IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

AUTOMATIC TEXT SUMMARIZATION AND DEADWOOD REMOVAL FOR PUNJABI LANGUAGE

Mohit Kumar*, Abhinash Singla

* Student, Mtech CSE Department, Bhai Gurdas Institute of Engineering & Technology, Sangrur.
Assistant Professor, CSE Department, Bhai Gurdas Institute of Engineering & Technology, Sangrur.

ABSTRACT

Due to large information over the internet it becomes very difficult and laborious task to get the useful information. Automatic text summarization is one of the techniques which give the shorthand information which is half as the original text. This paper proposes the system to generate summary by using different features and the by applying the deadwood rules. Deadwood is the words or phrases in the sentence which can be omitted without losing its original meaning. The first step is Pre-Processing using segmentation, tokenization and stop word removal. After that deadwood rules are implemented to shorten the text and then score is calculated over the seventeen features. In last the highest scored sentences are selected to form the summary.

KEYWORDS: Automatic text summarization, Deadwood word and Phrase, Sentence Extraction.

INTRODUCTION

Now a day's internet has made people’s life so easy. But still people want everything to be done quickly and more easily so in the age of this multimedia system one of the basic needs is to get the information quickly. Automatic text summarization is one of the fastest growing techniques which shorten into summary from document losing its original meaning.

There are a number of scenarios where this automatic text summarization is useful. For example, the data retrieval system could present an automatically made summary in its list of retrieval results, from where the user decides which data or news are interesting and worth opening for a closer look—this is how Google models to some degree with the snippets shown in its search results. some other areas where ATS can be employed are news articles, Medical history of patient, Education area, Research papers, Entertainment like movies, Sports like scoreboard showing scores of game, to mobile devices as SMS, Summarization of information for government officials, businessmen, researches, etc. Text summarization tasks can be classified into single-document and multi-document Summarization. In single-document summarization, the summary of only one document is to be built, while in multi-document summarization the summary of a whole collection of documents (such as all the particular news related are searched for a query) is built. In this thesis we have experimented with single-document summaries for the text data written in the Punjabi (Gurumukhi Script) Language.

LITERATURE SURVEY

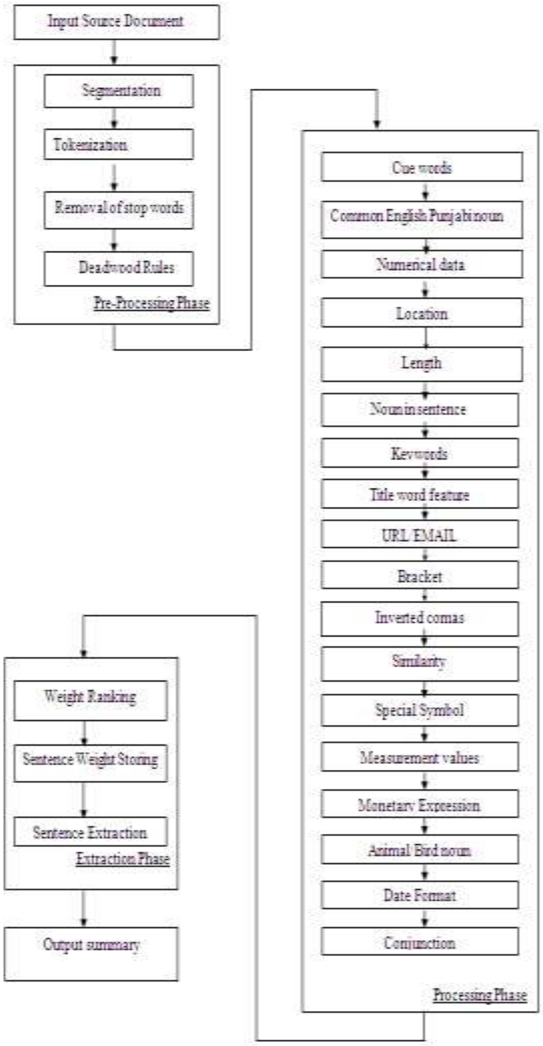
Gurmeet Singh and Karun Verma (2014) describe the new features in processing phase in Punjabi text extractive system which consist of two phases 1) Pre Processing 2) Processing. In this paper term preprocessing is defined as the phase which identifies the word boundary, sentence boundary, Punjabi stop words elimination and root word identification and in the processing phase, adds the new feature to which they are calculated and a weight is assigned to each sentence on the reference of which unwanted sentences are eliminated from the input text. It has been tested over twenty Gurmukhi Documents consists of news(achieves high rates over 90%) and for stories (achieves range above 94%) and for articles (the accuracy achieved was 82.04%,) taken randomly from Internet.

Mandeep Kaur and Jagroop Singh (2013) proposes a system in which first it will provide the summary by applying five different features. The weight is calculated for particular sentence and summary is generated. After that the deadwood rules applied on the generated summary to eliminate the deadwood words and phrases in the Punjabi text document. Deadwood means that the word or phrase which has no meaning. By omitting these words and phrases it will shorten the text length but the meaning remains same.

PROPOSED METHODOLOGY

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

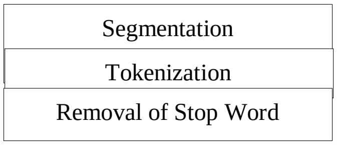
- Input Punjabi text will be in Unicode format.
- Single document summarization is performed.
- Summary for single theme is generated.



Proposed Architecture

A. Pre-processing

First of all the pre-processing on the input Punjabi text is done.it will shorten the unnecessary text or words from the sentences. Pre-processing involves:



Block diagram of Pre Processing Phase

B. Sentence Segmentation

Sentence segmentation consists of dividing or splitting the source text into sentences. The end markers used by us in this system are ?, ! and |.

C. Tokenization

Tokenization is the process of braking down the sentences into words. Sentences are tokenized by identifying the space and comma between the words. So the list of lists is created in which each list contains elements as words or also called tokens which are maintained for further processing.

D. Stop word identification or elimination

Stop words are those words which do not provide any specific meaning to the sentence. They just occupy the space rather than providing any benefit. Such words should be identified and removed, because they may increase the comparison time and may result in incorrect results. A database for such words is created which contain the list of stop words.

Table 1 Some stop words

ਦੀ	ਸੀ	ਹੋ	ਸਨ
ਨੂੰ	ਵੀ	ਉਹ	ਕੀਤੀ
ਹੈ	ਵਾਲੇ	ਉਸ	ਅਤੇ
ਹਨ	ਪਰ	ਕਰ	ਕਰਕੇ

RULES FOR IMPLEMENTATION OF DEADWOOD

After the pre-processing phase the deadwood rules has to be implemented with the survey of a large amount of data, rules are formed to eliminate Deadwood material from the paragraph. The rules for eliminating Deadwood are as follow:

Table 2 Rule of Deadwood Eliminated

Rule	**Effect**
If length of sentence>50	Remove the sentence from paragraph
If sentence is written under quotation marks	Remove the sentence from paragraph
If the sentence contain Deadwood word	Replace the word with the word in database which is not Deadwood
If the sentence contain the Deadwood phrase	Remove the phrase from the sentence
If 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

A. Deadwood words and phrases

In this research deadwood words and phrase implemented along with the adjective adjective and adjective adverb rule.

Table 3 Deadwood words

Deadwood	After removing Deadwood
ਿਬਲਕੁਲ ਠੀਕ	ਠੀਕ
ਸ਼ਵਾਲ ਪੁੱਛੋ	ਪੁੱਛੋ

ਇਸ ਕਾਰਨ ਕਰਕੇ	ਇਸ ਕਰਕੇ

Table 4 Deadwood Phrase

S.No.	Deadwood
1.	ਅੰਤ ਿਵਚ ਇਹੀ ਨਤੀਜਾ ਿਨਕਲਦਾ ਹੈ ਿਕ
2.	ਸਭ ਤ□ ਵੱਡੀ ਗੱਲ ਇਹ ਹੈ ਿਕ
3.	ਕਿਹਣਾ ਚਾਹੁੰਦਾ ਹ□ ਿਕ

Table 5 Deadwood Adjective rule

S.No.	Deadwood
1.	ਬਹੁਤ
2.	ਵੱਡਾ
3.	ਸੁੰਦਰ
4	ਮਹਿੰਗਾ

Table 6 Deadwood Adjective rule

S.No.	Deadwood
1.	ਅੱਜ
2.	ਕੱਲ੍ਹ
3.	ਅੱਜ-ਕੱਲ੍ਹ
4	ਭਲਕੇ

PROCESSING PHASE

In processing phase, feature value for every sentence is calculated. Some features in Punjabi language are different from other languages. Every sentence in the paragraph is assigned a weight according to certain features. These features are attributes that attempt to represent the data used for their task. We discover seventeen features for each sentence. Each feature is given a value or score. These features are as follows:

A. Identification of Title word (S1)

In this feature all the sentences in the Punjabi paragraph are identified which contains the title word. Then scores are identified for those sentences. The sentences containing the words which also occur in the title gives high scores.

$$\text{Score (S1)} = \frac{\text{Title words in the sentence}}{\text{No. of words in the title}}$$

B. Length (S2)

The number of words in the sentence is to be identified in this feature. The score for this feature can be calculated as the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document.

$$\text{Score(S2)} = \frac{\text{No. of words in Sentence}}{\text{No. of words in largest sentence}}$$

C. Cue words (S3)

The cue words are the index words which occur in the sentence increases the importance of the sentence. Cue words are ਨਤੀਜਾ,ਨਤੀਜੇ,ਨਿਚੋੜ,ਅੰਤਵਿੱਚ and ਸਿੱਟਾ

Score(S3) = $\underline{\text{ Cue Words in the sentence }}$ Length of the sentence

D. Common English Punjabi Noun Feature (S4)

English words are now commonly being used in Punjabi. Eg consider a Punjabi language sentence such as ਟੈਕਨਾਲੋਜੀ ਦੇ ਯੁੱਗ ਵਿੱਚ ਮੋਬਾਈਲ (Technology de yug vich mobile).This sentence contains ਟੈਕਨਾਲੋਜੀ (Technology) and ਮੋਬਾਈਲ (mobile) as English-Punjabi nouns. Also these should obviously not be coming in Punjabi dictionary.

$$\text{Score (S4)} = \frac{\text{Common English Punjabi in Sentence}}{\text{Ssentence length}}$$

E. Position (S5)

In this feature the place of the sentence is identified whether it is located at the beginning or end or middle of the paragraph. The sentences at the beginning and the end of the paragraph are given highest scores. If there are 3 sentences in the paragraph then the score for each sentence can be calculated as:

Score(S5) for 1st sentence -3/3

Score(S5) for 2nd sentence -2/3

Score(S5) for 3rd sentence -3/3

F. Numeric Data (S6)

In this feature sentences containing the numeric data are identified. Sentences containing numeric data are important and have higher probability to be extracted for the summary. E.g. ਦਿੱਲੀ ਵਿੱਚ 7 ਫਰਵਰੀ ਨੂੰ ਵਿਧਾਨ ਸਭਾ ਦੀਆਂ ਚੌਣਾ ਵਿੱਚ 70 ਸੀਟਾ ਵਿੱਚੇ ਆਪ ਨੂੰ 40 ਸੀਟਾ ਭਾਜ਼ਪਾ ਨੂੰ 25 ਅਤੇ ਕਾਂਗਰਸ ਨੂੰ 5 ਸੀਟਾ ਆਉਣ ਦਾ ਅਨੁਮਾਨ ਹੈ।In this sentence 7, 70, 40, 25 and 5 are numbers.

$$\text{Score (S6)} = \frac{\text{No. Numerical data in sentence}}{\text{Sentence Length}}$$

G. Nouns in Sentence (S7)

A noun is the name of the person, place or thing. Sentences containing nouns are important and have higher probability to be extracted for the summary. Nouns are ਮਨਵੀਰ, ਸੱਚੇਨਦਰਾ, ਯੌਗੀ, ਸਨਦੀਪ, ਮਯੂਰ, ਸ਼ਿਵਾਂਗੀ, ਚਾਰੂ

$$\text{Score (S7)} = \frac{\text{No. of Noun words in sentence}}{\text{Sentence Length}}$$

H. Presence of Brackets (S8)

Sometimes sentences may contain brackets such as ( ) parentheses, {} curly brackets etc. mostly braces contains material which could be omitted without destroying or altering sentence meaning. After doing analysis it has been found that brackets do not contain important information and has lower probability to be included for the summary.

$$\text{Score (S8)} = \frac{\text{Sentence Length - Total no of words within brackets in sentence}}{\text{Sentence Length}}$$

I. Presence of Inverted Commas (S9)

In Gurmukhi (“ ”,‘ ’) quotation marks or inverted comma surrounding quotations, direct speech, literal title or name etc. contains important information. After doing analysis it has been found that an inverted comma has higher probability to be included for the summary.

$$\text{Score (S9)} = \frac{\text{No. of words in quotation marks or inverted commas}}{\text{Sentence Length}}$$

J. Similarity between Two Sentences (S10)

Similarity between two sentences is used to determine whether two sentences are semantically equal or not. Root words are used for determining similarity between two Sentences. If two Sentences having maximum root words match then they have higher probability of being similar.

K. Presence of URL’s or Email Addresses (S11)

Internet is important and widely used application now days. Text document may have URL's or Email Addresses present in it, which provides more information about the document in process. After doing analysis of various newspapers and documents it has been found that this feature has very high importance than other and needs to be extracted for the summary.

In this case system will increase the weight of sentence by 1 for each URL or email address found in it.

L. Animal/Bird Noun (S12)

A noun may be the name of the animal and bird. Sentences containing nouns are important and have higher probability to be extracted for the summary. Nouns are ਚਿੜੀ, ਸ਼ੇਰ, ਚਿਤਾ, ਹਾਥੀ, ਮੋਰ etc.

$$\text{Score (S12)} = \frac{\text{No. of Noun words in sentence}}{\text{Sentence Length}}$$

M. Use of Special symbol (S13)

In this feature special symbols like $, %, & etc. been introduced and After doing analysis it has been found that an special symbols has higher probability to be included for the summary.

$$\text{Score (S13)} = \frac{\text{Special Symbol in sentence}}{\text{Sentence Length}}$$

N. Measurement Values (S14)

In this feature sentences containing the values like 50 km has high probability to introduce in the summary. The score for this feature can be calculated as the ratio total measurement values that occur in sentence over the sentence length. eg. 50 ਕਿਮੀ: and 10 ਲੀਟਰ

$$\text{Score (S14)} = \frac{\text{Total no. of measurement values in sentence}}{\text{Sentence Length}}$$

O. Monetary Expression (S15)

In this feature sentences containing the values like ਇਕ ਰੁਪਏਆ,ਸੌ ਡਾਲਰ, ਪੱਜਾਹ ਰੁਪੇ has high probability to introduce in the summary.

$$\text{Score (S15)} = \frac{\text{Total number of monetary expression in sentence}}{\text{Sentence Length}}$$

P. Date Format (S16)

Dates are very important for any historical documents. Presence of dates in the sentence increases the importance of the sentence. This feature is considered as an important feature for summarization of text. Rules for various types of date formats are to be developed to recognize the various dates' patterns.

$$\text{Score (S16)} = \frac{\text{No. of dates in sentence}}{\text{Sentence Length}}$$

Q. Conjunction (S17)

Conjunction is referred to as the combination of more than one sentence into a single unit. It is considered that if conjunction keywords are present in the sentence it is considered as an important sentence.

$$\text{Score (S17)} = \frac{\text{No. of conjunction keywords sentence}}{\text{Sentence Length}}$$

SENTENCE-EXTRACTION PHASE

In Sentence-Extraction phase firstly final weight of every sentence is calculated using Weight-Ranking equation. After calculating final weight of every sentence, extraction of sentences is done according to compression ratio required.

ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

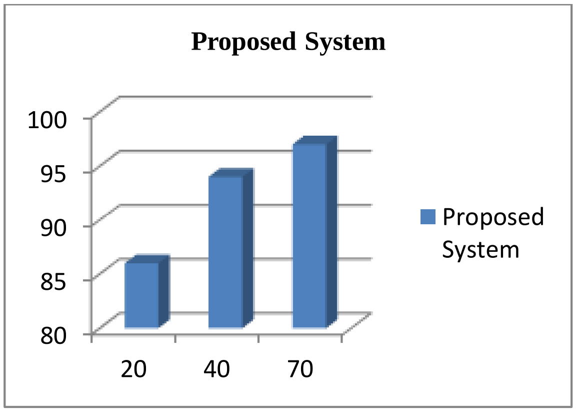
Sentence weight (Si) = S1 + S2 + S3….S17. Where Sentence Weight (*Si*) is a final weight of sentences (*Si*) and *f*1, *f*2.... *f17* are features which are computed above.

A. Selecting best sentences to include in the summary

The number of sentences to be included in the summary can be calculated as the ratio of the number of sentences in the Punjabi paragraph over the amount depends on the percentage selected by the user. The user can select sentences according to the weights to be included in the summary of the original text. Sentences to be included in the summary are those sentences which have highest scores.

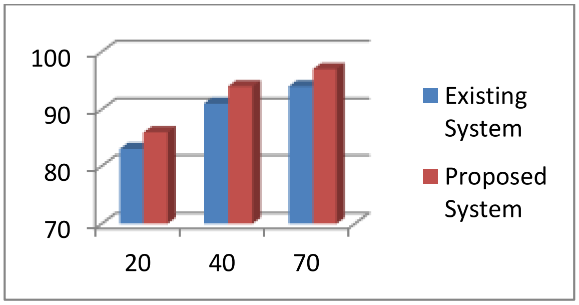
RESULTS AND DISCUSSION

The goal of our system is to get the most accurate and usefulness of the summary. The proposed system is tested on 30 documents consists of articles, news and stories which has been taken from internet randomly to evaluate the results. The proposed system has compared with the existing system and found that the proposed system has the more accuracy then the existing one as shown below.



Summary Accuracy Vs. Summary Size

It has been noticed that our system performs well and achieves high rates over 90% accurate. In case, when taken for comparison our system does really well. The compression ratio of 20% for the accuracy achieved was 86% in compare to 82.04% in news articles and for 40% and above it comes 94% in compare to 92% mark in existing system in terms of accuracy. For stories, existing system achieves highest accuracy for compression ratio 40% and above in the range above 94% and in our system it is 97%. So assuming this the proposed system is well defined with new features and the accuracy achievement is relative high then the existing system.



Summary Accuracy Vs. Summary Size of proposed system with existing system

CONCLUSION AND FUTURE SCOPE

This research work is to generate the best accurate summary for the system. As discussed, the proposed system generates the summary of the text based only on importance of the data on the basis of named entities extracted in the text paragraph. This System does not include the abstractive summary of the Punjabi text. Only Punjabi text single document generates the summary by the proposed system. In future system can be further extended by including the more deadwood rules and the semantic analysis of the text to the multi document from which the summary is to be generated. More corpus size will improve system performance. Deadwood rules can be improved by generating more corpuses. Summarization approaches can be implemented for other multimedia such as audio, video etc.

REFERENCES

[1] Gurmeet Singh and Karun Verma (2014), “A Novel Features Based Automated Gurmukhi Text Summarization System”,International Conference on Advance in Computing, Communication and Information Science, Elsevier, 2014 pp 424–432.

[2] Mandeep Kaur and Jagroop Singh (2013),”Deadwood Detection and Elimination in TextSummarization for Punjabi Language”, International Journal of Engineering Sciences, Vol. 8, Issue June 2013.

[3] Vishal Gupta and Gurpreet Singh Lehal (2013), “Automatic Text Summarization System forPunjabi Language”, Journal Of Emerging Technologies In Web Intelligence, VOL. 5, NO. 3.

[4] Vishal Gupta and Gurpreet Singh Lehal (2012), “Automatic Punjabi Text Extractive Summarization System”, Proceedings of COLING 2012, Mumbai, December 2012. pp 191–198.

[5] Ng Choon-Ching& Ali Selamat (2013), “Text Summarization Review”, http://comp.utm.my/

[6] ChetanaThaokar and Latesh Malik, “Test model for summarizing hindi text using extraction method”, In Information & Communication Technologies (ICT), 2013 IEEE Conference on, pages 1138–1143. IEEE,2013.

[7] Vishal Gupta and Gurpreet Singh Lehal (2012), “Complete Pre Processing phase of Punjabi Text Extractive Summarization System”, Proceedings of COLING 2012, Mumbai,pp199–206.

[8] R. C. Balabantaray, D. K. Sahooo, B. Sahoo and M. Swain, “Text Summarization using Term Weights’’, International Journal of Computer Applications, vol. 38, no. 1, pp. 10-14,2012.

[9] S Mangairkarasi and S Gunasundari, “Semantic based Text Summarization using Universal Networking Language”, International Journal of Applied Information Systems(IJAIS),\vol.3,No.8,2012

[10] Vishal Gupta andGurpreet Singh Lehal (2011), “Features Selection and Weight learning for PunjabiText Summarization”, International Journal of Engineering Trends and Technology- Volume2, Issue2- 2011.

[11] Vishal Gupta and Gurpreet Singh Lehal (2011), “Preprocessing Phase of Punjabi Language Text Summarization”,C. Singh et al. (Eds.): ICISIL 2011, CCIS 139, pp. 250–253,

[12] Vishal Gupta and Gurpreet Singh Lehal (2011), “ Automatic Keywords Extraction for Punjabi Language”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 5, No 3.