IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727,
PP 51-54
www.iosrjournals.org

Rule Based Question Generation for Marathi Text Summarization using Rule Based Stemmer

Deepali K. Gaikwad[1], Deepali Sawane[2] and C. Namrata Mahender[3]
[1,2,3]Department of Computer Science and Information Technology, Dr. BAMU, Aurangabad, Maharashtra, India.

Abstract: *One of the important applications of natural language processing is text summarization. But not much work has been done in Marathi language. The current paper, represent the question based text summarization system for Marathi language with rule based stemmer. The proposed rule based stemmer is used to stem Marathi word. This technique is used for generation of the appropriate question on given input/text.*
Keywords: *Text Summarization, Natural Language processing (NLP), Stemming.*

I. Introduction

Text summarization is application of natural language processing (NLP). Text summarization is the process of extracting important information from the source text and to present that information to the user in the form of summary. In various fields text summarization is used like education field, social media (news article's, twitter, facebook massages), biomedical field, government offices, researcher, etc. [1,2]. Text summarization work has been done in many languages namely Indian Languages like Hindi, Punjabi, Tamil, Telgu, Kannada, Bangali, Malayalam etc., and Non Indian languages like English, French, Italian, Arabic, Spanish, Japanese, china, Turkish etc., are available (Garg S. et. al., 2013). 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. Abstractive text summarization divided into structured approach and semantic approach. Structured based approach extract important information from the document through such schemes as templates, extraction rules and other structures such as tree, ontology, lead and body phrase structure. And Semantic based approach, focused on semantic representation of document as well as identify noun phrase and verb phrase by linguistic data processing. The multimodal semantic model, information item based method and semantic based methods fall under semantic based approaches [1,2,3].

Text summarization using questions works as guideline to pick the important aspect of the given text. The proposed method is to generate Question that accepts Marathi text as input and processes the input by applying POS tagging NER and stemming then generate the question as per the proposed rules. The answer of the generated question is the summary of the given input but this paper limits its discussion only on Marathi stemmer for nouns. Nouns as names of person and or name of place.

Stemming technique is used to convert word into their root form which is not their base form and is not found in dictionary. It is not necessary that stem word should be similar to the root of that word [4]. For example, 'भारतीय'(Bharatiya). If the exact word is not present in the dictionary it may cause unreliable result. With the help of stemmer, one can reduce the derived word into its stem. 'भारत' (Bhart) is stem word of previous word [5].

II. Stemming

A stemmer can perform operation of converting morphologically identical words to root word without performing morphological analysis of that term. For performing stemming of terms, very less work has been done for Indian languages. But much of work can be seen in English and other non-Indian languages. A typical English stemmer reduces the words 'fishing', 'fished', and 'fisher' to the root word, 'fish'. Not many resources are available for Marathi as research is still at its early stage for Marathi.

Stemming techniques are divided into two categories: Language Specific (Rule-Based) and Statistical (Corpus-Based) techniques.

2.1 Language Specific or Rule-Based stemmer

Language Specific or Rule-Based stemmer make use of certain pre-define rule according to language to map the morphological alternative of the word to its base form. This language related rules are created manually by the linguists. The output of rule-based stemmers is quite better than statistical stemmers because they not only strip the affixes from the word but can also change the complete word (‘ate’ to ‘eat’). The creation of rule-based stemmers is very time-consuming, and it requires linguistic experts and resources such as dictionaries, stem tables, etc.

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.

Table 1. Rule Based Stemming Techniques

Rule Based Stemming Techniques	**Description**
Table lookup (brute force stemming)	- These techniques make use of a lookup table that contains the root word corresponding to the inflected or derived words. -These algorithms are also called as dictionary based algorithms. -These techniques require various language resources, and they cannot handle the words outside the dictionary.
Affix stripping algorithms	-The prefix or suffix of the word is called affix. Affix removal algorithms delete suffix and/or prefix of the word according to specific rules or suffix list. -These techniques cannot handle variations caused due to compounding, spelling variations and produce a number of errors as the words produced after stripping of affixes are sometimes not real words.
Morphological stemmers	-These stemmers take into account the morphology of the language while stemming. -The development of these algorithms requires complete knowledge of the language and its morphology.

2.2 Statistical or Corpus-Based Techniques

It based on unsupervised learning of the language by analyzing the lexicon or finding the co-occurrence or context of the words in the corpus. These are also called corpus-based techniques. These algorithms also perform suffix stripping but after performing some statistical analysis on the corpus. The major advantage of statistical techniques is that it does not require any prior knowledge of the language or language resources which are useful for many languages where the resources are either not available or are incomplete to provide effective results.

Table2. Statistical Based Stemming techniques

Statistical (Corpus-Based) Techniques	**Description**
Lexicon analysis based Stemmer	-These stemmers understand the morphology of the languages by analyzing the lexicon of the language. -Word variants are identified from the lexicon using different methods such as computation of frequencies of substrings, string distances or similarities
Corpus analysis based Stemmer	These stemmers use the context or co-occurrences statistics of the corpus words to perform stemming.

[6].

In this paper, affix stripping algorithm of rule based or language specific stemmer is used for generating questions of noun referring named entities name of person.

Table3. Marathi Stem Word Example

Suffix	**Marathi Word**	**Root Word**	**Question**
ला	रामला	राम	कोणाला
चं	रामचं	राम	कोणाचं
चा	रामचा	राम	कोणाचा
ची	रामची	राम	कोणाची
चे	रामचे	राम	कोणाचे
च्या	रामच्या	राम	कोणाच्या
साठी	पुजाऱ्यासाठी	पुजारी	कोणासाठी
सोबत	पुजाऱ्यासोबत	पुजारी	कोणासोबत
सारखी	मुलीसारखी	मुलगी	कोणासारखी
सारखा	मुलासारखा	मुलगा	कोणासारखा
सारखे	मुलासारखे	मुल	कोणासारखे
मुळे	गीतामुळे	गीता	कोणामुळे
बाबत	मुलाबाबत	मुलगा	कोणाबाबत
बरोबर	मुलाबरोबर	मुलगा	कोणाबरोबर

जवळ	मुलाजवळ	मुलगा	कोणाजवळ
कडे	भारताकडे	भारत	कोणाकडे

III. Design of Stemmer for Question Generation

1. Acquiring the given text and splitting in into sentences.
2. Splitted sentences are passed to a POS tagger for word tagged information.
3. The nouns from each tagged sentences are further classified with noun referring to person and noun referring to location, other noun formation are not consider for this work.
4. The name entity especially person noun is replace with who and location is replace with where.
5. Apply rule based stemmer on question.
6. The generated questions are validated manually.
7. The answers of generated question group together to form summary according to rank of question.

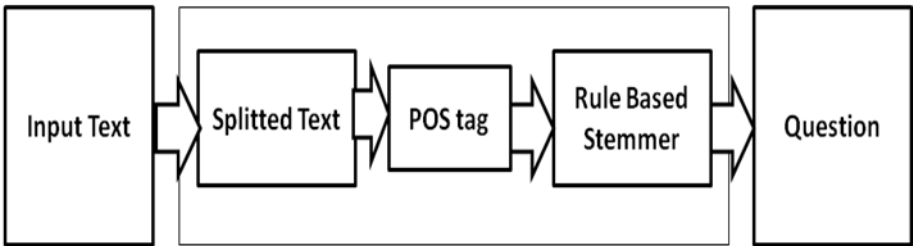

Fig. 1 Question Generation System using Stemmer

IV. Proposed Method

1. **T**ake Marathi sentence as input.

For example

“एकदा पुजाऱ्याच्या गावात पूर येतो .”

2. Apply POS tagger and NER on each word of sentence.

In this step, POS Tagger gets the noun (N/NN), pronoun (PPN), verb (VM/VX), adverb (ADV), Adjective (JJ), etc. for each word in sentence.

एकदा\\QO पुजाऱ्याच्या\\NN गावात\\NN पूर\\ NV येतो\\VM.

In this sentence noun or named entity referring to person name is identified i.e. “पुजाऱ्याच्या” and one noun or named entity referring to location is identified i.e. “गावात”.

3. If some word is not found in dictionary then stemming is used to convert word into root form. After stemming the word “पुजाऱ्याच्या” convert into “पुजारी” and “गावात” convert into “गाव”.
4. The name entity especially person noun is replaced with “कोण” and the name entity location is replaced with “कोठे”.

In above sentence noun or named entity as person name is “पुजाऱ्याच्या” is replace with “कोणाच्या” according to Marathi language grammar rule. For this we check the suffix of noun using stemming. Then the suffix of noun is added to the question like” कोण + ा + च्या” =”कोणाच्या”.

5. The output of sentence is

“एकदा कोणाच्या गावात पूर येतो”.

The answers thus generated are grouped together as paragraph and it’s the general summarized information. Yet at the present moment how to rank question, which to consider for forming the summary is not discussed but the answer of the generated question is the intermediate of original text to fully summarized text.

V. Result

In this paper, 28 sentences are collected to generate questions. The collected sentences in which found 10 noun stemmed words out of which 6 words correctly stemmed. So the result of all this study is as follows:

$$\text{Efficiency of Marathi stemmer} = \frac{\text{No. Of noun which are correctly stemmed}}{\text{Total no. Of stemmed word}} *100$$

Table 4. Stemmed word accuracy

Total no. Of stemmed word	**No. Of noun which are correctly stemmed**	**No. Of noun which are Incorrectly stemmed**
10	6	4

The Accuracy of Marathi stemmer is 60%.

VI. Conclusion

Text summarization collects important information from original document to present in the form short summary. Text summarization used in various fields like education, industry, government offices, medical, etc. techniques of text summarization is classified into two categories: Abstractive and extractive techniques. Abstractive text summarization consists of understanding the source text by using linguistic method to interpret the text and expressing it in natural language. Lots of work has been done in Hindi, Punjabi, Tamil, etc. languages. But not much work has been done in Marathi language. The present work on text summarization of Marathi text with question based system using rule based stemmer technique.

For generating question, we used rule based approach of abstractive text summarization and POS tagger, NER tools and rule based stemmer. Here Marathi text is taken as input, on it POS tagger is applied and then questions are generated for the given input as per Marathi language rules. At this stage we have framed rules of stemmer only for "कोण" type questions. Thus it can be extended to learning all Wh–type questions too.

References

[1] Atif, K. and Naomie, S., A review on abstractive summarization methods, Journal of Theoretical and Applied Information Technology, 59(1), 2014.

[2] Gaikwad D.K and C.Namrata Mahender, A Review Paper on Text Summarization, International Journal of Advanced Research in Computer and Communication Engineering. 3(5), 2016.

[3] Gupta, V. and Lehal, G.S., A survey of text summarization extractive techniques, Journal of emerging technologies in web intelligence, 2(3), 2010, pp.258-268.

[4] Gupta, V., Hindi Rule Based Stemmer for Nouns, International Journal of Advanced Research in Computer Science and Software Engineering, 4(1), 2014,pp.62-65.

[5] Mishra, U., Prakash C., MAULIK: An Effective Stemmer for Hindi Language, International Journal on Computer Science and Engineering (IJCSE), 5(4), 2012.

[6] Jasmeet Singh and Gupta, V., A systematic review of text stemming techniques, Artificial Intelligence Review, 2016, pp.1-61.

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