|  |  |
| --- | --- |
| question classification Using Rule-Based Model IN QA SYSTEM  Ibtihal khan[[1]](#footnote-1)  Department of Computing Science and Mathematics  Dundalk Institute of Technology  Dundalk, Ireland  [ibtihalkhan82@gmail.com](mailto:ibtihalkhan82@gmail.com)  ABSTRACT  Question Answering Systems are an important Field in Natural Language Processing and has wide range of real world applications. A typical QA system comprises of three major processes: Query Processing and Classification, Information retrieval and Answer Extraction and Ranking. Question Classification (QC) is one the primary tasks carried out in any Question Answering System (QAS). State-of-the-art question classification systems plays a significant role in identifying the answer type expected from a given query which can be utilized to achieve considerable improvement in the answer extraction phase. This study conducts a preliminary review of existing literature discussing the various approaches to the problem of question classification. The literature thus reviewed are analysed and categorized into different groups based on the general approach and the specific algorithms employed in the model. Further, a novel strategy is proposed to build a simple binary classification model which classifies the given query as either objective or descriptive  *Keyword-question answering system, question classification, rule based model* INTRODUCTION Question answering system leverage on Natural Language Processing and Other associated techniques to provide an answer to a user generated query from the data repository. In such system both the query and answer will be formulated in natural languages. A QA system attempts to extract a precise and exact answer for a given query as opposed to the general natural language-based search engines available. In other words, a QA system is simply an extension of a general Search engine. A QA system primarily consist of three major sub-processes:   1. Query Processing and classification 2. Information Retrieval 3. Answer Extraction       Figure -1) Question Answer Framework | Templates for DAFx-08, FINLAND, FRANCE |

Out of the given sequence of steps, Query classification phase processes the query and assigns an expected answer type to the query. This information can further be used for improving the answer extraction phase. The existing literature on the field emphasizes on the fact that, incorrect question classification is a major factor affecting the performance of such system. Hence, Question Classification module design in question answering systems is a major area of concern and an improvement in the phase could improve the overall performance of such systems.

This study attempts to understand the field better by conducting a preliminary review of seminal literature on such QA systems and specifically the question classification model in those systems. The collected literature will then be analysed on the basis of the techniques employed to tackle the problem including ML and non-ML techniques. Further, leveraging on the literature analysis thus conducted a simple rule-based question classification model is proposed to classify the given query as either descriptive or objective questions suitable for constructing a simple closed-domain QA system

The rest of the paper is organised as follows. Section 2.previous related work. Section. 3.implementation explained using CRISP-DM. Section. 4.discussion of hypothesis. Section. 5.statistical analysis. Section 6.Experiment and Result Section 7.conclusion and future work. Section 8.References

# RELATED SEARCH

Jinzhong,yanan Zhou and Yuan Wang [3] proposed a two level question classification based on vector machine and the question semantic similarity so that they can reduce the complexity of language understanding by improving the specific domain

Muhammad Arifur Rahan,vitalie Scurfu [11] proposed an approach for question classification by subsetting tree kernel using support vector machines so question classification can obtain higher classification accuracy

Liang Wang, Hui Zhang, Deqing Wang, Jia Huang [9]

Used support vector method in this method semantic grams are extracted from questions through semantic thesauruses and N-gram and used as the features for SVM

So that it can solve the problem of classification of Chinese question and propose a feature extracting method for Chinese question classification.

Muntaha Al-Asa d,nour Al-Khdour,Mutaz Bni Younes,Enas Khwaileh,Mahmoud Hammad,Mohammad AL-Smadi [8] they apply text classification to detect similar Arabic question using morphological,syntactic,semantic,and lexical features using XGBoost supervised machine learning model applied on a real dataset with the size so it can efficiently and accurately detect if two questions are similar.

Alaa Mohasseb, Mohamed Bader-el-den, Mihaela Cocea [4] and Han Liu [2] introduced a framework using hierarchical synthetic minority oversampling technique algorithm so that it can evaluate the impact of handling class imbalance in the classification accuracy

Payal Biswas,Aditi Sharan and Rakesh Kumar [5] had given a method in which they have confined the 6 course grain classes and 50 fine fine grain class introduced by Li and Routh into 3 broader categories: definition type question, descriptive type question and factoid type question so that its gives an precise answer

Mourad Sarrouti,Abdelmonaime Lachkar and Said El Alaoui [10] have defined the Syntactic Patterns of each Biomedical QTs. and classified the Biomedical Questions into three broad categories: Yes/No, Factoid and Summary Questions so that this approach can be used to effectively classify Biomedical Questions with higher accuracy

Md Moinul Hoque;Paulo Quaresma Proposed an hybrid approach that analyses a question placed in English language both syntactical and semantically. The approach proposes a semantic memory and a question understanding based classification method combined with a pattern based method so that this approach can develop more accurate question classifier

Refany Anhar,Teghu Bharat Adji and Noor Akhmad Setiawan [1] they conducted a study in which they compare Bi-LSTM with a multilayer perception for checking the accuracy rate in which they found that Bi-LSTM had a higher accuracy rate of 83% compared to the multilayer perception which had an accuracy of 78% so using this study result they can solve the problem of speed and accuracy of Chabot

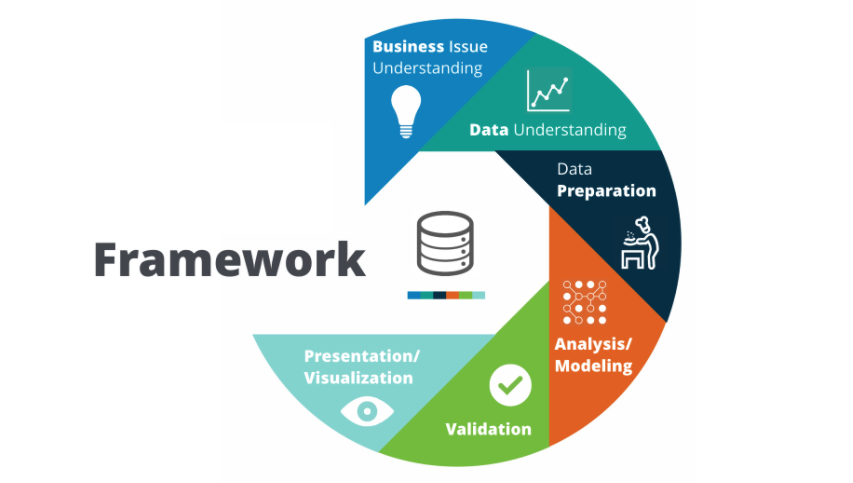
Bui Thanh Hung [7] follows machine learning based approach in which they use pre-train word embedding’s in bidirectional LSTM model for Vietnamese question classification and integrate automated question answering systems into Educational Support System so the user get an appropriate answer

Songjiang Hu,Yajun Du and Xianglin Luo [2] proposed question classification model(BAL)that combined with Long Short-Term Memory(LSTM) network which has memory capability compared to Conventional Neural Network(CNN) and can better process text data in sequential Forms and achieved better result in classification and text prediction and finally this method can get multi-level feature of the text and achieve good result

Tanmay Vakare ,Kshitij Verma and Vedant Jain [6] buit a correlation matrix for dependency relation tags and using topology representation of WordNet for semantic between similarity of words and their approach focuses on matching dependency relation between word of sentence with that of others and also calculate the semantic similarity between words of the sentences by calculating the distance between them in their topological semanting representation then the combination of these metrics generate a score which determine the semantics closeness of the sentences

Hengxun Li,Ning Wang,Guangjun Hu,Weiqing Yang [12], proposed a context-aware hybrid model based on PGM and word2vec to deal with question classification in a heuristic and semantic way so that search engine could work on a smaller dataset and the final answer could be more close to the user intention

**IMPLEMENTATION EXPLAINED USING CRISP\_DM**

****

**Business Understanding:** In the first phase of the implementation I tried to understand the problem of QA system by reviewing the literature view because this study aim to implement accurate QA system that gives precise answer to the user generated query

**Data Understanding** In the second phase I tried to search the dataset that contain question and answer pair for specific domain

**Data preparation:** In this phase I collected the data from the website containing question answer pair for specific domain and after collecting the data the question is labelled manually as descriptive or objective

**Analysis/Modeling:**In this phase firstly the dataset question is labelled as Descriptive or Objective using the constructed rule based model and then matched with the manually labelled classes and if they matched correctly then we assign 1 for the correctly labelled question and 0 for not correctly labelled and then the correctly mapped question is analysed and accuracy of QA system is detected and analysis is conducted by taking the dataset from the url after requesting the dataset

**Validation:** In this phase I can see how much the constructed rule based model is valid for improving the accuracy QA system after analysing and interpreting the model

**Visualisation:** In this phase of implementation I can tell how good the constructed model is, after visualizing the plots and tables and also describe how good the constructed rule based model work for improving the accuracy of QA system

**DISCUSSION OF HYPOTHESIS**

This study proposes a simple rule-based classification model for the task of question classification model. The model attempts to bucket a user-generated query into one of the pre-defined categories: descriptive and objective. Such a simple classification system would be suitable for a simple closed domain QA system. Descriptive category denotes those question whose answer is constitutes more than one passage whereas objective category is assigned to those questions whose answer is contained within a paragraph. Hence a two level classification model such as this would direct the answer extraction phase from the candidate passages returned

The proposed model follows a simple rule based algorithm and assigns the respective category according to the question word encountered in the question. Why type of question belong to Descriptive type Question and Which, When, Who, What and Where type question belong to Objective class of Question and and how type question belong to both objective and Descriptive class of classification. Let us understand this using few examples from the Dataset. Various conventions used in the Question Pattern are as follows:

QW: Question Word as Who,Where,what,when etc.

auxV:auxilary verb

***Why-****Question type : always descriptive*

*Question pattern : QW +[anything]*

*Example:*

* Why did Lincoln issue the Emancipation Proclamation?
* Why did Coolidge not attend law school?
* Why are ducklings particularly vulnerable?

***When-****Question type : always objective*

*Question pattern : QW +[anything]*

*Example:*

* When was the Six Day War?
* When did organized agriculture appear in Nile Valley?
* When do African elephants lie down?

***Where-****Question type : always objective*

*Question pattern : QW +[anything]*

*Example:*

* Where was the largest elephant ever recorded shot?
* Where is Finland located?
* Where is old Ghana in relation to present Ghana?

***Which-****Question type : always objective*

*Question pattern : QW +[anything]*

*Example:*

* Which spice originally attracted Europeans to Indonesia?
* Which county was Lincoln born in?
* Which countries established colonies in Canada?

***Who-****Question type : always objective*

*Question pattern : QW +[anything]*

*Example:*

* Who is Daffy Duck?
* Who is the most popular rock group in Finland?
* Who did Ford nominate for Vice President?

***how-****Question type : objective and descriptive*

*Question pattern : QW +[anything] – objective*

*QW + auxV + [anything]-descriptive*

*Example:*

*Objective -*

* How many Eagle Scouts were involved in Ford's funeral procession?
* How long is a leopard's tail?

*Descriptive –*

* How did Fillmore ascend to the presidency?

***what-****Question type : objective and descriptive*

*Question pattern : QW +[anything] – objective*

*Example:*

*Objective -*

* What did The Legal Tender Act of 1862 establish?
* What countries border Egypt?
* What did Alessandro Volta publish in 1800?

Algorithm [I] can be utilized to classify the given query into any one of the predefined categories utilizing the query word and synaptic structure of the question

Once we have classified the query into specific category we can then predict the expected answer type for the given query. Table [1]outlines the question, its possible categories and the expected type of answer in accordance with the category

Algorithm 1:

Case 1:question word=”why”

Result=”Descriptive”

Case 2:question word=”when”

Result=”Objective”

Case 3:question word=”where”

Result=”Objective”

Case 4:question word=”which”

Result=”Objective”

Case 5:question word=”how”

If question pattern=QW+auxV

Result=”Descriptive”

Else

Result=”Objective”

Case 6:question word=”who”

Result=”Objective”

Case 7:question word=”what”

Result=”Objective”

Default:

Result=”Unable to Evaluate”

|  |  |  |
| --- | --- | --- |
| Question | Category | Expected |
| How | Descriptive | paragraph |
| Objective | Numeric(quantity) |
| Why | Descriptive | paragraph |
| When | Objective | Numeric(date, time) |
| Where | Objective | Location |
| Which | Objective | Noun word after QW |
| Who | Objective | Person/Organization name |

Table 1.Question category and Expected

Answer type

**STATISTICAL ANALYSIS**

The data was collected containing question answer pair and question type

526 observations

7 variables

Article Title -Title of the article– ordinal categorical

Question - WH type question - ordinal categorical

Answer - answer of the given question – ordinal categorical

Question\_keyword - store WH keyword– ordinal categorical

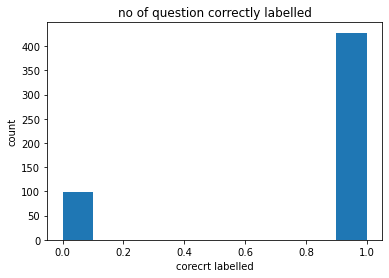
Rule label- question type –ordinal cat

Manually label – question type labeled manually – ordinal cat

Correctly\_labelled - correctly labeled question – nominal cat

**PLOT 1**

**-Checking the feasibility of rule-based model**

****

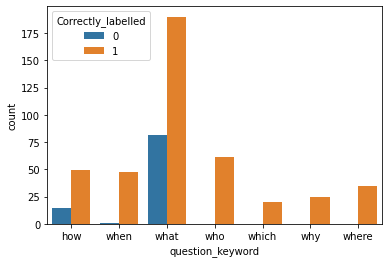
-Most of the question is correctly labeled with the constructed rule based model

-As 1 indicate YES for correctly labeled and 0 indicate NO for correctly labeled

Research question how questions keyword influences by correctly\_labelled (i.e. how much rule based model is feasible)

**PLOT 2**

**-Checking the feasibility of rule based model for each question word separately**



-Most of the Wh question is correctly labeled with the constructed rule based model

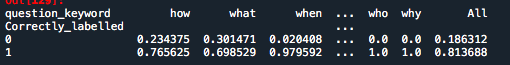
-As 1 indicate YES for correctly labeled and 0 indicate NO for correctly labeled

**Macintosh HD:Users:ibtihalkhan:Desktop:Screen Shot 2021-01-09 at 5.02.56 PM.png**

-by looking at the table all Wh question got correctly label except how,what and when in which 15,82,1 question are not correctly label in total respectively but still they has frequency for correctly label question

-As 1 indicate YES for correctly labeled and 0 indicate NO for correctly labeled

**#propostion table**

****

-Look like there is a relationship between question\_keyword and Correctly\_labelled

-As ALL WH question are almost correctly labeled with the constructed rule based model

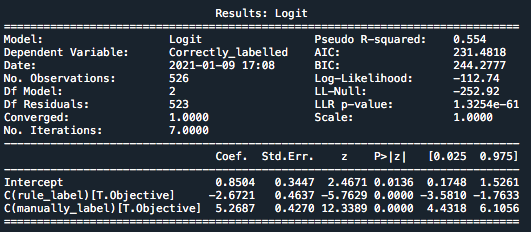
-Giving high accuracy of precise answer

**Perform logistic regression**

- H0: Null model is prefer

-H1: model with rule label and mannuly\_label is preferred

MODEL SUMMARY



-LRT comparing to null model

-As p-value=1.3254e-61>0.05,we reject Null Hypothesis i.e. model with rule label and manually label is preferred

-pi here is probability of correctly label

-logit(pi)=log(success odd of correctly label)=

0.8504-2.6721rule\_labelObjective+5.2687manually\_labelObjective

-np.exp(0.8504)

-est success odds for Correctly label is

2.3405828978634684 when we take rule label and manually label as objective

-np.exp(-2.6721)

-est success odds for correctly label change by a factor

is 0.06910694823051845 when we compare a Descritve to a Objective type for rule\_label - -np.exp(5.2687)

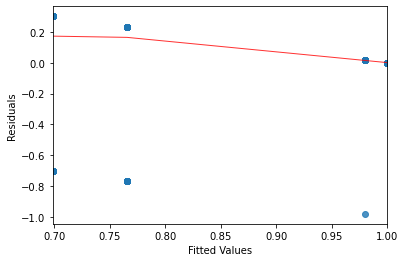
-est success odds for correctly label change by a factor

Are 194.16338590453654 when we compare a Descritve to an Objective type for manual label?

* All terms are significantly different 0 (i.e. they have a relationship with Correctly labeled )

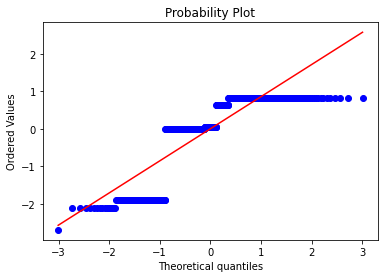
**-Perform one-way ANOVA test**

**PLOT 4**

****

-Spread of residuals not looks the same so not happy with equal vary assumption

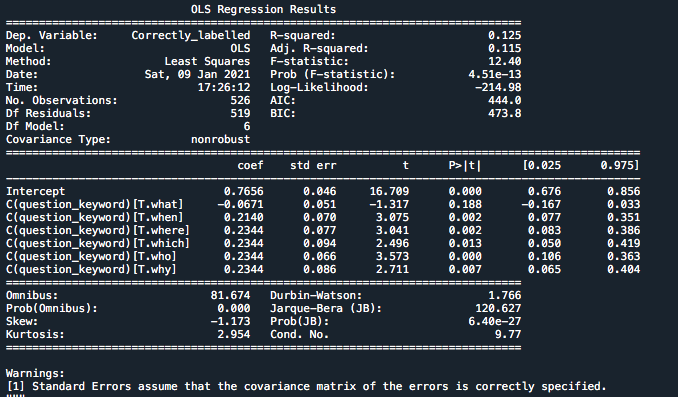
**PLOT 5**



-Evidence of strong deviations from normality as not follows the line

not happy with this assumption

MODEL SUMMARY

****

Looking at the summary we have line equation is

y=0.7656-0.0671question\_keywordwhat+0.2140question\_keywordwhen+0.2344question\_keywordwhere+0.2344 question\_keywordwhich +0.2344 question\_keywordwho+0.2344 question\_keywordwhy

-The model explain only 12.5% variance

**-As all variable are categorical that why only use logistic regression and anova test for analysis**

**EXPERIMENT AND RESULT**

For experimental purpose I have taken Question answer data set of 526 questions from the kaggle and then the rule-based algorithm is used to assigned classes to each of the questions in the dataset as either descriptive or objective and then matched with the manually labeled classes

The model is evaluated on the basis of correct classifications produced with respect to the actual labels. After evaluating the model we get high accuracy for all almost all the WH question as the accuracy result shown in Table [2]

|  |  |  |  |
| --- | --- | --- | --- |
| **Question** | **Total** | **Correctly classified** | **Accuracy** |
| why | 25 | 25 | 100% |
| where | 35 | 35 | 100% |
| which | 20 | 20 | 100% |
| what | 272 | 190 | 70% |
| when | 49 | 48 | 98% |
| how | 64 | 49 | 77% |
| who | 61 | 61 | 100% |

Table 2.performance measurement

By looking at the plot 2 it shows that almost all the WH question is correctly labelled using the constructed rule- based model as 0 indicates NO and 1 indicates YES and we get the high accuracy almost for all Wh question.

**CONCLUSION AND FUTURE WORK**

The problem of question classification is discussed in the study which is one of the most important steps in any Question Answer systems. A good question classification system can inform the system of the expectations of the user who has generated the query and hence assist in extracting the extracting the exact answer as expected by the user. The proposed model is suitable for a question classification module to be incorporated in a closed domain question answering system and The achivement in this work lies within the field that proposed model obtained is light in weight as compared to the exisisting ones but moreover decrease the question classification comlexity which in turn increases the accuracy of question classification and hence improve the QA system because In proposed model In spite of utilizing the large feature set for question classification I have employed the nature of the question and classified the Wh questions into just two categories that is Descriptive type and Objective Type. From the Experimental Results, we can see that the model proposed in this paper can improve the accuracy of question classification. In the follow up study, we will also optimize the algorithm and structure to further improve the classification effect of the model

**REFERENCES**

[1]Refany Anhar,Teghu Bharat Adji and Noor Akhmad Set tiawan, *Department of Electrical and Information Engineering Universitas Gadjah Mada*,2019

[2] Songjiang Hu,Yajun Du and Xianglin, *School of Computer and Software Engineering Xihua University Chengdu*,2019

[3] Jinzhong1,Yanan Zhou2 and Yuan Wang2,1 *School of Computer Science Zhongyuan University of Technology,2 College Of Computer and Information Engineering Henan University of Economics and Law,*2012

[4] Alaa Mohasseb1, Mohamed Bader-el-den1, Mihaela Cocea1 and Han Liu2, 1*School of Computing, University of Portsmouth, 2School of Computer Science and Informatics matics, Cardiff University* ,2018

[5] Payal Biswas,Aditi Sharan and Rakesh Kumar, *School of Computer and Systems Sciences Jawaharlal Nehru University*, 2014

[6] Tanmay Vakare ,Kshitij Verma and Vedant Jain,*Dept of Information Technology SVIT,*2018

[7] Bui Thanh Hung, *Data Analytics & Artificial Intelligence Laboratory Engineering Thu Dau Mot University,*2019

[8] Muntaha Al-Asa d1,nour Al-Khdour,Mutaz Bni Younes,Enas Khwaileh2,MahmoudHammad3,Mohammad AL-Smadi3, *1Dept. of Computer Science Jordan University of Science,2Technology-Irbid, Jordan and Dept. of Software Engineering Jordan University of Science, 3Technology-Irbid, Jordan,*2019

[9]Liang Wang1 , Hui Zhang1 , Deqing Wang1 , Jia Huang1,*Laboratory of Software Development Environment, Beihang University, Beijing,2009*

[10] Mourad Sarrouti1,Abdelmonaime Lachkar2 and Said El Alaoui Ouatik1,*1Department of Computer Science, FSDM, Sidi Mohamed Ben Abdellah University, Fez, Morocco 2Department of Electrical and Computer Engineering, ENSA, Sidi Mohamed Ben Abdellah University, Fez, Morocco*,2015

[11] Md Moinul Hoque;Paulo Quaresma, *Department of informatics,University of Evors,Protugual,*2015

[12] Hengxun Li1,Ning Wang1,Guangjun Hu1,Weiqing Yang2,1*First Research Institute of the Ministry of Public Security of PRC, Beijing,2* *Chinese Academy of Sciences, Beijing,*2017

1. [↑](#footnote-ref-1)