International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 3 | Issue 3 | ISSN : 2456-3307

Question Generation System for Marathi Text

Deepali K. Gaikwad*, C. Namrata Mahender
Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, Maharashtra, India

ABSTRACT

Text summarization is process of selecting important information of source document or data and produce short summary. Manually summarize large document is very difficult. Summarization has been done in various Indian as well as non- Indian languages. But, not much work has been done for Marathi language. The present research paper, represent question generation system for Marathi text with rule based approach. The rule based approach of abstractive text summarization is used for generate question on Marathi text for this POS tagger, Named Entity Recognition and Rule based stemmer techniques are applied for generate the question. After generating questions, question are classify and then questions are rank the according to frequency or priority and answer of the ranked question is summary of given input.

Keywords : POS Tagging, Named Entity Recognition, Stemming, Rule Based Approach.

I. INTRODUCTION

Text summarization means collecting essential information from original data and present in the form of short summary. The need of summarization in various fields like Biomedical, government offices, education, social media, researchers, etc. Text summarization most of the work has been done in Indian as well as non-Indian languages. But, not much work has been done in Marathi language. Text summarization approaches can be classified into two groups: extractive summarization and abstractive summarization. Abstractive summarization consists of understanding the source text by using linguistic method to interpret the text and expressing it in own language [1, 2].

In this present research paper, text summarization using questions works as rule to extract the important aspect of the given source text. The system transforms a declarative sentence into its interrogative counterpart. The proposed method is focus to generate Question that accepts Marathi text as input and processes the input by applying POS tagging, NER, stemming and rule based approach then generate the question as per the proposed rules. For further use, generated questions are classify then rank the according to frequency or priority. The answer of the ranked question is the summary of the given input.

II. PRE-PROCESSING TOOLS

A. Part-of-Speech (POS) Tagging

Part-of-speech (POS) tagging is a process of assigning the words in a text corresponding to a particular part of speech. A fundamental version of POS tagging is the identification of words as nouns, verbs, adjectives etc. The POS tagging, approaches can be divided into three categories; rule based tagging, statistical tagging and hybrid tagging. The rule based POS tagging approach that uses a set of handcrafted rules. A stochastic approach assigns a tag to word using frequency, probability or statistics. The main drawback of rule-based system is that it fails when the text is unknown.

B. Named Entity Recognition

Named entity recognition (NER) also called entity identification and entity extraction. It is locates and classify named entities in text into predefined categories such as location, person names, organization, expression of times, quantities, monetary values, percentages, etc. Its task to identifying such named entities. NER system relies on hand written rules [3]. NER is also needed to know which named entity extracted from given text and classifies them in proper categories i.e. person name or location etc. These can be useful to generate questions from given text. [4].

C. Stemming

A stemmer can perform operation of transform morphologically identical words to root word without performing morphological analysis of that term. Stemming techniques are divided into two categories: Language Specific (Rule-Based) and Statistical (Corpus-Based) techniques. Rule-based stemming methods are further divided into three categories: Table Lookup, Affix Stripping, and Morphological. We use one of them is Affix stripping Approach for stemming.

Affix (prefix or suffix) of the word. Affix removal algorithms delete suffix and/or prefix of the word as per specific rules or suffix list. Most of the work has been done on suffix stripping as contrast to prefix. For designing rule for the stemmer required language expert and resources [5,8].

Example- "गाडीवर" (gadivar) with the help of stemmer, one can reduce the derived word into its stem. "गाडी" (gadi) is stem word of previous word [6, 7].

D. Rule Based Approach

Rules are created to generate question from given Marathi text:

1. If the noun referring to any person name is found in the given sentence, then replace it with कोण) kon, Who) word.
2. If the noun referring to any location, city, country name or organisation is found in given input sentence, then replace it with कोठे) kothe, Where) word.
3. If any date format, year, time and the list of weeks or months or the months are found in words given input sentence, then replace it with केव्हा (kevha, When) or कधी,kadhi) When(word.
4. If any cardinal or ordinal or integer or the numbers in word is found in given input sentence, then it replace with किती (kiti, How much/How any) word.
5. If noun referring to the any animal name, things or abbreviation is found in given input sentence, then it replace with काय (kay, what) word.
6. If question has been generated, then the punctuation marks "."(पूर्णविराम, full-stop) replace with **"?"** (प्रश्न चिन्ह, Question mark) [2].

III. QUESTION GENERATION SYSTEM

Automatic question generation is sub field of natural language processing. The generate question automatically in Marathi language using rule based approach as used in structured abstractive text summarization. In rule based approach handwritten rules are created according to grammatical rules of Marathi language to generate question from given sentence. The system generate the question starts with the words कोठे (Where), कोण (who), किती (how), केव्हा (When) , काय (What), etc. The system cannot design rule to generate the questions like का (why), कसा (how), because why type question are opinion based and opinion varies person to person.

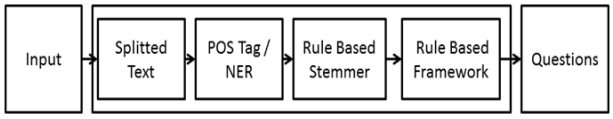

Figure 1. Question Generation System

By using above rules, for the given sentence, the system mainly tries to generate shallow questions with question words: "काय (kay, What)", "केव्हा (kevha, When)", "कोठे (kothe, Where)", "कोण / कोणी (kon / koni, Who/Whom)", "किती (kiti, How much/How many)", etc. [2, 9, 10].

IV. PROPOSED METHOD

1. Take Marathi text as input.

> एक १० वर्षाचा मुलगा आईस्क्रीमच्या दुकानात गेलातो . टेबलाजवळ बसला आणि त्याने वेटरला विचारलं, "आईस्क्रीम कोण केवढ्याला आहे?" वेटर म्हणाला, ५ रुपयेनंतर .तो मुलगा हातातील नाणी मोजू लागला . त्याने आईस्क्रीमचा लहान कप केवढ्याला, असं विचारलं . वेटरने त्रासिकपणे उत्तर दिलं, ४ रूपयेतो मुलगा . त्या मुलाने .म्हणाला मला आईस्क्रीमचा लहान कप द्या आईस्क्रीम खाल्ल, बिल दिलं आणि तो गेला.
> वेटर त्याचा रिकामा कप उचलायला गेला आणि त्याला जे दिसलं त्याने त्याचं मन हेलावून गेलं कपाजवळ . आईस्क्रीमची .रुपया त्याने टीप म्हणून ठेवला होता१ ऑर्डर देण्यापूर्वी त्या लहान मुलाने त्या वेटरचा विचार त्या मुलाने संवेदनक्षमता .केला होता आणि कदर करण्याची वृत्ती दाखवली होतीस्वतःचा विचार . मी प्रत्येक .करण्यापूर्वी त्याने इतरांचा विचार केला होता वेळी नाही म्हणतं पण कधीतरी दुसऱ्यांयचा विचार

[11].

2. Split paragraph into sentences.
3. Apply POS tagger and Named Entity recognition on each word of sentence.

Example: एक १० वर्षाचा मुलगा आईस्क्रीमच्या दुकानात गेला.

POS tag: एक//QC १०//CD वर्षाचा मुलगा//NNP आईस्क्रीमच्या//N दुकानात//NN गेला// VM. //SYM

In this sentence one noun or named entity referring to person name is identified i.e. मुलगा and one noun or named entity referring to location is identified i.e. दुकानात.

4. The name entity especially person noun i. e, is replaced with "कोण" and "गाव" is replaced with " काय ".

Input: "लोक गाव सोडून जायला सुरुवात करतात."

Output: "कोण गाव सोडून जायला सुरुवात करतात"?

Output: "लोक काय सोडून जायला सुरुवात करतात"?

Apply rule-based stemmer when some complex word is not found in dictionary. So convert this complex word into root form or stem word using rule-based stemmer.

Example: देवावर

Stemming: देव + ा + वर

5. Applied same process on each sentence of input paragraph, then it generate question for each i.e. for 'कोण', 'कोठे', 'किती', 'काय' and 'केव्हा'.

V. RESULT

In this research, from 5 different paragraph 80 sentences are collected and analysed it and generates 245 questions and these questions are related to 'कोण', 'कोठे', 'किती', 'काय' and 'केव्हा'. So the result of all this study is as follows:

> Accuracy of questions= (Correct Question / Total no. Of Question)*100

Table 2. Performance Of The System

Types of Question	Total no. of Question	Correct Question	Incorrect Question	Accuracy
कोण	89	72	17	80.89%
कोठे	27	25	02	92.59%
किती	25	24	01	96%
काय	92	87	05	94.56%
केव्हा	12	08	04	66.66%
Total	245	216	29	88.16%

VI. CONCLUSION

Text summarization extracts important information from huge data. Text summarization techniques classified into two categories: Abstractive and extractive text summarization. Summarize Marathi text is very difficult because Marathi corpus is not available. The present work on generate question for Marathi text using rule based approach.

For generating question, we used rule based approach of abstractive text summarization and rule based POS tagger, NER, rule based stemmer. First taken Marathi text as input, applied POS tag and Stemmer on it and then generate questions of wh-type i.e. कोण, कोठे, किती काय and केव्हा. The accuracy of question generation system for Marathi text is 88.16%. The generated questions are used for Marathi text summarization. For summarizing Marathi text by question, then rank the generated question according to frequency or priority. The answer of the ranked question is summary of the given input.

VII. REFERENCES

[1]. Deepali K.Gaikwad and C.Namrata Mahender.2016."A Review Paper on Text Summarization".International Journal of Advanced Research in Computer and Communication Engineering.(Mar.2016).Vol.5, Issue 3.

[2]. Shikha Grag and Vishal Goyal. 2013."System for Generating Questions Automatically From Given Punjabl Text".(2013). ITC.

[3]. Govilkar Sharvari, Bakal J.W and Rathod Shubhangi.2015."Part of Speech Tagger for Marathi Language".International Journal of Computer Applications.(2015) Volume 119-No.18.

[4]. Ramandeep Kaur and Shilpy Bansal.2014."A Review on various Techniques for Automatic Question Generation".International Journal of Emerging Trends & Technology in Computer Science (IJETTCS).(2014) Vol.3.Issue 6.

[5]. Payal Grag and Er.Charndeep Singh Bedi.2014."A review on Question Generation System Form Punjabi Text".International Journal of Emerging Trends and Technology in Computer Science (ITETTCS).(2014).

[6]. Jasmeet Singh and Vishal Gupta.2016."A Systematic Review of Text Stemming Techniques".Artif Intell Rev Springer Science Business Media Dordrecht.(2016).

[7]. Majgaonker Mudassar M.And Siddiqui Tanveer J, 2010."Discovering Suffixes : A case Study for Marathi Language". International Journal on Computer Science and Engineering.(2010).Vol.02.No.08.

[8]. Deepali K.Gaikwad, Deepali Sawane and C.Namrata Mahender. 2017."Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer" IOSR Journal of Computer Engineering (IOSR-JCE).(2017).Vol.3.pp 51-54.

[9]. Sheetal Rakangor, Y.R.Ghodasara.2015."Literature Review of Automatic Question Generation Systems".International Journal of Scientific and Research Publications.(Jan 2015) .

[10]. Deepali K.Gaikwad, Vaishali Kadam and C.Namrata Mahender."Question Based Text Summarization Under Rule Based Framework".(Presecnted in 2nd ICKE 2016, Dr.B.A.M.U, Aurangabad, MS, India).

[11]. http://www.m4marathi.net/forum/(-marathi-katha-marathi-goshti-marathi-bodh-katha)/Marathi-bodh-katha-on-opportunity-in-life.