Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 6, Issue. 6, June 2017, pg.207 – 216

Optimizing Accuracy of Document Summarization Using Rule Mining

Poonam Kolhe[1], Prof. Ashish Kumbhare[2]

[1]Department of Computer Science & Engineering, Shri Balaji Institute of Technology & Management, Betul, RGPV University , MP
[2] Department of Computer Science & Engineering, Shri Balaji Institute of Technology & Management, Betul, RGPV University, MP
[1] Pk5758@gmail.com, [2] ashishkumbhare99@gmail.com

Abstract :- The massive quantity of data available today on the Internet has reached unforeseen volumes; thus, it is humanly unfeasible to efficiently sieve useful information from it. New information is continuously being generated. The growing accessibility of online information has necessitated intensive research in the area of automatic text summarization within the Natural Language Processing (NLP) community. Often due to time constraints we are not able to consume all the data available. Therefore it is essential to be able to summarize the text so that it becomes easier to ingest, while maintaining the essence and understandability of the information. In this paper we aim to design an algorithm that can recognize the action word by abstraction and summarize the input document by extraction and attempting to modify this extraction using a NLP tools like WordNet. Our main goal is to form a shorter version of the source document, by preserving its meaning and information content.

Keywords- Automatic Summarization, Extraction, Abstraction, NLP, WordNet.

I. Introduction

The World Wide Web has brought us a vast amount of on-line information. Due to this fact, every time someone searches something on the Internet, the response obtained is lots of different Web pages with huge information, which is impossible for a person to read completely. As the information resources in both online and offline are increasing exponentially, the major challenge is to find relevant information from large amount of data. Text summarization is an effective technique that is used in combination with Information Retrieval and Information filtering systems to save the user time.

Text summarization is the process of creating a shorter version of one or more text documents. Automatic text summarization has become an important way of finding relevant information in large text libraries or in the Internet.

The paper presents the details of propose system. Objective of proposed system is defined in section 2. Literature view of propose system is discussed in section 3. Methodology of proposed system discussed in detail in section 4. with flowchart. Propose system algorithm in section 5. Finally, section 6.concludes the paper.

II. Objective of Proposed System

The increase in the performance and fast accessing of web resources has made a new challenge of browsing among huge data on Internet. Since digitally stored information is more and more available, users need suitable tools able to select, filter, and extract only relevant information. The demand for automatic tools which are able to "understand", index, classify and present information in a clear and concise way of text documents has grown drastically in recent years. One solution to this problem is using automatic text summarization (TS) techniques.

Text summarization is extremely helpful in tackling the information overload problems. It is the technique to identify the most important pieces of information from the document, omitting irrelevant information and minimizing details to generate a compact coherent summary document[3].

To address the problem discuss above, we propose the system that will improve the efficiency, accuracy and reduce the delay to identifying the text summary by improving the data mining techniques.

Text summarization is very sophisticated so, proper input data preprocessing and document clustering is very important. So many data mining techniques are available but in or proposed work included the NLP. Text summarization is the process of creating a shorter version of one or more text documents. Automatic text summarization has become an important way of finding relevant information in large text libraries or in the Internet.

Text Summarization focuses on getting the "meaning" *Extractive* and *Abstractive*. Extractive summaries produce a set of the most significant sentences from a document, exactly as they appear. Abstractive summaries attempt to improve the coherence among sentences, and may even produce new ones.

The proposed system having following key objectives:-

- Proposed system is efficient for reduce the delay in text summarization.
- Proposed system is able to improve the accuracy of text summarization.
- Proposed system is able for text summarization of the following languages-
    - English
    - Hindi
    - Marathi

III. Literature View

In this we overview the terms and tools used in the proposed system:-

Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or user) and task (or tasks). Text summarization approaches can be broadly divided into two groups: extractive summarization and abstractive summarization[1].

A. Extractive summarization-

In Extractive Summarization system important text segments of the original text are identified and presented as they are. The summary in extractive summarization contains the words and sentences of the original text[3].

B. Abstractive summarization-

In abstractive summarization original text is interpreted and is written in a condensed form so that the resulting summary contains the essence of the original text. The summary in extractive summarization contains the words and sentences of the original text. This may not happen in abstractive summarization system. These methods have the ability to generate new sentences, which improves the focus of a summary, reduce its redundancy and keeps a good compression rate[3].

C. Automatic summarization-

It is the process of condensing textual content into a concise form for easy digestion by human, using a computer program. Summaries can be produced from a single document or multiple documents; they should be short and preserve important information.

D. NLP (Natural Language Processing)-

The main aim of Natural Language Processing (NLP) is to convert the human language in to a formal representation that easy for computer to manipulate. This is used for preprocessing the data. NLP encompasses anything a computer needs to understand natural language (typed or spoken) and also generate the natural language.

1. *Natural Language Understanding (NLU):* The NLU task is understanding and reasoning while the input is a natural language. Here we ignore the issues of natural language generation.

2. Natural Language Generation (NLG): NLG is a subfield of natural language processing NLP. NLG is also referred to text generation[6].

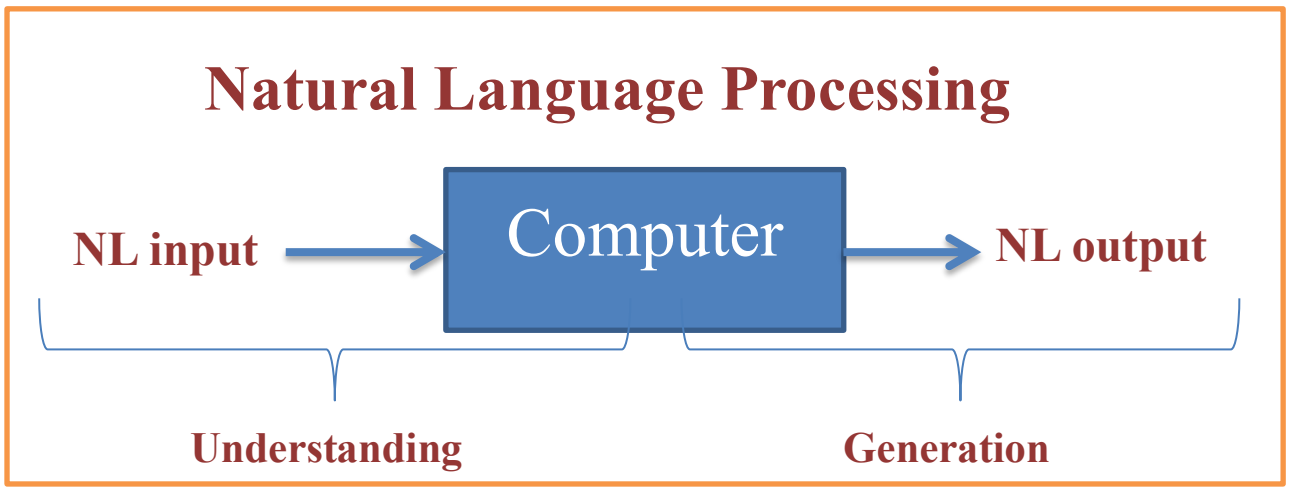

Fig.1 Process of NLP

E. R.I.-WordNet(English)

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser. WordNet is

also freely and publicly available for download. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet superficially resembles a thesaurus, in that it groups words together based on their meanings. However, there are some important distinctions. First, WordNet interlinks not just word forms—strings of letters—but specific senses of words. As a result, words that are found in close proximity to one another in the network are semantically disambiguated. Second, WordNet labels the semantic relations among words, whereas the groupings of words in a thesaurus does not follow any explicit pattern other than meaning similarity.

WordNet includes the lexical categories nouns, verbs, adjectives and adverbs but ignores prepositions, determiners and other function words. Words from the same lexical category that are roughly synonymous are grouped into synsets. Synsets include simplex words as well as collocations like "eat out" and "car pool." The different senses of a polysemous word form are assigned to different synsets[10].

WordNet does not include information about the etymology or the pronunciation of words and it contains only limited information about usage. WordNet aims to cover most of everyday English and does not include much domain-specific terminology.

WordNet has been used for a number of different purposes in information systems, including word-sense disambiguation, information retrieval, automatic text classification, automatic text summarization, machine translation and even automatic crossword puzzle generation. A common use of WordNet is to determine the similarity between words[12].

Fig.2 WordNet(English)

F. Hindi WordNet

The Hindi WordNet is a system for bringing together different lexical and semantic relations between the Hindi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Hindi WordNet is inspired by the famous English WordNet.

In the Hindi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Hindi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet. The Hindi WordNet deals with the content words, or open class category of words. Thus, the Hindi WordNet contains the following category of words- Noun, Verb, Adjective and Adverb.

Each entry in the Hindi WordNet consists of following elements:

1. Synset: It is a set of synonymous words.

2. Gloss: It describes the concept. It consists of two parts:

   a. Text definition: It explains the concept denoted by the synset.

   b. Example sentence: It gives the usage of the words in the sentence.

3. Position in Ontology: An ontology is a hierarchical organization of concepts, more specifically, a categorization of entities and actions. For each syntactic category namely noun, verb, adjective and adverb, a separate ontological hierarchy is present. Each synset is mapped into some place in the ontology. A synset may have multiple parents[11].

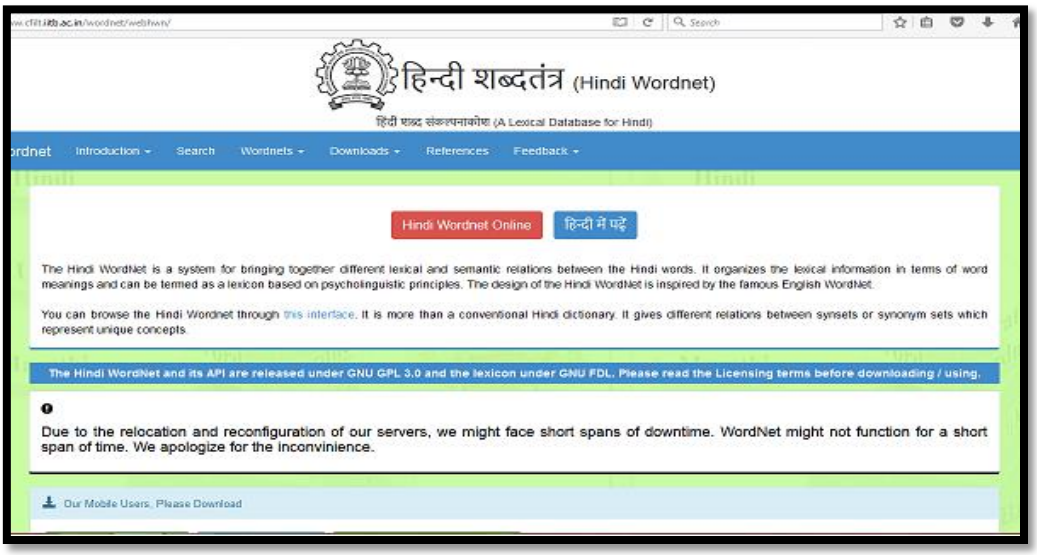

Fig.3 Hindi WordNet

G. Marathi WordNet

The Marathi WordNet is a system for bringing together different lexical and semantic relations between the Marathi words. It organizes the lexical information in terms of word meanings and can be termed as a lexicon based on psycholinguistic principles. The design of the Marathi WordNet is inspired by the famous English WordNet. In the Marathi WordNet the words are grouped together according to their similarity of meanings. Two words that can be interchanged in a context are synonymous in that context. For each word there is a synonym set, or synset, in the Marathi WordNet, representing one lexical concept. This is done to remove ambiguity in cases where a single word has multiple meanings. Synsets are the basic building blocks of WordNet.

The Marathi WordNet deals with the content words, or open class category of words. Thus, the Marathi WordNet contains the following category of words-Noun, Verb, Adjective and Adverb[7][11].

Fig.4 Marathi WordNet

IV. Methodology of Proposed System

This paper proposes a new approach to text summarization. The method ensures good coverage and avoids redundancy. In propose approach first read the document and then find the action words with the help of NLP. Then select the lines L1…….Ln, after line selection find the summary and check through propose algorithm given in next section. Final output of algorithm is our text summary. Propose work is applicable for three languages i.e. English, Hindi and Marathi.

A. Jaccard Similarity

$$\text{Jaccard's sim}(A,B)=P(A\cap B)|P(A\cup B) \text{ ---------1}$$

The determination of the association between two words with Jaccard coefficient. Jaccard index is a name often used for comparing similarity, dissimilarity, and distance of the data set. Measuring the Jaccard similarity coefficient between two data sets is the result of division between the number of features that are common to all divided by the number of properties as shown below[8].

Consider two sets A = {0,1,2,5,6} and B = {0,2,3,5,7,9}. How similar are A and B?

The Jaccard similarity is define

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \text{ ----------2}$$

$$= \frac{|\{0,2,5\}|}{|\{0,1,2,3,5,6,7,9\}|} = \frac{3}{8} = 0.375$$

More notation, given a set A, the cardinality of A denoted |A| counts how many elements are in A. The intersection between two sets A and B is denoted A ∩ B and reveals all items which are in both sets. The union between two sets A and B is denoted A ∪ B and reveals all items which are in either set.

Confirm that JS satisfies the properties of a similarity[9].

*1. With clusters***:** Another approach is to add clustering. We may have some items which basically represent the same thing. We place these represent-the-same-thing objects in clusters.

C1 = {0,1,2}, C2 = {3,4}, C3 = {5,6}, C4 = {7,8,9}

For instance, C1 might represent action movies, C2 comedies, C3 documentaries, and C4 horror movies.

Now we can represent Aclu = {C1,C3} and Bclu = {C1,C2,C3,C4} since A only contains elements from C1 and C3, while B contains elements from all clusters. The Jaccard distance of the clustered sets is now[9]

$$JS_{clu}(A,B) = JS(A_{clu},B_{clu})$$

$$= \frac{|\{C1,C2\}|}{|\{C1,C2,C3,C4\}|} = \frac{2}{4} = 0.5$$

B. Shingling-

The first option is the bag of words model, where each document is treatedas an unordered set of words. A more general approach is to shingle the document. This takes consecutive words and group them as a single object. A k-shingle is a consecutive set of k words. So the set of all 1-shingles is exactly the bag of words model. An alternative name to k-shingle is a k-gram. These mean the same thing.

D1 : I am Sam. D2 : Sam I am.

D3 : I do not like green eggs and ham. D4 : I do not like them, Sam I am.

The (k = 1)-shingles of D1∪D2∪D3∪D4 are: {[I], [am], [Sam], [do], [not], [like], [green], [eggs], [and], [ham], [them]}.

The (k = 2)-shingles of D1∪D2∪D3∪D4 are: {[I am], [am Sam], [Sam Sam], [Sam I], [am I], [I do], [do not], [not like], [like green], [green eggs], [eggs and], [and ham], [like them], [them Sam]}.

The set of k-shingles of a document with n words is at most n - k. The takes space O(kn) to store them all. If k is small, this is not a high overhead. Furthermore, the space goes down as items are repeated[9].

1.Character level: We can also create k-shingles at the character level. The (k = 3)-character shingles of D1 ∪ D2 are: {[iam], [ams], [msa], [sam], [ami], [mia]}.

The (k = 4)-character shingles ofD1∪D2 are: {[iams], [amsa], [msam], [sams], [sami], [amia], [miam]}[9].

C. Flowchart for propose system is as follows:

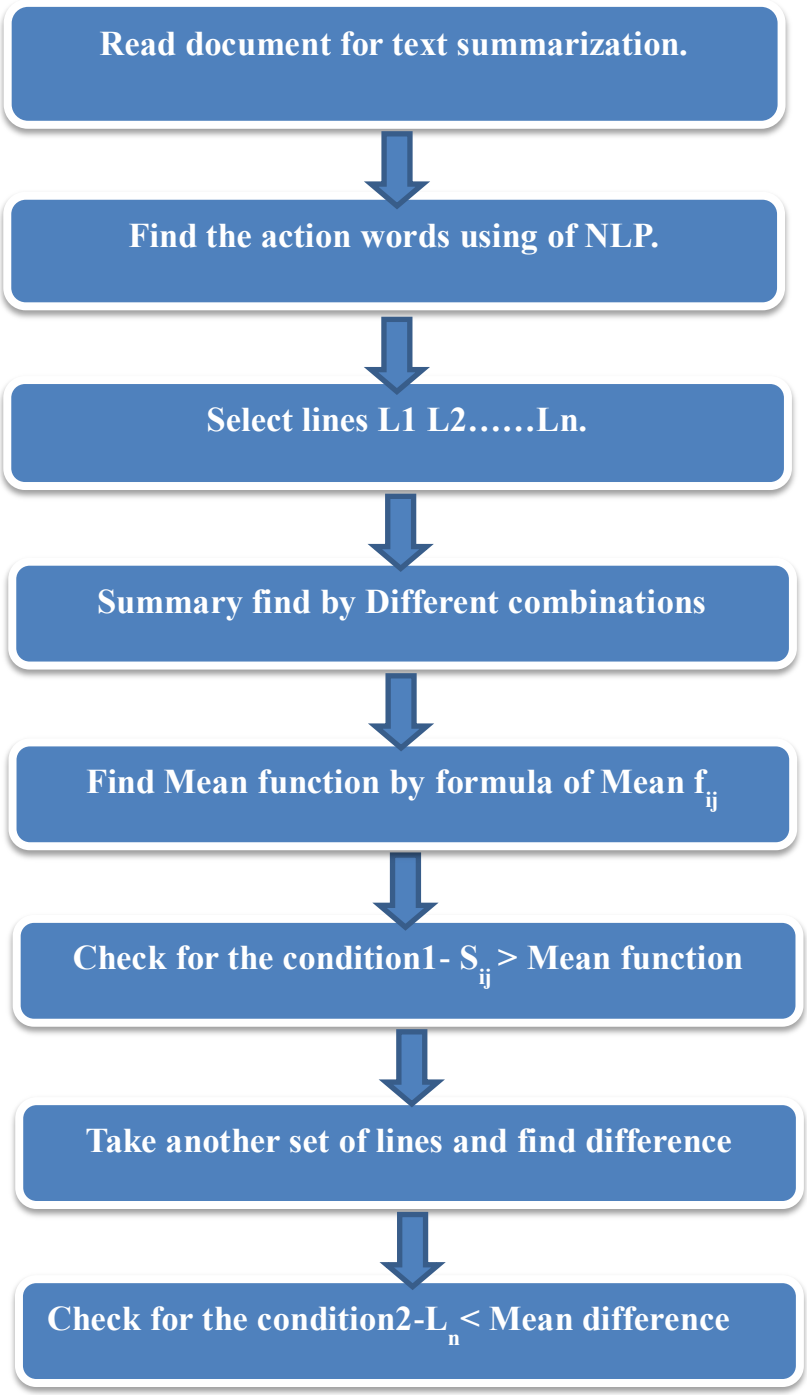

Fig.5 Flowchart for propose system

V. Algorithm for Proposed System

Algorithm for text summarization is as follows:

Step 1: Input document read for text summarization.

Step 2: Find the action words with the help of NLP.

Step 3: Select lines L1 L2……Ln.

Step 4: Suppose we have 5 lines L1…L5 then summary find by Different combinations as-

In 1st set we get-S_{12} S_{13} S_{14} S_{15}

In 2nd set we get-S_{23} S_{24} S_{25}

In 3rd set we get-S_{34} S_{35}

In 4th set we get-S_{45}

Step 5: Now we find Mean function-

Mean $f_{ij} = \sum S_{ij}$ / Total sum

Step 6: After finding mean function check for the condition1-

if S_{ij} > Mean function, then

Discard

else

document store

Step 7: Then take line L1 L3 L5. Find combination as-

L1L3 and L3L5, then we get difference by

Summary = 1- Sum = Difference

Step 8: After finding difference check for the condition2-

if L_n< Mean difference

Discard

else

output display.

VI. Conclusion

In this proposed system provides an efficient technique for Text Summarization, method ensures good coverage and avoids redundancy. Using proposed algorithm we get more efficient Text Summary. For this purpose we use the NLP as well as tools like WordNet for summarization of three different languages. This is the main advantages of this algorithm. This system should able to summarize the text of three languages i.e. English, Hindi and Marathi.

In future there is large scope for text summarization because huge number of information continuously grows on internet day after day. So, due to this large amount of data the demands for new systems for text summarization is always in scope with respect to needs of time and different changing situations in this area.

Acknowledgement

The author is thankful to Assistant Prof. Ashish Kumbhare, faculty of Computer science and Engineering, SBITM, Betul, RGPV University Bhopal, MP for providing necessary guidance to prepare this paper.

References

1) Elena Lloret "TEXT SUMMARIZATION : AN OVERVIEW " Dept. Lenguajes y Sistemas Informaticos Universidad de Alicante, Spain.
2) Nikita Munot, Sharvari S. Govilkar "Comparative Study of Text Summarization Methods" International Journal of Computer Applications (0975 – 8887) Volume 102– No.12, September 2014
3) Simran kaur1, wg.cdr anil chopra2 " Document Summarization Techniques" International Journal of Computer Science Engineering (IJCSE) Vol. 5 No.02 March2016

4) Archana AB, Sunitha. C "An Overview on Document Summarization Techniques" International Journal on Advanced Computer Theory and Engineering (IJACTE) Volume-1, Issue-2, 2013
5) George A. Miller (1995). "WordNet: A Lexical Database for English." Communications of the ACM Vol. 38, No. 11: 39-41. Christiane Fellbaum (1998, ed.) "WordNet: An Electronic Lexical Database." Cambridge, MA: MIT Press.
6) Bird, Steven, Edward Loper and Ewan Klein (2009), "Natural Language Processing with Python." O'Reilly Media Inc
7) Naik Ramesh R, C.Namrata Mahender, "Development of WordNet for Marathi Language", (NCAC'2013), 05-06 march 2013.
8) Suphakit Niwattanakul*, Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu "Using Jaccard Coefficient For Keywords similarity" Proceedings of the International MultiConference of Engineers and Computer Scientists 2013 Vol I, IMECS 2013, March 13 - 15, 2013, Hong Kong.
9) Instructor: Jeff M. Phillips, University of Utah. CS 6955 Data Mining; Spring 2013.
10) "English WorldNet," http://wordnet.princeton.edu
11) "Hindi and Marathi WordNet," http://www.cfilt.iitb.ac.in/
12) https://en.wikipedia.org/wiki/WordNet.

