Deadwood Detection and Elimination in Text Summarization for Punjabi Language

Mandeep Kaur[1] and Jagroop Kaur[2]
[1]Assistant professor Baba Farid College of Engineering & Technology,Bathinda, Punjab, Pin Code-151001,India
[2]Assistant Professor University College of Engineering, Punjabi University Patiala, Punjab, Pin Code-147002,India
mandythind88@gmail.com, jagroop_80@rediffmail.com

Abstract

As the internet is growing rapidly, this has resulted in large amount of information. Text summarization provides shorthand version for such information, which is no longer than half of the original text. This paper proposes a system for detection and removal of Deadwood in summaries for Punjabi language. Deadwood means word or phrase that can be omitted without loss in meaning. Removing it shortens and clarifies the summary. The first step in this process is preprocessing which consists of sentence segmentation and removal of Punjabi stop words and then in the second step weight is assigned to the sentences in the source text .We used five different features for the assignment of weight to the sentences. In the next step the highest scoring sentences are selected to form the summary. In the last step the Deadwood is eliminated and removed from the summary.

Keywords: **Deadwood, Phrase, summary**

1. INTRODUCTION

Summarization is a very interesting and useful task that gives support to many other tasks as well as it takes the advantage of techniques developed for related natural language processing tasks. Summarization is the process of condensing a source text into a shorter version preserving its information content [1, 2]. Input to the automatic stext summarization is source text and the output is summary text. This output summary contains same information of the original text and that is no longer than half of the original text. The product of this procedure still contains the most important points of the original text. Deadwood are those unnecessary words and phrases that if omitted will cause no change in the meaning of the summary. Deadwood refers to the thing that is no longer useful or productive. These words and phrases will just increase the size of the summary and do not provide any

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**

meaning to that. Text summarization addresses both the problems of selecting the most important portions of the text and problem of generating summaries [9, 10]. In early classic summarization systems like Luhn Summarization system (1958) and Rath et al (1961) [3, 4] only a few features were used for summarization like frequency of the text. Summarization techniques can be divided into extractive and abstractive summarization. The extractive method consists of selecting the sentences from the original text. The only thing considered by this method is to decide whether a particular sentence is to be included in the summary or not [5]. The abstractive method used for summarization follows the steps used by human beings for creating the summaries. It consists of understanding the source text and re-telling the source text in few words or lesser words than the source text [6].In the proposed system we used the extractive method for summary generation for the Punjabi text and removing the Deadwood matter from the Punjabi text.

2. ROLE OF DEADWOOD WORDS AND PHRASES

Deadwood means someone or something that is unwanted and unneeded, one that is burdensome or superfluous. In text Deadwood is the material that provides no meaning to the sentence or paragraph. These are those unnecessary words and phrases that if omitted will cause no change in the meaning of the summary. Deadwood refers to the thing that is no longer useful or productive. These words and phrases will just increase the size of the summary and do not provide any meaning to that. In summarization of the Punjabi text the summary created will still contain such words and phrases that are not providing any meaning to the summary and they are just increasing the length of the summary. Such words and phrases are considered to be deadwood and they should be identified and removed from the summary. Deadwood can be classified in two ways-Word level and Phrase level.

2.1. DEADWOOD AT PHRASE LEVEL

There are a large amount of phrases that can be considered as deadwood in Punjabi Text summarization. Consider following example

ਜੇ ਤੁਸੀਂ ਚਾਹੁੰਦੇ ਹੋ ਤਾਂਮੈਂ ਤੁਹਾਡੀ ਮਦਦ ਕਰਨ ਨੂੰ ਤਿਆਰ ਹਾਂ

In the above sentence if ਜੇ ਤੁਸੀਂ ਚਾਹੁੰਦੇ ਹੋ ਤਾਂ jē tusī chande hō tāṃ is omitted then the sentence will become:

ਮੈਂ ਤੁਹਾਡੀ ਮਦਦ ਕਰਨ ਨੂੰ ਤਿਆਰ ਹਾਂ

This is completely accurate sentence in Punjabi. Omission of phrase ਜੇ ਤੁਸੀਂ ਚਾਹੁੰਦੇ ਹੋjē tusī chande hō tā has not change the meaning of the sentence.

Like the above explained phrase there are a number of phrases that can be omitted from the text without losing the meaning. Some of them are shown in the table below:

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

S.No.	Phrase
1.	ਇੰਝ ਜਾਪਦਾ ਹੈ ਕਿ
2.	ਸਾਰੀਆਂ ਚੀਜ਼ਾਂ ਨੂੰ ਧਿਆਨ ਵਿਚ ਰੱਖਦੇ ਹੋਏ
3.	ਜਿਵੇਂ ਕਿ ਕਿਹਾ ਜਾਂਦਾ ਹੈ ਕਿ
4.	ਅੰਤ ਵਿਚ ਇਹੀ ਨਤੀਜਾ ਨਿਕਲਦਾ ਹੈ ਕਿ
5.	ਸਭ ਤੋਂ ਵੱਡੀ ਗੱਲ ਇਹ ਹੈ ਕਿ
6.	ਮੈਂ ਕਹਿਣਾ ਚਾਹੁੰਦਾ ਹਾਂ ਕਿ

2.2 DEADWOOD AT WORD LEVEL

There are certain words that add bulk to the sentence. Removal of such words cause no harm to the meaning of the sentence. Some of the examples are given below in the table:

Deadwood	After removing Deadwood
ਬਿਲਕੁਲ ਠੀਕ	ਠੀਕ
ਸ਼ਵਾਲ ਪੁੱਛੇ	ਪੁੱਛੇ
ਇਸ ਕਾਰਨ ਕਰਕੇ	ਇਸ ਕਰਕੇ
ਹੁਣ ਦੇ ਸਮੇਂ ਵਿਚ	ਹੁਣ
ਕਈ ਗੁਣਾ	ਬਹੁਤ
ਬਿਲਕੁਲ ਪੂਰਾ	ਪੂਰਾ
ਇਸ ਨਤੀਜੇ ਕਰਕੇ	ਇਸ ਕਰਕੇ
ਕੱਟਿਆ ਜਾਂ ਵੱਢਿਆ	ਕੱਟਿਆ
ਬਿਲਕੁਲ ਗ਼ਲਤ	ਗ਼ਲਤ
ਸਿਰੇ ਤੋ ਖਾਰਿਜ ਕਰਨਾ	ਖਾਰਿਜ ਕਰਨਾ
ਜ਼ਿਆਦਾ ਸਪੱਸ਼ਟ	ਸਪੱਸ਼ਟ
ਜ਼ਿਆਦਾ ਤਾਕਤਵਰ	ਤਾਕਤਵਰ
ਪੂਰਾ ਯਕੀਨ	ਯਕੀਨ

3. AUTOMATIC TEXT SUMMARIZATION SYSTEM

Automatic text summarization addresses the problem of overloading. It provides the users with the most important and relevant information. The goal of this research is to implement a system that can be able to eliminate the deadwood from the Punjabi text and that can assign

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

weights to the sentences and generate summary by selecting the best sentences. Following assumptions are made for developing the system:

- Input Punjabi text will be in Unicode format.
- Single document summarization is performed.
- Summary for single theme is generated.

Punjabi paragraph is entered as an input to the system. Pre-processing is done on this document. After pre-processing the weights are assigned to the sentences. Summary is created by selecting the best sentences. Then Deadwood is eliminated from the summary.

4. PREPROCESSING OF PUNJABI TEXT

To create a summary of the Punjabi Text, First of all pre-processing is done on the input Punjabi Text before further processing [6].Pre-processing consists of two steps:

1. Sentence Segmentation:

It consists of dividing or splitting the source text into sentences.Paragraph is divided into the sentences by using end markers. The end markers used by us in the system are ?,! and [.

2. Stop word identification or elimination:

These words are frequently present in the Punjabi text they do not provide any meaning to the sentence like hY{hai} [is], □□ {je} [if] etc.

5. FEATURES FOR ASSIGNMENT OF WEIGHT TO THE SENTENCES

After the preprocessing phase weight is assigned to each sentence in the source text. Five features are used in the Punjabi Text summarization to identify the important sentences in the text. The features are identified for each sentence in the Punjabi Text [5]. The result of every feature is considered to be the score. The sum of score of all the features becomes the total score of the sentence. [7, 8,].

Total/Final score for the sentence=Score of feature 1+Score of feature 2+…….+ Score of Feature n

The final score of the sentence marks the importance of the sentence in the Punjabi Text.

5.1 Identification of Punjabi Title word:

Title of the text plays an important role in Punjabi Text summarization. Sentence containing the title word will be considered important and will be the part of the summary. Score for this feature can be calculated as:

$$S(1) = \frac{\text{No. of title words in the sentence}}{\text{No. of words in the title}}$$

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

5.2 Identification of number of words in Punjabi Sentence:

Short sentences can be detected by using this feature. The short sentences are considered less important so they must be discarded, being not the part of Summay.These sentences are given very less score and they are not included in the summary. The score for this feature can be calculated as:

$$S(2) = \frac{\text{No. of words in S}}{\text{No. of words in largest sentence}}$$

5.3 Identification of place of sentence:

The sentences at the beginning and ending of the Punjabi Text are considered to be the important sentences to be considered in the summary.So, the place of the sentences in the Punjabi Text can be identified by using this feature. The sentences at the beginning and ending have high score. If there are 7 sentences in the paragraph then the score will be 7/7 for 1st sentence,6/7 for the 2nd sentence,5/7 for the 3rd sentence,4/7 for the 4th sentence,3/7 for the 5th sentence,2/7 for the 6th sentence and 1/7 for the 7th sentence.

5.4Identification of Frequent Punjabi words:

The words which are frequently occurring in every Punjabi sentence are considered to be the important words. The Punjabi sentences containing such words are considered to be the important and provided in the summary. The sentence containing the frequent words are assigned a score 1.

5.5 *Identification of Numbers:*

There can be a number of sentences in the Punjabi text which contain the numbers. These sentences must be identified as they are considered important to be the part of the summary. The score for this feature can be calculated as:

$$S(5) = \frac{\text{No. of Numerical data in}}{\text{Sentence Length}}$$

6. SELECTION OF BEST SENTENCE

When all the sentences are assigns a final score then the best sentences are selected to create a summary. The sentences with the highest scores are considered as the best sentences. The size of the summary will be n/3, where n is the number of the sentences in the source text.

7. DEADWOOD DETECTION AND ELIMINATION SYSTEM

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

In the last phase Deadwood is detected and eliminated from the summary created. A database for Deadwood words and phrases is created. Six rules are formed for the detection and the elimination of Deadwood from the Punjabi text. The decision table below shows the rules and their effect on the Punjabi text.

Rule	Effect
If length of sentence>50	Remove the sentence form the Paragraph
If sentence is written under the quotation marks	Remove the sentence form the paragraph
If the sentence contain Deadwood word	Replace the word with the word in the database which is not Deadwood
If the sentence contains the Deadwood phrase	Remove the phrase from the sentence
If the sentence contains the combination of adjective-adjective	Remove the successor adjective from the sentence
If the sentence contains the combination of adjective-adverb	Remove the adjective from the sentence

8. Experimental Setup

8.1 DATA

The test data consists of paragraphs written in Punjabi language in Unicode font. The data will be containing different themes collected from newspapers and books i.e. it is a heterogeneous collection. The paragraphs are classifies into four types:

- Stories

This data will consist of Punjabi paragraphs containing stories.

- Numeric

This data will contain those Punjabi paragraphs that consist of numbers.

- Articles

This data will consist of those paragraphs those Punjabi paragraphs which are providing information about some particular topic.

- Biographies

This data will contain the biographies of different persons.

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

8.2 DATA STRUCTURE

Data structures used in the implementation of the system are:

- A file containing data from various sources.
- An array to hold sentences of input data.
- Tables to hold stop words, deadwood words, deadwood phrases and scores of the sentences.

8.3 TEST CASES

There are three types of test cases in the current system depending upon the choice made by the user. The three test cases can be:

4.2.3.1 Test case 1-Summary without Deadwood Elimination

4.2.3.2 Test case 2-With Deadwood elimination on the input data.

4.2.3.3 Test Case 3-With Deadwood elimination on the output summary.

9. EVALUATION AND RESULTS

The test is conducted on 50 paragraphs of each type. The table below shows the percentage of presence of Deadwood in the Punjabi Paragraph for each test case.

Type of Data	Test Case 1	Test Case 2	Test Case 3
Stories	15.15%	0%	0%
Numeric	2.63%	0%	0%
Articles	1.72%	0%	0%
Biographies	1.75%	0%	0%

10. CONCLUSION AND FUTURE WORK

There is a lot of work left behind that can be done to improve the system. Following work can be done further to increase the accuracy of the system.

1. Semantic Analysis

The system implemented through this research uses only syntax analysis not the semantic analysis of the text. If there is semantic analysis then the meaning of the sentence can be known by the system so accuracy will increase highly.

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**
JAB

2. Adjective Removal Rule

This rule is not yet implemented. With the implementation of this rule the Deadwood can be eliminated on a large scale because Punjabi text contains such combinations frequently.

11. REFERENCES

[1] Barzilay, Regina and Michael Elhadad. Using Lexical Chains for Text Summarization. In Proceedings of the Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL Madrid, 1997.

[2] H. Gregory Silber, Kathleen F. McCoy, "Efficient Text Summarization Using Lexical Chains", University of Delaware.

[3] H. P. Luhn, "The Automatic Creation of Literature Abstracts" IBMJournal of Research and Development, vol. 2, pp.159-165. 1958.

[4] G. J. Rath, A. Resnick, and T. R. Savage, "The formation of abstracts by the selection of sentences" American Documentation, vol. 12, pp.139-143.1961.

[5] Hovy, E. and Lin, C-Y. 1999, "Automated Text Summarization in SUMMARIST". I. Mani and M.T. Maybury (eds.), Advances in Automatic Text Summarization, The MIT Press, pages 81-94.

[6] Sankar K, Sobha L, "An Approach to

Text summarization", Proceedings of CLIAWS3, Third International Cross Lingual Information Access Workshop, pages 53–60, Boulder, Colorado, June 2009.

[7] Ladda Suanmali, Naomie Salim, Mohammed Salem Binwahlan, "Automatic Text Summarization using Feature Based Fuzzy Extraction", University Technology Malaysia.

[8] Ladda Suanmali, Naomie Salim, Mohammed Salem Binwahlan, "Fuzzy Logic Based Method For Improving Text Summarization", International Journal of Computer Science and Information Security, Vol. 2, No. 1, 2009.

[9] M.A Fattah and fuji Ren,"Automatic Text Summarization" In proceedings of Word Academy of Science,Engineering and Technology Volume 27.pp192-195,Feburary 2008.

[10] G. Salton,"Automatic Text Processing: The Transformation Analysis, and Retrieval of Information by Computer" Addison-wesley Publishing Company.1989.

[11] C.Y Lin," Training a selection function of extraction" In proceedings of the eight international conference on information and knowledge management, Kansas City, Missouri, United States.pp.55-62.1999.

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**

[12] Kupiec, J. Pederson, J and Chen, F (1995).”A trainable document summarizer”.In proceedings of 22nd Conference on Uncertanity in Artificial Intelligence.

[13] Evans, D.K (2005),”Similarity –Based multilingual multi-Document Summarization”.Technical Report CUCS-104-05, Columbia University.

[14] Chuang T.W, Yang J.Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms.Proc. Of the ACL-04 Workshop.Barcelon, Espana, 2004.

* * * * *

© 2013 Journal Anu Books. **Authors are responsible for any plagiarism issues.**