Natural Language Question Answering

Umma Khatuna Jannat[1], Nirmala S Guptha[2], Anitha K[3]
[1]*Reva University Bengaluru*
[2]*Reva University Bengaluru*
[3]*Reva University Bengaluru*
(E-mail: ummakhatunajannat@gmail.com, nirmalaguptha@reva.edu.in. anitha.k@reva.edu.in)

***Abstract*—** As the amount of information available online is growing, the need to access it becomes increasingly important and the value of natural language processing applications becomes clear. Machine translation helps us conquer language barriers that we often encounter by translating technical manuals, support content or catalog at a significantly reduced cost. The challenge with machine translation technologies is not in translating words, but in understanding the meaning of sentences to provide a true translation. In this textual analysis can be one of the important areas where future behaviours can be predicted, Extraction of the information from any user-given passage or body of text plays a major role in Data Analysis. The user is not restricted by any general genre or category of data .The system can pose the questions based on the given textual contents. For the predicted answer system can assess the correct answer. A simple reading comprehension, while there are a lot of systems that run on machine learning algorithms, most if not all of them are closed domain with training from existing wiki articles or corpora. This work is one step closer to tackling this problem and provides a true understanding between man and machine.

Keywords— Machine learning, Textual analysis, Data analysis.

I. INTRODUCTION

The sheer measure of data in our regular day to day existence is totally stunning in the present day and age. To comprehend everything, manual extraction is administered simply in view of its sheer size, so it falls on us PC architects to figure out how to computerize the extraction of data in a way that is significant, simple and generally secure. That was the motivation to begin this undertaking, to gadget a framework that will read through a group of content and give you the best and most important data in a sorted out and justifiable to make it straightforward huge assortments of content without reading through the entire thing. Obviously, this is less demanding said than done and we encountered various difficulties attributable to the very idea of dialect. Be that as it may, we continued on and the outcome is a task we are glad to have made. Have a data extraction framework that is lightweight, versatile and simple to utilize. As said previously, the common of dialect is a sufficient test for this sort of attempted. Past this, the approach we took to accomplish the objective was additionally something of an issue. Our first break was attempting to actualize a profound learning model however that would restrict our client contribution to specific fields and that isn't the vision we had. Next time, we attempted straight up string coordinating with fluffy and likelihood, while this strategy certainly worked, it was not tried and true in its exactness. At long last, we settled on reg As said previously, the very idea of dialect is a sufficient test for this sort of attempted. Past this, the approach we took to accomplish the objective was likewise something of an issue. Our first split was endeavouring to actualize a profound learning model yet that would restrict our client contribution to specific fields and that isn't the vision we had. Next time, we attempted straight up string coordinating with fuzzy and likelihood, while this strategy unquestionably worked, it was not tried and true in its exactness. At last, we settled on consistent articulations to acclimatize data into something we can answers from. The contention could be improved that CFG was however this was only a less demanding idea for us to wrap our heads around and it works. The following test was to compose the syntax to get the greater part of the normally utilized sentences in our dialect which took a considerable measure longer than we anticipated. In general, we have made the most ideal rendition with our opportunity and restricted know-what about the area yet there is positively space for a considerable measure of changes. Normal articulations to acclimatize data into something we can answer from. The contention could be improved that CFG was yet this was only a less demanding idea for us to wrap our heads around and it works. The following test was to compose the syntax to get the vast majority of the regularly utilized sentences in our dialect which took a great deal longer than we anticipated. By and large, we have made the most ideal variant with our chance and restricted know- what about the space yet there is unquestionably space for a ton of enhancements.

II. LITERATURE SURVEY

In a system developed Athira P. M, Et.al [5], presented an architecture of ontology-based domain- specific natural language question answering that applies semantics and domain knowledge to improve both query construction and answer extraction. The web as a broad scope, auto-updating knowledge store, answer is mined automatically with a wide range of queries with much less work than is required by modern search engines. The system is able to filter semantically matching sentences and their relations effectively, it ranked the correct answers higher in the result list. Another system developed by Pragisha K. Et.al [6], described about the. It receives Malayalam natural language questions from the user and extracts most appropriate response by analyzing a collection of Malayalam documents. The

system handles four each question. The main answer extraction module is NER in Malayalam. The proposed system design and implementation of a QA system in Malayalam also covered the implementation of some linguistic resources classes of factual questions what, which, Where and which, it extracts precise answer and short answer for user queries in Malayalam. Research and reviews in question answering system developed by Sanjay K Dwivedi Et.al[7] propose taxonomy for characterizing Question Answer (QA) systems, survey of major QA systems described in literature and provide a qualitative analysis of them. It includes the QA system like Linguistic Approach, Statistical approach, pattern matching approach, Surface Pattern based, Template based etc, They observed that the choice of a technique is highly problem specific. Often a hybrid approach, blending evidently different techniques, provides improved results in the form of high speed, increased relevancy, and higher accuracy and recall measures. QA techniques based on linguistic approach, statistical approach and pattern based approach will continue to remain in sharp focus. In a System developed by Poonam Gupta Et.al [8] A Survey of Text Question Answering Techniques. Question answering is a difficult form of information retrieval characterized by information needs that are at least somewhat expressed as natural language statements or questions, and was used as one of the most natural type of human computer communication. In comparison with classical IR, where complete documents are considered similar to the information request, in question answering, and specific pieces of information are come back as an answer. The user is interested in a precise, understandable and correct answer, which may consult to a word, sentence, paragraph, image, audio fragment, or an entire document [9]. The main purpose of a QA system is to find out "HOW, WHY, WHEN, WHERE, HOW, WHAT, WHOM and WHO?"[10].QA systems combines the concepts of information retrieval (IR) with information extraction (IE) methods to identify a set of likely set of candidates and then to produce the final answers using some ranking scheme [11].Types of QA systems are Web Based Question Answering Systems.IR / IE Based Question Answering Systems. Restricted Domain Question Answering systems. Rule Based Question Answering Systems. Template Matching Automatic Answering System For natural languages questions proposed by Pachpind Priyanka Et.al [12], Frequently Asked QA System that replies with prestored answers to user questions asked in regular English, rather than keyword or sentence structure based retrieval mechanisms. Techniques: pattern matching technique Types of QA Systems are, closed-domain QA that deals with questions under a specific domain. Open domain QA that deals with questions about almost everything, and can rely only on general ontology and world knowledge. Main modules are: Pre-processing: (a) converting SMS abbreviations into common English words (b) removing stop words, and (c)removing vowels. Question template matching: The pre-processed text is coordinated against each and every pre stored template awaiting it finds the best template. Answering the matching answer will be returned to the end user.

III. QUESTION ANSWERING APPROACHES

A. Linguistic approach

An inquiry noting rationale contains AI based strategies that incorporate Natural Language handling (NLP) strategy and learning base. The learning data is composed as generation administer, rationale outlines, layouts, metaphysics and semantic systems; it is utilized amid the examination of QA match. Parsing, Tokenization, and POS labeling are semantic strategies, it actualized to clients address for detailing it into an exact inquiry that predetermined concentrate the particular reaction from basic database. In late work the restriction of learning base is acknowledged as the capacity to give a circumstance particular are Clark et al [1] introduced methodologies for increasing on the web content with information base inquiry noting capacity. Existing inquiry noting START [2], QA framework by chang Et.al [3] and mishra Et.al [4] have obtained web as their insight asset.

B. Statistical Approach

Significance of measurable approach is expanded by the sudden development of accessible online content archives. Factual methodologies are autonomous of SQL and can plan questions in common dialect shape. One burden of factual approach is it treats each term freely and neglects to recognize etymological highlights for a blend words or expression. Measurable methods effectively connected to the diverse phases of the QA framework. The system utilized for characterization reason for existing is Maximum entropy models, bolster vector machine (SVM) classifiers, Bayesian classifiers. The vital work in light of the measurable technique was IBM‟s factual QA [9] framework. It utilized most extreme entropy display for question/answer based different N-gram highlights.

C. Pattern Matching

Approach The example coordinating methodology utilizes the expressive energy of content examples. It replaces the complex handling engaged with other contending approaches. "World Cup 2014 held?" takes after the example "Where was held?" and its answer example will be "was held at". There are two methodologies: Surface Pattern based and Template based. A large portion of the examples coordinating QA frameworks utilize the surface content examples while some of them likewise depend on formats for reaction age.

Surface Pattern: Based It is either human made or naturally learned examples through illustrations. Answer sentences for instance, the inquiry "Where was Football" is extricated utilizing factual procedures or information mining measures. Example learned by in self-loader and the most good application region is little and medium size site.

Template based: This approach makes utilization of preformatted designs for questions. The fundamental focal point of this approach is more on exhibit as opposed to clarification of inquiries and answers. The formats set is worked keeping in mind the end goal to contain the ideal number of layouts secure that it adequately cover the space of issue, and every one of its individuals speaks to an extensive

variety of inquiries of their own sort. The substance spaces of Templates, which are missing components bound to the idea of the inquiry that must be filled to create the question layout to recover the relating reaction from the database. The reaction returned by inquiry will be crude information; it is back pedalling to the client.

IV. SYSTEM ANALYSIS AND DESIGN

The system is completely based in Python with no extra dependencies except the modules included in the code itself. As of this version of the project, nothing else is needed or used to execute it. An overview of the working of the system :-Take the input in the form of input, Spell-check and prepare it for the grammar, Parse it through the custom made grammar with the regexpparser from nltk,Parsed tree is traversed to check labels against required context, Labels are either made into keys or values of a dictionary that serves as the knowledge base, Take the input of question from the user, Depending on the type of question (who, what, where and when only), the respective modules are executed, Each module is tailor made to answer that specific question word, If the system is not able to answer the question or the data is insufficient to determine a probable outcome, the system will ask for choices, User can put up to 4 options, one must be correct then the Answer is displayed.

V. IMPLEMENTATION

After having moderate success with string matching, we now use regular expressions to extract relevant information. Each sentence is parsed through the RegexpParser() from the nltk library. Ideally, we would be able to predict every possible pattern but that is unlikely and a majority will have to do.We use the parts of speech tags of each word to associate them.

A few code snippets –

```
pattern = """

            P3: {<NNP>+ <CC>* <VB.*>*}
            P8: {<P3><TO|IN|DT|R.*>*
<VB.*>*
<JJ.*|NN|NNS>* <VB.*>* <TO|IN|DT|R.*>* <P3>*
<NN|NNS>* }
            P2: {<P8><TO>*      <NNP>? }
            C1: {<P8><P7>}
            P1:     {<NNP><VB.*>*      <NNP>?
<TO|IN|DT|R.*>* <VB.*>* <NNS|NN>* <NNP>?}
            P6 :    {<CD>*    <NN.*>*      <VB.*>*
<NN.*>?}
            P7 :    {<P6><R.*>*    <CC>*    <TO>*
<VB.*>* <NN.*>?}
            C1: {<P8|P2><P7>}
            P5: {<DT>* <NN.*>* <VB.*>* <IN>*
<VB.*>* <JJ>* <NN>* <NNS>* <NNP>? }

            P4: {<NNP>+ <VB.*>* <TO>* <VB.*>*
<NN.*>?}
            P7:    {<P6><R.*>*        <CC>*
<VB.*>*  <J1><R.*>* <J1><VB.*>* <J1><NN.*>?}
"""

#This is the custom made grammar to catch patterns
tags=nltk.pos_tag(nltk.regexp_tokenize(text,"[\w//,]+ "))
chunker=nltk.RegexpParser(pattern)
tree=chunker.parse(tags)

#Parsing with the regexpparser from nltk

for sub in tree.subtrees():
words=[]
ifsub.label() in label:
fori in sub.leaves():
words.append(i[0])
count=-1
fori in words:
count=count+1
ifi in ent:
phrase[' '.join(words)]=words[count]
nnpc=1
break
#Assimilation of information using dictionaries
ifnnpc==0:
count=-1
fori,j in nltk.pos_tag(words):
count=count+1
if j=='NN' or j=='NNS':
phrase[' '.join(words)]=words[count]
break
```

VI. RESULTS

Following is an excerpt from our system taken as is –

Enter text –

the college was beautiful, Sam was running home, Sham was in class, Tom was after Jerry, the wind was blowing

Enter question –

What was beautiful?

College

Enter question –

Who was running?

Sham

Where was Sham running?

Home

Who was after Jerry?

Tom

What was the wind doing?

Blowing

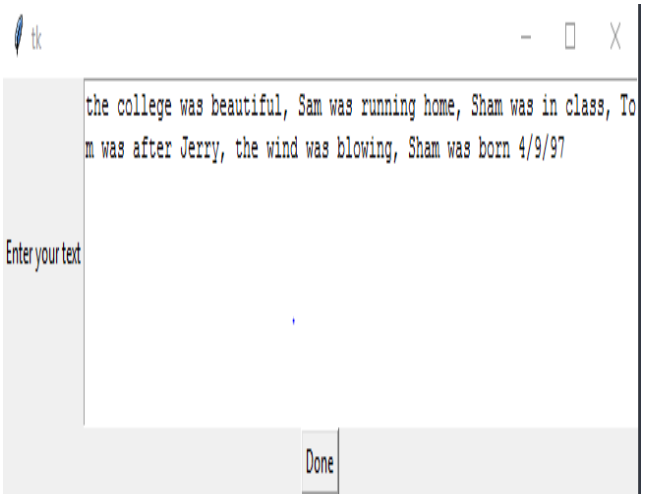

Figure 1: Entering the text

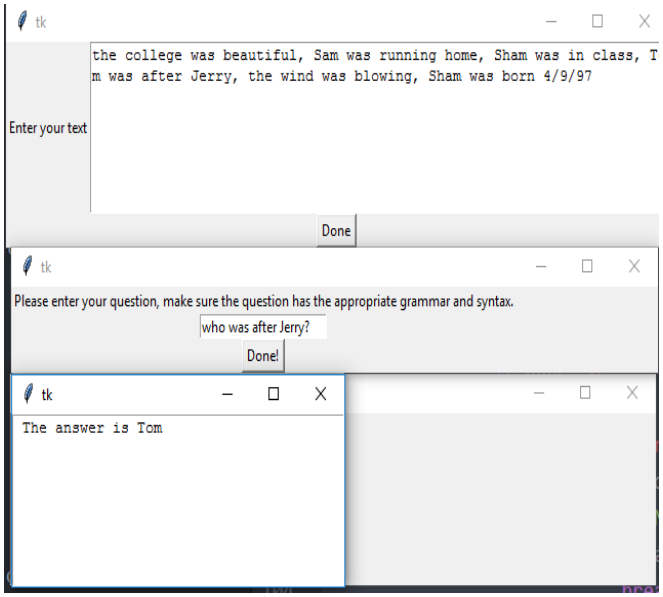

Figure 2: One example for who type question

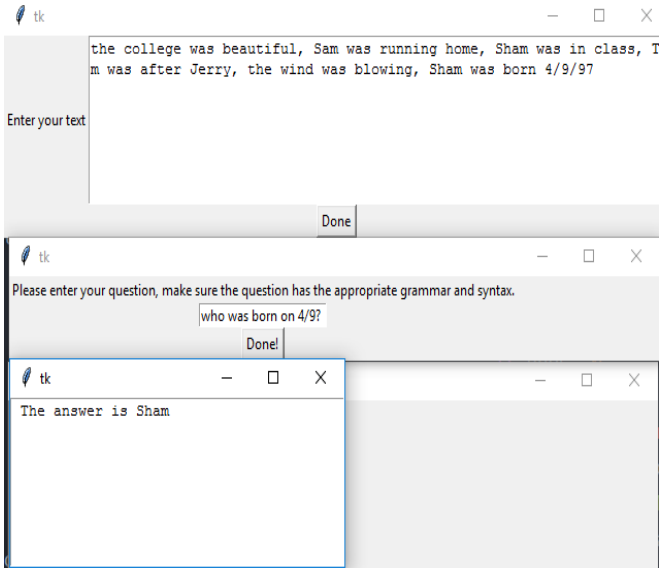

Figure 3: Another example for who type question.

VII. CONCLUSION

Our project is a start towards something truly special where natural language is more than just a byte stream for a machine. As the amount of information available online is growing, the need to access it becomes increasingly important and the value of natural language processing applications becomes clear. Machine translation helps us conquer language barriers that we often encounter by translating technical manuals, support content or catalogs at a significantly reduced cost. The challenge with machine translation technologies is not in translating words, but in understanding the meaning of sentences to provide a true translation. We hope our program can take us one step closer to tackling this problem and provide a true understanding between man and machine.

ACKNOWLEDGMET

We would like to convey our sincere thanks to School of Computing & Information Technology, Reva University, Bengaluru for their inspiring guidance and invaluable suggestions that they have given through out the research.

REFERENCES

[1] Clark P, Thompson J, and Porter B. "A knowledgebased approach to question answering". In Proceedings of AAAI"99 Fall Symposium on Question-Answering Systems, 1999, pp. 43-51.

[2] Katz B. "Annotating the World Wide Web using natural language". In Proceedings of the 5th RIAO conference on Computer Assisted Information Searching on the Internet, 1997, pp. 136-159.

[3] Chung H, Song YI, Han KS, Yoon DS, Lee JY, and Rim HC. "A practical QA System in Restricted Domains. In Workshop on Question Answering in Restricted Domains". 42nd Annual Meeting of the Association for Computational Linguistics (ACL), 2004, pp. 39-45.

[4] Cai D, Dong Y, Lv D, Zhang G, Miao X."A Web- based Chinese question answering with answer validation". In Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering, pp. 499-502, 2005.

[5] Athira P. M., Sreeja M. and P. C. Reghuraj"Architecture of an Ontology-Based Domain-Specific Natural Language Question Answering System."Department of Computer Science and Engineering, Government Engineering College, Sreekrishnapuram, Palakkad, Kerala, India, 678633.

[6] Pragisha K. "design and implementation of a QA system in Malayalam".

[7] Sanjay K Dwivedi, Vaishali Singh. "Research and reviews in question answering system" Department of Computer Science, B. B. A. University (A Central University) Luck now, Uttar Pradesh,226025,India.

[8] Poonam Gupta, Vishal Gupta Assistant Professor, Computer Science & Engineering Department University Institute of Engineering & Technology Panjab University, Chandigarh.

[9] Kolomiyets, Oleksander. And Moens, Marie- Francine. "A survey on question answering technology from an information retrieval perspective". Journal of Information Sciences 181, 2011.5412-5434. DOI: 10.1016/j.ins.2011.07.047. Elsevier.

[10] Moreda, Paloma.,Llorens Hector., Saquete, Estela. And Palomar, Manuel. "Combining semantic information in question answering systems" Journal of Information Processing and Management 47, 2011. 870- 885. DOI: 10.1016/j.ipm.2010.03.008. Elsevier.

[11] Ko, Jeongwoo., Si, Luo., and Nyberg Eric. "Combining evidence with a probabilistic framework for answer ranking and answer merging in question answering" Journal: Information Processing and Management 46, 2010 541-554. DOI: 10.1016/j.ipm.2009.11.004. Elsevier.

[12] Pachpind Priyanka P, Bornare Harshita N, Kshirsagar Rutumbhara B, Malve Ashish " An Automatic Answering System Using Template Matching For Natural Language Questions".D BE Comp S.N.D COE & RC, YEOLA.

Ms.Umma Khatuna Jannat

Student

Ms. Umma Khatuna Jannat holds M.Tech degree in Data Engineering and Cloud Computing at Reva University, Bengaluru. I have an experience on research over 5 years in Information Technology. I have published 3 journal on Digital communication. My areas of interest are Software engineering, Fuzzy logic, Machine learning, IOT, System design

Qualification:BSc , M. Tech

Dr. Nirmala S. Gupta

Associate Professor

Mrs. Nirmala S. Gupta holds M. Tech. in Computer Science and Engineering. She is pursuing her Ph. D. from REVA University. She has 16 years of teaching, 3 years of Research and 3.5 years of Industry Experience. Her areas of specialization are Simultaneous Localization and Mapping, Robotics, Data Mining and Big Data and Analytics. She has guided 2 research scholars at PG level. She has presented and published 1 research paper in international conference and 10 research papers in National conferences.

Qualification:B.E.,M.Tech.,Ph. D.

Mrs. Anitha K.

Assistant Professor

Mrs. K Anitha holds M.Tech degree in Digital Communication Engineering at MSRIT, Bengaluru. She has an experience spanning over 6 years in both industry and teaching. She has published one journal on Digital communication. Her areas of interest and teaching include Digital communication, embedded system, IOT, Electronic circuits, Logic design, Microprocessor and Microcontroller.

Qualification:B. E., M. Tech