ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Web-Based Arabic Question Answering System using Machine Learning Approach

Waheeb Ahmed
Department of Information Technology
Kannur University
Kannur, India

Ajusha PV
Department of Information Technology
Kannur University
Kannur, India

Babu Anto P
Department of Information Technology
Kannur University
Kannur, India

Abstract: Question Answering (QA) systems are complex software capable of answering a question in natural language. The source of information for these systems is a given corpus or, as assumed here, the Web. To generate the exact answer, these systems carry out several subtasks among which the question classification and answer extraction. The main goal of this paper is to employ machine learning techniques for question classification and answer selection tasks. This study presents a supervised support vector machine (SVM) classifier for question classification and answer selection. The question classifier is trained to identify the answer type of the submitted question and directs the answer extraction module to re-rank the answer candidates retrieved from the initial retrievers. A set of flexible features are used , such as lexical features, and syntactic features. Similarly, the answer selection stage is using SVM classifier with a set of extracted features and trained on a set of question-answer pairs. We assess the performance of answer selection task using the Mean Reciprocal Rank(MRR).

Keywords: Question Answering System, Information Retrieval, Information Extraction, Machine Learning, Natural Language Processing

I. INTRODUCTION

In order to get information from the web, we usually use search engines. Traditional Search engines have played a crucial role in assisting the users to search the required information from the massive amount of information available in the web. The result returned by the search engine will be a list of links to documents. The user has to look for the answer by browsing the links to the documents. Recent information demand suggests the need for better mechanisms capable of satisfying information needs of users in a more natural way. Question Answering (QA) systems have emerged as a feasible solution for the development of such mechanisms [1][2].Question-Answering systems are gaining high popularity nowadays. The main advantage of such QA systems is that the user can supply a question in natural language and he/she will get a concise and appropriate answer rather than just a displayed list of links to documents.

There are several Approaches used in Question answering systems that are based on different purpose namely linguistic-based approach, statistical-based approach and pattern matching approach. Several issues are present which plays vital role in question answering system such as Question Classification, Question Processing, knowledge Sources for QA, Answer Extraction, Answer Selection[3][4][5].

Current QA techniques use a variety of linguistic resources to assist in understanding the questions and the text documents. The most common linguistic tools/resources are: parsers, morphological analyzers, part-of-speech taggers, named entity extractors, dictionaries, and WordNets [6,7,8 , 9, 10, 11]. Despite the promising results of these approaches, they have two main disadvantages. First, development of such linguistic resources is a very complex task. Second, their performance rates are usually not optimal.

To overcome the load of human crafted rules in building a QA system, the machine learning approach came into existence as an alternative. Furthermore, employing learning techniques is the best way to automatically adapt a QA system to some specificities, such as: the language used (hand-made rules are considered language-specific) and the domain of knowledge (some rules/pattern may also be specific to a domain) for specialized corpora. Some works have utilized the machine learning approach in the question classification for English [12][13][14].

In this paper, we present a QA system that can answer factoid questions. This system is based on a full machine learning approach that requires a minimum knowledge about the lexicon and the syntax of the language. It is developed on the idea that the questions and their answers are commonly expressed using the same set of terms. It applies a supervised approach instead of a statistical method for answering factoid questions. The main advantage of the proposed Question answering system is that it is web-based, which is not restricted to a specific domain. For a state-of-the art QA system, it should use machine learning approach rather than rule-based approach. The main disadvantage of the rule-based approach for question answering is that it requires rule-writing skills and deep understanding of the language syntax.

II. QA SYSTEM ARCHITECTURE

Our QA system consists of three components: Question Analysis module, which consists of some linguistic tools including tokenizer, stemmer, stop-word removal and question class identifier. Next is document retriever and Passage Retrieval(PR) modules. The document retriever will return the top 10 documents ranked by Google search engine. The PR returns a small number of relevant passages from the document collection. where the passages with a high probability of containing the answer are returned from the document collection; Last component is Answer Extraction(AE) and answer selector module, which extracts and returns ranked and exact answers from the retrieved passages where candidate answers are selected using a machine-learning approach, and the final answer recommendation of the system is produced. The following sections describe each of these modules..

A. Question Analysis module

Question analysis module consists of two subtasks namely: question preprocessing, and question classification.

Question Preprocessing. To preprocess a question, first the question characters are normalized by removing the diacritics(special marks on letters), then question is split into tokens. The stemming and part of speech of each token would be specified. Stemming is a task of converting different morphological variations of the same word into their original form, called a stem. To perform the stemming process, a stemmer based on a set of predefined word translation rules is used. In this paper, Shereen Khoja's stemming algorithm [15] is applied to remove the suffix of the original word in the question item to convert the word to its valid stem. With the help of stemming process, the size and complexity of the feature vector that expresses a question item can be further reduced and therefore the performance of the question classifier can be improved. After performing the stemming, the stem words sets from each question in the training question set are merged together and the repeated stems are eliminated. Finally, this step produces an initial feature set, i.e., the whole feature set[16].

Question Classification. Question classification is performed in a similar way to document classification .However, question classification is considered extremely challenging comparing to document classification in getting a reasonable accuracy while classifying questions. The reason is in question classification there exist no much information or words and this may not be enough to efficiently classify questions as compared to document classification. This will have impact on discriminating power in classifying questions. Some of these works were accomplished for question answering system and information retrieval system[16] , [17], [18].Question classification has been carried out using different approaches. Some of these approaches use linguistics, pattern matching or machine learning for performing this task depending on their resources and size of domain. The pattern matching has been successfully applied to the closed domain question answering and emerged as quiet simple and effective approach. Machine learning approach is considered more effective since what you will need is to annotate your training data in a way a classifier can read. For example, Support Vector Machine(SVM) is considered as most effective classifier in machine learning for question classification tasks. Hence, in our work we adopted SVM for performing the classification task. For *machine learning section*, first it needs to find an appropriate classifier and then extracting suitable features from question in order to create a feature vector. Most of the works in the literature which have used machine learning approaches from question classification literature, they used Support Vector Machine(SVM) classifier [18, 19, 20, 21,22,23,24,25,26] and they got the better results in comparison to other classifiers and approaches. So we chose to use SVM classifier for question classification. SVM classifier has several kernels , we used linear kernel since better results are obtained by using this kernel as compared to other kernels. When using high number of features it is better to use linear kernels[27]. In our problem as each word of a question and its linguistic tags can be considered as a single feature, the number of features are very high. The classification effect based on multiple features is better than based on single feature[28].Therefore, we used different features for getting better view of a question in order to find its class or answer type.

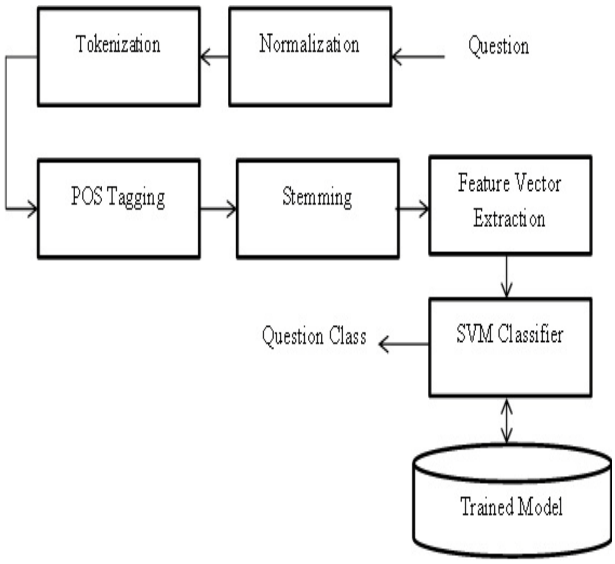

Figure 1. Architecture for question classification module

As shown in figure 1. the question would be normalized and tokenized and then some features would be extracted which can be used in the machine learning. POS tagging will be applied to identify the question head word and other parts of speech in the question. Stemming is applied to get the root of words. Next, features are extracted to be used by the SVM classifier. The output would be a label assigned to the given question.

Question Type Taxonomy. The semantic classes of a question after classification is defined by question taxonomy. We adopted the two layered question taxonomy approach of coarse and fine grained classes proposed by Li and Roth et al. [6].

Table I. Question classes: fine and coarse grain

Coarse Class	Fine Class
ABBREV	Abbreviation, explanation
ENTITY	product, religion, sport, substance, symbol, technique, other, term, vehicle, word , animal, body, color, currency, event, food, instrument, language, letter, plant,
DESCRIPTION	Definition, description, manner, reason
HUMAN	Group, individual, title, description
NUMERIC	Code, count, date, distance, money, order, period, percent, speed, temp, size, weight, other
LOCATION	City, country, mountain, state, other
ORGANIZATION	Organization or institute, group or committee

Feature Extraction. Various types of features are available which are currently chosen for question classification. We are extracting two types of features based on the kinds of linguistic information: lexical, and syntactical features.

1. Lexical Features. The lexical features are considered as the context words presenting in the question. In question classification, a question is represented in a similar way to the representation of document in the vector space model, i.e., a given question can be represented as the vector:

$q = (q1, q2, \ldots, q_N)$

where qi is referring to the frequency of term ith in question q and N is the total number of terms. only non-zero valued features are stored in the feature vector. Next, the given question q is expressed in the form:

$q = \{(t1, f1), \ldots, (tp, fp)\}$

where fi is referring to the frequency of the term ith in the given question q. The collection of these features is called bag-of-words features or unigrams features. In this context, unigrams is a special case of the n-gram. For the purpose of extracting n-gram features, any n sequential words in a question is considered as a feature. For example, the lexical features of the sample question Given the question “What did Edward Binney and Howard Smith invent in 1903?” . Lexical information such as question words (i.e: who, how, when, what) is listed in Table 2 (Note: all questions have been translated into English for this paper).

2. Syntactic Features. These types of features are extracted from the syntactical form of a question. The two common types of syntactic features which are used for question classification are tagged unigrams and head words. Tagged Unigrams refers to the part-of-speech tag for each word in a question like NN (Noun), NP (Noun Phrase), VP (Verb Phrase), JJ (adjective), and etc.

Table II. Lexical features example

Feature Space	Features extracted
Unigram	{(What,1)(did,1)(Edward,1)(Binney,1)(and,1)(Howard,1)(Smith,1)(invent,1)(in,1)(1903,1)(?,1)}
Bigram	{(What-did,1)(did-Edward,1)(Edward-Binney,1)(Binney-and,1)(and-Howard,1)(Howard-Smith,1)(Smith-invent,1)(invent-in,1)(in-1903,1)(1903-?,1)}
Trigram	{(What-did-Edward,1)(did-Edward-Binney,1)(Edward-Binney-and,1)(Binney-and-Howard,1)(and-Howard-Smith,1)(Howard-Smith-invent,1)(Smith-invent-in,1)(invent-in-1903,1)(in-1903-?,1)}
Wh-Word	{(What,1)}
Word-Shapes	{(lowercase,4)(mix,5)(digit,1)(other,1)}

Table III. Given questions with their headwords and classes

Questions	Class
Who developed the <u>Macintosh computer</u>?	HUM:ind
What is a micron?	DESC:def
When is the official first <u>day</u> of summer?	NUM:date
What <u>river</u> in the US is known as the Big Muddy?	LOC:other
What did Edward Binney and Howard Smith invent in 1903?	ENTY:other
What <u>currency</u> does Argentina use?	ENTY:currency
What is the <u>date</u> of Mexico 's independence?	NUM:date
What <u>color</u> is a giraffe 's tongue ?	ENTY:color
Which <u>country</u> gave New York the Statue of Liberty ?	LOC:country
Who was the <u>abolitionist</u> who led the raid on Harper 's Ferry in 1859 ?	HUM:ind
What is the name of the <u>satellite</u> that the Soviet Union sent into space in 1957 ?	ENTY:product

For example, the head word for the question “Which country gave New York the Statue of Liberty?” is “country”. The presence of the word “country” in this question can highly contribute to the classifier for classifying this question as “LOC:country”. Table 3 lists sample questions from Text Retrieval Conference(TREC) dataset together with their class labels. The head words are identified by being underlined. To determine the head word of a sentence, a syntactic parser is required. For sentences written in Arabic language, we used the Stanford parser for Arabic[6].

B. Document Processing Module

This module returns the top 10 documents from the web that are likely to contain answers to the user's question. It performs a query reformulation and send it to the Google search engine. The query reformulation subtask extracts terms from the user's question and creates a query from these terms. This query is passed to the Google search engine, which retrieves a set of documents which are likely to contain these terms.

Passage Retrieval.. This subtask retrieves passages which are most likely to contain the answer, rather than simply returning the passages which share some words with the question. When a user submits a question, the passage retrieval method returns the passages with the relevant terms (excluding stop-words) from the set of documents returned by the search engine using a traditional information retrieval technique based on the vector space model. Then, it computes the similarity between the sets *of n*-gram of the passages and the question submitted by the user in order to obtain the new scores/weights for the passages. The weight of a passage is associated with the largest *n*-gram structure of the question that can be found in the passage itself. The passage with the larger the *n*-gram structure, means the greater the weight of the passage. Consequently, it retrieves the passages with highest scores/weights.

Answer Extraction. The aim at this stage is to extract all the possible answer candidates. However, it is not practical to consider all the keywords of a passage as candidates for computation time issue so we adopted semantic and syntactic features to eliminate some of them. Considering factual questions which ask for specifications about an entity or an event, answers can be confined to the noun phrases that are modifiers of a noun or a verb (the *focus*). Since this relation is not always marked in the passages, all nominal groups could be candidates. However their number remains too important, so we employed semantic criteria to limit/restrict this set. Some questions may require a named entity as an answer (e.g. *Who is the rector of the Harvard University?* which expects a name of person in response) or not, and are differentiated by the question analysis module. For questions that requires a named entity in response, all named entities of the expected answer type annotated in the passages are extracted, in addition to the proper nouns in unmarked noun phrases.

Feature Vector Extraction. The flow of typical QA system is done by question processing, document retrieval, passage retrieval, and candidate answer extraction and selection. While analyzing the question , the question terms are extracted, question class, question focus. Then, the documents containing some or all of the question terms are retrieved. Named Entities that match the question class are extracted from the returned documents. Features that are extracted from question analysis phase and from the extracted candidate answers(resulted from processing the retrieved documents) are used for selection of answer candidates. The extracted features are shown in Table 4.

Table IV. Extracted features from question and candidate answers

Category	Feature
Question Terms(QT)	All QT match Part-Of-Speech(POS) match Head-word match Stem match
Candidate Answer(CA)	Length of word Normalized position in the document Matching with Part-Of-Speech(POS) Matching with attached function words
Question Class(QC)	Pair of QC and NE of CA
Question Focus(QF)	Pair of focus terms in question and CA

We used SVM using a feature vector consisting of all features mentioned in Table 4 and linear kernel.

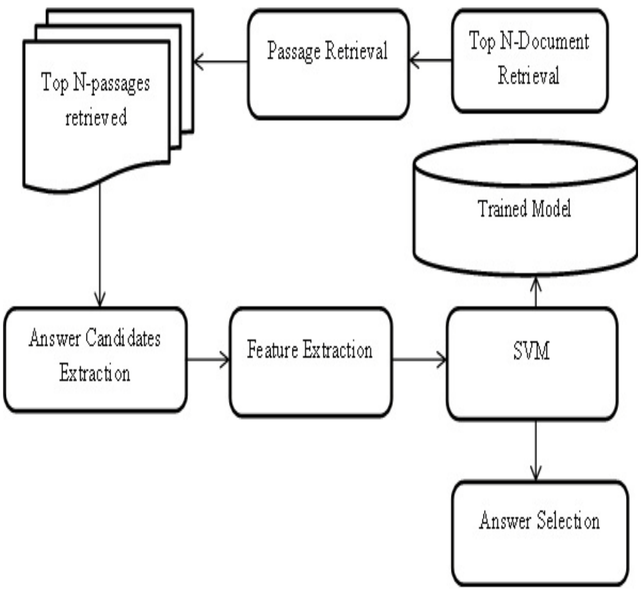

Fig.ure 2. Answer Extraction and selection module

Answer Selection. Answer selection using machine learning is defined as the process of training and classifying answer candidates into correct answers and incorrect answers for input question. To use SVM , we need to prepare a set of training examples with feature vector of candidate answers. For any given question , the question analysis module will analyze the question , retrieve the relevant documents , process them and list the candidate answers. The parameters computed during the question processing are stored and utilized for the purpose of creating feature vectors for each candidate answer.

We adopted SVM for the question classification. Similarly, the SVM is used as the classification algorithm in the answer selection stage. We used the same convention/ judgment set of TREC [29] which labels the selected answer as +1 (answer) or -1 (non-answer). The reference/judgment set consists of a set of answer candidates that have been manually annotated. Each answer candidate that contains the answer focus terms and question class is treated as a positive example, and otherwise will be considered as a negative example. The training set consists of a translated version of 1000 TREC-10 question/answer pairs fed to the SVM. In figure 2 top 10 documents which contain some or all of the question terms are retrieved. The documents are segmented into passages. Answer extraction techniques are applied on the N retrieved passages. Features have been extracted from the answer candidates. The feature vector is fed into the SVM. The SVM re-ranks the answer candidates based on a trained model of question-answer pair. The answer with the highest score is selected as the final answer.

III. EVALUATION AND RESULT

For question classification we combine all the individual feature sets, that includes: Unigram, Bigram , Wh-Word, Head-Word , and Question- Class. Table 5 shows the results corresponding with different training data sets.

Table V. The accuracy of using SVM classifier with several features combined

Question Type	Precision	Recall	F-measure
PERSON	97.2	84.6	90.4
LOCATION	91	79	84.6
ORGANIZATION	94.1	82	87.6
NUMERIC	86	95.7	91
AVERAGE	92.075	85.325	88.4

The SVM performed very well in classifying the different categories of questions with average of 88.4 for the f-measure/accuracy. For evaluation of answer selection task, we used the same method adopted by TREC-QA Track and it is called answer ranking. For answer ranking, the evaluation measure is the Mean Reciprocal Rank(MRR) . The score is calculated as the rank position of the first correct answer. If the rank of the first correct answer is n, the resulted score is 1/n. The testing set consists of 434 translated questions from TREC-QA Track. All the questions in the evaluation set have at least one correct answer in the retrieved documents. The question set for each category is given in Table .

Table VI. Question set for each category

Question Class	Number of questions
PERSON	110

LOCATION	112
ORGANIZATION	105
NUMERIC	107
Total	434

The results of answer selection are shown in figure 3.

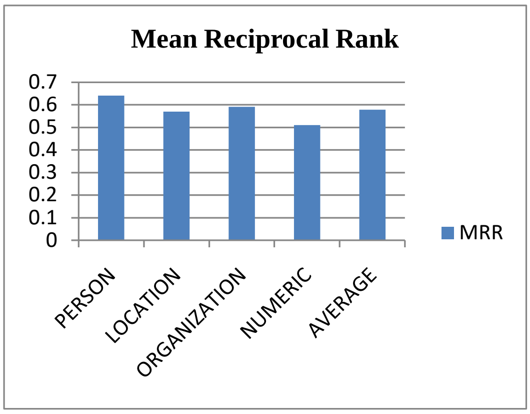

Fig.ure 3. The MRR values for the different question categories

All questions used in the evaluation have at least one correct answer in the returned documents. The total number of questions for each question category used is shown in Table 6. As shown in figure 3 the question set of category "PERSON" got the highest MRR rank. Since this type of questions are easy and questions of this category can be identified just using keywords. The average MRR score recorded is 0.577 which is a good result comparing to other works on answer selection made on English language[30][31][32][33][34].

IV. CONCLUSION

The technology of Question-answering mostly requires extensive human-made knowledge, many language resources, such as parsers, named entity extractors, dictionaries, WordNets and a domain-restricted knowledge base. To overcome the burden of human-made knowledge and external sources, this paper presented another machine learning-based question-answering system for Arabic language that classifies the input question and selects the answer candidate through the proposed machine learning approach. Our QA system consists of several components: question analyzer, document retriever, passage retriever and answer extractor and selector. For the question classifier, our system proves that using several features extracted can improve classification accuracy compared to a bag of word approach. As for the question answering result, we also obtained a good MMR score considering the matter that no high quality language tools involved in the process. A training set of translated questions along with their labels from TREC-10 are used to train the SVM classifier for the answer selection module. The experimental results show that the QA system performed with average MRR score of 0.577 using answer selector. Hence, for future research, more training data should be added to increase the accuracy of the system.

V. REFERENCES

[1] K. Höffner, S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A. N. Ngomo, "Survey on Challenges of Question Answering in the Semantic Web," Semantic Web Journal, pp. 1–26, 2016.

[2] A. Bouzianea, D. Bouchiha, N. Doumiand and M. Malki, "Question Answering Systems: Survey and Trends," In Proc. Of The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT), Procedia Computer Science, Vol. 73, pp. 366 – 375, 2015.

[3] A. Pundge, S.A Khillare., and C. N. Mahender, "Question Answering System, Approaches and Techniques: A Review," International Journal of Computer Applications, Vol. 141, No.3, 2016.

[4] U. kumar, L. Sindhu, "A survey of Natural Language question answering system," International journal of computer applications , vol. 108 ,No. 15, pp. 0975-8887, 2014.

[5] K. Sanjay, and V. Singh, "Research and reviews in question answering system," International Conference on Computational Intelligence: Modeling Techniques and Applications (CIMTA), ,Procedia Technology 10 , pp. 417 – 424 ,2013.

[6] Stanford Arabic Parser, http://nlp.stanford.edu/projects/arabic.shtml, last visited, January 2017.

[7] Buckwalter Arabic Morphological Analyzer. [Online]. Available: https://catalog.ldc.upenn.edu/ldc2004l02. (Current visit January 2017)

[8] Stanford POS Tagger for Arabic, Available:. [Online] . Available: http://nlp.stanford.edu/software/tagger.shtml (Current January 2017)

[9] Arabic Named Entity Recognizer. [Online]. Available: visithttp://qatsdemo.cloudapp.net/farasa/(Current January 2017)

[10] Arabic NLP, RDI's Arabic Lexical Dictionaries. [Online]. Available: http://www.rdi-eg.com/technologies/arabic_nlp.htm. (Current visit January 2017)

[11] Arabic WordNet, http://globalwordnet.org/arabic-wordnet/ (Current visit January 2017)

[12] D. Zhang and W. S. Lee, "Question Classification using Support Vector Machine," In Proceedings of the 26rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 26-32, 2003.

[13] Z. Yu , L. Su , L. Li , Q. Zhao, C. Maoa, and J. Guo, "Question classification based on co-training style semi-supervised learning," Pattern Recognition Letters 31, Vol. 31, pp. 1975–1980, 2010.

[14] M. Skowron and K. Araki, "Effectiveness of Combined Features for Machine Learning Based Question Classification," Journal of Natural Language Processing, Information Processing Society of Japan, Vol. 6, 63-83, 2005.

[15] Khoja , S. Stemming Arabic Text. [Online]. Available: http://zeus. cs.pa cificu.edu/ shereen/ research.htm, (Current January 2017)

[16] A. Sangodiah, R. Ahmad and W. F. Ahmad, "A review in feature extraction approach in question classification using Support Vector Machine", IEEE International Conference on Control System, Computing and Engineering (ICCSCE), pp. 536 - 541 , 2014.

[17] X. Li. and D. Roth, "Learning question classifiers," Proc. 19th International Conference on Computational Linguistics, ACM, 2002.

[18] Z. Huang, T. Marcus and Z. Qin, " Question classification using head words and their hypernyms,", In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08), pp. 927–36,2008.

[19] M. M. Hoque, and P. Quaresma, "A semantic-Based Technique for Question Classification in Question Answering Systems- A hybrid approach", 18th IEEE International Conference on Computer and Information Technology(ICCIT), pp. 38-43 , 2015.

[20] J. Xu, Y. Zhou, & Y. Wang, "A Classification of Questions Using SVM and Semantic Similarity Analysis." In proceeding of Sixth International Conference on Internet Computing for Science and Engineering (ICICSE), IEEE, pp. 31-34, 2012.

[21] Z. Huang, T. Marcus and C. Asli. "Investigation of question classifier in question answering.", In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP '09), pp. 543–50, 2009.

[22] K. Vijay, D. Sujatha and C. Soumen, "Enhanced answer type inference from questions using sequential models," In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05), pp. 315–22, 2005.

[23] L. Wang; H.Zhang; D. Wang; and J. Huang, "Chinese Question Classification Based on Semantic Gram and SVM", International Forum on Computer Science-Technology and Applications, Vol. 1, pp. 432 – 435, 2009.

[24] L. Xia; Z. Teng, and F. Ren, "Question classification in chinese restricted-domain based on SVM and domain dictionary", In Proceedings Of the International Conference on Natural Language Processing and Knowledge Engineering, NLP-KE '08, pp. 1-6,2008.

[25] V. Singh; S. K. Dwivedi, "An integrated pattern matching and machine learning approach for question classification", 1st IEEE International Conference onNext Generation Computing Technologies (NGCT), pp. 762-767, 2015.

[26] S. F. Kusuma, D. Siahaan and U. L. Yuhana, "Automatic Indonesia's Questions Classification Based On Bloom's Taxonomy Using Natural Language Processing", IEEE International Conference on Information Technology Systems and Innovation (ICITSI), 2015.

[27] E. Sherkat, and M. Farhoodi, "A Hybrid Approach for Question Classification in Persian Automatic Question Answering Systems", In proceedings of the 4th International Conference on Computer and Knowledge Engineering (ICCKE), 279-284,2014.

[28] Y. Liu, X. Yi ,R. Chenand, and Y. Song, "A Survey on Frameworks and Methods of Question Answering", In Proceedings of te3rd International Conference on Information Science and Control Engineering (ICISCE), pp. 115-119, 2016.

[29] E.M. Voorhees, "Overview of the TREC 2001 question answering track," In: Proceedings of the 10th Text Retrieval Conference, pp. 42–52, 2001.

[30] E. Hovy, U. Hermjakob, C.Y. Lin, "The use of external knowledge in factoid QA", In: Proceedings of the 10th Text Retrieval Conference, pp. 644–652, 2001.

[31] A. Ittycheriah, M. Franz, and S. Roukos, "IBM's statistical question answering system TREC-10", In: Proceedings of the 10th Text Retrieval Conference, pp. 258–264,2001.

[32] M. M. Soubbotin, "Patterns of potential answer expressions as clues to the right answers," In: Proceedings of the 10th Text Retrieval Conference, pp. 293–302, 2001.

[33] S-J Yen, Y-C Wub, J-C Yang, Y-S Lee, C-J Lee, J-J Liu, "A support vector machine-based context-ranking model for question answering", Information Sciences, pp. 77–87, 2013.

[34] A. Grappy, B. Grau, M-H. Falco, A-L. Ligozat, I. Robba, A. Vilnat, "Selecting answers to questions from Web documents by a robust validation process", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 55-62, (2011)