Volume 5, Issue 8, August 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Word Summarization from a Paragraph Using Word Sense Disambiguation

Kirtipreet Kaur*, Er. Deepinderjeet Kaur
Department of CSE & Desh Bhagat University,
Punjab, India

Abstract— Now-a-days, with the increasing demand of the internet there is a huge volumes of information available on the internet. Users find it difficult to get information precisely. There should be a system from where users get main content instead of lengthy, redundant information. Text summarization is one of the techniques where we can get compressed version of original document. But so many problems realized in text summarizers such as redundancy, ambiguity (such as wrong spellings), and meaningless information. After having such problems in text summarization we have made a summarizer where a summary might not contain same text explicitly present in the original but still it will provide meaningful summary with no redundancy and ambiguity.

Keywords— word summarization, text summarization, word sense disambiguation, redundancy, ambiguity.

I. INTRODUCTION

Internet is widely used by people to come across information using proficient information retrieval (IR) tools, such as Google, Yahoo, AltaVista, etc., where findings are abundant. In most of the cases, users feel bore with the very tedious and time consuming job. Automatic Text Summarization will help the users to find the relevant information rapidly. Automatic Text summarization is the process of producing a condensed version of original document. This condensed version should have important content of the original document. Research is being done since many years to generate coherent and indicative summaries using different techniques.

Text summarization can be broadly classified into two types: Single document summarization and multi-document summarization. Single document summarization is the process of creating a summary from a single text document. Multi-document summarization shortens a collection of related documents into single summary. But this paper focuses on single document summarization.

The text summarization can be categorized into extractive and abstractive based on the nature of text representation in the summary. Extraction means to select the phrases or sentences having the highest score from the original text and combined to obtain the new shorter text without changing the source text. Abstraction means to probe and interpret the text by using linguistic methods. Mostly extraction method is used to produce the summary in automated text summarization system. But we are combining both techniques such as we are extracting the summary from the database and summary will not contain any original text entered as an input.

Similarly **word summarization** is also just like a text summarization in which we will give the text as a paragraph and from that text we can find the word summary using Word Sense Disambiguation technique.

Word Sense Disambiguation is a challenging technique in Natural Language Processing. There are some words in the natural languages which can cause ambiguity about the sense of the word. Those words are called polysemous words. Word sense disambiguation (WSD) is the solution to the problem. Word Sense Disambiguation is a task of automatically assigning a correct sense to the words which are polysemous in a particular context. WSD is mainly used in Information Retrieval (IR), Information Extraction (IE), Machine Translation (MT), Content Analysis, Word Processing, Lexicography and Semantic Web**.**

There are two approaches that are followed for Word Sense Disambiguation (WSD): Knowledge Based approach and Machine-Learning Based approach. In Knowledge based approach, it requires external lexical resources like Word Net, dictionary, thesaurus etc. In Machine learning-based approach, systems are trained to perform the task of word sense disambiguation. In English, Words likes Bark, Lie, book, etc. can be considered example of polysemous words. Human beings are blessed with the learning power. They can easily find out what is the correct meaning of a word in a context. But for computer it is a difficult task. So, we need to develop an automatic system which can perform like humans do i.e. the system which can find out the correct meaning of the word in particular context.

II. REVIEW OF RELATED WORK

Automated text summarization is an old eminent research area and dates back to the 1950s. As a result of the information overloading on the web there is large-scale interest in automatic text summarization during these days.

Alok Ranjan Pal, Diganta Saha [2] said that the text Summarization is the procedure by which the significant portions of a text are retrieved. Most of the approaches perform the summarization based on some hand tagged rules, such as

© 2015, IJARCSSE All Rights Reserved

format of the writing of a sentence, position of a sentence in the text, frequency of few particular words in a sentence etc. But according to different input sources, these predefined constraints greatly affect the result. The proposed approach performs the summarization task by unsupervised learning methodology. The importance of a sentence in an input text is evaluated by the help of Simplified Lesk algorithm.

Ms.Pallavi D.Patil, Prof.N.J.Kulkarni [4] presented a algorithm using fuzzy logic. In this new generation, where the tremendous information is available on the internet, it is difficult to extract the information quickly and most efficiently. There are so many text materials available on the internet, in order to extract the most relevant information from it, we need a good mechanism .This problem is solved by the Automatic Text Summarization mechanism. This paper focuses on the Fuzzy logic Extraction approach for text summarization.

A.R.Kulkarni, S.S.Apte [6] presented a better algorithm such that this paper proposes a better approach for text summarization using lexical chaining and correlation of sentences. Lexical chains are created using Wordnet. The score of each Lexical chain is calculated based on keyword strength, Tf-idf & other features. The concept of using lexical chains helps to analyze the document semantically and the concept of correlation of sentences helps to consider the relation of sentence with preceding or succeeding sentence. In this paper they discuss a summarization method, which combines lexical chaining with correlation of sentences in which relation of a sentence with the preceding sentence is considered.

Anjal R.Deshpande, Lobo L. M. R. J [8] presented a paper according to them; a summarization system consists of reduction of a text document to generate a new form which conveys the key meaning of the contained text. Due to the problem of information overload, access to sound and correctly-developed summaries is necessary. Text summarization is the most challenging task in information retrieval. Data reduction helps a user to find required information quickly without wasting time and effort in reading the whole document collection. This paper presents a combined approach to document and sentence clustering as an extractive technique of summarization.

Manisha Prabhakar, Nidhi Chandra [11] presented a paper based on pragmatic analysis .In this paper, text summarization technique is designed for the documents having the fixed format. The proposed system generates the summary of the fixed format documents by analyzing all the different parts of the documents. The system consists of five stages. In first stage each sentence is partitioned into the list of tokens and stop words are removed. In second stage, frequency usage is counted for each word. In third stage, assign POS tag for each weighted term and Word sense disambiguation is done. In the fourth stage, pragmatic analysis is performed. After Pragmatic Analysis, summarized sentences will be store in a database.

III. OBJECTIVES

1. To generate error free summary such as to detects spelling mistakes.
2. To generate a meaningful word summary with correct sense of word related to the text entered.
3. Word Summary can be a single sentence and can be produced from a single word and also from a paragraph
4. Summary should be effective.
5. Quality of a summary should be good.

IV. PROPOSED METHODOLOGY

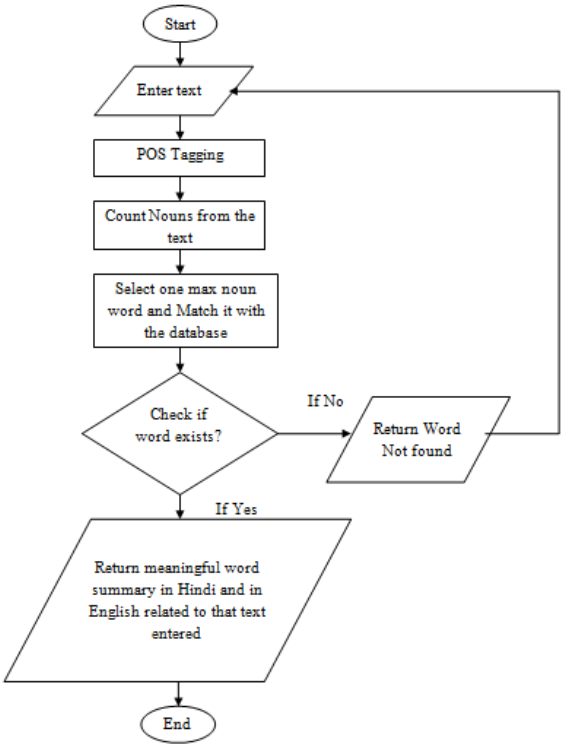

Fig. 1 : Proposed Methodology

© 2015, IJARCSSE All Rights Reserved

The following figure represents the diagram of the propose system. Proposed model has the following stages:-

1. First the user has to enter the text/paragraph to be summarized.
2. Then POS Tagging is performed.
3. By doing POS Tagging we can get the noun, adverb, adjective, verb etc.
4. From this we can count the number of times, the nouns occur. And from them we can get the maximum noun word occurred.
5. After this that noun word match with the database which can be a dictionary database and then the system will check if a word exists in the database or not. If the system finds the same word in the database, then it will give meaningful sentences which depict the proper meaning of that word related to the text entered in English as well as in Hindi language. If the system does not find a word/text being entered by the user in the database, then it will show an error called 'Word Not Found' and the user will have to again enter the particular text whose summary or meaning he/she wants to know.

V. IMPLEMENTATION AND RESULTS

In this paper we are comparing our summarizer with online summarizer (autosummarizer.com). Results are given below:-

1. Suppose we have a single word to summarize. Online summarizer will either give null summary or give same output as input.

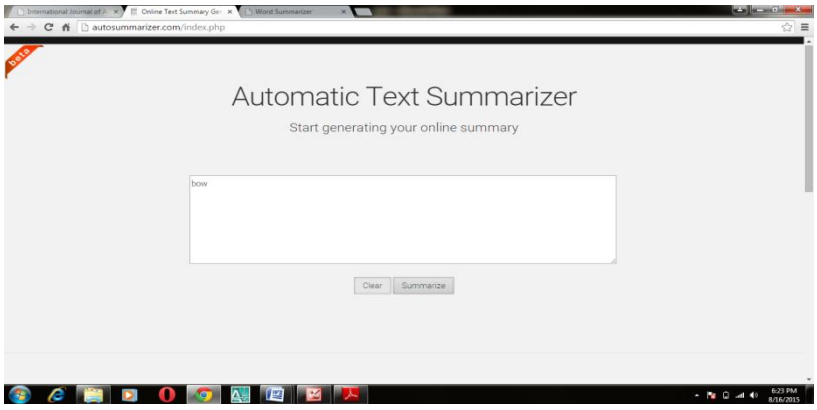

Fig. 2

Similarly our word summarizer gives meaningful word summary from a single word also which is not detected by online summarizer in Fig. 3.

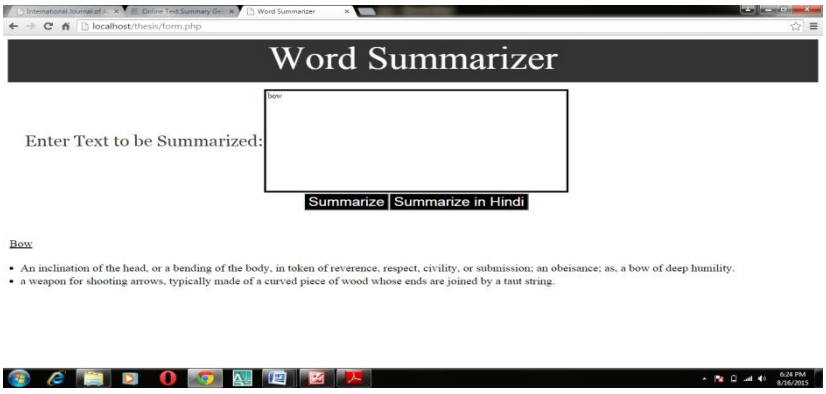

Fig. 3

2. If we are giving a text/paragraph then online summarizer give redundant, random and meaningless sentences as shown below:-

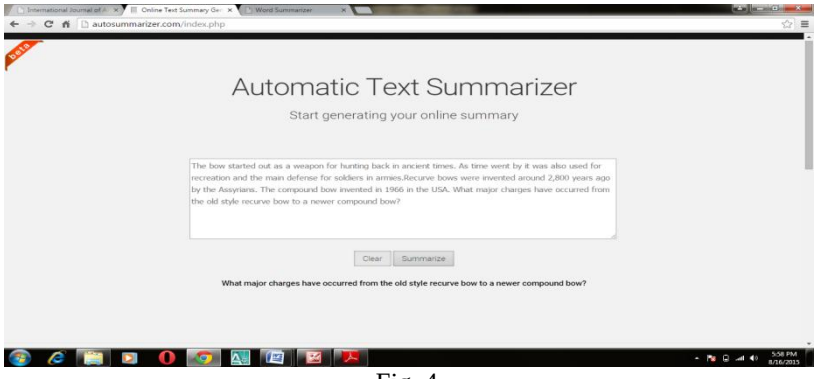

Fig. 4

© 2015, IJARCSSE All Rights Reserved

Our summarizer gives single word meaningful summary from a text/paragraph in English as well as in Hindi. As we know bow is an ambiguous word having two meanings. Our summarizer will give summary related to the meaning of a text entered which is not be detected by online summarizer. The Fig. 5 showing the summary of the first meaning of the bow.

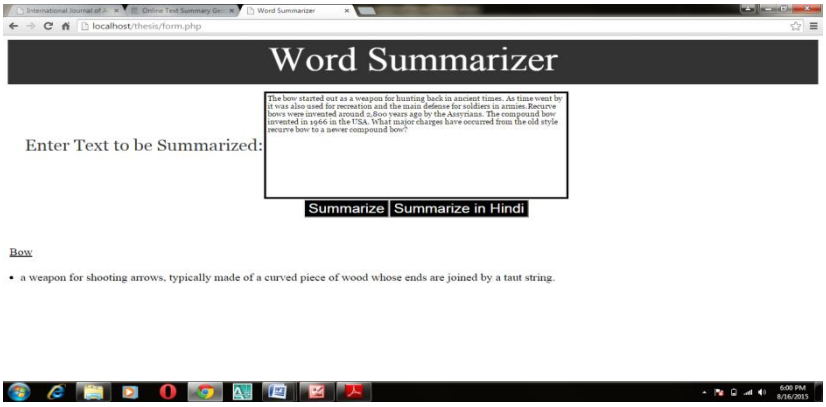

Fig. 5

The Fig.6 shows the summary of a word bow in Hindi language.

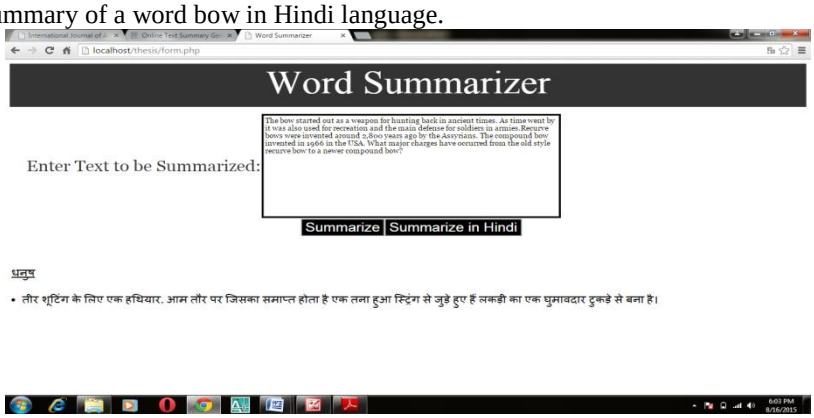

Fig. 6

The Fig.7 shows the summary of a word bow (which is its second meaning) related to the text.

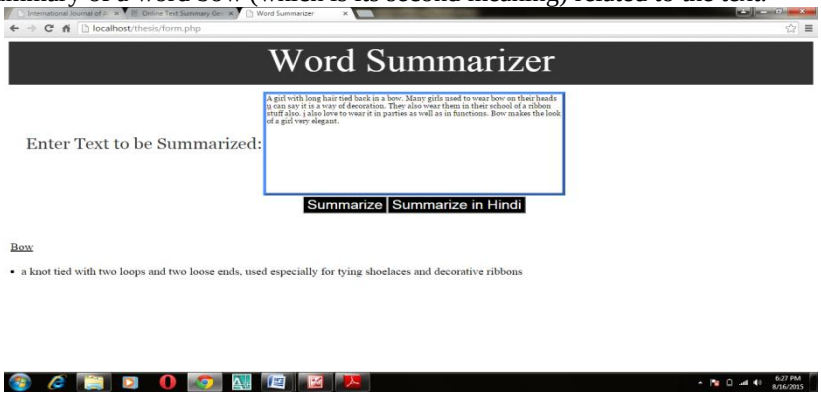

Fig. 7

3. Online summarizer also gives summary which is full of ambiguous words such as wrong spelling shown in the Fig. 8:-

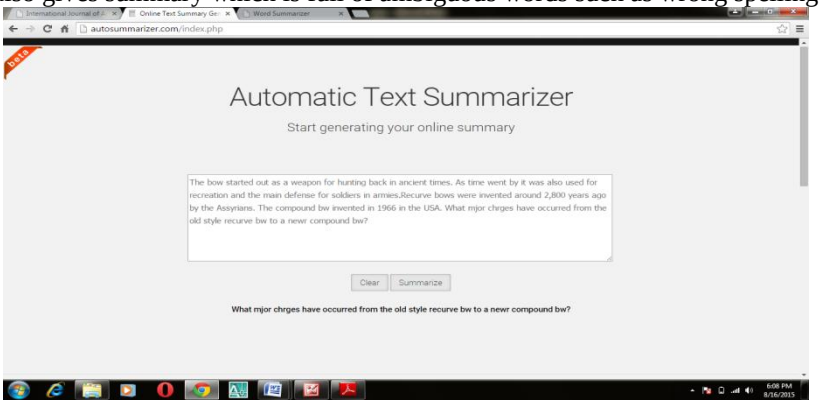

Fig. 8

© 2015, IJARCSSE All Rights Reserved

This problem is solved by our summarizer that it will not give summary with ambiguous words and if our summarizer finds any wrong spelling as shown in the Fig. 9, its result would be word not found.

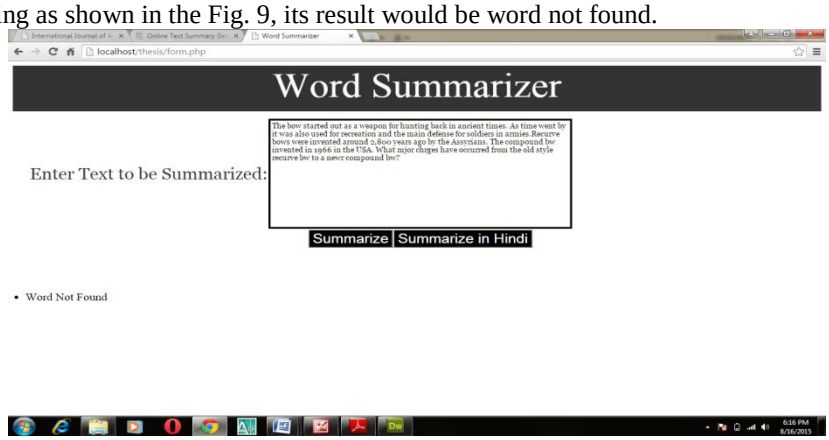

Fig. 9

VI. CONCLUSION

In this paper, we proposed a hybrid system where we have combined word summarization with knowledge based approach of word sense disambiguation, having dictionary and implemented it in PHP. As we know there are many words such as bank, bat, bow etc which have multiple meanings, users find it difficult to get the proper sense of a word related to the text entered so we have used word sense disambiguation approach, in which the meaning of the text is obtained and the word summary is generated which is a meaningful summary with proper sense of the text or paragraph entered as an input. We are also producing a word summary from a single word because many online summarizers produced the same word entered as an input to the output which is again a kind of redundancy. We also produce summary in another Natural language such as Hindi language as the language is making hindrances in the advantages of Information Technology revolution in India. So, there is the need of the adequate measures to perform natural language processing (NLP) through computer processing so that computer based system can be interacted by users through natural language like Hindi. The result shown above concludes that the system we have implemented will produce a high quality compressed summary than the other text summarizers.

REFERENCES

[1] E. Padma Lahari, D.V.N. Siva Kumar, S. Shiva Prasad , "*Automatic Text Summarization with Statistical and Linguistic Features using Successive Thresholds*", IEEE International Conference on Advanced Communication Control and Computing Technologies, ISBN No. 978-1-4799-3914-5/14 ©2014.

[2] Alok Ranjan Pal, Diganta Saha, "*An Approach to Automatic Text Summarization using WordNet*", IEEE International Conference on Advanced Communication Control and Computing Technologies, 978-1-4799-2572-8/14©2014.

[3] Vishal Patil, Mahalakshmy Krishnamoorthy, Parag Oke, Prof. M. Kiruthika, " *A Statistical Approach for Document Summarization*", International Journal of Advance Computer Technology | Vol. 2, No. 2.

[4] Ms.Pallavi D.Patil, Prof.N.J.Kulkarni, "*Text Summarization Using Fuzzy Logic*", International Journal of Innovative Research in Advanced Engineering (IJIRAE) Volume 1 Issue 3 (May 2014).

[5] D.Y. Sakhare, Dr. Raj Kumar, "*Syntactic and Sentence Feature Based Hybrid Approach for Text Summarization*", I.J. Information Technology and Computer Science, 2014, 03,Page no- 38-46

[6] A.R.Kulkarni, S.S.Apte, "*An Automatic Text Summarization using lexical cohesion and correlation of sentences*", International Journal of Research in Engineering and Technology, Volume: 03 Issue: 06 | Jun-2014.

[7] Dipti Y. Sakhare, Dr.Rajkumar, "*Neural Network Based Approach To Study The Effect Of Feature Selection On Document Summarization*", ISSN: 0975-4024 Vol. 5 No. 3 Jun-Jul 2013.

[8] Anjali R. Deshpande, Lobo L. M. R. J., "*Text Summarization using Clustering Technique*", International Journal of Engineering Trends and Technology (IJETT) - Vol.4 Issue8- August 2013.

[9] Shallu, Vishal Gupta, "*A Survey of Word-sense Disambiguation Effective Techniques and Methods for Indian Languages*", Journal of Emerging Technologies in Web Intelligence, Vol. 5, No. 4, November 2013.

[10] Mohsen Pourvali and Mohammad Saniee Abadeh, "*Automated Text Summarization Base on Lexical Chain and graph Using of WordNet and Wikipedia Knowledge Base*", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[11] Manisha Prabhakar, Nidhi Chandra, "*Automatic Text Summarization Based On Pragmatic Analysis*", International Journal of Scientific and Research Publications, Volume 2, Issue 5, May 2012.

© 2015, IJARCSSE All Rights Reserved