Complete Pre Processing phase of Punjabi Text Extractive Summarization System

Vishal GUPTA[1] Gurpreet Singh LEHAL[2]
(1) UIET, Panjab University Chandigarh, India
(2) Department of Computer Science, Punjabi University Patiala, Punjab, India
vishal@pu.ac.in, gslehal@gmail.com

ABSTRACT

Text Summarization is condensing the source text into shorter form and retaining its information content and overall meaning. Punjabi text Summarization system is text extraction based summarization system which is used to summarize the Punjabi text by retaining the relevant sentences based on statistical and linguistic features of text. It comprises of two main phases: 1) Pre Processing 2) Processing. Pre Processing is structured representation of the original Punjabi text. In Processing, final score of each sentence is determined using feature-weight equation. Top ranked sentences in proper order are selected for final summary. This paper concentrates on complete pre processing phase of Punjabi text summarization system. Pre processing phase includes Punjabi words boundary identification, Punjabi sentences boundary identification, Punjabi stop words elimination, Punjabi language stemmer for nouns and proper names, allowing input in proper format and elimination of duplicate sentences.

KEYWORDS : Punjabi Text Summarization System, Pre Processing Phase, Punjabi stemmer for nouns and proper nouns, Natural Language Processing

1 Introduction

Automatic text summarization (Kyoomarsi et al., 2008) is reducing the source text into a shorter form retaining its information content and overall meaning. The goal of automatic text summarization is to present most important contents from information source to the user in a shorter version. Text Summarization (Gupta & Lehal, 2010) methods can be classified into abstractive and extractive summarization. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. Extractive summary deals with selection of important sentences from the original text. The importance of sentences is decided based on statistical and linguistic features of sentences. Text Summarization Process can be divided into two phases: 1) Pre Processing phase (Gupta & Lehal, 2011a) is structured representation of the original text. 2) In Processing (Fattah & Ren, 2008; Kaikhah, 2004; Neto, 2000) phase, final score of each sentence is determined using feature-weight equation. Top ranked sentences in proper order are selected for final summary. This paper concentrates on complete pre processing of Punjabi text extractive summarization system. Punjabi text Summarization system is text extraction based summarization system which is used to summarize the Punjabi text by retaining the relevant sentences based on statistical and linguistic features of text. Punjab is one of Indian states and Punjabi is its official language. Punjabi is spoken in India, Pakistan, USA, Canada, England, and other countries with Punjabi immigrants. Punjabi is written in 'Gurmukhi' script in eastern Punjab (India), and in 'Shahmukhi' script in western Punjab (Pakistan). For some of Indian languages like Hindi, Bengali etc. a number of automatic text summarization systems are available. For Punjabi, the only text summarization system available is online:http://pts.learnpunjabi.org/default.aspx and no other Punjabi summarizer is available in the world. Pre processing phase includes Punjabi words boundary identification, Punjabi sentences boundary identification, Punjabi stop words elimination, Punjabi language stemmer for nouns and proper names, allowing input in proper format, elimination of duplicate sentences and normalization of Punjabi noun words in noun morph.

2 Complete Pre Processing Phase of Punjabi Text Summarization System

Various Sub phases for complete pre processing of Punjabi text summarization system are given below:

2.1 Punjabi language stop words elimination

Punjabi language stop words (Gupta & Lehal, 2011a) are most frequently occurring words in Punjabi text like: ਦੇ dē, ਹੈ hai, ਨੂੰ nūṃ and ਨਾਲ nāl etc. We have to eliminate these words from the original text otherwise, sentences containing them can get influence unnecessarily. We have made a list of Punjabi language stop words by creating a frequency list from a Punjabi corpus. Analysis of Punjabi corpus taken from popular Punjabi newspaper Ajit has been done. This corpus contains around 11.29 million words and 2.03 lakh unique words. We manually analyzed these unique words and identified 615 stop words. In the corpus of 11.29 million words, the frequency count of these stop words is 5.267 million, which covers 46.64% of the corpus.

Sample input sentence:-

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

gharēlū gais dī samssiā pahil dē ādhār tē hall hōvēgī-thind

Sample output sentence:-

ਘਰੇਲੂ ਗੈਸ ਸਮੱਸਿਆ ਪਹਿਲ ਆਧਾਰ ਹੱਲ-ਥਿੰਦ

gharēlū gais samssiā pahil ādhār hall –thind

As we can see from the sample input and output of Punjabi stop words elimination sub phase that four stop words (ਦੀ dī, ਦੇ dē, ਤੇ tē, ਹੋਵੇਗੀ hōvēgī) have been eliminated from the sample output sentence.

2.2 Punjabi language stemmer for nouns and proper nouns

The purpose of stemming (Islam et al., 2007; Kumar et al., 2005; Ramanathan & Rao, 2003) is to obtain the stem or radix of those words which are not found in dictionary. If stemmed word is present in dictionary (Singh et al., 1999) , then that is a genuine word, otherwise it may be proper name or some invalid word. In Punjabi language stemming (Gupta & Lehal, 2011b; Gill et al., 2007; Gill et al., 2009) for nouns and proper names, an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi noun morph and proper names list. An in depth analysis of corpus was made and the eighteen possible noun and proper name suffixes were identified like ◌ੀਆਂ īāṃ, ◌ਿਆਂ Iāṃ, ◌ੁਆਂ ūāṃ, ◌ਾਂ āṃ, ◌ੀਏ Īē, ◌ੇ Ē and ◌ੀਓ Īō etc.

Proper names are the names of person, place and concept etc. not occurring in Punjabi dictionary. Proper names play an important role in deciding a sentence's importance. From the Punjabi corpus, 17598 words have been identified as proper names. The percentage of these proper names words in the Punjabi corpus is about   13.84 %. Some of Punjabi language proper names are ਅਕਾਲੀ akālī, ਲੁਧਿਆਣਾ ludhiāṇā, ਬਾਦਲ bādal and ਪਟਿਆਲਾ paṭiālā etc.

Algorithm of Punjabi language stemmer for nouns and proper names is as below:

The algorithm of Punjabi language stemmer (Gupta & Lehal, 2011b) 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 current Punjabi word ends with ◌ੀਆਂ īāṃ, ◌ਿਆਂ iāṃ, ◌ੁਆਂ ūāṃ then remove ਆਂ āṃ from end.

Step 2: Else If current Punjabi word ends with ◌ੀਏ īē then remove ਏ ē from end.

Step 3: Else If current Punjabi word ends with ◌ੀਓ Īō then remove ਓ ō from end.

Step 4: Else If current Punjabi word ends with ◌ੀਆ ā, ਈਆ īā then remove ਆ ā from end.

Step 5: Else If current Punjabi word ends with ਈ ī, ਵਾਂ vāṃ, ◌ਾਂ āṃ, ◌ੋਂ ōṃ, ◌ੀਂ īṃ and
ਜ/ਜ਼/ਸ ja/z/s then remove the corresponding suffix from end.

Step 6: Else If current Punjabi word ends with ◌ੇ ē, ◌ਿਓ iō, ◌ੋ ō, ◌ਿਉਂ iuṃ and ◌ਿਆ iā then
remove the corresponding suffix and add kunna at the end.

Step 7: Current Punjabi Stemmed word is checked against Punjabi noun morph or Proper names
list. If found, It is Punjabi noun or Punjabi Proper name.

Algorithm Input: ਫੁੱਲਾਂ phullāṃ (Flowers) and ਲੜਕੀਆਂ laṛkīāṃ (Girls)

Algorithm Output: ਫੁੱਲ phull (Flower) and ਲੜਕੀ laṛkī (Girl)

An in depth analysis of output of Punjabi language stemmer for nouns and proper names has been done over 50 Punjabi documents of Punjabi news corpus of 11.29 million words. The efficiency of Punjabi language noun and Proper name stemmer is 87.37%, which is tested over 50 Punjabi news documents of corpus and is ratio of actual correct results to total produced results by stemmer.

2.3 Normalization of Punjabi nouns in noun morph

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 noun words **ਤਿੱਬਤੀ** tibbtī, **ਥਾਂ** thāṃ and **ਦਸ਼ਾ** dasaā can also be written as **ਤਿਬਤੀ** tibtī, **ਥਾ** thā and **ਦਸਾ** dasā respectively. To overcome this problem Punjabi noun morph has been normalized for different spelling variations of same Punjabi noun words.

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 occurences 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.

Step 5 : 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

An exhaustive analysis has been done on fifty Punjabi news documents for normalization of Punjabi nouns and it is discovered that very less spelling variations are found. Only 1.562% noun words show the variations in their spellings. TABLE 1 shows that out of these 1.562% words, percentage of words having one, two or three variations:

Number of Variants	Words Frequency (%)	Example
1	99.95	ਪੰਜਾਲ਼ੀ pañjālī, ਪੰਜਾਲੀ pañjālī
2	0.046	ਉੱਖਲ਼ੀ ukkhlaī, ਉਖਲ਼ੀ ukhlaī , ਉੱਖਲੀ ukkhlī
3	0.004	ਅੰਗਰੇਜੀ aṅgrējī, ਅੰਗਰੇਜ਼ੀ aṅgrēzī, ਅੰਗ੍ਰੇਜੀ aṅgrējī, ਅੰਗ੍ਰੇਜ਼ੀ aṅgrēzī

TABLE 1 – Percentage Word Occurrence with Spelling Variations Count

Thus, above table represents that, the variations found for majority of the words is just 1 and in worst case, it can go up to 3. And no case has been found with more than three spelling variants.

2.4 Allowing input restrictions to input text

Punjabi Text Summarization system allows Unicode based Gurmukhi text as input. Gurmukhi is the most common script used for writing the Punjabi language. Punjabi Text Summarization system can accept maximum upto 1,00000 characters as input otherwise it will give error message. Majority of input characters should be of Gurmukhi otherwise error will be printed.

Algorithm:

Step 1 : If Uploaded input file is in Unicode based .txt format then calculate input character length and go to step 2, otherwise display the error message “Can not accept input of this type!!!”

Step 2 : If input character length> 1,00000 characters then display error message “Input length exeeds the limit” otherwise go to step 3.

Step 3 : From the input text, calculate length of Gurmukhi characters, Punctuation mark characters, numeric characters, English characters and other characters.

If Gurmukhi characters length is less than equal to Punctuation character length or numeric characters length or English Characters length or other characters length then display error message “Can not accept the input!!!”

Else If English characters length or other characters length is greater than equal to 10% of total input characters length then display error message “Can not accept the input!!!”

Else Go to Stop words elimination phase.

2.5 Elimination of duplicate sentences from Punjabi input text

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. Some summarization systems delete the duplicate sentences in the output summary and other systems delete them in the input itself. It is desirable to delete the duplicate sentences from input because numbers of input sentences are reduced and processing phase takes less time. Punjabi text summarization system eliminates the duplicate sentences from the input Punjabi text. An exhaustive analysis has been done on fifty Punjabi news documents for determining the frequency of duplicate sentences and it is discovered 9.60% sentences are duplicate. Minimum frequency of a duplicate sentence in a single Punjabi news document is two, maximum frequency is four and average frequency is three. 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. 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.

3 Pre processing algorithm for Punjabi text summarization system

The algorithm for complete Pre Processing Phase (Gupta & Lehal 2011a) proceeds by checking input Punjabi text into proper format and segmenting it into sentences and words. Set the scores of each sentence as 0. Normalize the Punjabi noun morph for different spelling variations of nouns. For each word of every sentence follow step 1 and step 2:

Step 1 : If current Punjabi word is stop word then delete all the occurrences of it from current sentence.

Step 2 :If current Punjabi word is not present in Punjabi dictionary, Punjabi noun morph, common English-Punjabi nouns list, Punjabi proper nouns list then apply Punjabi Noun and proper noun Stemmer for the possibility of Punjabi noun or proper noun.

Step 3: Delete redundant (duplicate) sentences from input text, to prevent them occurring in final summary and produce output of preprocessing phase.

TABLE 2 shows sample input and output sentences, for pre processing algorithm.

Input Punjabi sentence	Output Punjabi sentence
ਮੁੱਖ ਮੰਤਰੀ ਨੇ ਕਿਹਾ ਕਿ ਉਹ ਅੱਜ ਕਾਂਗਰਸ ਉਮੀਦਵਾਰ ਭਰਤ ਸਿੰਘ ਬੈਲੀਵਾਲ ਲਈ ਵੋਟਾਂ ਦੀ ਅਪੀਲ ਕਰਨ ਲਈ ਆਏ ਹਨ ਪਰ ਵੋਟ ਪਾਉਣ ਤੋਂ ਪਹਿਲਾਂ ਪਾਰਟੀ ਦੀ ਨੀਤੀ, ਨਿਯਤ ਤੇ ਨੇਤਾ ਬਾਰੇ ਜਰੂਰ ਵਿਚਾਰ ਦੀ ਲੋੜ ਹੈ। mukkh mantrī nē kihā ki uh ajj kāṅgras umīdvār bharat siṅgh bailīvāl laī vōṭāṃ dī apīl karan laī āē han par vōṭ pāuṇ tōṃ pahilāṃ pāraṭī dī nītī, niyat tē nētā bārē jarūr vicār dī lōṛ hai.	ਮੁੱਖ ਮੰਤਰੀ ਅੱਜ ਕਾਂਗਰਸ ਉਮੀਦਵਾਰ ਭਰਤ ਸਿੰਘ ਬੈਲੀਵਾਲ ਵੋਟ ਅਪੀਲ ਵੋਟ ਪਹਿਲ ਪਾਰਟੀ ਨੀਤੀ ਨਿਯਤ ਨੇਤਾ ਜਰੂਰ ਵਿਚਾਰ ਲੋੜ mukkh mantrī ajj kāṅgras umīdvār bharat siṅgh bailīvāl vōṭ apīl vōṭ pahil pāraṭī nītī niyat nētā jarūr vicār lōṛ

TABLE 2 – Pre processing algorithm sample input and output sentences

A through analysis of result of pre processing phase has been done on fifty Punjabi news documents and stories and it is discovered, that with pre processing phase there is gain in 32% efficiency of Punjabi Text Summarization system at 50% compression ratio.

Conclusion

Punjabi text summarization system is first of its kind Punjabi summarizer and is available online at http://pts.learnpunjabi.org/default.aspx. In this paper, we have discussed the complete pre processing phase for Punjabi text summarization system. Most of the lexical resources used in pre processing such as Punjabi stemmer, Punjabi nouns normalizer, Punjabi proper names list, common English-Punjabi nouns list, Punjabi stop words list etc. had to be developed from scratch as no work was done previously in that direction. For developing these resources an in-depth analysis of Punjabi corpus, Punjabi dictionary and Punjabi morph had to be carried out using manual and automatic tools. This is first time that these resources have been developed for Punjabi and these can be beneficial for developing other NLP applications for Punjabi.

References

Ananthakrishnan Ramanathan and Durgesh Rao, (2003). A Light Weight Stemmer for Hindi. In Workshop on Computational Linguistics for South-*Asian Languages, EACL'03*.

Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy. (2008). Optimizing Text Summarization Based on Fuzzy Logic. In: proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, pages 347-352, University of Shahid Bahonar Kerman, UK.

Gurmukh Singh, Mukhtiar S. Gill and S.S. Joshi, (1999). Punjabi to English Bilingual Dictionary. Punjabi University Patiala, India.

Khosrow Kaikhah, (2004). Automatic Text Summarization with Neural Networks. In Proceedings of IEEE international Conference on intelligent systems, pages 40-44, Texas, USA.

Mandeep Singh Gill, Gurpreet Singh Lehal and S. S. Gill, (2007). A full form lexicon based Morphological Analysis and generation tool for Punjabi. International Journal of Cybernatics and Informatics, pages 38-47, Hyderabad, India.
http://www.advancedcentrepunjabi.org/punjabi_mor_ana.asp

Mandeep Singh Gill, Gurpreet Singh Lehal and S.S. Joshi, (2009). Part of Speech Tagging for Grammar Checking of Punjabi, Linguistic Journal, 4(1): 6-21, Road Town, Tortola British Virgin Islands.

Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan, (2007). 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), pages 19-23, Irbid, Jordan.

Mohamed Abdel Fattah and Fuji Ren (2008). Automatic Text Summarization. In: Proceedings of World Academy of Science Engineering and Technology, volume 27, pages 192-195.

Joel L. Neto, (2000). Document Clustering and Text Summarization. In: Proceedings of 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, pages 41-55, London.

Praveen Kumar, S. Kashyap, Ankush Mittal and Sumit Gupta, (2005). : A Hindi question answering system for E-learning documents. In Proceedings of International Conference on Intelligent sensing and Information processing, pages 80-85, Banglore, India

Vishal Gupta and Gurpreet Singh Lehal, (2010). A Survey of Text Summarization Extractive Techniques. In International Journal of Emerging Technologies in Web Intelligence, 2(3): 258-268.

Vishal Gupta and Gurpreet Singh Lehal, (2011a). Preprocessing Phase of Punjabi Language Text Summarization. In Proceedings of International conference on Information Systems for Indian Languages Communications in Computer and Information Science ICISIL2011, pages 250–253, Springer-Verlag Berlin Heidelberg.

Vishal Gupta and Gurpreet Singh Lehal, (2011b). Punjabi language stemmer for nouns and proper nouns. In proceedings of the 2nd Workshop on South and Southeast Asian Natural Language Processing (WSSANLP) IJCNLP, pages 35-39, Chiang Mai, Thailand.