Text Summerization Techniques

Anjana Tiha
Department of Computer Science
University of Memphis, Memphis, TN
Email: atiha@memphis.edu

Abstract—This paper is focused with text summerization and currently popular text summerization techniques. This paper surveys on recent research on text smmerization.

Index Terms—Text Summerization, Machine Learning, Trending Text Summerization Techniques, Deep Learning.

I. Introduction

Automatic text summarization is the process of generating concise and condensed, representation from one or more text documents data that fluently captures the core meaning and concept of the original text. There has been exponential growth of the world-wide web and social media usage causing dramatically increase in speed and the scaling of information dissemination. With such vast amount of accessible text documents on the Internet, it has become imperative to use text summarization in order to save time and effort in finding the right information. Summarization can be used for text, document, image and video. In image summarization the summarization system finds the most representative and important images. For videos, system tries to extract the important events from the long-time frame and uneventful context. Text summarization is a sub field of machine learning and data mining. There are mainly two approaches for generating automatic text or document summarization. Most common one is using extractive style, which extracts sentences essential to preserve the core idea of the text and combines them together to generate comprehensive text summery. This technique often use machine learning to score word, sentence or paragraph. Abstractive approach focuses more on the semantic meaning of the text and generates summery using natural language processing techniques.

II. Preprocessing

Preprocessing phase often requires sentence boundary identification, special character, digits, lower case, stop word removal and stemming.

III. Summerization Techniques

A. Extraction-based summarization

In this summarization task, the automatic text summarization system extracts objects like texts, words, paragraphs from single document or the entire collection and generates summery by scoring objects based on selected features and combines them together. This approach does not modify the objects themselves. Similarly, in image summarization, the system extracts images from the collection.

1) Word Based Scoring: There are different techniques for generating summery using word based scoring. In word based scoring, each sentence in a document is scored based on some selected features. After scoring based on features, cumulative score for each sentence is selected for ranking. Highest scored sentences are selected to be in summary. Some of the common features for word based scoring includes word frequency, TF-IDF, upper case letter, proper noun and numerical data inclusive sentence.

Word Frequency:
Frequent words are assumed to be of greater importance and more reflective of the contents and hence given higher scores.

TF/IDF:
Text is first preprocessed by removing, hyperlinks, special characters, digits and stop words. Then words are stemmed to reduce to root words. Finally, TF-IDF is calculated for each sentence. Sentences with highest cumulative scores are selected for summarization.

Lexical Similarity:
Scores assuming that important sentences are identified by strong chains.

Upper Case:
Sentences with more importance are assumed to have more uppercase letters containing crucial information. Therefore, sentences are assigned higher scores that contains words with one or more upper case letters.

Proper Noun:
Hypotheses on the concept that sentences with higher number of proper nouns are more important.

2) Sentence-based Scoring: This approach is based on the features of the sentence itself.

Cue-phrases
Some cue phrases can contain key information in document. For example, the sentences started by ”in

summary", "in conclusion", "the paper describes" often contains the summery. Also, emphasizes are given by phrases such as "the best", "the most important", "significantly", "in particular". Phrases like "according to the study", "in our finding" contains important factual information. By identifying these key phrases, it is possible to extract key information from document. Also, domain-specific phrases or terms can be good indicators of significant content of a text.

Sentence Position:
The relative position of the sentence may indicate its importance. For example, the most important sentences tend to come at the beginning of a document. Also, for domain specific document summarization, domain specific knowledge can be used.

Sentence Resemblance To The Title:
Sentence resemblance to the title is the vocabulary overlap between this sentence and the document title. Words in the title is considered important and contains key terms that contain the central concept of the document. Sentence with high resemblance with title is where individual sentences contains words of the title, or have an overlapping vocabulary with title. Sentences with higher resemblances to the title are deemed to be more important than other sentences. The concept centrality embeded in title vocabulary can help retain important sentences. Sentence Length: This approach penalizes sentences that are either too short or long.

Sentence Inclusion of Numerical Data:
Sentences containing numerical data is considered important and is more likely to be included in the document summary.

3) Graph-based Scoring:: In graph-based methods the score is generated by the analyzing relationship among sentences in a document. When a sentence refers to another it generates a link with an associated weight between them. The weights are used to generate the scores of a sentence.

Text is preprocessed by removing digits, special characters, stop words and converting them to lower case. After initial preprocessing, stemming is done to reduce to root words. In the graph based approach, all the sentences are considered as a node in a graph. Based on shared or similar information like overlapping vocabulary, two nodes are connected. The nodes or sentences with higher cardinality is considered more important for summarization.

Text Rank:
It extracts the important keywords from a text document and determines the weight of the importance of words within the entire document by using a graph-based model.

Cluster based method:
Often documents have an inherent structural design and key information's are placed strategically in a document. Therefore, some part of the document contains for cumulative information than other parts. Documents are usually organized in a pattern by using the structural nature of document or text, summarization can be developed. In this approach, each document is divided into multiple segments or clusters and objects (word, sentence) in each segment are given weights based on term importance, TF-IDF or other features. Then each segment is ranked by calculating the overall importance score/other score by generating cumulative score for all the objects present in a segment. Segments with highest scores are selected for summary.

4) Chalanges Of Extractive Summarization: Problem with extractive summarization is that it can lead to lack of coherency. Also, summery especially unit sentences can be too long to understand the concept accurately. Often document can contain counter information. Presenting them in summery can lead to ambiguity and confusion beating the original motivation behind summarization. Also, key information could get ignored due to selection of scoring method.

B. Abstraction-based summarization

In abstraction based approach, summarization system paraphrases segments of the source document. abstraction based technique can generate more condensed text than extraction technique. But, this requires usage of natural language generation technology, which is still in growth stage. While there has been some recent work in abstractive summarization, the majority of summarization systems are using extractive methods.

C. Machine Learning approach:

Sentences are scored based on some selected features of the text document.

References

[1] Gupta, Vishal, and Gurpreet Singh Lehal. "A survey of text summarization extractive techniques." Journal of emerging technologies in web intelligence 2.3 (2010): 258-268.

[2] Ferreira, Rafael, et al. "Assessing sentence scoring techniques for extractive text summarization." Expert systems with applications 40.14 (2013): 5755-5764.

[3] Rush, Alexander M., Sumit Chopra, and Jason Weston. "A neural attention model for abstractive sentence summarization." arXiv preprint arXiv:1509.00685 (2015).

[4] Gupta P, Pendluri VS, Vats I. Summarizing text by ranking text units according to shallow linguistic features. InAdvanced Communication Technology (ICACT), 2011 13th International Conference on 2011 Feb 13 (pp. 1620-1625). IEEE.

[5] Balahur A, Lloret E, Boldrini E, Montoyo A, Palomar M, Martínez-Barco P. Summarizing threads in blogs using opinion polarity. InProceedings of the workshop on events in emerging text types 2009 Sep 17 (pp. 23-31). Association for Computational Linguistics.
[6] Barrera A, Verma R. Combining syntax and semantics for automatic extractive single-document summarization. Computational Linguistics and Intelligent Text Processing. 2012:366-77.
[7] Fattah MA, Ren F. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. Computer Speech & Language. 2009 Jan 31;23(1):126-44.
[8] Shardan R, Kulkarni U. Implementation and evaluation of evolutionary connectionist approaches to automated text summarization.
[9] Lloret E, Palomar M. Text summarisation in progress: a literature review. Artificial Intelligence Review. 2012 Jan 1;37(1):1-41.