Question Classification using a Rule based Model

Aarthi D, Viswanathan V, Nandhini B, Ilakiyaselvan N

Abstract: ***Question Answering is one of the most common applications for data acquisition. Although the majority of text-mining applications strive to improve the user experience and the tools used to find appropriate answers, the problems still exist because the web content is constantly increasing. The Questions Classification (QC) task is one of the main tasks in improving the classification system is to classify types of questions in the text mining application. A large number of QC methods are introduced to help resolve classification problems, most of which are bag of words approaches. In this project, we propose a QC system that uses Parts of Speech (POS) Tagger and Named Entity Recognition (NER) Tagger from the Stanford core Natural Language Processing (NLP) to classify the questions correctly. We started by cleaning the data by removing the available labels in the questions then we proceed by tagging the questions by splitting words and tagging each and every words in the input question with the POS Tagger. After this step, we will convert them into a pattern without changing the structure of the question. Then we proceed by tagging the question with NER Tagger. Finally, we will do confirmation process for certain question types which is performed by confirming question type module to make the system work efficiently.***

Keywords: ***Natural Language processing, Text mining, Question answering system, Data mining.***

I. INTRODUCTION

Classification of questions is a basic role in Question and Answering Systems (QAS). Identifying the exact type of question facilitates the extraction of more accurate answers. Continuous growth in the amount of web content, however, makes it difficult to find relevant answers. The most difficult type of question to identify is the factoid questions.

Various methods have been suggested with the aim of improving the detection and classification of questions according to the question types. Most of these are focused on linguistic characteristics and bag of words. Various taxonomies were suggested in [1–4] and from Li and Roth's categories [5] are the common classification taxonomy of factoid (' wh-') queries. The taxonomy of two layers consists of a collection of six coarse grained categories that are Abbreviation, Person, Definition, Individual, Location and Numeric meaning, and fine grained classes such as Language, Manner, Color, Event and City. The features are the key to obtaining a perfect question classifier and linguistic features does a perfect role in the development of a question classifier with good accuracy [6].

Revised Manuscript Received on November 06, 2019.

***** Correspondence Author

Aarthi D, Viswanathan V, Nandhini B, Ilakiyaselvan N, School of Computing Science and Engineering, Vellore Institute of Technology, Chennai, India.

Many studies have categorized user questions by various features such as bag-of-words [7] [8] [9] and others like uni-gram and word form features [10].

In [2] form of query is suggested as a specific category of semantics and is characterized by some common properties.

In addition, several previous studies used machine learning algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), and Decision Tree (DT) are used to identify queries. SVM is one of the most widely used algorithms. In [8] and [9] says about using SVM in certain machine learning algorithms such as NB, K-Nearest Neighbors (KNN) and DT The integration of features such as syntactic, lexical and semantic attributes with an SVM classifier increases the performance of question classification.

The classification of 'wh' questions is more difficult to classify them into proper semantic categories than to classify other styles in the answering systems of questions [11]. Moldovan et.al [12] says the majority of errors occur as a result of the misclassification of QAS questions is because of the misclassification of the users question type. Y. Hao et.al in [3] have combined matching patterns and machine learning algorithms for problem identification, while [13] have categorized questions by their predicted answer styles. The classification in [14] is done by using Semi-Supervised Learning and in [15] it is done by using some machine learning algorithms. The POS [16] and NER [17] taggers are used to tag the question strings to make the classification process easy.

In this paper, we have proposed a new method of classifying questions according to its type. We are doing POS tagging on the question string using Stanford NLP Maxent Tagger. We also recognize the Named Entities and tags to each word. Using Stanford NLP RegexNER. It uses a user-defined RegexNER pattern. We have used an algorithm to classify the question in one of the following question types they are “WHAT”,” WHEN”,” WHO”,” WHERE”,” WHICH”,” WHY”,” HOW” and "AFFIRMATION". These taggers and algorithms are used in order to improve the performance of the question classification system.

In this paper, Section 2 discussed about the description of the previous work in question classification. Section 3 explains the methodology suggested. Section 4 shows the framework. Section 5 describes the experimental set-up and explains the results with the chart. Ultimately, the paper is concluded by Section 6.

II. QUESTION CLASSIFICATION AND ITS METHODS

In this section, we are analyzing the work related using Li and Roth's problem categories on QC methods and machine learning algorithms.

P. Le-Hong et.al [6] suggested a compact set of features using typed dependencies as semantic features.X. Li et.al in [7] used composite statistics and rule based classifiers with various classifiers and hybrid methods for multiple classifiers.

Three separate classifiers were proposed in [9] problem classification process. One of the six fine classifiers and a coarse classifier identify the problem sequentially twice. In addition, the coarse classifier and fine classifiers were used with various machine learning algorithms. Finally, by marking head nouns, researchers listed what-type questions in [11].

In addition, various features have been incorporated, like regional syntactic features, semantic feature and category dependence. Unlabeled questions were used for semi-supervised learning in [14] researchers in conjunction with marked questions. Therefore, tri-training was chosen to improve the accuracy of the function of identification of queries. Therefore, in [15], a hierarchical classifier with two levels was proposed for problem classification. A method has been proposed in [18] that are using the feature selection algorithm to evaluate correct features for various types of questions.Though researchers in [19] proposed a SVM-based classifier. In addition; a problem identification method based on SVM was proposed in [20].

III. PROPOSED APPROACH

A. Question Features:

In this we have done analysis of factoid questions. "wh" questions have uniquefeatures, structures, and characteristics that helps us to identify and characterize them according to their question types. The major feature of a factoid questions is the existence of question words like what, where, why, how, who, when and what. We also have questions with how, how many, how often, how far, how much, how long, how old, etc... In addition, we have also classifies affirmation questions. e.g. "Do you know?"

Apart from this, these question structures start with a preposition. e.g." In which year did India got Independence?" or "at what time did u left the college?". In some cases, we can find the question word in the questions middle part. e.g." The cerebellum is in what part of the body?".

Most of the factoid questions are related to facts, events, suggestions and ideas e.g. "How do you make a ball?" other questions may contain two question words. e.g." What does extended definition mean and how would one write a paper on it?" We can expect any kind of response answer in factoid questions.

B. Question Classification:

The proposed methodology uses four main features for the classification of the questions according to its question types. They are Cleansing Data, POS Tagging, NER Tagging and Confirming Question Types. These features help us to maintain the structure of the question and help us to classify them based on the tagging.

- **Cleansing Data:**

The major work of this module is to perform cleansing on the input dataset. In this it cleanses the data by removing the label from each and every question in the data set. This done by using Apache spark map transformation. Map transformation is nothing but it takes one element and it process the element according to the custom code and produces one element. In this our code will be splitting the question string into words and the words will be joined after removing the label. The map transformation will result in the cleansed dataset.

- **POS (Parts of Speech) Tagging:**

In this module the question string will be tagged with the POS tagging. The POS Tagging is done by Stanford NLP Maxent Tagger. The question string is tagged by creating the input stream for each line and tokenizing them. Then followed by tagging the tokenized question string.

- **NER (Named Entity Recognition) Tagging:**

The NER module is used to recognize the named entities in the question string and it will tag them to each word. It uses Stanford NLP RegexNER. In this the system used user defined Regexner Pattern for NER tagging. The tagged string will help in classifying them according to the question types using the named entities tagged.

- **Confirming Question Type:**

This module is used to confirm two question type. They are "when" and "affirmation" types. They classify and return the question type using a testing format. The format they use is parts of speech and their position match. With this format they classify the question with "when" type or "affirmation" type.

- **Question Classifier Algorithm:**

This classifier starts by preparing the input dataset. For this we take sample input text which has some random questions and cases not available in our "li and Roth" dataset. Then we cleanse the "li and Roth" dataset. Finally, we combine these two into one dataset. First we take the input string and check if it has "WHAT" if it is true then we use POS tagging on the question string, we check it for "/WP" pattern, if it's true then we will add them to "WHEN" else we will see for "/WDT" pattern if it has, then we do POS tagging for it and if it contains "/WHENTYPE" if it has then it will be in "WHEN" type question. Similarly, it checks for "WHICH" and "WHERE" type if it falls under them then it will be in that question type. Then, we use our first question string and we make use of confirm question type module and analyze the question for "AFFIRMATION" type based on the result it will be in affirmation and then we again use the input string and search for other question types "WHO"," WHEN"," WHERE"," WHICH"," WHY"," HOW" if they have any one of these question words they will be classified under their type. If the question does not fall under any of these types, then we will label them unknown. The detailed process of question classification is given in the algorithm.

QUESTION CLASSIFICATION ALGORITHM

Input the question string
While (question string is not null)
If (Question string contains "WHAT")
//Now use POS Tagging
If (tagged string contains "/WP" Pattern)
//" WHAT" Type
Else If (tagged string contains "/WDT" Pattern)
//Now use NER

Tagging
If (NER tagged line has "/WHENTYPE" & check with type confirm module)

//"WHEN" Type
Else
//"UNKNOWN" Type
Else If (line contains "WHO")
//" WHO" Type
Else If (line contains "WHEN")
//" WHEN" Type
Else If (line contains "WHERE")
//" WHERE" Type
If (line contains "WHY")
//" WHY" Type
Else If (line contains "WHICH")
//" WHICH" Type

Else If (line contains "HOW")

//" HOW" Type
Else If (check with confirm module for affirmation type)
//" AFFIRMATION" type

IV. DESIGN

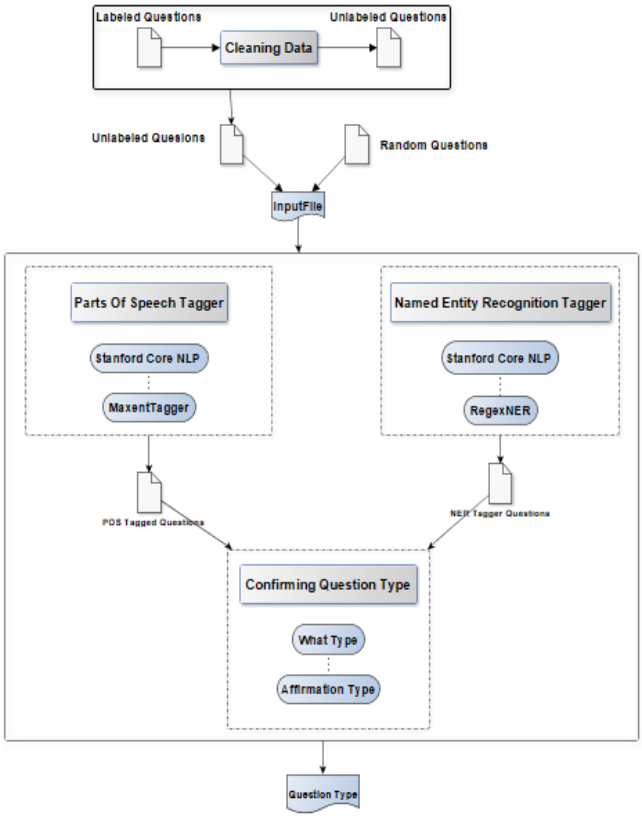

Figure 1: Question Classification Framework

The classification process starts with cleaning the question that is removing the labels from the questions. Then the unlabeled questions are combined with some random questions and given as the input file. The questions are then tagged by POS tagger from Stanford NLP by splitting them and combining them after tagging. The next step is doing NER tagging on the POS tagged question string and for confirming when and affirmation type question. We have the last module after that we will have the questions classified according to their question types. Figure 1 shown the complete design flow of Question Classification framework.

V. RESULTS AND DISCUSSION

In this experiment we have classified questions according to their question types. We used maxent tagger for POS tagging and RegexNERSequenceClassifier for named entity recognition. We used spark framework for cleansing data. The dataset used is Li and Roth data set with 1000 questions and some sample questions that are not present in the previous data set.

The result of the experiment is shown in the Figure 2. The chart shows the comparison of the total questions and the question type found by the classifier. The questions that the classifier could not classify is taken as unknown.

The question classification system uses the taxonomy of classification of factoid question types that is Li and Roth' Categories. We are concentrating on their classification of questions. The collection of coarse grained categories of their two-layer taxonomy consists of Location, Numeric Value, Description, human, Entity, Abbreviation and fine grained classes such as Manner, Color, City, Expression, Event. The data set had different question types with different number of questions. It has total of 591 "WHAT" question type but the classifier found only 331 and classified the other questions in other question type or under "UNKNOWN" type. The dataset had total of 42 "WHEN" question type and our classifier found all of them and classified them correctly under "WHEN" type. We totally have 107 "WHO" question type but our classifier found 107 questions and they classified others under "UNKNOWN" type and we had 58 "WHERE" type and we found 52 and classified them. Then we had 28 "WHICH" type and 17 "WHY" types and our classification system found them and classified them with better accuracy. Initially we had total of 131 "HOW" types and classified them correctly. We could also see that the "AFFIRMATION" types of questions are total of 5 and 3 questions are classified correctly.

The chart shows that the classifier could find" WHY"," WHICH"," WHEN"," HOW"," WHO" questions correctly. It has classified "AFFIRMATION"," WHERE" and "WHAT" with some deviation and it classified 3.3% of the questions as "UNKNOWN". So from the results we can see our question classification system works with 80.8% accuracy.

The results of the classifier are given in the Table 1 with the percentages. We can see that our classifier classified "Affirmation" type of question 60% as it could not find the question which has questions at the end of the question. E.g.) "You are Tom, aren't you?". Similarly, the classifier could classify only 56% of "WHAT" type and 96.20% of "WHERE" type and that is

because the classifier could not find the type when the question words appear at the middle of the question.

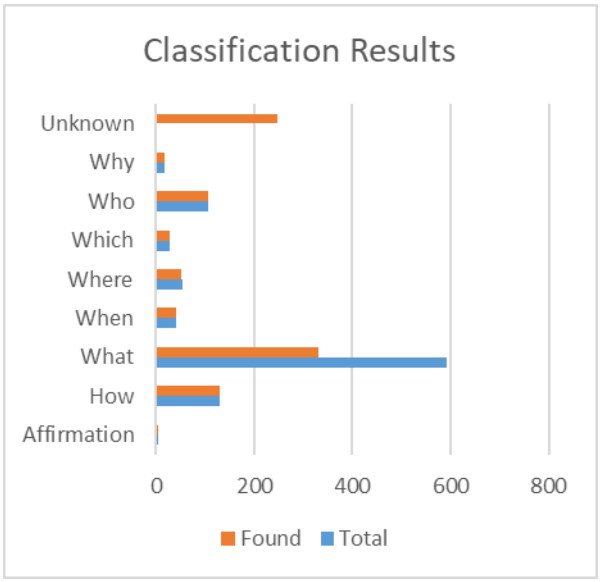

Figure 2: Question Classification Results

Question Type	Percentage
Affirmation	60%
How	100%
What	56%
When	100%
Where	96.20%
Which	100%
Who	100%
Why	100%

Table 1: Question Classification Results with Percentages

Thus, the problem arises with the classifier when the position of the question words changes.

VI. CONCLUSION

In our approach, the classification done by using Parts of Speech and Named Entity Recognition tagger for better classification and this approach helped to maintain the question structure without changing it. The use of Maxent tagger and NER from Stanford core NLP improved the classification of the questions. The results show that the classifier could classify the questions properly. Only we need to concentrate on the "WHAT", "AFFIRMATION" and "WHERE" types of questions. The approach has to be made flexible so that our classifier could able to find even these question types question completely that results in good performance.

In the future, we can add some steps in our approach and compare it with other machine learning algorithm for getting better results. In addition, we can test the proposed classifier in other domains, with other question types etc...

REFERENCE

1. A. Mohasseb, M. Bader-El-Den, M. Cocea, Question categorization and classification using grammar based approach, Information Processing and Management, 2018.
2. O. Kolomiyets, M.-F. Moens, A survey on question answering technology from an information retrieval perspective, Inf. Sci. 181 (24) (2011) 5412–5434.
3. F. Bu, X. Zhu, Y. Hao, X. Zhu, Function-based question classification for general qa, in: Proceedings of the 2010 conference on empirical methods in natural language processing, Association for Computational Linguistics, 2010, pp. 1119–1128.
4. J. Bullington, I. Endres, M. Rahman, Open ended question classification using support vector machines, in: MAICS, 2007.
5. X. Li, D. Roth, Learning question classifiers: the role of semantic information, Nat. Lang. Eng. 12 (03) (2006) 229–249.
6. P. Le-Hong, X.-H. Phan, T.-D. Nguyen, Using dependency analysis to improve question classification, in: Knowledge and Systems Engineering, Springer, 2015, pp. 653–665.
7. X. Li, X.-J. Huang, L.-d. Wu, Question classification using multiple classifiers, in: Proceedings of the 5th Workshop on Asian Language Resources and First Symposium on Asian Language Resources Network,2005.
8. D. Metzler, W.B. Croft, Analysis of statistical question classification for fact-based questions, Inf. Retr. 8 (3) (2005) 481–504.
9. D. Zhang, W.S. Lee, Question classification using support vector machines, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, ACM,2003, pp. 26–32.
10. M. Mishra, V.K. Mishra, H. Sharma, Question classification using semantic, syntactic and lexical features, Int. J. Web Semant. Technol. 4 (3) (2013) 39.
11. Z. Huang, M. Thint, Z. Qin, Question classification using head words and their hypernyms, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2008, pp. 927–936.
12. F. Li, X. Zhang, J. Yuan, X. Zhu, Classifying what-type questions by head noun tagging, in: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, Association for Computational Linguistics, 2008, pp. 481–488.
13. D. Moldovan, M. Pa¸sca, S. Harabagiu, M. Surdeanu, Performance issues and error analysis in an open-domain question answering system, ACM Trans. Inf. Syst. 21 (2) (2003) 133–154.
14. F. Benamara, Cooperative question answering in restricted domains: the webcoop experiment, 2004.
15. T.T. Nguyen, L.M. Nguyen, A. Shimazu, Using Semi-Supervised Learning for Question Classification, vol. 3, Information and Media Technologies Editorial Board, 2008, pp. 112–130.
16. T.T. Nguyen, L.M. Nguyen, Improving the accuracy of question classification with machine learning, in: Research, Innovation and Vision for the Future, 2007 IEEE International Conference on, IEEE, 2007, pp. 234–241.
17. Márquez, Lluís, and Lluis Padró. "A flexible POS tagger using an automatically acquired language model." Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997.
18. Speck, R., Ngonga Ngomo, A.-C.: Ensemble learning for named entity recognition. In: Mika, P., et al. (eds.) ISWC 2014, Part I. LNCS, vol. 8796, pp. 519–534. Springer, Heidelberg (2014)
19. N. Van-Tu, L. Anh-Cuong, Improving question classification by feature extraction and selection, Indian J. Sci. Technol. 9 (17) (2016).
20. D. Metzler, W.B. Croft, Analysis of statistical question classification for fact-based questions, Inf. Retr. 8 (3) (2005) 481–504.
21. S. Xu, G. Cheng, F. Kong, Research on question classification for automatic question answering, in: Asian Language Processing (IALP), 2016 International Conference on, IEEE, 2016, pp. 218–221.

AUTHORS PROFILE

Aarthi D is pursuing her research in School of computing science and engineering, VIT Chennai. She received a Bachelor's degree in Information Technology at Anna University and a Master's degree in computer science and engineering at Anna University of Technology. Her research interests in the areas of Natural language processing, Semantic web and Data mining.

Viswanathan V is a professor in the School of Computing Science and Engineering at Vellore Institute of Technology, Chennai. He completed his Doctoral degree from Anna University, Chennai, India, by contributing his ideas to the field of Semantic Web Technologies and Social media marketing. He has a teaching experience of over 20 years in the field of Computer Science. His research interests include Data mining, Semantic Web, and Social Network Analysis. He has authored articles in Semantic web technologies for renowned publications.

Nandhini B is presently doing his M.Tech software Engineering (Integrated) at Vellore Institute of Technology, Chennai. She has research interest in the field of Natural Language Processing, Software Engineering and Semantic Web. She is a motivated technologist and interested in developing new projects.

N. Ilakiyaselvan is currently working as Assistant Professor (Senior) in the School of Computing Science and Engineering at Vellore Institute of Technology, Chennai. He completed his Bachelor's degree in Information Technology and Master's degree at Anna University. He has 9 years of teaching experience and research interests in the field of Software Engineering, Natural language Processing, and Biomedical signal analysis.