Volume 9, No.5, May 2018

Journal of Global Research in Computer Science

REVIEW ARTICLE

Available Online at www.jgrcs.info

RULE BASED TEXT SUMMARIZATION FOR MARATHI TEXT

Deepali K. Gaikwad [*1] and Namrata Mahender C. [2]

[*1, 2] Department of Computer Science and Information Technology, Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad, Maharashtra, India

deepa.gaikwad76@gmail.com[1] nam.mah@gmail.com[2]

***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 regional languages. But, not much work has been done for Marathi language. The present research paper, represent Marathi text summarization with rule based approach using question generation system. The rule based approach of abstractive text summarization is used for generate question on Marathi text for this Rule based 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 noun priority and answer of the higher ranked and noun priority group together is summary of given input.

Keywords: Rule Based Framework; Question Generation System; Question Ranking; Answer Extraction;

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 work has been done in Indian regional languages like Hindi, Punjabi, Tamil, Telgu, Kannada, Malayalam, etc. But not much work has been done in Marathi language text summarization. Text summarization approaches can be classified into two groups: extractive summarization and abstractive summarization. Extractive summary collect important sentences from source text and group together without changing its meaning. Abstractive summarization consists of understanding the source text by using linguistic method to interpret the text and expressing it in own language [1, 2].

The most of regional language summarized text using extractive method. But extractive methods can fail to capture the relations between concepts in a text. For this reason summarize Marathi text using rule based structured abstractive summarization. The abstractive summary produced highly coherent, cohesive, rich information and less redundant summary [1, 2, 3].

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 declarative sentences 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 and noun based priority. The answer of the higher ranked question and noun based priority questions are group together which is the summary of the given input. This is the new approach of text summarization.

RULE BASED FRAME WORK

Rule based frame work contain conditional statements which are used to generate the questions from text. Accuracy of the rule based system depends upon the rules that are created according to the language morphology to generate the questions [4]. If questions properly generated, further processing leads to more precision. So it is an important unit. In rule based approach rules are created to generate the questions from the existing text written in Marathi language. For example if a Person name is found in the sentence then question with “कोण (who)” word can be generated, if location name is found in the sentence then question with “कोठे (where)” word can be generated, etc. The system is not designed to generate the question “का (why)”, “कसा (how)” etc. question due to Why? is opinion based or reason based or condition based and opinion or reason or condition varies person to person, situation to situation. A lots of variation with not much stability. 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 organization 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) 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 replace with “?” (प्रश्न चिन्ह , Question mark).

QUESTION GENERATION SYSTEM

A question is a lingual expression used to make a requirement for information and information may be provided with an answer [5]. 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 much/How any), 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.

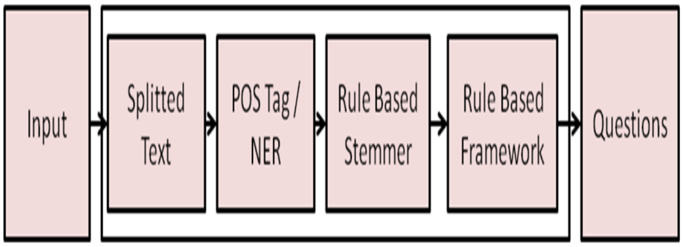

Figure 1. Question Generation System

By following above rules, for the given sentence, the system mainly tries to generate shallow questions. We observed that different types of question generated on one particular sentence.
e. g. Input Sentence:

श्री जवाहरलाल नेहरु हे भारताचे पहिले पंतप्रधान होते.

Question:

i) कोण भारताचे पहिले पंतप्रधान होते?

ii) श्री जवाहरलाल नेहरु हे भारताचे पहिले कोण होते?

iii) श्री जवाहरलाल नेहरु हे कुठले पहिले पंतप्रधान होते?

iv) श्री जवाहरलाल नेहरु हे भारताचे कितवे पंतप्रधान होते?

After generating question, classify question according to Bloom's Taxonomy. In the English language, there are six types of questions i.e. Who, Where, How Many/ How Much, What, When, Why [6].

QUESTION RANKING

Questions are generated for text summarization i.e. for collection of important content from given text. In this present research, for this purpose ranked the generated question as per priority i.e. given to the noun only and frequency of generated question i.e., different types of question occurs on one particular sentences.

ANSWER EXTRACTION

Extracted of the answer of the higher ranked question i.e. of higher frequency and question of noun priority based. The following points got noted that, Noun based priority decision is yielding appropriate information as we can see that stories too that.

SUMMARY

The answer of the higher ranked question and priority, are arranged manually to get the order of the subject of summary and finally the combined answer are the final output i.e., summarized data.

RESULT

In this research, from 5 different paragraph 118 sentences are collected and analysed it and generates 245 questions and these questions are related कोण (who), कोठे (Where), किती (how much/How any), केव्हा (When), काय (What). Out of these 216 questions are correct and 29 questions are incorrect. The precision of question generation system is 88.16% for wh- type question is shown in Table I.

Table I. Precision of the Question Generation System

Types of Question	Total no. of Question	Correct Question	Incorrect Question	Precision
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%**

For Marathi text summarization, first generate question using rule based approach and then ranked the generated question according to frequency and noun priority based. The compression ratio calculated story wise in table II.

Table II. Evaluation of Summary

Story No.	Length of Full Story (No. of sentences)	Length of Summary (No. of sentences in Summary)	Compression Ratio (CR)
I	24	18	0.75
II	22	16	0.72
III	21	17	0.80
IV	20	16	0.80

V	31	23	0.74
	118	**90**	**90/118=0.76**

The compression ratio of Marathi text summarization is 0.76.

CONCLUSION

Text summarization extracts important information from huge data. Text summarization techniques classified into two categories: extractive and abstractive text summarization. Summarize Marathi text is very difficult because Marathi corpus is not available. In this present work, targeted information is extracted by rule based text summarization question generation system for Marathi language with the help of 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 कोण (who), कोठे (Where), किती (how much/How any), केव्हा (When), काय (What). The precision 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 noun priority. The answer of the higher ranked question and noun priority based question group together is summary of the given input.

REFERENCES

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

[2] Shikha Grag and Vishal Goyal, “System for Generating Questions Automatically From Given Punjabl Text”, ITC, 2013.

[3] Atif khan and naomie salim,” A review on abstractive summarization Methods”, Journal of Theoretical and Applied Information Technology, Vol. 59 No. 1, 2014.

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

[5] Kaur Jaspreet and Bathla Ashok Kumar, “ A Review on Automatic Question Generation System from a given Hindi Text”, International Journal of Research in Computer Application and Robotics, Vol.3 Issue.6. pp. 87-92, 2015.

[6] Khillare Sunil Arun, Development of Question Answering System in Devnagari Script: In Context of Examination (Doctoral Dissertation), Ph. D. Thesis, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad, India, 2016.