Automatic 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 relevant sentences based on statistical and linguistic features of text. Punjabi text summarization system is available online at website: http://pts.learnpunjabi.org/default.aspx It comprises of two main phases: 1) Pre Processing 2) Processing. Pre Processing is structured representation of original Punjabi text. Pre processing phase includes Punjabi words boundary identification, Punjabi sentences boundary identification, Punjabi stop words elimination, Punjabi language stemmer for nouns and proper names, applying input restrictions and elimination of duplicate sentences. In processing phase, sentence features are calculated and final score of each sentence is determined using feature-weight equation. Top ranked sentences in proper order are selected for final summary. This demo paper concentrates on Automatic Punjabi Text Extractive Summarization System.

KEYWORDS : Punjabi Text Summarization System, Pre Processing Phase, Processing Phase, Punjabi stemmer for nouns and proper nouns, Punjabi Named Entity Recognition, Punjabi Keywords Identification

1 Introduction

Automatic text summarization (Kyoomarsi et al., 2008; Gupta & Lehal, 2010) is reducing the source text into a shorter form retaining its information content and overall meaning. Text Summarization (Lin, 2009) Process can be divided into two phases: 1) Pre Processing phase (Gupta & Lehal, 2011a) is structured representation of the original text. 2) Processing (Fattah & Ren, 2008; Kaikhah, 2004; Neto et al., 2000) phase determines the final score of each sentence using feature-weight equation and top ranked sentences in proper order are selected for final summary. This paper concentrates automatic Punjabi text 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. For Punjabi language, Punjabi text summarization system is the only text summarizer and is available online: http://pts.learnpunjabi.org/default.aspx . 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 restrictions to input text, elimination of duplicate sentences and normalization of Punjabi noun words in noun morph. In processing phase, various features influencing the relevance of sentences are decided and calculated. Some of statistical features are sentence length feature, keywords selection feature (TF-ISF approach) and number feature etc. Some of linguistic features that often increase the candidacy of a sentence for inclusion in summary are: sentence headline feature, next line feature, noun feature, proper noun feature, cue phrase feature and presence of headline keywords in a sentence etc. Final score of each sentence is determined using feature-weight equation. 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.

2 Pre Processing Phase of Punjabi Text Summarization System

Various sub phases of 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, the frequency count of these stop words is 5.267 million, which covers 46.64% of the corpus.

2.2 Punjabi language stemmer for nouns and proper names

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 possible eighteen noun and proper name suffixes were identified like ◌ੀਆਂ īāṃ, ◌ਿਆਂ iāṃ, ◌ੂਆਂ ūāṃ and ◌ਾਂ āṃ etc. and various rules for Punjabi noun stemming have been generated. The algorithm of Punjabi language stemmer for nouns and proper names has been published in (Gupta & Lehal, 2011b). The efficiency of this 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. Some results of Punjabi language stemmer for nouns and Proper names for various possible suffixes are ਫੁੱਲਾਂ phullāṃ “flowers” → ਫੁੱਲ phull “flower” with suffix ◌ਾਂ āṃ, ਮੁੰਡੇ muṇḍē “boys”→ ਮੁੰਡਾ muṇḍā “boy” with suffix ◌ੇ ē and ਫਿਰੋਜ਼ਪੁਰੋਂ phirōzpurōṃ → ਫਿਰੋਜ਼ਪੁਰ phirōzpur with suffix ◌ੋਂ ōṃ etc.

2.3 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. 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!!!”.

2.4 Elimination of duplicate sentences from Punjabi input text

Punjabi Text Summarization system eliminates the duplicate sentences from the input Punjabi text. 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.

2.5 Normalization of Punjabi nouns in noun morph and input text

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 “tibbati”, ਥਾਂ thāṃ “place” and ਦਸ਼ਾ dasaā “condition”, ਬ੍ਰਿਗੇਡ brigēḍ “brigade” can also be written as ਤਿਬਤੀ tibtī “tibbati”, ਥਾ thā “place” and ਦਸਾ dasā “condition”, ਬਰਿਗੇਡ barigēḍ “brigade” respectively. To overcome this problem, input Punjabi text and Punjabi noun morph has been normalized 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, ਖ਼, ਗ਼, ਜ਼, ਫ਼, and ਲ਼. For doing normalization of Punjabi noun morph and Punjabi input text, replace all the occurences of ◌ੱ aadak , ◌ਂ bindi at top, ◌਼ bindi at foot with null character and replace all the occurences of Punjabi foot character ◌੍ with suitable ਰ ra or ਵ v or ਹ ha characters.

3 Processing Phase of Punjabi Text Summarization System

Various sub phases for processing phase of Punjabi text summarization system are given below:

3.1 Punjabi sentence relative length feature

This feature is calculated as published in (Fattah & Ren, 2008). Very short sentences are avoided for including in final summary as often they contain less information. On the other hand lengthy Punjabi sentences might contain lot of information. This feature is calculated by dividing number of words in a sentence with word count of largest sentence. Its value will be always less than or equal to 1.

3.2 Punjabi Keywords/ Title Keywords identification

Punjabi keywords identification system is first of its kind system available and is implemented by us as published in (Gupta & Lehal, 2011c). Prior to it no other Punjabi keywords identification system was available. Keywords are thematic words containing important information. Punjabi keywords are identified by calculating TF-ISF (Term Frequency-Inverse Sentence Frequency) (Neto et al., 2000) score. The TF-ISF measure of a noun word w in a sentence s, denoted TF-ISF(w,s), is computed by: TF-ISF(w,s)= TF(w,s)* ISF(w) where the term frequency TF(w,s) is the number of times that noun word w occurs in sentence s, and the inverse sentence frequency ISF(w) is given by the formula: ISF(w) = log(|S|/ SF(w)) , where the sentence frequency SF(w) is the number of sentences in which the noun word w occurs. Top scored Punjabi noun words (Top 20%) with high value of TF-ISF scores are treated as Punjabi keywords.

3.3 Numeric data identification

Numerical data (Fattah & Ren, 2008) is important and it is most probably included in the document summary. The sentence that contains numerical data (Digits, Roman and Gurmukhi numerals) is important and it is most probably included in the document summary. The score for this feature is calculated as the ratio of the number of numerical data in sentence over the sentence length.

3.4 Punjabi named entity recognition

Rules based Punjabi named entity recognition system is first of its kind system available and is implemented by us for identifying proper nouns as published in (Gupta & Lehal, 2011d). Prior to it, no other rule based Punjabi named entity system was available. It uses various gazetteer lists like prefix list, suffix list, middle name list, last name list and proper name lists for checking whether the given word is proper name or not. After doing aanalysis of Punjabi corpus, various gazetteer lists have been developed. The Precision, Recall and F-Score for condition based NER approach are 89.32%, 83.4% and 86.25% respectively.

3.5 Punjabi sentence headlines and next lines identification

In single/multi news documents, headlines are most important and are always included in the final summary. Line just next to headline might contain very important information related to summary and is usually included in summary. In Punjabi news corpus with 957553 sentences, the frequency count of these headlines/next lines is 65722 lines which covers 6.863% of the corpus.

In Punjabi headlines detection system, if current sentence does not ends with punctuation marks like '।' vertical bar 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 etc. but does not ends with enter key or new line character then set the next line flag to true for that line. Those Punjabi sentences which belong to headlines are always given highest score equal to 10 and their headline flags are set to true. The accuracy of Punjabi headline identification system and next line identification is 97.43% and 98.57% respectively which is tested over fifty Punjabi single/multi news documents. Next lines are always given very high weight equal to 9 and their next line flags are set to true.

3.6 Punjabi nouns and Proper names identification

Those Punjabi sentences containing nouns (Neto et al., 2002) and proper names are important. Input words are checked in Punjabi noun morph for possibility of nouns. Punjabi noun morph is having 37297 noun words. Proper nouns are the names of person, place and concept etc. not occurring in Punjabi Dictionary. From the Punjabi news corpus, 17598 words have been identified as proper nouns. Punjabi nouns and proper noun feature score is calculated by dividing number of Punjabi nouns/ proper names in a sentence with length of that sentence.

3.7 Punjabi Cue Phrase identification

Cue Phrases are certain keywords like in conclusion, summary and finally etc. These are very much helpful in deciding sentence importance. Those sentences which are beginning with cue phrases or which contain these cue phrases are generally more important than others. Firstly a list of Punjabi Cue phrases has been developed and then those sentences containing these Cue phrases are given more importance.

3.8 Calculation of scores of sentences and producing final summary

Final scores of sentences are determined from sentence-feature-weight equation. $w_1f_1+w_2f_2+w_3f_3+\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots w_nf_n$ where $f_{1,}$ $f_{2,}$ $f_{3\ldots\ldots\ldots\ldots\ldots}f_n$ are different features of sentences calculated in the different subphases of Punjabi text summarization system and $w_{1,}$ $w_{2,}$ $w_{3\ldots\ldots\ldots\ldots\ldots}w_n$ are the corresponding feature weights of sentences. Mathematical regression (Gupta & Lehal, 2011e ; Fattah & Ren, 2008) has been used as model to estimate the text features weights for Punjabi text summarization. Three most important features of Punjabi text summarization system are Punjabi headline identification feature, Punjabi next line identification feature and number identification feature. 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.

4 Results and Discussions

Punjabi text summarization has been tested over fifty Punjabi news documents (with 6185 sentences and 72689 words) randomly taken from Punjabi Ajit news corpus having 11.29 million words and fifty Punjabi stories (with 17538 sentences and 178400 words) randomly taken from www.likhari.org website. We have applied four intrinsic measures of summary evaluation 1) F-Score 2) Cosine Similarity 3) Jaccard Coefficient and 4) Euclidean distance for Punjabi news documents and stories. Gold summaries (reference summaries) are produced by including most common sentences of manually produced summaries by three human experts at 10%, 30% and 50% compression ratios. For Punjabi news documents, value of average F-Score is 97.87%, 95.32% and 94.63% at 10%, 30% and 50% compression ratios respectively and value of average Cosine similarity is 0.9814, 0.9629 and 0.9522 at 10%, 30% and 50% compression ratios respectively. For Punjabi stories, value of average F-Score is 81.78%, 89.32% and 94.21% at 10%, 30% and 50% compression ratios respectively and value of average Cosine similarity is 0.8226, 0.8838 and 0.9432 at 10%, 30% and 50% compression ratios respectively. The results of intrinsic summary evaluation show that for Punjabi news documents, Punjabi text summarization system performs very well at 10% compression ratio, because at 10% compression ratio usually headlines and next lines are extracted which are enough to describe the whole text but for Punjabi stories, performance of Punjabi text summarization system is not good at 10% compression ratio, because headlines are not present in stories and only few lines are extracted in summary which are not enough to describe the sense of complete story. 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 news documents and Punjabi stories. For Punjabi news documents, the accuracy of question answering task is 78.95%, 81.38% and 88.75% at 10%, 30% and 50% compression ratios respectively. The results of question answering task show that for Punjabi news documents, performance of Punjabi text summarization system is low at 10% compression ratio because news documents are usually short and at 10% compression ratio, mainly headlines and next lines are extracted which are not sufficient to give all answers of question-answering task. For Punjabi stories, the accuracy of question answering task is 80.65%, 84.26% and 90.72% at 10%, 30% and 50% compression ratios respectively. For Punjabi news documents, the accuracy of keywords association task is 80.13%, 92.37% and 96.32% at 10%, 30% and 50% compression ratios respectively. For Punjabi stories, the accuracy of keywords association task is 84.29%, 90.68% and 95.16% at 10%, 30% and 50% compression ratios respectively. For Punjabi news documents and stories, the accuracy percentage for the task of keywords association is very well at 50% compression ratio because at 50% compression ratio, summary produced is enough to cover majority of gold keywords. Both intrinsic and extrinsic summary evaluation methods show that at 50% compression ratio, Performance of Punjabi text summarization system is good for both Punjabi news documents and Punjabi stories because summary produced is enough to describe the whole text.

Conclusion

Punjabi Text Summarization system is first of its kind Punjabi summarizer and is available online at http://pts.learnpunjabi.org/default.aspx . Most of the lexical resources used in pre processing and processing such as Punjabi stemmer, Punjabi nouns normalizer, Punjabi proper names list, common English-Punjabi nouns list, Punjabi stop words list, Punjabi suffix and prefix 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 Natural language processing 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.

Jimmy Lin (2009). Summarization. *Encyclopedia of Database Systems*, *Springer-Verlag* Heidelberg, Germany.

Joel larocca Neto, Alex A. Freitas and Celso A.A.Kaestner,(2002). Automatic Text Summarization using a Machine Learning Approach. *In Book: Advances in Artificial Intelligence: Lecture Notes in computer science*, Springer Berlin / Heidelberg, volume 2507, pages 205-215.

Joel Larocca Neto, Alexandre D. Santos, Celso A.A. Kaestner and Alex A. Freitas, (2000). Document Clustering and Text Summarization. *In Proceedings of 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining*, pages 41-55, London.

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.

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.

Vishal Gupta and Gurpreet Singh Lehal (2011c). Automatic Keywords Extraction for Punjabi Language. *International Journal of Computer Science Issues*, 8(5) : 327-331.

Vishal Gupta and Gurpreet Singh Lehal (2011d). Named Entity Recognition for Punjabi Language Text Summarization. *In International Journal of Computer Applications,* 33(3): 28-32.

Vishal Gupta and Gurpreet Singh Lehal (2011e). Feature Selection and Weight Learning for Punjabi Text Summarization. *In International Journal of Engineering Trends and Technology,* 2(2): 45-48.