Preprocessing Phase of Punjabi language Text Summarization

Vishal Gupta[1]and Gurpreet Singh Lehal[2]
[1]Assistnt Professor, Computer Science & Engineering,
University Institute of Engineering & Technology,
Panjab University Chandigarh, India, vishal@pu.ac.in

[2]Professor, Department of Computer Science,
Punjabi University Patiala, Punjab, India, gslehal@yahoo.com

Abstract. Punjabi Text Summarization is the process of condensing the source Punjabi text into a shorter version, preserving its information content and overall meaning. It comprises two phases: 1) Pre Processing 2) Processing. Pre Processing is structured representation of the Punjabi text. This paper concentrates on Pre processing phase of Punjabi Text summarization. Various sub phases of pre processing are: Punjabi words boundary identification, Punjabi language stop words elimination, Punjabi language noun stemming, finding Common English Punjabi noun words, finding Punjabi language proper nouns, Punjabi sentence boundary identification, and identification of Punjabi language Cue phrase in a sentence.

Keywords: Punjabi text summarization, Pre Processing, Punjabi Noun stemmer

1 Introduction to Text Summarization

Text Summarization[1][2] is the process of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. Abstractive Text Summarization is understanding the original text and retelling 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 [2] is structured representation of the original text. Various features influencing the relevance of sentences are calculated. 2)In Processing [3][4][13] phase, final score of each sentence is determined using feature-weight equation. Top ranked sentences are selected for final summary. This paper concentrates on Pre processing phase, which has been implemented in VB.NET at front end and MS Access at back end using Unicode characters [5].

2. Pre Processing phase of Punjabi Text Summarization

2.1 Punjabi Language Stop Word Elimination

Punjabi language Stop words are frequently occurring words in Punjabi text. We have to eliminate these words from original text, otherwise, sentences containing them can get influence unnecessarily. We have made a list of Punjabi language stop

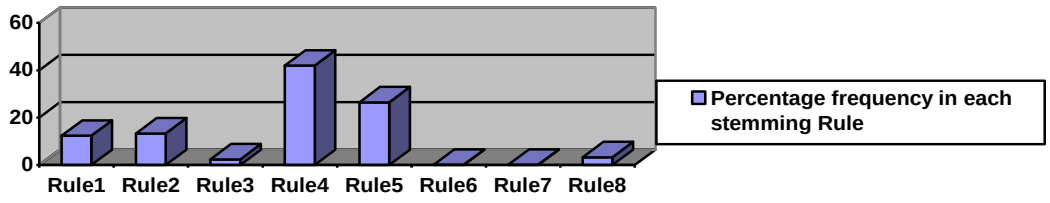
words by creating a frequency list from a Punjabi corpus. Analysis of Punjabi corpus taken from popular Punjabi newspapers has been done. This corpus contains around 11.29 million words and 2.03 lakh unique words [11]. We manually analyzed unique words and identified 615 stop words. In corpus of 11.29 million words, the frequency count of stop words is 5.267 million, which covers 46.64% of the corpus. Some commonly occurring stop words are ਦੀ dī, ਤੋਂ tōṃ, ਕਿ ki ,ਅਤੇ atē, ਹੈ hai, ਨੇ nē etc.

2.2 Punjabi Language Noun Stemming

The purpose of stemming [6][7] is to obtain the stem or radix of those words which are not found in dictionary. If stemmed word is present in dictionary, then that is a genuine word, otherwise it may be proper name or some invalid word. In Punjabi language noun stemming[9][10][14], an attempt is made to obtain stem or radix of a Punjabi word and then stem or radix is checked against Punjabi dictionary [8], if word is found in the dictionary, then the word's part of speech is checked to see if the stemmed word is noun. An in depth analysis of corpus was made and various possible noun suffixes were identified like ੀਆਂ īāṃ, ਿਆਂ iāṃ, ੂਆਂ ūāṃ, ਾਂ āṃ, ੀਏ īē etc. and the various rules for noun stemming have been generated. Some rules of Punjabi noun stemmer are ਫੁੱਲਾਂ phullāṃ → ਫੁੱਲ phull with suffix ਾਂ āṃ, ਲੜਕੀਆਂ laṛkīāṃ → ਲੜਕੀ laṛkī with suffix ੀਆਂ īāṃ, ਮੁੰਡੇ muṇḍē → ਮੁੰਡਾ muṇḍā with suffix ੇ ē etc.

An In depth analysis of output is done over 50 Punjabi documents. The efficiency of Punjabi language noun stemmer is 82.6%. The accuracy percentage of correct words detected under various rules of stemmer are: ੀਆਂ īāṃ rule1 86.81%, ਿਆਂ iāṃ rule2 95.91%, ੂਆਂ ūāṃ rule3 94.44%,ਾਂ āṃ rule4 92.55%, ੇ ē rule5 57.43%, ੀਂ īṃ rule6 100%, ੋਂ ōṃ rule7 100% and ਵਾਂ vāṃ rule8 79.16%. Errors are due to rules violation or dictionary errors or due to syntax mistakes. Dictionary errors are those errors in which, after noun stemming, stem word is not present in noun dictionary, but actually it is noun. Syntax errors are those errors, in which input Punjabi word is having some syntax mistake, but actually that word falls under any of stemming rules. Overall error % age, due to rules voilation is 9.78%, due to dictionary mistakes is 5.97% and due to spelling mistakes is 1.63%.
Graph1 depicts the percentage usage of the stemming

Graph1. Percentage Frequency of various Stemming Rules



2.3 Finding Common English-Punjabi noun words from Punjabi Corpus

Some English words are now commonly being used in Punjabi. Consider a sentence such as ਟੈਕਨਾਲੋਜੀ ਦੇ ਯੁੱਗ ਵਿਚ ਮੋਬਾਈਲ *Technology de yug vich mobile.* It contains ਟੈਕਨਾਲੋਜੀ *Technology* and ਮੋਬਾਈਲ *mobile* as English-Punjabi nouns. These should obviously not be coming in Punjabi dictionary. These are helpful in deciding sentence importance. After analysis of Punjabi corpus, 18245 common English-Punjabi noun words have been identified. The percentage of Common English-Punjabi noun words in the Punjabi Corpus is about 6.44 %. Some of Common English Punjabi noun words are ਟੀਮ team, ਬੋਰਡ board, ਪ੍ਰੈੱਸ press etc.

2.4 Finding Punjabi language Proper Nouns from Punjabi Corpus

Proper nouns are the names of person, place and concept etc. not occurring in dictionary. Proper Nouns play important role in deciding a sentence's importance. From the Punjabi corpus, 17598 words have identified as proper nouns. The percentage of these proper noun words in the Punjabi corpus is about 13.84 %. Some of Punjabi language proper nouns are ਅਕਾਲੀ akālī, ਅਜੀਤ ajīt, ਭਾਜਪਾ bhājapā etc.

2.5 Identification of Cue Phrase in a sentence

Cue Phrases [12] 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 Some of commonly used cue phrases are ਅੰਤ ਵਿੱਚ/ ਅੰਤ ਵਿਚ ant vicc/ant vic, ਕਿਉਕੀ Kiukī, ਸਿੱਟਾ siṭṭā, ਨਤੀਜਾ/ਨਤੀਜੇ natījā/natījē etc.

3. Pre Processing Algorithm for Punjabi Text Summarization

Pre Processing phase algorithm proceeds by segmenting the source Punjabi text into sentences and words. Set the scores of each sentence as 0. For each word of every sentence follow following steps:

- Step1: If current Punjabi word is stop word then delete all the occurrences of it from current sentence.
- Step2:If Punjabi word is noun then increment the score of that sentence by 1.
- Step3: Else If current Punjabi word is common English-Punjabi noun like ਹਾਊਸ house then increment the score of current sentence by 1
- Step4: Else If current Punjabi word is proper noun like ਜਲੰਧਰ jalndhar then increment the score of current sentence by 1.
- Step5: Else Apply Punjabi Noun Stemmer for current word and go to step 2.

Sample input sentence is ਤਿੰਨਾਂ ਸ਼ਰਤਾਂ ਤੇ ਜੋ ਪੂਰਾ ਉਤਰਦਾ ਹੈ ਉਸ ਨੂੰ ਹੀ ਵੋਟ ਦਿੱਤਾ ਜਾਣਾ ਚਾਹੀਦਾ ਹੈ। tinnāṃ shartāṃ 'tē jō pūrā utradā hai us nūṃ hī vōṭ dittā jāṇā cāhīdā hai.

Sample output sentence is ਤਿੰਨ ਸ਼ਰਤ ਉਤਰਦਾ ਵੋਟ ਚਾਹੀਦਾ tinn sharat utradā vōṭ cāhīdā with Sentence Score is 2 as it contains two noun words.

4. Conclusions

In this paper, we have discussed the various pre-processing operations for a Punjabi Text Summarization System. Most of the lexical resources used in pre-processing 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 indepth analyis of Punjabi 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 other NLP applications in Punjabi.

References

1. Berry, Michael W, “Survey of Text Mining: Clustering, Classification and Retrieval”, Springer Verlag, LLC, New York, 2004.
2. Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, 347-352, 2008.
3. Mohamed Abdel Fattah and Fuji Ren, "Automatic Text Summarization", Proceedings of World Academy of Science, Engineering and Technology, Vol 27,ISSN 1307-6884, 192-195, Feb 2008.
4. Khosrow Kaikhah, "Automatic Text Summarization with Neural Networks", in Proceedings of second international Conference on intelligent systems, IEEE, 40-44, Texas, USA, June 2004.
5. Internet: http://www.tamasoft.co.jp/en/general-info/unicode-decimal.html
6. Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan, A light weight stemmer for Bengali and its Use in spelling Checker”, Proc. 1st Intl. Conf. on Digital Comm. and Computer Applications (DCCA 2007), Irbid, Jordan, March 19-23, 2007.
7. Praveen Kumar, Ankush Mittal and Sumit Gupta, “A query answering system for E-learning Hindi documents”, South Asian Language Review, VOL.XIII, Nos 1&2, Jan-June, 2003.
8. Gurmukh Singh, Mukhtiar S. Gill and S.S. Joshi, “Punjabi to English Bilingual Dictionary”, Punjabi University Patiala, 1999.
9. Mandeep Singh Gill, G.S. Lehal and S.S. Joshi, “Part of Speech Tagging for Grammar Checking of Punjabi”, The Linguistic Jornal Volume 4 Issue 1, 6-21, 2009.
10. Internet: http://www.advancedcentrepunjabi.org/punjabi_mor_ana.asp
11. Punjabi Unique word Corpus.
12 The Corpus of Cue Phrases, Internet:http://www.cs.otago.ac.nz/staffpriv/alik/papers/apps.ps
13. Neto, Joel al., "Document Clustering and Text Summarization", Proc. 4th Int. Conf. Practical Applications of Knowledge Discovery and Data Mining, 41-55, London, 2000.
14. Ananthakrishnan Ramanathan and Durgesh Rao, “ALightweight Stemmer for Hindi”, Workshop on Computational Linguistics for South-Asian Languages, EACL, 2003.