Automatic Text Summarization System for Punjabi Language

Vishal Gupta
UIET, Panjab University, Chandigarh, India
Email: vishal@pu.ac.in

Gurpreet Singh Lehal
Department of Computer Science, Punjabi University, Patiala, India
Email: gslehal@gmail.com

Abstract— **This paper concentrates on single document multi news Punjabi extractive summarizer. Although lot of research is going on in field of multi document news summarization systems but not even a single paper was found in literature for single document multi news summarization for any language. It is first time that this system has been developed for Punjabi language and is available online at: http://pts.learnpunjabi.org/. Punjab is one of Indian states and Punjabi is its official language. Punjabi is under resourced language. Various linguistic resources for Punjabi were also developed first time as part of this project like Punjabi noun morph, Punjabi stemmer and Punjabi named entity recognition, Punjabi keywords identification, normalization of Punjabi nouns etc. A Punjabi document (like single page of Punjabi E-news paper) can have hundreds of multi news of varying length. Based on compression ratio selected by user, this system starts by extracting headlines of each news, lines just next to headlines and other important lines depending upon their importance. Selection of sentences is on the basis of statistical and linguistic features of sentences. This system comprises of two main steps: Pre Processing and Processing phase. Pre Processing phase represents the Punjabi text in structured way. In processing phase, different features deciding the importance of sentences are determined and calculated. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences. Scores of sentences are determined from sentence-feature-weight equation. Weights of features are determined using mathematical regression. Using regression, feature values of some Punjabi documents which are manually summarized are treated as independent input values and their corresponding dependent output values are provided. In the training phase, manually summaries of fifty news-documents are made by giving fuzzy scores to the sentences of those documents and then regression is applied for finding values of feature-weights and then average values of feature-weights are calculated. High scored sentences in proper order are selected for final summary. In final summary, sentences coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios. This extractive Punjabi summarizer is available online.**

Index Terms—**Punjabi text summarizer, extractive summarization, named entity recognition, keywords identification, headlines identification**

I. Introduction

Automatic text summarization [1] [2] deals with reducing the source-text into a shorter version preserving its contents and overall meaning. Generally there are two phases of text summarization [3] systems: 1) Pre-Processing-phase [4] represents the source text in structured way. 2) In Processing phase [5] [6] [7] different features deciding the importance of sentences are determined and calculated. Scores of sentences are determined using equation of feature weights and high scored sentences in proper order as of input text are extracted for final summary. This paper describes single document multi news Punjabi extractive summarizer. It is text extraction based summarization system which is used to summarize the single Punjabi document with multi news by retaining the relevant sentences based on statistical and linguistic text features. Punjab is one of Indian states and Punjabi is its official language. For Punjabi language, it is the only summarizer available as no other Punjabi summarizer exists. This summarizer is available online at: http://pts.learnpunjabi.org/ and has two phases. 1) Pre processing phase [4][13] includes finding boundary of Punjabi sentences, Elimination of Punjabi-stop-words, Stemmer for Punjabi nouns and proper names, Allowing input restrictions to input text, Elimination of duplicate sentences and normalization of Punjabi noun words in noun morph. 2) In processing phase, different features deciding the importance of sentences are determined and calculated. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification

doi:10.4304/jetwi.5.3.257-271

of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences etc. Sentence-feature-weight equation is applied for finding the final-scores of sentences. Weights of each feature are calculated using weight learning methods. Top ranked sentences in proper order are selected for final summary at selective compression ratios.

There is very complex derivational morphology for English language but not in case of Punjabi. As compared to English, Punjabi has rich system of inflectional morphology. Usually an English verb has five distinct inflectional forms. Different forms of a verb 'go' in English are go, gone, going, goes and went. But a Punjabi verb can have an average forty eight forms based on gender, tense, aspect value, number and person in any sentence. Moreover there are up to two causative forms for some of Punjabi verbs and further there will be on an average forty eight forms for each such causative form. Punjabi language is entirely different from other languages of world based on its syntax and grammar. For Punjabi summarizer, there is need to develop lexical resources for Punjabi because these resources are not available. The architectural diagram of Punjabi summarizer is given in Figure 1.

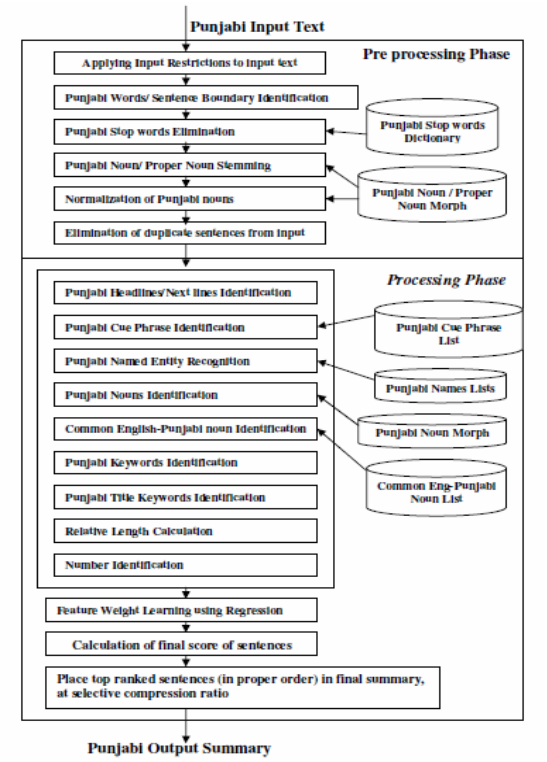

Figure 1. Overall architecture of Punjabi summarizer

II. PRE PROCESSING PHASE

Pre-Processing-phase represents the source text in structured way. Pre processing phase [4][13] includes finding boundary of Punjabi sentences, elimination of Punjabi-stop-words, stemmer for Punjabi nouns and proper names, allowing input restrictions to input text, elimination of duplicate sentences and normalization of Punjabi noun words in noun morph. The architectural diagram for Pre Processing Phase is given in Figure 2.

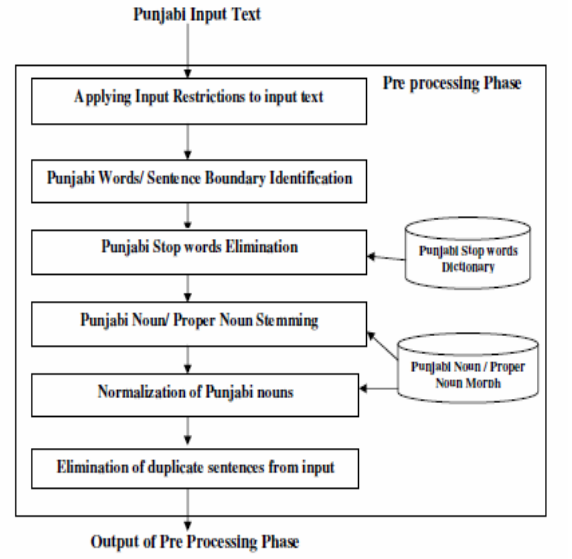

Figure 2. Pre processing phase of Punjabi summarizer

Different sub phases of pre-processing phase of Punjabi text summarization system are given below:

A. Boundary Identification Punjabi Words and sentence

From the Punjabi text, remove the punctuation mark characters like; . " " : - -- space character, tab space and so on for finding individual Punjabi words and sentence boundary is identified by presence of vertical bar ।, question mark ?, exclamation sign !, enter key, new line character etc at the end of sentence.

B. Applying Input Restrictions

Punjabi Text Summarization system allows Unicode based Gurmukhi text as input. Gurmukhi is the most common script used for writing the Punjabi language. Majority of input characters should be of Gurmukhi, otherwise error will be printed. From the input text, calculate length of Gurmukhi characters, punctuation mark characters, numeric characters, English characters and other characters. If number of Gurmukhi characters are less than equal to number of punctuation characters or number of numeric characters or number of English characters or number of other characters then error message is produced, otherwise if number of English characters or number of other characters are greater than equal to 10% of total input characters length, then error is produced "Can not accept the input!!!".

C. Punjabi Stop Words Elimination

Punjabi stop words are high frequency words appearing in Punjabi text like: ਹੈ hai "is", ਨੂੰ nūṃ "to",

ਨਾਲ nāl "with", **ਤੋਂ** tōṃ "from" and **ਦੇ** dē "of" etc. We need to delete these stop words from the source text, otherwise, those sentences which contain them may get importance unnecessarily. We have prepared Punjabi-stop-words-list by developing frequency-list from Punjabi-corpus. Punjabi corpus is taken from popular Punjabi newspaper Ajit and its thorough analysis is done. There are around 11.29 million words and 2.03 lakh unique words in this corpus. We have found 615 Punjabi stop words after analyzing the unique words of Punjabi corpus. The frequency of Punjabi stop words in corpus is 5.267 million words, which is equal to 46.64% of the corpus. Sample input and output for stop words elimination phase:

ਘਰੇਲੂ ਗੈਸ ਦੀ ਸਮੱਸਿਆ ਪਹਿਲ ਦੇ ਆਧਾਰ ਤੇ ਹੱਲ ਹੋਵੇਗੀ-ਥਿੰਦ

"Problem of domestic gas will be solved on priority basis-Thind"

In the input text **ਦੀ**, **ਦੇ**, **ਤੇ** and **ਹੋਵੇਗੀ** are Punjabi stop words. Sample output text after removing the stop words is:

ਘਰੇਲੂ ਗੈਸ ਸਮੱਸਿਆ ਪਹਿਲ ਆਧਾਰ ਹੱਲ -ਥਿੰਦ

"Problem domestic gas solved priority basis-Thind"

D. Punjabi Stemmer for Nouns/Names

The objective of any stemmer [19] [20] is to get the root of those words which are not in their basic forms and are not present in morph/dictionary. After stemming, if word is found in morph/dictionary [21], then it is correct word, otherwise it can be name or some incorrect word. In case of Punjabi stemmer [4][12][13] for nouns/ names, objective is to find root words and then root words are checked in Punjabi morph for nouns and in Punjabi names dictionary. After analyzing the Punjabi corpus, 18 suffixes were found for Punjabi nouns/names like ◌ਾਂ āṃ , ◌ਿਆਂ iāṃ, ◌ੂਆਂ ūāṃ and ◌ੀਆਂ īāṃ etc. and different rules for Punjabi noun/name stemming have been developed. Some outputs of stemmer for Punjabi nouns/names for different suffixes are:

ਲੜਕੀਆਂ laṛkīāṃ "girls" → **ਲੜਕੀ** laṛkī "girl" with suffix **◌ੀਆਂ** īāṃ, **ਮੁੰਡੇ** muṇḍē "boys" → **ਮੁੰਡਾ** muṇḍā "boy" with suffix **◌ੇ** ē, **ਫਿਰੋਜ਼ਪੁਰੋਂ** phirōzpurōṃ → **ਫਿਰੋਜ਼ਪੁਰ** phirōzpur with suffix **◌ੋਂ** ōṃ and **ਫੁੱਲਾਂ** phullāṃ "flowers" → **ਫੁੱਲ** phull "flower"with suffix **◌ਾਂ** āṃ etc.

The algorithm of Punjabi language stemmer [12] for nouns and proper names proceeds by segmenting the source Punjabi text into sentences and words. For each word of every sentence follow following steps:

Step 1: If suffix of current-Punjabi-word is **◌ਆਂ** āṃ (in case of **◌ੂਆਂ** ūāṃ, **◌ਿਆਂ** iāṃ and **◌ੀਆਂ** īāṃ), **◌ੇ** ē (in case of **◌ੀਏ** īē), **◌ੋ** ō (in case of **◌ੀਓ** Īō), **◌ਾ** ā (in case of **◌ੀਆ** ā, **◌ੀਆ** īā), **◌ਾਂ** vāṃ, **◌ੀ** ī , **◌ਾਂ** āṃ, **◌ੀਂ** īṃ, **ਜ/ਜ਼/ਸ** ja/z/s and **◌ੋਂ** ōṃ then delete the respective suffix from end and then go to Step 4.

Step 2: Else If current-word ends with **◌ੇ** ō, **◌ਿਓ** iō, **◌ੇ** ē, **◌ਿਆ** iā and **◌ਿਉਂ** iuṃ then delete the respective suffix and add kunna at the end and then go to Step 4.

Step3: Else current word is some unknown name or incorrect word.

Step 4: Stemmed Punjabi word is searched in Punjabi-noun morph/ names-dictionary. If it is found, It is Punjabi noun or Punjabi-name.

Algorithm Input: **ਮੁੰਡੇ** muṇḍē "boys" and **ਫੁੱਲਾਂ** phullāṃ "flowers"

Algorithm Output: **ਮੁੰਡਾ** muṇḍā "boy and **ਫੁੱਲ** phull "flower"

Punjabi stemming algorithm for nouns/names has been tested over fifty single-document-multi-news documents of Punjabi news corpus and its accuracy is 87.37%. This overall accuracy of Punjabi stemmer is ratio of correctly stemmed words to the total stemmed words by stemmer. Similarly the accuracy of each individual rule of stemmer is ratio of correct results under that rule to total results produced under that rule. Three types of errors can occur in case of Punjabi stemmer: 1) Dictionary errors 2) Violation of stemming-rules 3) Syntax mistakes. In case of dictionary errors, after stemming, root word is not found in Punjabi noun-morph/names dictionary, but in reality it is Punjabi noun/ proper name. In syntax errors, there is some syntax mistake while typing the Punjabi word, but actually it lies under any of stemming-rules. Overall stemming-errors, due to spelling mistakes is 0.45%, due to dictionary mistakes is 2.4% and due to rules violation is 9.78%.

Examples of errors due to rules violation are as follows:
Punjabi word **ਹਲਕੇ** halkē "light weight" is adjective and **ਬਦਲੇ** badlē "in lieu of" is adverb. These words are not found in Punjabi noun-morph/ names dictionary, but they lie under **◌ੇ** ē stemming-rule which treats them noun after stemming, but it is not true.
Examples of dictionary errors are as follows:
Some Punjabi words like **ਪ੍ਰਦੇਸਾਂ** pradēsāṃ "foreign"and **ਮੁਨਾਫ਼ਿਆਂ** munāphaiāṃ "profits" are actually nouns but are not present in noun morph or Punjabi dictionary. These words lie under **◌ਾਂ** āṃ rule and **◌ਿਆਂ** iāṃ rule of Punjabi stemmer and after performing noun stemming their root

words ਪ੍ਰਦੇਸ pradēs “foreign”and ਮੁਨਾਫ਼ਾ munāphā “profit” are also missing from Punjabi noun morph or Punjabi dictionary due to which these words are not considered as nouns by Punjabi stemmer.

Examples of syntax errors are as follows:
Some times we wrongly type the spellings of certain Punjabi noun words like ਚਿੜੀਆ chiṛīā “sparrow”and ਆਕ੍ਰਤੀ ākrtī “shape”but their correct spellings are ਚਿੜੀਆ chiṛīā “sparrow”and ਆਕ੍ਰਿਤੀ ākritī “shape” respectively, due to this these words are not found in Punjabi noun morph or Punjabi dictionary.

E. Normalization of Punjabi Nouns

This sub phase works on spelling normalization issues for Punjabi nouns, thereby resulting in multiple spelling variants for the same noun word. It is first time that Punjabi Normalizer [13] has been developed for Punjabi nouns. Problem with Punjabi is the non-standardization of Punjabi spellings. Many of the popular Punjabi noun words are written in multiple ways. For example, the Punjabi words ਚੰਡੀਗੜ੍ਹ chaṇḍīgaṛh “chandigarh”, ਪ੍ਰਕਾਸ਼ prakāsh “light”, ਜ਼ਿਲ੍ਹਾ jailhā “district” and ਖ਼ਿਆਲ khaiāl “idea” can also be written as ਚੰਡੀਗੜ chaṇḍīgaṛ “chandigarh”, ਪਰਕਾਸ਼ parkāsh “light”, ਜ਼ਿਲਾ zilā “district” ਜਿਲਾ jilā “district” ਜਿਲ੍ਹਾ jilhā “district” and ਖਿਆਲ khiāl “idea” respectively. To overcome this problem, input Punjabi text and Punjabi noun morph has been normalized for different spelling variations of Punjabi noun words. Punjabi noun morph is having 37297 noun words. The text has been normalized for the various for the various characters like ◌ੱ aadak, ◌ੰ bindi at top, Punjabi foot character ◌੍ for ਰ ra, ਵ v and ਹ ha and ◌਼ bindi at foot for ਸ਼ sha, ਖ਼ k̲ha, ਗ਼ ġa, ਜ਼ za, ਫ਼ fa, and ਲ਼ ḷa.

The algorithm for normalization of Punjabi nouns proceeds by copying noun_morph into another table noun_morph_normalized. For each noun word in table noun_morph_normalized follow the following steps:

Step 1 : Replace all the occurrences of ◌ੱaadak with null character.

Step 2 : Replace all the occurrences of ◌ੰ Bindi at top with null character.

Step 3 :Replace all the occurrences of ◌੍ Punjabi foot characters with any of suitable ਰ (ra) or ਵ (v) or ਹ (ha) characters.

Step 4 :Replace all the occurrences of ◌਼ bindi at foot with null character.

Step5:Noun_morph_normalized is now normalized

Step 6: End of algorithm

Algorithm Input: ਟੱਬ ṭabb , ਰਕਮੀਂ rakmīṃ, ਆਕ੍ਰਿਤੀ ākritī and ਖ਼ਿਆਲ khaiāl

Algorithm Output: ਟਬ ṭab, ਰਕਮੀ rkamī, ਆਕਰਿਤੀ ākritī and ਖਿਆਲ khiāl

After exhaustive analysis of Punjabi news corpus it is found that there is very less spelling variation in Punjabi nouns. Only 1.562% nouns show variation in their spellings. Out of these 1.562% words, percentage of words having one, two or three variations in the Punjabi news corpus are 99.95%, 0.046 % or 0.004% respectively. Figure 3 shows that rules for Bindi at foot and Aadak have maximum applicability. The least used rule is for Bindi at top and rule for foot characters is having usage of 22% in standardization of Punjabi nouns.

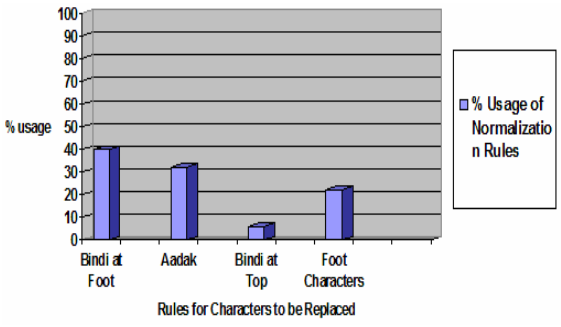

Figure 3. %Usage of normalization rules

F. Elimination of Duplicate Sentences

Duplicate sentences are the redundant sentences which need to be deleted otherwise these can get the influence unnecessarily and due to which certain other important sentences will not be displayed in the summary. In our system, duplicate sentences are deleted from input by searching the current sentence in to the sentence list which is initially empty. If current sentence is found in sentence list then that sentence is set to null otherwise it is added to the sentence list being the unique sentence. This elimination prevents duplicate sentences from appearing in final summary. An exhaustive analysis has been done on fifty Punjabi multi news documents for determining the frequency of duplicate sentences and it is discovered 9.60% sentences are duplicate. Average frequency of a duplicate sentence in a Punjabi document is three and maximum frequency is four. Out of 9.6% duplicate sentences from fifty Punjabi news documents, there are 5.4% sentences with minimum frequency two, 2.29% sentences with average frequency three and 1.91% sentences with maximum frequency four.

III. PROCESSING PHASE

In processing phase [22], different features deciding the importance of sentences are determined and calculated. Feature-weight equation is applied for finding the final-scores of sentences. Weights of each feature are

calculated using regression based weight learning method. Figure 4 describes the processing-phase of Punjabi text summarizer.

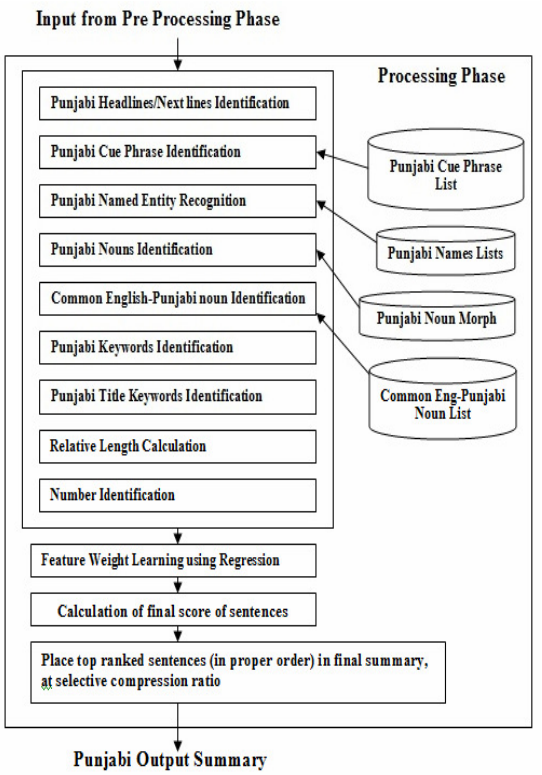

Figure 4. Processing phase

Features are of two types statistical and linguistic features. Some of the statistical features are Punjabi keywords identification, relative sentence length feature and numbered data feature. Various linguistic features for selecting important sentences in summary are: Punjabi-headlines identification, identification of lines just next to headlines, identification of Punjabi-nouns, identification of Punjabi-proper-nouns, identification of common-English-Punjabi-nouns, identification of Punjabi-cue-phrases and identification of title-keywords in sentences etc. Scores of sentences are determined from sentence-feature-weight equation:

$w_1f_1+w_2f_2+ w_3f_3+\ldots\ldots\ldots\ldots....w_nf_n$

Where f_1, f_2, f_3…… f_n are different features of sentences calculated in the different sub phases of Punjabi text summarization system and w_1, w_2, w_3…… w_n are the corresponding feature weights of sentences. Weights of features are determined using mathematical regression. Using regression, feature values of some Punjabi documents which are manually summarized are treated as independent input values and their corresponding dependent output values are provided. In the training phase, manually summaries of fifty news-documents are made by giving fuzzy scores to the sentences of those documents and then regression is applied for finding values of feature-weights and then average values of feature-weights are calculated. High scored sentences in proper order are selected for final summary. In final summary, sentences coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios. The sub phases for Processing-phase [22] are as follows:-

A. Identification of Headlines and Next lines

It is first time that an automatic system for identification of multi news headlines and lines just next to headlines of a single document has been developed for Punjabi language. Headlines are highly important in news documents and are always part of final summary. There can be very important information in the next-line to headline, so next-line usually becomes part of final summary. In Punjabi-news-corpus with 957553 sentences, the frequency-count of these headlines/next lines is 65722 lines, which is 6.863% of the news-corpus. In Punjabi a sentence usually ends with '**।**' vertical bar, ?, or ! etc. and in Punjabi headlines identification system, if current sentence does not ends with punctuation marks like '**।**' vertical bar, ?, or ! etc. but ends with enter key or new line character then set the headline flag for that line to true. If the next subsequent line of this headline ends with punctuation marks like '**।**' vertical bar, ? or ! etc. but does not ends with enter key or new line character then set the next line flag to true for that line. An in depth analysis of results of headlines detection system and next lines identification system has been done over fifty Punjabi news documents taken randomly from Punjabi news corpus. Headline sentences are assigned very high score equal to 10 and their headline flags are set to true. The accuracy of Punjabi headline identification system is 97.43%. This accuracy is tested over 50 Punjabi single/multi-news documents. There are 2.57% errors due to the reason that some times name of author and location name are written in second line after the headline with enter key as last character, so it may be wrongly picked as second headline which is wrong. For example consider the following single news document as input:

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ

ਉਜਾਗਰ ਸਿੰਘ (ਬਰਨਾਲਾ)

"Eighty girls were taught the sewing work

Ujagar Singh (Barnala)"

In the above news, second line containing author name and location name **ਉਜਾਗਰ ਸਿੰਘ (ਬਰਨਾਲਾ)** "Ujagar Singh (Barnala)" will also be treated as headline along with first line because it also ends with enter key, but this line is not important and should not come in summary. The accuracy of next lines identification system is 98.57% which is tested over fifty Punjabi single/multi news documents. Errors of 1.43% are due to the same reason of multiple headlines may come in a single news, due to which, some times next line may be missing from summary. Suppose the case in which there are multiple headlines in single news and we need only two lines in summary at 10% compression ratio and further suppose

second headline is not important as it is only containing name of author and location. But in this case next line will be missing from summary and second headline will be wrongly placed in summary. Next-lines are always assigned higher weight-age equal to 9 and their next line flags are set to true.

B. Punjabi Cue Phrase Identification

Cue Phrases are some important terms in text documents like: finally, conclusion, conclude, summary and summarize etc. Sentences containing cue-phrases are given more weight-age for summary than other sentences. We have prepared a list of Punjabi cue-phrases for assigning more weight-age to sentences containing these cue-phrases. Problem with Punjabi is non-standardization of Punjabi spellings. Many of the popular Punjabi words are written in multiple ways. As for example, the word **ਵਿਚ** "in" can be written both with and without addak, so both of these forms have been included in cue phrases. For example **ਅੰਤ ਵਿੱਚ/ ਅੰਤ ਵਿਚ** "in the end" and **ਸੰਖੇਪ ਵਿੱਚ/ ਸੰਖੇਪ ਵਿਚ** "in brief" etc. For those sentences containing cue phrase/cue phrases, their cue phrase flag is set to true. The frequency count of cue phrases is 58708 words in Punjabi news corpus which covers 0.52% of this corpus.

C. Punjabi Named Entity Recognition

Punjabi rule based named entity recognition system is first of its kind developed and implemented for identifying proper names in Punjabi text [9]. There was no other Punjabi rule based NER system was available prior to our NER. Different gazetteer lists are used in it like prefix-list, suffix-list, middle-name-list, last-name-list and names-list for checking whether the given Punjabi word is name or not. Gazetteer lists are developed by doing analyses of Punjabi news-corpus.

For checking if next-word in Punjabi-name or not, Prefix-list includes different prefixes of Punjabi names. like **ਸ਼੍ਰੀ.** "Mr.", **ਸ਼੍ਰੀਮਤੀ** "Mrs.", **ਸ.** "sardar", **ਪ੍ਰਿ.** "Prin." and **ਡਾ:** "Dr." etc. There are fourteen prefixes identified from the Punjabi-news-corpus. We have developed prefix-list by making freq-list from corpus. The freq-count of prefix-words is 17,127 which includes 0.15% of the corpus. Suffix-list includes various suffixes of names like **ਪੁਰੀ** "puri", **ਪੁਰਾ** "pura", **ਜੀਤ** "jit" and **ਪੁਰ** "pur" etc for checking if current-Punjabi-word is name or not. There are fifty suffixes identified from the Punjabi-news-corpus. We have developed suffix-list by making freq-list from corpus. The freq-count of suffix-words is 225306 which includes 1.99% of the corpus.

Punjabi-middle-names-list includes various middle-names of persons like **ਕੁਮਾਰੀ** "kumari", **ਕੌਰ**"kaur" and **ਕੁਮਾਰ** "kumar" etc for checking if that word is name or not. There are 08 middle-names identified from Punjabi-news-corpus. We have developed middle-names-list by making freq-list from corpus. The freq-count of middle-name words is 97907 which includes 0.8672% of the corpus. Punjabi-last-names-list includes various last-names of persons like **ਗੋਇਲ** "goel", **ਗੁਲਾਟੀ** "gulati" and **ਖੁਰਾਨਾ** khurānā "khurana" etc for checking if that word is name or not. There are 310 last-names identified from Punjabi-news-corpus. We have developed last-names-list by making freq-list from Punjabi-corpus. The freq-count of last-name words is 69268 which includes 0.6135% of the corpus. For finding importance of sentences, proper names are very much useful. There are 17598 proper-names identified from the Punjabi-news-corpus. Punjabi-proper- names-list covers 13.84% of words from Punjabi-news-corpus. The value of Punjabi-names-feature is calculated by taking ratio of number of Punjabi-names in a sentence to the length of that sentence and value of this feature for a sentence lies between 0 to 1.

The Algorithm for rule based Punjabi NER has been published in [9] and it increments the NER score of current sentence by 01 if current Punjabi-word matches with any word from any of prefix-list or suffix-list or names-list or middle-names-list or last-names-list. Punjabi NER has been tested over fifty Punjabi-news-documents with Precision=89.32%, Recall=83.4%, F-score=86.25% and 13.75% errors. There are no errors in prefix rule. There are 1% errors in suffix rule for example **ਕਰਿਆਣਾ** "grocery" and **ਅਫਸਰ** aphsar "officer" are not found in Punjabi nouns-morph/dictionary but both of them fall under suffix-rule which treats them as Punjabi-names which is false. There are 0.25% errors in middle-name-rule for example in a proper-name **ਕੌਰ ਸਿੰਘ** "Kaur Singh" both middle-names are together as a single name, but they lie under middle-name-rule which makes NER score equal to 2 in this case. There are 10% errors in last-name-rule for example in case of a Punjabi-names **ਕਰਤਾਰ ਸਿੰਘ ਜੰਗੀਆਣਾ** "Kartar Singh Jangiana" and **ਬੰਤਾ ਸਿੰਘ ਬੰਟੀ** "Banta Singh Banti" their last names **ਜੰਗੀਆਣਾ** "Jangiana" and **ਬੰਟੀ** "Banti" are not in last-names-list but are part of proper-names-list so their NER score will be wrongly incremented to 2 in each case. There are 0.25% errors in proper-names-rule for example a Punjabi word **ਨਿਹਾਲ** "Nihal" some times is treated as Punjabi-name and some times is treated in different sense, but because it is part of proper-names-list so it will be always treated as a proper-name by Punjabi NER. Remaining 2.25% errors are because of those Punjabi-names which do not fall under any of NER rules for example **ਗਰੀਣ ਐਵਿਨਿਊ** "Green Avenue" or because of those Punjabi-words which are some times treated as Punjabi-names and some times treated as Punjabi-nouns for example a Punjabi word

ਬਹਾਦਰ "brave" is some times treated as Punjabi-name and some times treated as noun.

D. Punjabi Nouns/ Common English-Punjabi Nouns Identification

Sentences Possessing Punjabi-Nouns [1] are given more weight-age. Punjabi words are searched in noun morph or stemming is done for possibility of nouns. There are 37297 nouns in Punjabi-noun-morph [10]. The score of this feature is ratio of number of Punjabi-nouns in a sentence to length of that sentence. The range of value of this feature is between 0 to 1. The frequency of Punjabi nouns is 16.56% of words in Punjabi-news-corpus. The accuracy of Punjabi noun identification phase is 98.43% which is tested over 50 Punjabi-news-documents of Punjabi-corpus. Errors of 1.57% are due non existence of certain Punjabi nouns in noun morph and due to stemming errors of Punjabi noun stemmer [4].

In these days, there is common usage of English-words in Punjabi text. Consider a Punjabi sentence "**ਟੈਕਨਾਲੋਜੀ ਦੇ ਯੁੱਗ ਵਿਚ ਮੋਬਾਈਲ**" "In the era of mobile technology" This sentence includes **ਮੋਬਾਈਲ** "mobile" and **ਟੈਕਨਾਲੋਜੀ** "technology" as common English-Punjabi nouns. Majority of such terms are not found in Punjabi dictionary or Punjabi noun morph. In text summarization, Sentences containing common English-Punjabi nouns are assigned more weight-age. A small offline module has been developed to generate the database for common English-Punjabi nouns by analyzing the Punjabi news corpus along with frequency of these common English-Punjabi nouns into Punjabi news corpus. Punjabi-words are searched in Common-English-Punjabi-nouns-dictionary for possibility of common English-Punjabi nouns. This value of this feature is calculated by ratio of number of common-English-Punjabi-nouns in a sentence to the length of that sentence. The range of value for this feature lies from 0 to 1 for a sentence. From Punjabi news corpus, frequency count of common-English-Punjabi-nouns is 18245, which covers 6.44% of Punjabi-corpus. The accuracy of common-English-Punjabi-noun identification phase is 95.12% which is tested over 50 Punjabi news-documents of Punjabi-corpus. Errors of 4.88% are due non existence of certain common English-Punjabi nouns in database of common English-Punjabi nouns.

E. Punjabi Keywords/Title Keywords Identification

Keywords are thematic words containing important information. Keywords are helpful in deciding the sentence importance. Punjabi keywords identification system is first of its kind system developed and implemented as prior to it no other Punjabi keywords identification system was available. Algorithm for Punjabi Keywords Identification [7] [11]:

Step 1:- Set noun flag to true for those words of input text which are found in Punjabi noun morph.

Step 2:- For each Punjabi word w, find its TF-ISF-Score which is calculated by multiplying TF(w,s) with ISF(w). Where TF(w,s) is the frequency of word w in sentence s, and the inverse sentence frequency ISF(w) = log(|Sl/ SF(w)). Sentence-frequency SF(w) is the frequency of sentences containing word w. Store top ranked words (with high TF-ISF-Scores) with Punjabi_noun_flag= true in a priority queue.

Step 3:- Delete top 20% of Punjabi-noun-words from the priority queue, which are candidates for keywords in this phase.

The Precision, Recall and F-Score of Punjabi keywords identification system are 80.4%, 90.6% and 85.2% respectively which are calculated by analyzing the results of keywords identification system over fifty Punjabi news documents. Errors of 14.8% are because of absence of some Punjabi-nouns in noun-morph or dictionary errors or syntax mistakes in input text or due to violation of stemming-rules. In case of dictionary errors, after stemming, root word is not found in Punjabi noun-morph/names dictionary, but in reality it is Punjabi noun/ proper name. In syntax errors, there is some syntax mistake while typing the Punjabi word, but actually it lies under any of stemming-rules. Examples of errors due to rules violation are: Punjabi word **ਹਲਕੇ** halkē "light weight" is adjective and **ਬਦਲੇ** badlē "in lieu of" is adverb. These words are not found in Punjabi noun-morph/ names dictionary, but they lie under ੇ ē stemming-rule which treats them noun after stemming, but it is not true.

Title lines are the headlines of single/multi news documents. Sentences containing Title-keywords [5] are given more weight-age. For obtaining Title-keywords, stop words are removed from title-lines with headline_flag= true. This feature-score is calculated as ratio of unique title-keywords in a sentence to the total number of title-keywords. The efficiency of Punjabi title keywords identification is 97.48% which is calculated over fifty Punjabi single/multi news documents of Punjabi corpus. Errors of 2.52% are due to the reason that some of stop words may be left in the title line as Punjabi stop words list is not exhaustive and contains 615 Punjabi stop words.

F. Punjabi Sentence Relative Length Feature

Short Punjabi sentences are avoided for including in final summary as often they contain less information [5]. But lengthy sentences can have lot of important information. This value of this feature is calculated as ratio of frequency of words in current sentence to the words frequency of largest sentence. The value of this feature is always less than or equal to one.

Punjabi-Sentence-Length-feature-Score= frequency of words in current sentence / words frequency of largest sentence.

G. Numeric Data Identification Feature

For text summarization, those sentences containing contain numeric data [5] are assigned more weight-age Numeric digits, Gurmukhi and Roman numerals are considered as numeric data. The value for this feature is determined by dividing the frequency of numeric data in current sentence by the length of that sentence.
Number-feature-score= Frequency of numeric data in current sentence/ Sentence Length

H. Calculation of Scores of Sentences and Producing Final Summary

Final scores of sentences are determined from sentence-feature-weight equation.

w1f1+w2f2+ w3f3+……wnfn Where f1, f2, f3……fn are different features of sentences calculated in the different sub phases of Punjabi text summarization system and w1, w2, w3……wn are the corresponding feature weights of sentences. We have applied regression [5] [8] model for estimating the weights of text features for Punjabi text summarization system. A relation between inputs and outputs is established. Regression can be represented in the matrix notation as below:

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

Where
$[Y]$ is fuzzy output vector having values between 0 to 1 based on importance of sentences given manually for fifty documents.
$[X]$ is the input matrix (feature parameters) for different features having values between 0 to 1.
$[w]$ is weight matrix of system (with weights w1, w2…….w10 in the given equation)
In the training corpus, m denotes total number of sentences.
Weight w of a particular feature k (k=1 to 10) with input matrix x and fuzzy output matrix y can be calculated as follows:-

$$w = \frac{\sum_{i=01 \text{ to } m} (x_i - \text{mean}(x))(y_i - \text{mean}(y))}{\sum_{i=01 \text{ to } m} (x_i - \text{mean}(x))^2}$$

From the above equation, weights of each of ten features of Punjabi text summarization have been calculated. Table I shows the results of weight learning using regression.

TABLE I.
WEIGHT LEARNING RESULTS USING REGRESSION

Features	Learned weights
Sentence relative length feature	0.31
Punjabi Keywords identification feature	0.29
Number feature	2.54
Headline feature	10
Lines just next to headline feature	9
Punjabi noun feature	0.42
Punjabi proper noun feature	0.75
Common English-Punjabi noun feature	1.29
Punjabi Cue phrase feature	1
Punjabi Title keywords feature	1.8

From results of weight learning in Table I, we concludes that three most important features of Punjabi Text Summarizer are identification of Punjabi-headlines, identification of next-lines and identification of numeric data. Top ranked sentences in proper order are selected for final summary. In final summary, sentence coherence is maintained by properly ordering the sentences in the same order as they appear in the input text at the selective compression ratios.

Algorithm for single document multi news Punjabi Text Summarization System:-
Algorithm starts by splitting the input Punjabi text into sentences and words. Initially scores of every sentence is set to 0.

Step I: Delete the duplicate sentences from input text by searching the current sentence in the sentence list which is initially empty. For each sentence check the following condition: If current sentence is found in sentence list then Current sentence is set to null being the duplicate sentence. Else Current sentence is added to the sentence list being the unique sentence. Follow steps II to step XII for every word in sentences.

Step II: Delete stop words from every sentence in input.

Step III: Calculate the noun-score of sentence, if current Punjabi-word is noun.

Step IV: Calculate common-English-Punjabi-noun score of sentence, if current Punjabi-word is common-English-Punjabi noun.
Step V: Calculate proper-name-score of sentence, If current Punjabi-word is proper-name.

Step VI: Apply stemmer [12] for Punjabi Noun/Proper Names for those words which are not found in nouns-morph/ common-English-Punjabi-noun-list/ proper-names-dictionary and go to step III.

Step VII: Calculate numeric-feature-score of sentence, if current Punjabi-word is any number like 45.

Step VIII: Set the headline-flag= true for current Punjabi-word, if it is part of headline.
Step IX: Set the next-line-flag= true for current Punjabi-word, if it is part of line just next to headline.

Step X: Calculate the score of keyword feature for current Punjabi-word using TF-ISF technique.

Step XI: Set cue-phrase-flag for current Punjabi-word to true, if it is cue-phrase.

Step XII: Calculate title-keyword-feature-score of current word, if it is title keyword.

Step XIII: Calculate the relative-length-feature-score of all the sentences.

Step XIV: Calculate the weight-age of each feature by applying regression using sentence-feature-weight-equation.

Step XV: Calculate final-scores of all the sentences by applying sentence-feature-weight-equation.

Step XVI: Select the top scored sentences at given compression ratios i.e. at 10%, 30%, 50% C.R. etc.

Step XVII: Final summary is formed by arranging top scored sentences in ascending order of their position in input text at selective compression ratios. In this step coherence of sentences is maintained in final summary.

Algorithm Input-

First News: -

ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ 'ਚ ਵਿਕਾਸ ਕਾਰਜਾਂ 'ਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ-ਭਾਨਾ

ਸ਼ਹਿਣਾ, 8 ਜਨਵਰੀ (ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਹਲਕਾ ਵਿਧਾਇਕ ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਹੇਠ ਹਲਕੇ ਦੇ ਵਿਕਾਸ ਕਾਰਜਾਂ ਵਿਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ ਹੈ। ਇਹ ਸ਼ਬਦ ਭਗਵਾਨ ਸਿੰਘ ਭਾਨਾ ਯੂਥ ਆਗੂ ਤੇ ਸੰਮਤੀ ਮੈਂਬਰ ਨੇ ਪਿੰਡ ਨਾਨਕਪੁਰਾ ਵਿਖੇ ਸ਼ਗਨ ਸਕੀਮ ਦੇ ਚੈਕ ਦੇਣ ਸਮੇਂ ਸੰਬੋਧਨ ਕਰਦਿਆਂ ਆਖੇ। ਭਗਵਾਨ ਸਿੰਘ ਨੇ ਕਿਹਾ ਕਿ ਸ਼ਗਨ ਸਕੀਮ ਲਈ ਰਹਿੰਦੇ ਪਰਿਵਾਰਾਂ ਲਈ ਛੇਤੀ ਹੀ ਬਾਕੀ ਦੀ ਰਾਸ਼ੀ ਜਾਰੀ ਕੀਤੇ ਜਾਣ ਦੀ ਹਲਕਾ ਵਿਧਾਇਕ ਨੇ ਹਾਮੀ ਭਰੀ ਹੈ ਅਤੇ ਸੰਮਤੀ ਰਾਹੀਂ ਵੀ ਪਿੰਡਾਂ ਲਈ ਸਬਮਰਸੀਬਲ ਪੰਪ ਤੇ ਗਰਾਂਟਾਂ ਦਿੱਤੀਆਂ ਜਾ ਰਹੀਆਂ ਹਨ। ਇਸ ਸਮੇਂ ਸੁਖਦੇਵ ਸਿੰਘ ਸਰਪੰਚ ਨਾਨਕਪੁਰਾ, ਪਵਨ ਕੁਮਾਰ, ਗੁਰਚਰਨ ਸਿੰਘ ਜ਼ੈਲਦਾਰ, ਕੇਰ ਸਿੰਘ ਪੱਖੋਕੇ, ਗੁਰਤੇਜ ਸਿੰਘ ਘੋਨਾ, ਜੰਗ ਸਿੰਘ ਪ੍ਰਧਾਨ ਟੈਕਸੀ ਯੂਨੀਅਨ, ਜਗਸੀਰ ਸਿੰਘ, ਕਰਤਾਰ ਸਿੰਘ ਪੱਖੋਕੇ ਆਦਿ ਆਗੂ ਵੀ ਹਾਜ਼ਰ ਸਨ।

Second News:-

ਅਧਿਆਪਕ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਨੂੰ ਸ਼ਰਧਾਂਜਲੀਆਂ ਭੇਟ

ਧਨੌਲਾ, 8 ਜਨਵਰੀ (ਨਿੱਜੀ ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਸੁਖਵਿੰਦਰ ਸਿੰਘ ਵੜੈਚ ਦੇ ਹੋਣਹਾਰ ਅਧਿਆਪਕ ਪੁੱਤਰ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਦੀ ਅੰਤਿਮ ਅਰਦਾਸ ਗੁਰਦੁਆਰਾ ਪਾਤਸ਼ਾਹੀ ਨੌਵੀਂ ਵਿਖੇ ਹੋਈ। ਇਸ ਮੌਕੇ ਵੱਖ-ਵੱਖ ਸਖ਼ਸ਼ੀਅਤਾਂ ਨੇ ਸ਼ਰਧਾ ਦੇ ਫੁੱਲ ਭੇਟ ਕੀਤੇ। ਜਥੇਦਾਰ ਸਾਧੂ ਸਿੰਘ ਰਾਗੀ ਸਾਬਕਾ ਚੇਅਰਮੈਨ ਮਾਰਕੀਟ ਕਮੇਟੀ ਭਦੌੜ ਨੇ ਕਿਹਾ ਕਿ ਸਾਡੇ ਕੋਲ ਅਜਿਹੀ ਦੁਖਦਾਈ ਮੌਤ 'ਤੇ ਮਾਪਿਆਂ ਕੋਲ ਭਾਣਾ ਮੰਨਣ ਨੂੰ ਕਹਿਣ ਲਈ ਸ਼ਬਦ ਵੀ ਨਹੀਂ। ਅਜਿਹੀ ਹੋਣਹਾਰ ਔਲਾਦ ਦਾ ਬੇ-ਵਕਤ ਚਲੇ ਜਾਣਾ ਬਹੁਤ ਦੁਖਦਾਈ ਹੈ। ਜਗਤਾਰ ਸਿੰਘ ਜੰਗੀਆਣਾ ਪ੍ਰਚਾਰਕ ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ ਅੰਮ੍ਰਿਤਸਰ ਨੇ ਕਿਹਾ ਕਿ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਤੇ ਧਨੌਲਾ ਵਾਸੀਆਂ ਨੂੰ ਹੀ ਨਹੀਂ, ਬਲਕਿ ਨਾਨਕੇ ਪਿੰਡ ਜੰਗੀਆਣਾ ਨੂੰ ਬਹੁਤ ਮਾਣ ਸੀ। ਉਸ ਵਿਚ ਨਿਆਣਿਆਂ ਵਾਲੀ ਚੰਚਲਤਾ ਘੱਟ ਸੀ ਅਤੇ ਸਿਆਣਿਆਂ ਵਾਲੀ ਲਿਆਕਤ ਵਧੇਰੇ ਸੀ। ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਹਲਕਾ ਵਿਧਾਇਕ ਭਦੌੜ, ਸ: ਭੋਲਾ ਸਿੰਘ ਵਿਰਕ ਮੈਂਬਰੀ ਕੌਮੀ ਜਨਰਲ ਕੌਂਸਲ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ, ਜਥੇਦਾਰ ਬਲਦੇਵ ਸਿੰਘ ਚੂੰਘਾਂ, ਜਥੇਦਾਰ ਅਮਰ ਸਿੰਘ ਬੀ.ਏ. ਮੈਂਬਰ ਸ਼੍ਰੋਮਣੀ ਗੁਰਦੁਆਰਾ ਪ੍ਰਬੰਧਕ ਕਮੇਟੀ ਅੰਮ੍ਰਿਤਸਰ, ਜਥੇਦਾਰ ਭਰਪੂਰ ਸਿੰਘ ਧਨੌਲਾ ਸਾਬਕਾ ਚੇਅਰਮੈਨ, ਗੁਰਵੀਰ ਸਿੰਘ ਗੁਰੀ ਯੂਥ ਕਾਂਗਰਸੀ ਆਗੂ, ਇਕਬਾਲ ਸਿੰਘ ਜੰਗੀਆਣਾ ਸੰਮਤੀ ਮੈਂਬਰ, ਗਮਦੂਰ ਸਿੰਘ ਮਾਨ ਸਰਪ੍ਰਸਤ ਆੜ੍ਹਤੀਆ ਐਸੋਸੀਏਸ਼ਨ ਧਨੌਲਾ, ਗੁਰਨਾਮ ਸਿੰਘ ਸਿੱਧੂ ਸਾਬਕਾ ਪ੍ਰਧਾਨ ਨਗਰ ਕੌਂਸਲ ਧਨੌਲਾ, ਗੁਰਚਰਨ ਸਿੰਘ ਕਲੇਰ ਪ੍ਰਧਾਨ ਆੜ੍ਹਤੀਆ ਐਸੋਸੀਏਸ਼ਨ, ਭਰਪੂਰ ਸਿੰਘ ਸਾਬਕਾ ਐਮ.ਸੀ. ਸੁਰਿੰਦਰ ਸਿੰਘ ਸੱਦੋਵਾਲੀਆ ਜ਼ਿਲ੍ਹਾ ਪ੍ਰਧਾਨ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ ਅੰਮ੍ਰਿਤਸਰ, ਚੇਤਨ ਸਿੰਘ ਮੂੰਮ, ਲਾਭ ਸਿੰਘ ਮੱਝੂਕੇ, ਜੰਗ ਸਿੰਘ ਜੰਗੀਆਣਾ, ਕਰਮਜੀਤ ਸਿੰਘ ਨੀਟਾ ਮੈਂਬਰ ਜ਼ਿਲ੍ਹਾ ਪ੍ਰੀਸ਼ਦ, ਹਰਨੇਕ ਸਿੰਘ ਸਾਬਕਾ ਪ੍ਰਧਾਨ ਸਹਿਕਾਰੀ ਸਭਾ ਜੰਗੀਆਣਾ, ਮਨਮੋਹਨ ਸਿੰਘ, ਗੁਰਤੇਜ ਸਿੰਘ ਸਰਪੰਚ, ਰਾਜ ਸਿੰਘ ਨੈਣੇਵਾਲੀਆ, ਗੁਰਪ੍ਰੀਤ ਸਿੰਘ ਕਲੱਬ ਆਗੂ, ਰਾਮ ਸਿੰਘ ਢੀਂਡਸਾ, ਗੁਰਮੀਤ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਬੂਟਾ ਸਿੰਘ ਬੁਰਜ, ਭਗਵਾਨ ਸਿੰਘ ਭਾਨਾ, ਹਰਵਿੰਦਰ ਸਿੰਘ, ਯਾਦਵਿੰਦਰ ਸਿੰਘ ਵਾਲੀਆ ਐਮ.ਸੀ., ਜਗਤਾਰ ਸਿੰਘ ਕਲੇਰ ਪ੍ਰਧਾਨ ਸਹਿਕਾਰੀ ਸਭਾ ਧਨੌਲਾ, ਗੁਰਪ੍ਰੀਤ ਸਿੰਘ ਚੀਮਾ ਆਦਿ ਨੇ ਹਾਜ਼ਰੀ ਲਵਾਈ।

Third News:-

ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਹੋਈ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਗਰੀਨ ਐਵੀਨਿਊ ਦੇ ਪਾਰਕ ਨੇੜੇ ਨਾਨਕਸਰ ਗੁਰਦੁਆਰਾ ਵਿਖੇ ਹੋਈ, ਜਿਸ ਵਿਚ ਸਰਬ ਸੰਮਤੀ ਨਾਲ ਬੂਟਾ ਸਿੰਘ ਚੌਹਾਨ ਅਤੇ ਸੁਰਜੀਤ ਸਿੰਘ ਦਿਹੜ ਸਰਪ੍ਰਸਤ, ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਪ੍ਰਧਾਨ, ਡਾ: ਅਮਨਦੀਪ ਸਿੰਘ ਟੱਲੇਵਾਲੀਆ ਅਤੇ ਕੁਲਵੰਤ ਸਿੰਘ ਧਿੰਗੜ ਮੀਤ ਪ੍ਰਧਾਨ, ਜਨਰਲ ਸਕੱਤਰ ਪਾਲ ਸਿੰਘ ਲਹਿਰੀ, ਸਹਾਇਕ ਜਨਰਲ ਸਕੱਤਰ ਲੈਕਚਰਾਰ ਸੁਖਮਿੰਦਰ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਜਥੇਬੰਦਕ ਸਕੱਤਰ ਬਿੰਦਰ ਖੁੱਡੀ ਕਲਾਂ, ਸੁਦਰਸ਼ਨ ਗੁੱਡੂ ਤੇ ਅਵਤਾਰ ਸਿੰਘ ਸੰਧੂ, ਪ੍ਰਚਾਰ ਸਕੱਤਰ ਅਸ਼ੋਕ ਭਾਰਤੀ ਅਤੇ ਬੰਤ ਸਿੰਘ ਬਰਨਾਲਾ ਵਿੱਤ ਸਕੱਤਰ ਲਵਪ੍ਰੀਤ ਚਾਸ ਮਸ਼ਾਰਿਕ ਤੇ ਸਹਾਇਕ

ਸਕੱਤਰ ਲਛਮਣ ਦਾਸ ਮੁਸਾਫ਼ਿਰ ਤੇ ਸਹਾਇਕ ਵਿੱਤ ਸਕੱਤਰ ਬਲਵਿੰਦਰ ਸਿੰਘ ਠੀਕਰੀਵਾਲਾ ਚੁਣੇ ਗਏ। ਚੋਣ ਉਪਰੰਤ ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਨੇ ਦੱਸਿਆ ਕਿ ਇੱਕੀ ਮੈਂਬਰੀ ਕਾਰਜਕਾਰਨੀ ਦਾ ਐਲਾਨ ਅਗਲੀ ਸੂਚੀ ਵਿਚ ਕੀਤਾ ਜਾਵੇਗਾ।

Fourth News:-

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ-ਸਿੱਧੂ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਮਾਲਵਾ ਸੱਭਿਆਚਾਰਕ ਅਤੇ ਵੈਲਫੇਅਰ ਕਲੱਬ ਬਰਨਾਲਾ ਵੱਲੋਂ ਚਲਾਏ ਜਾ ਰਹੇ ਸਿਲਾਈ ਸੈਂਟਰ ਦੀਆਂ ਦਸ ਵਿਦਿਆਰਥਣਾਂ ਨੂੰ ਸਿਖਲਾਈ ਸਰਟੀਫਿਕੇਟ ਵੰਡੇ ਗਏ। ਇਸ ਮੌਕੇ ਬੋਲਦਿਆਂ ਟਰੱਸਟ ਦੇ ਚੇਅਰਮੈਨ ਗੁਰਜਿੰਦਰ ਸਿੰਘ ਸਿੱਧੂ ਪ੍ਰਧਾਨ ਸਾਬਕਾ ਸੈਨਿਕ ਵਿੰਗ ਸ਼੍ਰੋਮਣੀ ਅਕਾਲੀ ਦਲ ਅਤੇ ਸੈਂਟਰ ਸੰਚਾਲਕ ਜਗਸੀਰ ਸਿੰਘ ਚੌਹਾਨ ਨੇ ਦੱਸਿਆ ਕਿ ਨਿਸ਼ਕਾਮ ਤੌਰ 'ਤੇ ਇਹ ਸੈਂਟਰ ਪਿਛਲੇ ਦਸ ਸਾਲ ਤੋਂ ਚੱਲ ਰਿਹਾ ਹੈ ਅਤੇ 70 ਵਿਦਿਆਰਥਣਾਂ ਉਕਤ ਵਿਦਿਆਰਥਣਾਂ ਤੋਂ ਇਲਾਵਾ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿੱਖ ਕੇ ਆਪਣੀ ਰੋਜ਼ੀ-ਰੋਟੀ ਕਮਾਉਣ ਦੇ ਯੋਗ ਹੋ ਸਕੀਆਂ ਹਨ। ਸਮਾਗਮ ਵਿਚ ਉਚੇਚੇ ਤੌਰ 'ਤੇ ਪੁੱਜੇ ਟਰੱਕ ਯੂਨੀਅਨ ਬਰਨਾਲਾ ਦੇ ਪ੍ਰਧਾਨ ਕੁਲਵੰਤ ਸਿੰਘ ਕੰਤਾ ਨੇ ਸੰਸਥਾ ਦੇ ਕੰਮਾਂ ਦੀ ਸ਼ਲਾਘਾ ਕੀਤੀ ਅਤੇ 21 ਸੌ ਰੁਪਏ ਸਹਾਇਤਾ ਲਈ ਵੀ ਦਿੱਤਾ। ਇਸ ਮੌਕੇ ਨਗਰ ਕੌਂਸਲ ਬਰਨਾਲਾ ਦੇ ਪ੍ਰਧਾਨ ਸ: ਪਰਮਜੀਤ ਸਿੰਘ ਢਿੱਲੋਂ, ਕੌਮੀ ਤਰਕਸ਼ੀਲ ਆਗੂ ਬਲਵਿੰਦਰ ਬਰਨਾਲਾ, ਬੌਬੀ ਬਾਂਸਲ ਸਮਾਜ ਸੇਵੀ, ਸਾਬਕਾ ਚੇਅਰਮੈਨ ਸੁਖਮਹਿੰਦਰ ਸਿੰਘ ਸੁੱਖੀ, ਜਥੇਦਾਰ ਜਰਨੈਲ ਸਿੰਘ ਭੋਤਨਾ ਅਤੇ ਹਰਪਾਲਇੰਦਰ ਸਿੰਘ ਰਾਹੀ, ਸੁਖਜੀਤ ਕੌਰ ਸੁੱਖੀ, ਮਾਰਕੀਟ ਕਮੇਟੀ ਬਰਨਾਲਾ ਦੇ ਚੇਅਰਮੈਨ ਕਰਨੈਲ ਸਿੰਘ ਠੁੱਲੀਵਾਲ, ਸੈਨਿਕ ਵਿੰਗ ਦੇ ਸਰਕਲ ਪ੍ਰਧਾਨ ਕੈਪਟਨ ਬੂਟਾ ਸਿੰਘ ਸਹੋਤਾ, ਕੈਪਟਨ ਮਹਿੰਦਰ ਸਿੰਘ ਮਾਨ, ਪੰਜਾਬੀ ਗਾਇਕ ਜੈਸੀ ਬਾਜਵਾ ਤੋਂ ਇਲਾਵਾ ਹੋਰ ਬਹੁਤ ਸਾਰੀਆਂ ਸੰਸਥਾਵਾਂ ਦੇ ਆਗੂਆਂ ਨੇ ਆਪਣੀ ਹਾਜ਼ਰੀ ਲਵਾਈ। ਇਸ ਵੇਲੇ ਤਿੰਨ ਗਰੀਬ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਮਸ਼ੀਨਾਂ ਵੀ ਭੇਟ ਕੀਤੀਆਂ ਗਈਆਂ।

The English Translation of above four multi news of single document is given below:-

First News:-

Under leadership of Ghunnas development activities are highly accelerated-Bhana

Shhina, 8 January (motivational letters) – Under the leadership of local MLA Sant Balbir Singh Ghunnas development activities in the constituency are highly accelerated. These words are spoken by Bhagwan Singh Bhana youth leader and Committee member in the village Nankpura while distributing the cheques of shagun scheme. Bhagwan Singh said that local MLA has agreed to release the remaining amount soon to rest of families for shagun scheme and submersible pumps & grants are also given to the villages through Committee. On this occasion the Nankpura Sarpanch Sukhdev Singh, Pawan Kumar, Gurcharan Singh jaildar, Kaur Singh Pakhoke, Gurtej Singh Ghona, Jang Singh head taxi union, Jagsir Singh, Kartar Singh pakhoke etc. leaders were also present.

Second News:-

Tributes paid to teacher Gurdeep Singh Vdaich

Dhnaula, January 8 (private motivational letters) - The final prayer for outstanding teacher Gurdeep Singh Vdaich son of Sukhwinder Singh Vdaich was held at Gurudvara ninth Kingdom. On this occasion, different dignities presented flowers of worship. Jathedar Sadhu Singh Ragi former chairman market committee Bhadaur said that they had no words to console his parents for this painful death other than obeying the God's will. Untimely demise of such a promising child is very painful. Jagtar Singh Jangiana preacher Shiromani Gurdwara Parbandhak Committee Amritsar said not only Dhnaula residents but also the maternal village Jangiana were proud of Gurdeep Singh Vdaich. He was having less restlessness of children and more wisdom like elders. Sant Balvir Singh Ghunnas local MLA Bhadaur, sardar Bhola Singh Virk member general council Shiromani Akali Dal, Jathedar Baldev Singh Chungan, Jathedar Amar Singh B.A. member committee Shiromani Gurdwara Parbandhak Committee Amritsar, Jathedar Bharpoor Singh Dhnaula ex chairman, Gurvir Singh Guri youth congress leader, Iqbal Singh Jangiana committee member, Gamdoor Singh Mann leader broker association Dhnaula, Gurnam Singh ex head city council Dhnaula, Gurcharan Singh Kaler broker association, Bharpoor Singh former M.C. Surinder Singh Saddowalia district head Shiromani Akali Dal Amritsar, Chetan Singh mumm, Labh Singh Majjhuke, Jang Singh Jangiana, Karamjit Singh Neeta member district prishad, Harnek Singh former head co-operative assembly Jangiana, Manmohan Singh, Gurtej Singh chairman of panchayat, Raj Singh Nainewalia, Gurpreet Singh club leader, Ram Singh Dhindsa, Gurmeet Singh Shhina, Boota Sungh Burj, Bhagwan Singh Bhana, Harwinder Singh, Yadwinder Singh Walia M.C. Jagtar Singh Kaler head co-operative assembly Dhnaula,Gurpreet Singh Cheema etc. were present.

Third News:-

Eelection held for literature discussion forum Barnala

Barnala, January 8 (Staff Reporter) – Election of literature discussion forum held at Gurdwara Nanksar near the park of Green Avenue, in which unanimously Buta Singh Chauhan and Surjit Singh Dehd patron, Dr Ujagar Singh Mann head, Dr: Amandeep Singh Tallewalia, Kulwant Singh Dhingad circle head, General Secretary Pal Singh Lahiri, assistant general secretary lecturer Sukminder Singh Shhina, organisational secretary Binder Khuddi kalan, Sudarshan Guddu, Avtar Singh Sandhu, publicity secretary Ashok Bharti, Bant Singh Barnala, finance secretary Lakshman Das Musafir and assistant financial Secretary Balwinder Singh Thikriwala were elected. After election Dr Ujagar Singh Mann said that executive list of twenty one members will be announced in the next list.

Fourth News:-

Eighty girls were taught the sewing work-Sidhu

Barnala, January 8 (Staff Reporter) – the sewing centre run by malwa cultural and welfare club Barnala distributed the training certificates to ten students. While speaking on this occasion, trust charman Gurjinder Singh Sidhu head former military wing Shiromani Akali Dal and centre administrator Jagsir Singh Chauhan said that this centre has been running for the past ten years without desire for reward and besides the above mentioned students 70 more students have been able to earn their living after learning the sewing work. Specially reached on this occasion the head of truck union Barnala Kulwant Singh Kanta praised the tasks of organization and gave rupees twenty one hundred for help. On this occasion, other than Barnala city council's head Sardar Paramjit Singh Dhillon, the national logical leader Balwinder Barnala, Boby Bansal social servant, former chairman Sukmhinder Singh Sukhi, Jathedar Jarnail Singh Bhotna, Harpalinder Singh Rahi, Sukhjit Kaur Sukhi, chairman of market committee Barnala Karnail Singh Thuthiwal, military wing circle leader Captain Boota Singh Shota, Captain Mahender Singh Maan, Punjabi singer Jassy Bajwa many more leaders of organizations were present. During this sewing machines presented to three poor girls."

Algorithm Output at 30% Compression Ratio:-

ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਚ ਵਿਕਾਸ ਕਾਰਜਾਂ ਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ-ਭਾਨਾ

ਸ਼ਹਿਣਾ, 8 ਜਨਵਰੀ (ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਹਲਕਾ ਵਿਧਾਇਕ ਸੰਤ ਬਲਵੀਰ ਸਿੰਘ ਘੁੰਨਸ ਦੀ ਅਗਵਾਈ ਹੇਠ ਹਲਕੇ ਦੇ ਵਿਕਾਸ ਕਾਰਜਾਂ ਵਿਚ ਬੇਹੱਦ ਤੇਜੀ ਆਈ ਹੈ।

ਅਧਿਆਪਕ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਨੂੰ ਸ਼ਰਧਾਂਜਲੀਆਂ ਭੇਟ

ਧਨੇਲਾ, 8 ਜਨਵਰੀ (ਨਿੱਜੀ ਪੱਤਰ ਪ੍ਰੇਰਕ)-ਸੁਖਵਿੰਦਰ ਸਿੰਘ ਵੜੈਚ ਦੇ ਹੋਣਹਾਰ ਅਧਿਆਪਕ ਪੁੱਤਰ ਗੁਰਦੀਪ ਸਿੰਘ ਵੜੈਚ ਦੀ ਅੰਤਿਮ ਅਰਦਾਸ ਗੁਰਦੁਆਰਾ ਪਾਤਸ਼ਾਹੀ ਨੌਵੀਂ ਵਿਖੇ ਹੋਈ।

ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਹੋਈ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਸਾਹਿਤ ਚਰਚਾ ਮੰਚ ਬਰਨਾਲਾ ਦੀ ਚੋਣ ਗਰੀਨ ਐਵੀਨਿਊ ਦੇ ਪਾਰਕ ਨੇੜੇ ਨਾਨਕਸਰ ਗੁਰਦੁਆਰਾ ਵਿਖੇ ਹੋਈ, ਜਿਸ ਵਿਚ ਸਰਬ ਸੰਮਤੀ ਨਾਲ ਬੂਟਾ ਸਿੰਘ ਚੌਹਾਨ ਅਤੇ ਸੁਰਜੀਤ ਸਿੰਘ ਦਿਹੜ ਸਰਪ੍ਰਸਤ, ਡਾ: ਉਜਾਗਰ ਸਿੰਘ ਮਾਨ ਪ੍ਰਧਾਨ, ਡਾ: ਅਮਨਦੀਪ ਸਿੰਘ ਟੱਲੇਵਾਲੀਆ ਅਤੇ ਕੁਲਵੰਤ ਸਿੰਘ ਧਿੰਗੜ ਮੀਤ ਪ੍ਰਧਾਨ, ਜਨਰਲ ਸਕੱਤਰ ਪਾਲ ਸਿੰਘ ਲਹਿਰੀ, ਸਹਾਇਕ ਜਨਰਲ ਸਕੱਤਰ ਲੈਕਚਰਾਰ ਸੁਖਮਿੰਦਰ ਸਿੰਘ ਸ਼ਹਿਣਾ, ਜਥੇਬੰਦਕ ਸਕੱਤਰ ਬਿੰਦਰ ਖੁੱਡੀ ਕਲਾਂ, ਸੁਦਰਸ਼ਨ ਗੁੱਡੂ ਤੇ ਅਵਤਾਰ ਸਿੰਘ ਸੰਧੂ, ਪ੍ਰਚਾਰ ਸਕੱਤਰ ਅਸ਼ੋਕ ਭਾਰਤੀ ਅਤੇ ਬੰਤ ਸਿੰਘ ਬਰਨਾਲਾ, ਵਿੱਤ ਸਕੱਤਰ ਲਛਮਣ ਦਾਸ ਮੁਸਾਫ਼ਿਰ ਤੇ ਸਹਾਇਕ ਵਿੱਤ ਸਕੱਤਰ ਬਲਵਿੰਦਰ ਸਿੰਘ ਠੀਕਰੀਵਾਲਾ ਚੁਣੇ ਗਏ।

ਅੱਸੀ ਲੜਕੀਆਂ ਨੂੰ ਸਿਲਾਈ ਦਾ ਕੰਮ ਸਿਖਾਇਆ-ਸਿੱਧੂ

ਬਰਨਾਲਾ, 8 ਜਨਵਰੀ (ਸਟਾਫ਼ ਰਿਪੋਰਟਰ)-ਮਾਲਵਾ ਸੱਭਿਆਚਾਰਕ ਅਤੇ ਵੈਲਫੇਅਰ ਕਲੱਬ ਬਰਨਾਲਾ ਵੱਲੋਂ ਚਲਾਏ ਜਾ ਰਹੇ ਸਿਲਾਈ ਸੈਂਟਰ ਦੀਆਂ ਦਸ ਵਿਦਿਆਰਥਣਾਂ ਨੂੰ ਸਿਖਲਾਈ ਸਰਟੀਫਿਕੇਟ ਵੰਡੇ ਗਏ।

The English Translation of above output is as follows:

Under leadership of Ghunnas development activities are highly accelerated-Bhana

Shhina, 8 January (motivational letters) – Under the leadership of local MLA Sant Balbir Singh Ghunnas development activities in the constituency are highly accelerated.

Tributes paid to teacher Gurdeep Singh Vdaich

Dhnaula, January 8 (private motivational letters) - The final prayer for outstanding teacher Gurdeep Singh Vdaich son of Sukhwinder Singh Vdaich was held at Gurudvara ninth Kingdom.

Eelection held for literature discussion forum Barnala

Barnala, January 8 (Staff Reporter) – Election of literature discussion forum held at Gurdwara Nanksar near the park of Green Avenue, in which unanimously Buta Singh Chauhan and Surjit Singh Dehd patron, Dr Ujagar Singh Mann head, Dr: Amandeep Singh Tallewalia, Kulwant Singh Dhingad circle head, General Secretary Pal Singh Lahiri, assistant general secretary lecturer Sukminder Singh Shhina, organisational secretary Binder Khuddi kalan, Sudarshan Guddu, Avtar Singh Sandhu, publicity secretary Ashok Bharti, Bant Singh Barnala, finance secretary Lakshman Das Musafir and assistant financial Secretary Balwinder Singh Thikriwala were elected.

Eighty girls were taught the sewing work-Sidhu

Barnala, January 8 (Staff Reporter) – the sewing centre run by malwa cultural and welfare club Barnala distributed the training certificates to ten students."

As can be seen from output of algorithm of Punjabi summarizer, at 30% compression ratio, mainly the headlines and next lines have been retrieved and at 50% compression ratio, more detailed summary is produced including headlines, lines just next to head lines and other important lines.

IV. RESULTS AND DISCUSSIONS

Punjabi summarization system has been tested over fifty Punjabi multi news documents (Data set containing 6185 sentences and 72689 words) from Punjabi news-corpus. We have applied four Intrinsic measures of summary evaluation 1) F-Score 2) Cosine Similarity 3) Jaccard Coefficient and 4) Euclidean distance and two extrinsic measures of summary evaluation 1) Question Answering Task and 2) Keywords Association Task for

Punjabi multi news documents. Firstly we have produced gold summaries (reference summaries) of these 50 Punjabi multi news articles. For making the gold summaries, three human experts have been assigned the task of producing the manual summaries separately of these 50 documents at 10%, 30% and 50% compression ratios. Finally gold summaries (reference summaries) are produced by including mostly common sentences of three manual summaries produced by three human experts at their respective compression ratios.

As First measure of intrinsic summary evaluation, we have calculated F-Score [23] at respective compression ratios 10%, 30% and 50% for Punjabi news documents and Punjabi stories as follows:

$$\text{F-Score} = \frac{(2 * \text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})}$$

$$\text{Recall} = \frac{\text{Number of correct sentences retrieved by system}}{\text{Total number of sentences retrieved by human expert}}$$

$$\text{Precision} = \frac{\text{Number of correct sentences retrieved by system}}{\text{Total number of sentences retrieved by system.}}$$

As second measure of intrinsic summary evaluation [24], we have calculated Cosine Similarity between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. Using Cosine-similarity-measure, documents are treated as term-vectors and the similarity of two documents corresponds to correlation between the vectors. Given two documents and vectors A and B are the term frequency vectors of these documents for term set T= $\{t_1,\ldots\ldots t_m\}$ Cosine similarity between two vectors is calculated as follows:

$$\text{COSINE\_SIMILARITY}(A, B) = \text{Cos}(\Theta) = (A \cdot B) / (|A|\,|B|) = \sum A_i \times B_i / \sqrt{\sum (A_i)^2} \times \sqrt{\sum (B_i)^2}$$

where i= 1 to n

Each dimension denotes the term with its frequency in the document and is non negative. The value of cosine similarity is non-negative and lies from 0 to 1. If cosine-similarity for two documents is closer to one, it means these two documents are very much similar to each other. For dissimilar type of documents cosine similarity is approaching towards zero. We have computed Cosine-similarity between our gold summary (reference summary) and summary produced by our Punjabi summarization system.

As third measure of intrinsic summary evaluation, we have calculated Jaccard-coefficient between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. The Jaccard-coefficient measures similarity as the intersection divided by the union of the objects. Given two documents and vectors A and B are the term frequency vectors of these documents over the term set T= $\{t_1,\ldots\ldots, t_m\}$ then Jaccard-coefficient is calculated as follows:

$$\text{Jaccard Coefficient} = \text{SIM}(A, B) = (A \cdot B) / (|A|^2 + |B|^2 - A \cdot B) = (A \cdot B) / (\sqrt{\sum (A_i)^2} \times \sqrt{\sum (A_i)^2} + \sqrt{\sum (B_i)^2} \times \sqrt{\sum (B_i)^2} - A \cdot B)$$ Where i= 1 to n

Where each dimension represents a term with its frequency in the document. The value of Jaccard-coefficient-measure ranges from 0 to 1. If value of Jaccard-coefficient is approaching towards one then two documents are almost similar. If value of Jaccard-coefficient is approaching towards zero then two documents are dissimilar.

As fourth measure of intrinsic summary evaluation, we have calculated Euclidean distance between our system produced summary and gold summary at respective compression ratios for Punjabi news documents and Punjabi stories. Measuring distance between text documents, given two documents with their key term frequency vectors X_{ik} and X_{jk} respectively, where k= 1 to n key terms. The Euclidean distance of the two documents is defined as follows:

$$\text{Euclidean distance}(X_{ik}, X_{jk}) = \left(\sum (X_{ik} - X_{jk})^2 \right)^{1/2}$$ for k=1 to n key terms.

The results of intrinsic summary evaluation are shown in Table II. and Figure 5 at respective compression ratios.

TABLE II.
RESULTS OF INTRINSIC SUMMARY EVALUATION

Compression Ratio (In %)	Intrinsic Summary Evaluation for Punjabi News Documents			
	Avg. F-score (In %)	Avg. Cosine Similarity	Avg. Jaccard Coeff.	Avg. Euclidean Distance
10%	97.87	0.98	0.97	0.12
30%	95.32	0.96	0.95	0.32
50%	94.63	0.95	0.94	0.56

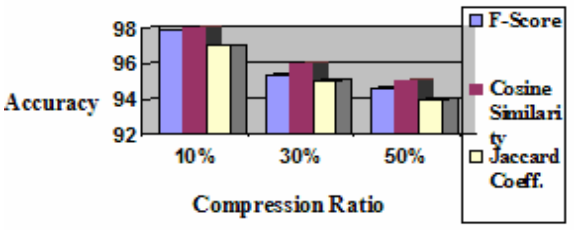

Figure 5. Intrinsic summary evaluation task

As can be seen from Table II and Figure 5, at 10% compression ratio, Average F-Score, Average Cosine Similarity and Average Jaccard Coefficient values are very high and Average Euclidean distance is very low because at 10%, usually few important sentences are extracted including headlines and next lines. Headlines and next lines are sufficient to describe the complete news document. The values of average F-Score, average cosine similarity and average Jaccard Coefficient are in descending order of compression ratios for Punjabi news

documents. Average value of Euclidean distance is in ascending order of compression ratios for Punjabi news documents. Few errors are due to presence of those sentences which contain many names of persons, but actually these sentences are not important.

Extrinsic measures [25] of summary evaluation are task oriented. We have performed question answering task and keywords association task as extrinsic measures of summary evaluation at compression ratios 10%, 30% and 50% respectively for Punjabi multi news documents. For performing the task of question answering, firstly three human experts have been given fifty multi news documents and then they jointly prepared five questions for each of fifty documents. Then answers of these questions are looked into system produced summary. For each correct answer, counter for number of correct answers is incremented by one for that document. Accuracy for performing task of question answering is calculated as follows:

Accuracy= No. of correct answers/ Total No. of questions asked

In keywords association task, keywords are the key terms which can represent the theme of whole document. For performing this task, firstly five keywords (gold keywords) have been extracted from source text by human experts and then these gold keywords have been associated with the summary produced by summarization system. Accuracy for performing task of keyword association is calculated as follows:

Accuracy= No. of gold keywords present in summary/ Total No. of gold keywords

The results of extrinsic summary evaluation are shown in Table III and Figure 6 at respective compression ratios.

TABLE III.
RESULTS OF EXTRINSIC SUMMARY EVALUATION

Compression Ratio (In %)	Extrinsic summary evaluation for Punjabi multi news documents	
	Accuracy of Question Answering Task (In %)	Accuracy of Keywords Association Task (In %)
10%	78.95	80.13
30%	81.38	92.37
50%	88.75	96.32

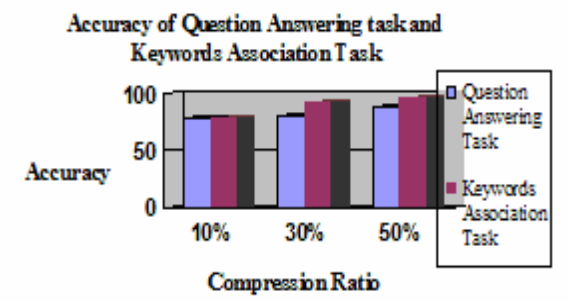

Figure 6. Extrinsic summary evaluation task

As can be seen from Table III and Figure 6 that at 10% compression ratio, Performance of multi news Punjabi Text Summarization System is low because news documents are usually short and at 10% CR, mainly headlines and lines just next to headlines are extracted which are not sufficient to give all answers of question-answering task. At 30% compression ratio, Punjabi Text Summarization System is able to give answers of 81.38% questions for Punjabi news documents. At 50% compression ratio, Punjabi Text Summarization System is able to give answers of 88.75% questions. Task of question answering is performed very well at 50% compression ratio with summary produced by Punjabi Text Summarization system because summary produced is enough to give answers of majority of questions.

At 10% compression ratio, average of 80.13% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. At 30% compression ratio, average of 92.37% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. At 50% compression ratio, average of 96.32% gold keywords are found in summary produced by Punjabi Text Summarization System for fifty Punjabi news documents. The accuracy percentage for the task of keywords association is low at 10% compression ratio because at 10% compression ratio, summary usually contains headlines and next lines and only few gold keywords are found in headlines and next lines. But at 50% compression ratio, the task of keywords association is performed very well because summary produced is enough to cover majority of gold keywords. The snap shot of Single document multi news Punjabi summarization system is given in Figure 7.

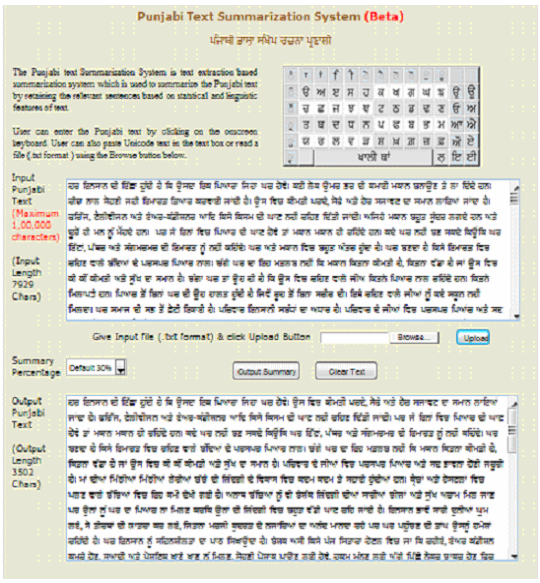

Figure 7. Web based online Punjabi text summarization system

V. COMPARISON OF PUNJABI TEXT SUMMARIZER WITH EXISTING INDIAN SUMMARIZERS

TABLE IV.
PERFORMANCE COMPARISON

Summarization Systems	Performance Comparison	
	Accuracy (In %)	Test Used
Single document multi news Punjabi Summarization System	For Multi News Single documents: F-Score = 95.32% Cosine Similarity= 96% Question Answering task with accuracy=81.38% (At 30% Compression Ratio)	Intrinsic and Extrinsic Summary Evaluation
Bengali Summarizer using Textual Images [14]	56%	Efficiency
Bengali Summarizer using Text Extraction [15]	84% (At 40% Compression Ratio)	Efficiency
Topic based Bengali Opinion Summarizer[16]	69.65%	F-Score
Multi Lingual Summarizer for English, Hindi, Gujarati & Urdu [17]	82%	Efficiency
Document Summarizer for Kannada [18]	For Literature: 70% For Entertainment: 80% For Sports: 76%	Efficiency

We can seen from Table IV, that performance of Punjabi Text Summarizer is reasonably good as compared with performance of other existing summarizers for Indian languages.

VI. CONCLUSONS

Single-document multi-news Punjabi Summarization system is first of its kind Punjabi summarizer and is available online at http://pts.learnpunjabi.org/. We have developed a number of lexical resources from scratch used in Punjabi text summarization such as Punjabi stemmer, Punjabi nouns normalizer, Punjabi named entity recognition, Punjabi Keywords Identification, Punjabi proper names list, common English-Punjabi nouns list, Punjabi stop words list, Punjabi suffix and prefix list, Punjabi cue phrase list etc. We have done thorough analysis of Punjabi corpus, Punjabi dictionary and Punjabi noun-morph for developing these resources. These Punjabi resources have been developed for the first time and these might be helpful for developing other NLP applications for Punjabi language.

REFERENCES

[1] F. Kyoomarsi, H.Khosravi, E. Eslami, P.K. Dehkordy, "Optimizing text summarization based on fuzzy logic", *In Seventh IEEE/ACIS International Conference on Computer and Information Science*, University of Shahid Bahonar Kerman, UK,pp. 347-352, 2008.

[2] V. Gupta and G.S. Lehal, "A Survey of Text Summarization Extractive Techniques", *International Journal of Emerging Technologies in Web Intelligence* vol.2, no.3, pp. 258-268, 2010.

[3] J. Lin, "Summarization", *In Encyclopedia of Database Systems*, Springer-Verlag Heidelberg, Germany, 2009.

[4] V. Gupta and G.S. Lehal, "Pre processing Phase of Punjabi Language Text Summarization", *In International Conference on Information Systems for Indian Languages Communications in Computer and Information Science*, Springer-Verlag Berlin Heidelberg, pp. 250–253, 2011.

[5] M.A. Fattah and F. Ren, "Automatic Text Summarization" *In World Academy of Science Engineering and Technology* vol. 27, pp.192-195, 2008.

[6] K. Kaikhah, "Automatic Text Summarization with Neural Networks", *In IEEE international Conference on intelligent systems*, Texas, USA, pp.40-44, 2004.

[7] J.L. Neto, A.D. Santos, C.A.A. Kaestner and A.A. Freitas, "Document Clustering and Text Summarization", *In 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, London, pp.41-55, 2000.

[8] V. Gupta and G.S. Lehal, "Features Selection and Weight learning for Punjabi Text Summarization", *In International Journal of Engineering Trends and Technology*, vol. 2, issue.2, pp. 45-48, 2011.

[9] V. Gupta and G.S. Lehal, "Named Entity Recognition for Punjabi Language Text Summarization", *In International Journal of Computer Applications*, vol.33 no.3, pp.28-32, 2011.

[10] M.S. Gill and G.S. Lehal, "Part of Speech Tagging for Grammar Checking of Punjabi", I*n The Linguistic Journal* vol.4, no.1, pp.6-21, 2009.

[11] V. Gupta and G.S. Lehal, "Automatic Keywords Extraction for Punjabi Language", *In International Journal of Computer Science,* vol. 8, issue 5, pp.327-331, 2011.

[12] V. Gupta and G.S. Lehal, "Punjabi language stemmer for nouns and proper nouns", *In the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) IJCNLP*, Chiang Mai, Thailand, pp.35-39, 2011.

[13] V. Gupta and G.S. Lehal, "Complete Pre processing Phase of Punjabi Language Text Summarization", *In International Conference on Computational Linguistics COLING-2012*, IIT Bombay, India, pp.199-205, 2012.

[14] U. Garain, A.K. Datta, U. Bhattacharya and S.K. Parui, "Summarization of JBIG2 Compressed Indian Textual Images", *In Proceeding of 18th International Conference on Pattern Recognition (ICPR'06)*, IEEE, Kolkata, India, 2006.

[15] K. Sarkar, "Bengali text summarization by sentence extraction", *In Proceedings of International Conference on Business and Information Management ICBIM'12*, NIT Durgapur, pp.233-245, 2012.

[16] A. Das and S. Bandyopadhyay, "Topic-Based Bengali Opinion Summarization", *COLING'10*, Beijing, China, pp.232–240, 2010.

[17] A. Patel, T. Siddiqui and U.S. Tiwari, "A language independent approach to multilingual text summarization", *In Proceedings of IEEE international conference RIAO2007*, Pittsburgh PA, U.S.A, 2007.

[18] R. Jayashree, M.K. Srikanta, K. Sunny, "Document Summarization in Kannada using Keyword Extraction", *In Proceedings of AIAA'11,CS & IT 03*, pp.121–127, 2011.

[19] M.Z. Islam, M.N. Uddin and M. Khan, "A light weight stemmer for Bengali and its Use in spelling Checker", *In Proceedings of 1st International Conference on Digital Comm. and Computer Applications (DCCA 2007)*, Irbid, Jordan, PP.19-23, 2007.

[20] A. Ramanathan and D. Rao, "A Lightweight Stemmer for Hindi", *In Workshop on Computational Linguistics for South-Asian Languages, EACL'03*, 2003.
[21] G. Singh, M.S. Gill and S.S. Joshi, "Punjabi to English Bilingual Dictionary", Punjabi University Patiala, India, 1999.
[22] V. Gupta and G.S. Lehal, "Automatic Punjabi Text Extractive Summarization System", *In International Conference on Computational Linguistics COLING-2012*, IIT Bombay, India, pp.191-198, 2012.
[23] H. Nanba and M. Okumura, "Some Examinations of Intrinsic Methods for Summary Evaluation Based on the Text Summarization Challenge", *In Proceedings of international conference on language resources and evaluation LREC'02*, pp.739-746, 2002.
[24] A. Huang, "Similarity Measures for Text Document Clustering", *In the Proceedings of New Zealand Computer Science Research Conference*, Christchurch New Zealand, pp.49-56, 2008.
[25] M. Hassel, "Evaluation of Automatic Text Summarization", *Licentiate Thesis*, Stockholm, Sweden, pp.1-75, 2004.

AUTHORS' INFORMATION

Dr. Vishal Gupta is Senior Assistant Professor in Computer Science & Engineering, University Institute of Engineering & Technology, Panjab University Chandigarh, India. He did his BTech. in Computer Science & Engineering from Shaheed Bhagat Singh College of Engineering & Technology, Ferozepur, Punjab in 2003. He did his M.Tech. and completed his Ph. D. in Computer Science & Engineering from Punjabi University Patiala in 2005 and 2013 respectively. He is among University toppers. He is winner of Young Scientist Award-2013 in Engineering & Technology at Punjab Science Congress. He has written around 41 research papers in reputed international and national journals and conferences. He has developed a number of research projects in field of NLP including synonyms detection, automatic question answering and text summarization etc. One of his research paper on Punjabi language text processing was awarded as best research paper by Dr. V. Raja Raman at an International Conference at Panipat. He is also a merit holder in 10^{th} and 12^{th} classes of Punjab School education board.

Professor Gurpreet Singh Lehal received undergraduate degree in Mathematics in 1988 from Panjab University, Chandigarh, India, and Post Graduate degree in Computer Science in 1995 from Thapar Institute of Engineering & Technology, Patiala, India and Ph. D. degree in Computer Science from Punjabi University, Patiala, in 2002. He joined Thapar Corporate R&D Centre, Patiala, India, in 1988 and later in 1995 he joined Department of Computer Science at Punjabi University, Patiala. He is actively involved both in teaching and research. His current areas of research are- Natural Language Processing and Optical Character recognition. He has published more than 25 research papers in various international and national journals and refereed conferences. He has been actively involved in technical development of Punjabi and has to his credit the first Gurmukhi OCR, Punjabi word processor with spell checker and various transliteration software. He was the chief coordinator of the project "Resource Centre for Indian Language Technology Solutions- Punjabi", funded by the Ministry of Information Technology as well as the coordinator of the Special Assistance Programme (SAP-DRS) of the University Grants Commission (UGC), India. He was also awarded a research project by the International Development Research Centre (IDRC) Canada for Shahmukhi to Gurmukhi Transliteration Solution for Networking.