Features Selection and Weight learning for Punjabi Text Summarization

Vishal Gupta[#1], Gurpreet Singh Lehal[*2]

[#]*University Institute of Engineering & Technology, Panjab University*
Chandigarh,India
[1]vishal_gupta100@yahoo.co.in

[*]*Department of Computer Science, Punjabi University Patiala,*
Punjab, India

***Abstract*— This paper concentrates on features selection and weight learning for Punjabi Text Summarization. Text Summarization is condensing the source text into a shorter version preserving its information content. It is the process of selecting important sentences from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. For Punjabi language text Summarization, some of statistical features that often increase the candidacy of a sentence for inclusion in summary are: Sentence length feature, Punjabi Keywords selection feature (TF-ISF approach) and number feature. Some of linguistic features that often increase the candidacy of a sentence for inclusion in summary are: Punjabi sentence headline feature, next line feature, Punjabi noun feature, Punjabi proper noun feature, common English-Punjabi noun feature, cue phrase feature and presence of title keywords in a sentence. Mathematical regression is used to estimate the text feature weights based on fuzzy scores of sentences of 50 Punjabi news documents.**

***Keywords*— Summarization features, Statistical features, Linguistic features, Weight learning**

I. INTRODUCTION

Text summarization [1][2][3] has become an important and timely tool for assisting and interpreting text information in today's fast-growing information age. It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the Internet, however, usually the Internet provides more information than is needed. Therefore, a twofold problem is encountered: searching for relevant documents through an overwhelming number of documents available, and absorbing a large quantity of relevant information. Text Summarization is a useful tool for selecting relevant texts, and for extracting the key points of each text. The goal of automatic text summarization is to take an information source, extract content from it, and present the most important content to the user in a condensed form.

Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered. In the practice of automatic text summarization, selection-based approach [4][5] has so far been the dominant strategy. In this approach, summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency and location etc. to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favourably positioned" content. Such an approach thus avoids any efforts on deep text understanding. They are conceptually simple, easy to implement.

A good summary system should extract the diverse topics of the document while keeping redundancy to a minimum. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word's AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. A summary can be employed in an indicative way as a pointer to some parts of the original document, or in an informative way to cover all relevant information of the text. In both cases the most important advantage of using a summary is its reduced reading time. Summary generation by an automatic procedure has also other advantages: (i) the size of the summary can be controlled (ii) its content is determinist and (iii) the link between a text element in the summary and its position in the original text can be easily established.

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Producing abstractive summary is very difficult at present. It may take some time to reach a level where machines can fully understand documents. An extractive summary, in contrast, is composed with a selection of important sentences from the original text. Extractive text summarization process [6] can be divided into two steps: 1) Pre Processing step and 2) Processing step. Pre Processing is structured representation of the original text. In Processing step, features influencing the relevance of sentences are decided and calculated and then weights are assigned to these features using weight learning method. Final score of each sentence is determined using Feature-weight equation. Top ranked sentences are selected for final summary.

Some of features [4][7][8] that often increase the candidacy of a sentence for inclusion in summary are:

Keyword-occurrence: Selecting sentences with keywords that are most often used in the document usually represent theme of the document.

Title-keyword: Sentences containing words that appear in the title are also indicative of the theme of the document

Location heuristic: In Newswire articles, the first sentence is often the most important sentence; in technical articles, last couple of sentences in the abstract or those from conclusions is informative of the findings in the document.

Indicative phrases: Sentences containing key phrases like "this report …"

Short-length cutoff: Short sentences are usually not included in summary.

Upper-case word feature: Sentences containing acronyms or proper names are included.

This paper concentrates on features selection for Punjabi Text Summarization. For Punjabi Text Summarization some of statistical features that often increase the candidacy of a sentence for inclusion in summary are: Sentence length feature, Punjabi Keywords selection feature (TF-ISF approach) and number feature. Some of linguistic features that often increase the candidacy of a sentence for inclusion in summary are: Punjabi sentence headline feature, next line feature, noun feature, proper noun feature, common English-Punjabi noun feature, cue phrase feature and presence of headline keywords in a sentence.

II. EARLY HISTORY

Interest in automatic text summarization, arose as early as the fifties. An important paper of these days is the one by Luhn [9][10] in 1958, who suggested to weight the sentences of a document as a function of high frequency words, disregarding the very high frequency common words. Edmundson [10][11] in 1969 implemented a automatic text summarization system, which, additionally to the standard keyword method (i.e., frequency depending weights), also used the following three methods for determining the sentence weights:

1. Cue Method: This is based on the hypothesis that the relevance of a sentence is computed by the presence or absence of certain cue words in the cue dictionary.
2. Title Method: Here, the sentence weight is computed as a sum of all the content words appearing in the title and (sub-) headings of a text.
3. Location Method: This method is based on the assumption that sentences occurring in initial position of both text and individual paragraphs have a higher probability of being relevant. The results showed, that the best correlation between the automatic and human-made extracts was achieved using a combination of these three latter methods.

The Trainable Document Summarizer by Kupiec [10][12] in 1995 performs sentence extracting task, based on a number of weighting heuristics. Specially, the following features for sentence scoring, some of them resembling those employed by (Edmundson, 1969), were used and evaluated:

1. Sentence Length Cut-O Feature: sentences containing less than a pre-specified number of words are not included in the abstract
2. Fixed-Phrase Feature: sentences containing certain cue words and phrases are included
3. Paragraph Feature: this is basically equivalent to Edmundson's Location Method
4. Thematic Word Feature: the most frequent words are defined as thematic words. Sentence scores are functions of the thematic words' frequencies
5. Uppercase Word Feature: upper-case words (with certain obvious exceptions) are treated as thematic words, as well.

Kupiec in 1995 used a corpus which contains 188 document/summary pairs from 21 publications in a scientific/technical domain. The summaries were produced by professional experts and the sentences occurring in the summaries were aligned to the original document texts, indicating also the degree of similarity as mentioned earlier, the vast majority (about 80%) of the summary sentences could be classified as direct sentence matches.

The ANES text extraction system by Brandow [10][13] in 1995 is a system that performs automatic, domain-independent condensation of news data. The process of summary generation has four major constituents:

1. Corpus analysis: this is mainly a calculation of the tf*idf -weights for all terms
2. Statistical selection of signature words: terms with a high tf*idf-weight plus headline-words
3. Sentence weighting: summing over all signature word weights, modifying the weights by some other factors, such as relative location.
4. Sentence selection: Selecting high scored sentences.

III. FEATURE SELECTION FOR PUNJABI TEXT SUMMARIZATION

For Punjabi language Text Summarization, feature selection [5] is done based on statistical and linguistic features of sentences of Punjabi text documents. Some of statistical features that often increase the candidacy of a sentence for inclusion in summary are: Sentence length feature, Punjabi Keywords selection feature (TF-ISF approach) and number feature. Some of linguistic features that often increase the candidacy of a sentence for inclusion in summary are: Punjabi sentence headline feature, next line feature, noun feature, proper noun feature, common English-Punjabi noun feature, cue phrase feature and presence of headline keywords in a sentence.

A. Punjabi Sentence Length Feature

Those Punjabi sentences, which are very short [5], are not candidates for summary sentences. On the other hand lengthy Punjabi sentences might contain lot of information. This feature is calculated by dividing number of words in a sentence with word count of largest sentence. Its value will be always less than or equal to 1.

Punjabi Sentence Length feature Score= number of words in a sentence/ word count of largest sentence.

B. Punjabi Keywords Selection Phase

Keywords are set of significant words in a document that give high-level description of the content for investigating readers and are useful tools for many purposes. They are used in academic articles to give an insight about the article to be presented. In a magazine, they give clue about the main idea about the article so that the readers can determine whether the article is in their area of interest. For Punjabi language keywords selection, TF-ISF approach is used. The basic idea of TF-ISF [14] score is to evaluate each word in terms of its distribution over the document. Indeed, It is obvious that words occurring in many sentences within a document may not be useful for topic segmentation purposes. It is used to evaluate the importance of a word within a document based on its frequency within a given sentence and its distribution across all the sentences within the document. The TF-ISF measure of a word w in a sentence s, denoted TF-ISF(w,s), is computed by:

TF-ISF(w,s)= TF(w,s)* ISF(w) where the term frequency TF(w,s) is the number of times that word w occurs in sentence s, and the inverse sentence frequency ISF(w) is given by the formula:

ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the word w occurs. Top scored Punjabi words (Top 20%) with high value of TF-ISF scores

C. *Number Feature*

Numerical data [5] is important and it is most probably included in the document summary. The sentence that contains numerical data is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data in sentence over the sentence length.
Sentence Number feature Score= Number of numerical data in a sentence/ Length of that sentence.

D. *Punjabi Sentence Headline Feature*

Headlines are most important and always included in the final summary. In Punjabi language, headlines are usually ended with enter key or question mark (?) or exclamation sign (!). Punjabi headline detection system has been implemented with 100% accuracy. Those Punjabi sentences which belong to headlines are always given highest score.

E. *Punjabi Sentence Next line Feature*

Lines next to headlines might contain very important information related to summary. Next lines are also given more importance as compared to other features. In Punjabi Text summarization, next line is always given very high weight equal to 9. This weight has been determined using regression as weight learning method over 50 Punjabi news documents.

F. *Punjabi Noun Feature*

Those Punjabi sentences containing nouns [2] are important. Input words are checked in Punjabi noun morph for possibility of nouns. Punjabi noun morph is having 74592 noun words. Examples of Punjabi nouns are shown in TABLE I.

TABLE I

PUNJABI NOUNS LIST

Punjabi Noun	**Punjabi Noun**	**Punjabi Noun**
ਪਹੀਆ (Pahīā)	ਟੱਬਰ (ṭabbar)	ਸਿੰਗ (siṅg)
ਪਰਛਾਂਵਾਂ (parchāṃvāṃ)	ਘਰ (ghar)	ਹੱਥ (hatth) and so on…

Punjabi noun stemmer has also been implemented with efficiency 82.6%. This efficiency is calculated over 50 Punjabi news documents. Punjabi noun feature score is calculated by dividing number of Punjabi nouns in a sentence with length of that sentence. The value of this feature for a sentence will be from 0 to 1.

G. *Common English Punjabi Noun Feature*

English words are now commonly being used in Punjabi. For example consider a Punjabi language sentence such as ਟੈਕਨਾਲੋਜੀ ਦੇ ਯੁੱਗ ਵਿਚ 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. Common English-Punjabi noun words are helpful in deciding sentence importance. After analysis of Punjabi news corpus of 11.29 million words, 18245 common English-Punjabi noun words have been identified. The percentage of these Common English-Punjabi noun words in the Punjabi Corpus is about 6.44 %. Some of Common English Punjabi noun words are given in TABLE II.

TABLE II

COMMON ENGLISH PUNJABI NOUNS

Common English Punjabi Nouns	**Common English Punjabi Nouns**	**Common English Punjabi Nouns**
ਟੀਮ (team)	ਰਿਪੋਰਟ (report)	ਪ੍ਰੈੱਸ (press)
ਯੂਨੀਵਰਸਿਟੀ (university)	ਬੋਰਡ (board)	ਯੂਨੀਅਨ (union) and so on

Common English-Punjabi noun feature score is calculated by dividing number of common English-Punjabi nouns in a sentence with length of that sentence. The value of this feature for a sentence will be from 0 to 1.

H. *Punjabi Proper Noun Feature*

Proper nouns [2] [5] are the names of person, place and concept etc. not occurring in Punjabi Dictionary. Proper nouns play an important role in deciding a sentence's importance. From the Punjabi news corpus of 11.29 million words, 17598 words have been identified as proper nouns. The percentage of these proper nouns words in the Punjabi corpus is about 13.84 %. Some of Punjabi language proper names are given in TABLE III. Punjabi proper noun feature score is calculated by dividing number of Punjabi proper nouns in a sentence with length of that sentence. The value of this feature for a sentence will be from 0 to 1.

TABLE III.
PUNJABI PROPER NOUNS

Punjabi Proper noun	**Punjabi Proper noun**	**Punjabi Proper noun**
ਅਕਾਲੀ (akālī)	ਜਲੰਧਰ (jalndhar)	ਲੁਧਿਆਣਾ (ludhiāṇā)
ਬਾਦਲ (bādal)	ਮਨਪ੍ਰੀਤ (manprīt)	ਪਟਿਆਲਾ (paṭiālā) and so on

I. *Punjabi Cue Phrase Feature*

Cue Phrases [15] are certain keywords like In Conclusion, Summary and Finally etc. These are very much helpful in deciding sentence importance. Those sentences which are beginning with cue phrases or which contain these cue phrases are generally more important than others. Firstly a list of Punjabi Cue phrases has been made and then those sentences containing these Cue phrases are given more importance. Some of commonly used Punjabi cue phrases have been given in TABLE IV.

TABLE IV.
PUNJABI CUE PHRASE LIST

Punjabi Cue Phrase	**Punjabi Cue Phrase**	**Punjabi Cue Phrase**
ਨਤੀਜਾ/ਨਤੀਜੇ (natījā/natījē)	ਸਿੱਟਾ (siṭṭā)	ਉਦਾਰਨ (udāran)

ਸਿੱਟੇ ਵੱਜੇ (siṭṭē vajjō)	ਵਿਆਖਿਆ (viākhiā)	ਨਤੀਜਾ/ਨਤੀਜੇ (natījā/natījē)

Punjabi Cue phrase feature score is calculated by dividing number of cue phrases in a Punjabi sentence with length of that sentence.

J. Punjabi Title Keywords Feature

Those Punjabi sentences containing title keywords [5] are important. Title keywords are obtained after removing stop words from title line. The score of this feature is calculated by dividing number of title keywords in a sentence with length of that sentence.

IV. FEATURE WEIGHT LEARNING USING REGRESSION

Mathematical regression [5] has been used as model to estimate the text features weights for Punjabi text summarization. In this model, a mathematical function can relate output to input. The feature parameters of many manually summarized English documents are used as independent input variables and corresponding dependent outputs are specified in training phase. A relation between inputs and outputs is established. In matrix notation we can represent regression as follow:

$$\begin{bmatrix} Y_0 \\ Y_1 \\ . \\ Y_m \end{bmatrix} = \begin{bmatrix} X_{01} & X_{02} & .... & X_{010} \\ . & . & .... & . \\ . & . & .... & . \\ X_{m1} & X_{m2} & .... & X_{m10} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ . \\ w_m \end{bmatrix}$$

Where

$[Y]$ is output vector.

$[X]$ is the input matrix (feature parameters)

$[w]$ is linear statistical model of system (the weights $w_1, w_2 \ldots\ldots\ldots w_{10}$ in the equation)

m is total number of sentences in the training corpus

V. RESULTS AND CONCLUSIONS

Fifty Punjabi news documents are manually summarized by assigning fuzzy scores to the sentences of these documents and then mathematical regression [5] has been used to estimate the text features weights and then average weights are taken in the final results. Results of weight learning for different features are shown in TABLE V.

TABLE V
WEIGHT LEARNING RESULTS USING REGRESSION

Features	**Learned weights**
Sentence length feature	0.3068
Punjabi Keywords selection feature (TF-ISF approach)	0.2932
Number feature	2.5444
Punjabi sentence headline feature	10
Punjabi sentence next line feature	9
Punjabi noun feature	0.4221
Punjabi proper noun feature	0.7486
Common English-Punjabi noun feature	1.2942
Punjabi Cue phrase feature	1
Punjabi Title keywords feature	1.8

From above results, we can conclude that three most important features are Punjabi headline feature, Punjabi next line feature and Punjabi Cue phrase feature. Most of the lexical resources used such as Punjabi stemmer, Punjabi proper name list, English-Punjabi noun list etc. had to be developed from scratch as no work had been done in that direction. For developing these resources an in depth analysis of Punjabi news corpus, Punjabi dictionary and Punjabi morph had to be carried out using manual and automatic tools. This the first time some of these resources have been developed for Punjabi and they can be beneficial for developing many Natural Language Processing applications in Punjabi.

REFERENCES

[1] Karel Jezek and Josef Steinberger, Automatic Text summarization, Vaclav Snasel (Ed.): Znalosti , pp.1-12, ISBN 978-80-227-2827-0, FIIT STU Brarislava, Ustav Informatiky a softveroveho inzinierstva, 2008.

[2] Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner, Automatic Text Summarization using a Machine Learning Approach, Book: Advances in Artificial Intelligence: Lecture Notes in computer science, Springer Berlin / Heidelberg, Vol. 2507, pp205-215, 2002.

[3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, Tapping into the Power of Text Mining, Journal of ACM, Blacksburg, 2005.

[4] Fang Chen, Kesong Han and Guilin Chen, An Approach to sentence selection based text summarization, In Proceedings of IEEE TENCON02, pp489-493, 2002.

[5] Mohamed Abdel Fattah and Fuji Ren, Automatic Text Summarization, In Proceedings of World Academy of Science, Engineering and Technology, Vol. 27, pp192-195, 2008.

[6] Vishal Gupta and Gurpreet Singh Lehal, A Survey of Text Summarization Extractive Techniques, Journal of Emerging Technologies in Web Intelligence, Vol. 2, No. 3, pp258-268, 2010.

[7] Madhavi K. Ganapathiraju, Overview of summarization methods, Self-paced lab in Information Retrieval, 2002

[8] Rasim M. Alguliev and Ramiz M. Aliguliyev, Effective Summarization Method of Text Documents, in Proceedings of IEEE/WIC/ACM international conference on Web Intelligence (WI'05), pp1-8, 2005.

[9] H. P. Luhn, The Automatic Creation of Literature Abstracts, Presented at IRE National Convention, New York, pp159-165, 1958. [10] F. Samaria and S. Young, HMM based architecture for face identification, Image Vision Computing, Vol.12, No.8, pp.537–583, 1994.

[10] Klaus Zechner, A Literature Survey on Information Extraction and Text Summarization, Computational Linguistics Program, 1997.

[11] H. P. Edmundson., New methods in automatic extracting, Journal of the ACM, 16(2): pp264-285, 1969

[12] J. Kupiec, J. Pedersen, and F. Chen, A trainable document summarizer, In Proceedings of the 18th ACM-SIGIR Conference, pp68-73, 1995.

[13] Ronald Brandow, Karl Mitze, and Lisa F. Rau, Automatic condensation of electronic publications by sentence selection. Information Processing and Management, 31(5): pp675-685, 1995.

[14] Neto, Joel et al., Document Clustering and Text Summarization." In N. Mackin, editor, Proc. 4th International Conf. Practical Applications of Knowledge Discovery and Data Mining , pp41--55, London, 2000.

[15] The Corpus of Cue Phrases, http://www.cs.otago.ac.nz/staffpriv/alik/papers/apps.ps