Summarization of “Terms and Conditions” based on machine learning.

Prof. Ashwini Khairkar[1], Namrata Dhaigude[2], Priyanka Kumari[3], Sanjana Thakur[4] , Tripti Chauhan[5]

[12345]Information Technology Department, Savitribai Phule Pune University
Address

[12345] Bharati Vidyapeeth's College Of Engineering For Women,Pune

Abstract

In today’s digital world, the amount of data getting generated is in abundance .To deal with such large amount of information there is a need to summarize the data. Automatic Text summarization helps in dealing with this vast information which is generated on web and deep web. In this paper, a genetic clustering algorithm called SENCLUS is used to perform the summarization. Each cluster is formed by checking the value of fitness function for each token. Two functions called the fitness function and the SENLCUS scoring function are used for checking relevance of tokens to the cluster and for scoring the tokens in each cluster respectively. After scoring, each token is ranked according to the score given. Finally the summary is formed with the sentences including the most highly ranked tokens according to the size of the summary to be generated.

Keywords: *Automatic Text Summarization, NLTK, Genetic clustering, SENCLUS, Fitness function, Scoring function*

1 Introduction

The Automatic Text Summarization (ATS) is the process of reducing a text document with a computer program to create a summary that maintain the most relevant points of the original document. Technologies which can obtain a coherent summary take into account variables such as length, writing style and syntax. ATS is part of machine learning as well as data mining. The main objective of summarization is to find a representative subset of the data, which consists of the information of the entire set. Summarization technologies are usually used in a large number of sectors in industry today. Common example of the use of summarization technology is in search engines such as Google.

Natural Language Toolkit (NLTK) is a tool for processing the natural language .It is a tool which provides interfaces to many corpora and lexical resources that are suitable for text processing .Text processing libraries are provided for performing various operations like classification ,tokenization, stemming etc. It is one of the best tools to build Python program which is used to work with the processing of human language data.

NLTK has many libraries that can be directly imported in the python program.These libraries have many inbuilt functions to perform text processing NLTK has a Punkt module which is used to perform sentence tokenization.Porter’s algorithm is used to perform stemming operation. Both these algorithms are applied using NLTK.

2 Literature Survey

Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization[2]

Using this model, we can partition the clustering problem in a modular way across neighborhoods, solve each part individually using a distributed K-means variant, then successively combine clusterings up the hierarchy where increasingly more global solutions are computed. Disadvantage is that K-means generate hard clusters.

Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure[3]

Abstractive cross-language summarization framework. First, the source language documents are translated into target language with a machine translation system. Then, the method constructs a pool of bilingual concepts and facts represented by the bilingual elements of the source-side predicate-argument structures (PAS) and their target-side counterparts. Finally, new summary sentences are produced by fusing bilingual PAS elements with the integer linear programming (ILP) algorithm to maximize both of the salience and translation quality of the PAS elements.
Disadvantage is that Abstractive summarization techniques tend to mimic the process of ‘paraphrasing’ from a text than just simply summarizing it. Texts summarized using this technique looks more human-like and produce more condensed summaries. These techniques are much harder to implement than extractive summarization techniques in general.

Evaluation of Summarization Methods by Manually Summarized Texts[4]

In this paper, a new evaluation method for text summarization of log of communication is proposed. Recently, many web sites which manually survey and summary their communication log are managed. These web sites pick up interesting conversation on log of BBS and modified it more attractively. We think that summaries on these web sites are very useful and more reliable than traditional questionnaire because a lot of pair of an original log and its handmade summary can be obtained. In this paper, we show the result of F-measure between summaries generated by common text summarization techniques and the BBS survey sites’ summaries .As a result, we can say that undiscovered human criteria for summarization should be found. Disadvantage is that when humans produce summaries of documents, they do not simply extract sentences and concatenate them, rather they create new sentences that are grammatically incorrect.

3 System Overview

The plugin detects the URL of the particular “Terms and Conditions” page the user is visiting .The document is extracted through the plugin and given to a Sentence Boundary Disambiguation (SBD) algorithm.
This algorithm performs the sentence tokenization. Then Porter’s algorithm is used to perform stemming and tokenization. We used Natural Language Toolkit (NLTK) to perform sentence tokenization as well as stemming.
NLTK has built-in modules to perform these operations. Sentence tokenization is performed by PunktSentenceTokenizer.

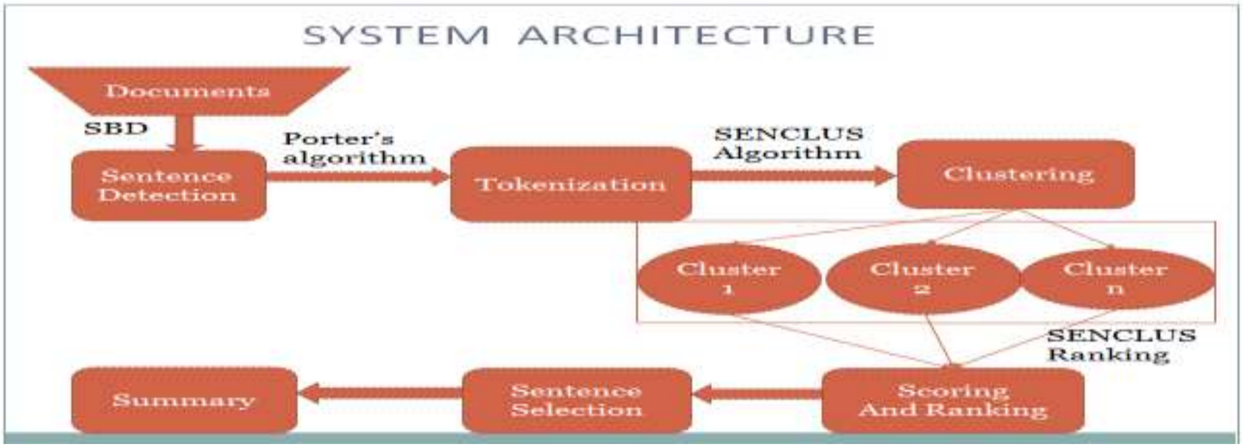

Fig -2: Proposed architecture of the system.

Stemming is done using NLTK module for Porter’s algorithm.

Main approach for automatic extractive text summarization called SENCLUS is being used. Using a genetic clustering algorithm, SENCLUS[1] clusters the sentences as close representation of the text topics using a fitness function based on redundancy and coverage, and applies a scoring function to select the most relevant sentences of each topic to be part of the extractive summary. SENCLUS generates the soft clusters. It does not need feature independence assumptions. It does not require a very huge dataset to have good performance metrics.

Finally, the top few sentences are selected from each cluster to generate the final summary which will be selected according to the size of the summary.

4. Conclusion

Thus we have proposed a system for the text summarization of terms and condition is for creating a plug-in which will give the summary of the "Terms and Conditions". Whenever the user registers for any website, they are made to agree to their "Terms and Conditions", our plug-in extracts those "Terms and Conditions" and displays the summary in a pop-up window in the plugin.

5.REFERENCES

[1].Genetic Clustering Algorithm for Extractive Text Summarization. 2015 IEEE Symposium Series on Computational Intelligence IEEE DOI 10.1109/SSCI.2015.139 2015.

[2].Hierarchically Distributed Peer-to-Peer Document Clustering and Cluster Summarization, Khaled M. Hammouda and Mohamed IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 5, MAY 2009.

[3] Abstractive Cross-Language Summarization via Translation Model Enhanced Predicate Argument Structure Fusing.Jiajun Zhang, Member, IEEE, Yu Zhou and Chengqing Zong, Senior Member, IEEE. Paper- 2329-9290

[4].Evaluation of Summarization Methods by Manually Summarized Texts on BBS Survey Web Sites , SCIS-ISIS 2012,Masao KUBO and Tatsuro ISHII and Hiroshi SATO Kobe, Japan, November 20-24, 2012