Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 10; April-June, 2015 pp. 100-103
© Krishi Sanskriti Publications
http://www.krishisanskriti.org/acsit.html

Automatic Text Summarization

Ankita Rajkhowa

M.Tech Assam Don Bosco University
E-mail: ankitarajkhowa@gmail.com

Abstract—*In today's fast growing information world, text summarization has become an important matter for interpreting text information. It is the process of summarizing a source text to a shorter version containing all its information and overall meaning. There are two methods of text summarization- Extractive and Abstractive summarization. This paper describes the extractive features of text summarization. Extractive summarization is the method of selecting sentences or paragraphs from the source document and concatenating them into shorter forms while abstractive summarization is the method of understanding the source document and generate its meaning into a shorter form. It also presents the evaluation measures of a text summarizer.*

Index Terms*: Text Summarization, extractive summarization, abstractive summarization.*

1. INTRODUCTION

Automatic text summarization summarization plays an important role in information retrieval. It is the process by which the important portion of the text is retrieved. It was first studied by Luhn over 50 years ago [5] and has continued to be a steady subject of research. Automatic text summarization has achieved a wide popularity in Natural Language Processing. It is the process of taking a source document to present the most important content in a shorter form and within a short span of time. The process must give the most important sentences that deliver the exact idea provided in the source text. Text understanding and text generation are the two processes that are directly associated with producing the summaries.

The summaries that were created are of two types- Extractive summaries and Abstractive summaries. Extractive summaries are those where the most important sentences are selected based on some scores given to each sentences and then arranged in a proper order to create the final summary. In abstractive summaries the original text is studied and then its meaning is condensed in a fewer words. Abstractive methods are a tough problem as the system has to understand the point of a text and they require the use of natural language generation technology which by itself is a growing field. Many works have been done on extractive summaries while very less work has been done in abstractive summaries. The reason is that, abstractive summaries require semantic analysis and grouping of the content using world knowledge. However, the system is not able to do it without a great deal of world knowledge.

Extractive summarization is the method of selecting sentences or paragraphs from the source document and concatenating them into shorter forms while abstractive summarization is the method of understanding the source document and generate its meaning into a shorter form. Extractive summaries are based on statistical and linguistic features of sentences while abstractive summaries are based on linguistic characteristics to find the most promising information from the input text document.

2. RELATED WORK

In paper [3], they discuss the most relevant approaches both in the area of single-document and multi-document summarization. For single document summarization they present two approaches- machine learning methods and Deep Natural language Analysis methods. Machine learning methods include Naïve-Bayes Methods, Rich Features and Decision Trees, Hidden Markov Models, Log-Linear Models and Neural Networks and Third Party Features. For multi document summarization, the Abstraction and Information Fusion, Topic driven summarization and MMR, Graph-spreading activation, Centroid-based summarization and Multilingual Multi-document summarization are presented. They also discuss the other approaches to summarization such as Short Summaries, Sentence Compression, and Sequential document representation. Special attention is devoted to automatic evaluation of summaries system as future research on summarization is strongly dependent on progress in this area.

The survey paper [2] is concentrating on extractive summarization methods. They also give some features for extractive text summarization. These features are important as, a number of methods of text summarization are using them. These features are covering statistical and linguistic characteristics of a language. Some of the extractive summarization methods are discussed which aims at picking out the most relevant sentences in the document and also maintaining a low redundancy in the summary.

In [8], first of all the authors distinguish three relatively independent tasks performed by all the summarizers: creating an intermediate representation of the input which captures only the key aspects of the text, scoring sentences based on that representation and selecting a summary consisting of several sentences. It also presents some of the most topic representation approaches, as well as those that have been gaining popularity because of their recent successes. They also point out some of the peculiarities of the task of summarization which have posed challenges to machine learning approaches for the problem and suggested solutions.

3. STEPS FOR EXTRACTIVE SUMMARIZATION

Extractive text summarization process can be divided into 3 steps: **(1) Pre-processing step**; **(2) Processing step**; and **(3) Summary generation.**

In the pre-processing step a structured representation of the original text is obtain [4]. It usually includes:

a) Sentence Boundary Identification: In English, sentence boundary identification with presence of dot at the end of the sentence.
b) Stop Word Removal: Common words with no semantics and which do not aggregate relevant information to the task are eliminated.
c) Stemming: The purpose of stemming is to obtain the stem or radix of each word, which emphasize its semantics.
d) Sentence Segmentation: Segments the input character sequences into tokens. Tokens are usually words, punctuations, numbers, etc.

The processing phase mainly consists of the following steps:

a) Sentence Feature Calculation: Features influencing the relevance of sentences are decided and calculated and then weights are decided and calculated and then weights are assigned to these features using weight learning.
b) Sentence Scoring: Final score of each sentence is determined using feature weight equation. The score measures how relevant the sentence is to the “understanding” of the text as a whole.
c) Selecting sentences in Proper Order: Top ranked sentences are selected for the final summary.

After going through the pre-processing and the processing stages the final summary of the document text is generated.

4. SOME FEATURES FOR EXTRACTIVE SUMMARIZERS

There are different features for extractive text summarization. Some of the word features and sentence features are discussed as follows:

1. Content Word Feature: Content words or keywords are basically noun, verb, adjective, and adverb. Sentences consisting of keywords have greater chances of including in summary. Its measure is determined by tf*idf measure.
2. Title Word Feature: Sentences containing words that appear in the title are given the most preferences and they have greater chances to include in the summary.
3. Cue Phase Feature: Cue phase like “in conclusion”, “this letter”, “result”, “this report”, “argue”, “summary”, “develop”, “attempt” etc, are most likely to be included in the summary.
4. Font-based Feature: Words appearing in uppercase, bold, italics, or underlined fonts usually contain important information. So sentences containing such words are taken for summary sentences.
5. Biased Word Feature: If a word appearing in the biased list, which is previously defined, then the sentence is considered to be important. Biased list contains domain specific words.
6. Proper Noun Feature: Sentences containing proper nouns such as names of places, person, concept etc, have greater chance of included in the summary.
7. Pronouns: Pronouns like “she, he, it” cannot be included in the summary. They have to be expanded into their corresponding noun.
8. Sentence Position Feature: Sentences are hierarchically organized with crucial information at the beginning and the end. So, sentences at the beginning and end of a paragraph are considered more important as they contain idea that can be useful for generating the summary.
9. Sentence Length Feature: Very large sentences and very small sentences are not usually included in the summary.
10. Paragraph Location Feature: Similar to the sentence location feature, paragraphs are also hierarchically organized. So, paragraphs at the beginning and end of a document contain crucial information and are generally included in the summary.
11. Sentence Similarity: The sentence similarity feature contributes to the sentence centrality feature, where keywords of one sentence are compared to other sentences.
12. Sentence Similarity: Similar to the sentence similarity feature, the keywords of the title sentence are compared to other sentences.

5. EVALUATION MEASURES

The most important task of a text summarization system is the evaluation of the summary. We must be concern about selecting the most appropriate methods and types of evaluation. Evaluation methods are useful in evaluating the usefulness and trustfulness of the summary [6]. Evaluating the qualities like comprehensibility, coherence, and readability is really difficult. The most common approach of system evaluation is to invite human-experts who compare different summaries and choose the best out of it. But this approach has some drawbacks. Firstly, the individuals who perform the evaluation task might have different views and opinion of what a good summary should contain. Secondly, it is very time consuming. Another approach for summary evaluation is Automatic system evaluation which is still an open research topic [9].

The evaluation methods can be broadly classified into two categories: extrinsic evaluation and intrinsic evaluation. Extrinsic evaluation is mostly concern with determining the utility and usability of the automatic summaries in the context of the specific tasks that make use of them [10]. Intrinsic evaluation is concern with determining the internal quality of the generated summaries.

Summarization evaluation methods which judge the quality of the summaries based on how they affect the completion of some other tasks are define as extrinsic evaluation methods. This method mainly assesses the quality of summaries indirectly through the performance of some task using the summaries. It is based on comparisons with the source document.

Summarization evaluation methods which judge the quality of summaries by direct analyses in terms of some set of norms is define as intrinsic evaluation methods. This method mainly assesses the performance of a text mining system component as an isolated unit connected to the other system components. Intrinsic evaluation can be categorized into two groups: content evaluation and text quality evaluation [7]. Content evaluation measures the ability to identify the key topics, whereas text quality evaluation collects the readability, grammar and coherence of automatic summaries.

Since there is not a base standard for evaluating summaries, different criteria are being used for evaluation [6]. The two most practical evaluation measures are Precision and Recall, which are used for specifying the similarity between the summary which is generated by the system versus the one generated by human. Precision (P) is calculated as the number of sentences occurring in both the system summary and human generated summary divided by the number of sentences in the system summary. Recall (R) is calculated as the number of sentences in the both system and human generated summaries divided by the number of sentences in the human generated summary. F-score is a composite measure that combines precision and recall. The basic way to compute the F-score is to count a harmonic average of precision and recall:

$$F = (2*P*R)/P+R$$

The most complex formula for measuring F-score is as given below:

$$F = (\beta^2 + 1)*P*R/\beta^2*P+R$$

where, β is a weighting factor that favors Precision when $\beta>1$ and favors Recall when $\beta<1$.

There are two other criteria for evaluating summary. These are Compression Ratio and Retention Ratio [6] where,

Compression Ratio: CR = Length S/Length T

Retention Ratio: RR = Information in S/Information in T

where, S is the summarized text and T is the main text. From this it is conclude that a good summary is the one with low CR and high RR.

REFERENCES

[1] H. P. Luhn, "The Automatic Creation of Literature Abstracts", *Presented at IRE National Convention*, New York, pp. 159-165, 1958.

[2] Vishal Gupta and Gurpreet Singh Lehal, "A Survey of Text Summarization Extractive Techniques", *Journal of Emerging Technologies in Web Intelligence*, Vol. 2, No. 3, August 2010.

[3] Dipanjan Das and Andre F.T. Martins, "A Survey on Automatic Text Summarization", November 21, 2007.

[4] Joel Larocca Neto, Alex A.Freitas, Celso A.A.Kaesther, “Automatic text Summarization using Machine Learning approach”, *(PUCPR)*, 1155.

[5] Krysta M. Svore, Lucy Vanderwende and Christopher J.C. Burges "Enhancing Single Document Summarization by Combining RankNet and Third –party Sources”, *Available: http://www.wikipedia.org.*

[6] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi and Bahareh Gholamzadeh, “A Comprehensive Survey on Text Summarization Systems”, 978-1-4244-4946-0/09/$25.00 © 2009 IEEE.

[7] Josef Steinberger and Karel Jeˇzek, "Evaluation measures for Text Summarization", *Computing and Informatics,* Vol.28, 2009, 1001-1026, V 2009-Mar-2.

[8] Ani Nenkova and Kathleen McKeown "A Survey of Text Summarization Techniques", in C.C. Aggarwal and C.X. Zhai (eds) *Mining Text Data*, DOI 10.1007/978-1-4614-3223-4_3, © Springer Science+Business Media LLC 2012, pg. 43-76.

[9] Hercules Dalianis, Martin Hassel, Koenraad de Smedt, Anja Liseth, Till Christopher Lech and Wedekind, "Porting and Evaluation of Automatic Summarization", 2003.

[10] Sonia Haiduc, Jairo Aponte, Andrian Marcus, “Supporting Program Comprehension with the Source Code Summarization”, *ICSE’10*, May 2-8, 2010, Cape Town, South Africa.

[11] Madhavi K. Ganapathiraju, “Overview of Summarization Methods”, 11-742: Self paced lab in Information Retrieval, November 26, 2002.

[12] Tadashi Nomoto and Tuji Matsumoto, “An Experimental Comparison of Supervised and Unsupervised Approaches to Text Summarization”, 0-7695-1119-8/01 $17.00 ©2001 *IEEE*.

[13] Fang Chen, Kesong Han, Guilin Chen, “An Approach to Sentence-Selection Based Text Summarization”, *Proceedings of IEEE TENCON’02*, pg 489-493, 2002.

[14] Kathleen McKeown, Julia Hirschberg, Michel Galley and Sameer Maskey, “From Text to Speech Summarization”, *Proceedings of IEEE ICASSP’05*, 0-7803-8874-7/05 $20.00 ©2005 IEEE.

[15] Nowshath Kadhar Batcha, Ahmed. M. Zaki, “Algebraic Reduction in Automatic Text Summarization- The State of the Art”, *International Conference on Computer and Communication Engineering (ICCCE 2010)*, 11-13 May 2010, Kuala Lumpur, Malaysia, 978-1-4244-6235-3/10/$26.00 ©2010 IEEE.

[16] M. Esther Hannah, Dr. Saswati Mukherjee, K. Ganesh Kumar, "An Extractive Text Summarization Based on Multivariate Approach", *2010 3rd International Conference on Advanced Computer Theory and Engineering {ICACTE),*IEEE 2010, V3: 157-161.

[17] Suneetha Manne, Shaik Mohammed Zaheer Pervez, Dr. S. Sameen Fatima, "A Novel Automatic Text Summarization System with Feature Terms Identification".

[18] Vipul Dalal, Dr. Latesh Malik, "A Survey of Extractive and Abstractive Automatic Text Summarization Techniques", *2013 6th International Conference on Emerging Trends in Engineering and Technology*, ©IEEE 2013, pg. 109-110.