector Space Model in Question Answering System", Procedia Computer Science, 2015 <1%

Publication

20	Submitted to University of Malaya Student Paper	<1%
21	repub.eur.nl Internet Source	<1%
22	C Juliane, A A Armant, H S Sastramihardja, I Supriana. "Question-answer pair templates based on bloom's revised taxonomy", IOP Conference Series: Materials Science and Engineering, 2018 Publication	<1%
23	docplayer.net Internet Source	<1%
24	Kittiphong Sengloiluean, Ngamnij Arch-int, Somjit Arch-int, Theerayut Thongkrau. "A semantic approach for question answering using DBpedia and WordNet", 2017 14th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2017 Publication	<1%
25	www.mdpi.com Internet Source	<1%
26	arxiv.org Internet Source	<1%
27	F Henry, D E Herwindiati, S Mulyono, J	<1%

Hendryli. "Sugarcane Land Classification with Satellite Imagery using Logistic Regression Model", IOP Conference Series: Materials Science and Engineering, 2017
Publication

28	Syopiansyah Jaya Putra, Muhammad Zidny Naf'an, Muhamad Nur Gunawan. "Improving the Scoring Process of Question Answering System in Indonesian Language Using Fuzzy Logic", 2018 International Conference on Information and Communication Technology for the Muslim World (ICT4M), 2018 Publication	<1%
29	Advances in Intelligent Systems and Computing, 2014. Publication	<1%

Exclude quotes	On	Exclude matches	Off
Exclude bibliography	On		