Impact Factor (SJIF): 5.301

International Journal of Advance Research in Engineering, Science & Technology

e-ISSN: 2393-9877, p-ISSN: 2394-2444

Volume 5, Issue 2, February-2018

A Summary on Text Summarization Techniques

Ms. Pranjali Yadav-Deshmukh [1], Mrs. Madhuri Bidwe[2]

[1]pranu313@gmail.com, [2]madhuribidwe@gmail.com

[1]Dr.D.Y Patil Polytechnic,Akurdi
[2]Dr.D.Y Patil Polytechnic,Akurdi

Abstract — In recent years, there has been a explosion in the amount of text data from a variety of sources. This volume of text is an invaluable source of information and knowledge which needs to be effectively summarized to be useful. In this review, the main approaches to automatic text summarization are described. We review the different processes for summarization and describe the effectiveness and shortcomings of the different methods..

Keywords - text summarization, knowledge bases, topic models

I. INTRODUCTION

With the dramatic growth of the Internet, people are overwhelmed by the tremendous amount of online information and documents. This expanding availability of documents has demanded exhaustive research in the area of automatic text summarization. a summary is defined as "a text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually, significantly less than that". Automatic text summarization is the task of producing a concise and fluent summary while preserving key information content and overall meaning. In recent years, numerous approaches have been developed for automatic text summarization and applied widely in various domains. For example, search engines generate snippets as the previews of the documents . Other examples include news websites which produce condensed descriptions of news topics usually as headlines to facilitate browsing or knowledge extractive approaches . Automatic text summarization is very challenging, because when we as humans summarize a piece of text, we usually read it entirely to develop our understanding, and then write a summary highlighting its main points. Since computers lack human knowledge and language capability, it makes automatic text summarization a very difficult and non-trivial task. Automatic text summarization gained attraction as early as the1950s. An important research of these days was for summarizing scientific documents. Luhn et al. introduced a method to extract salient sentences from the text using features such as word and phrase frequency. They proposed to weight the sentences of a document as a function of high frequency words, ignoring very high frequency common words. Edmundson et al. described a paradigm based on key phrases which in addition to standard frequency depending weights, used the following three methods to determine the sentence weight:

1) Cue Method: The relevance of a sentence is calculated based on the presence or absence of certain cue words in the cue dictionary.

2) Title Method: The weight of a sentence is computed as the sum of all the content words appearing in the title and headings of a text.

3) Location Method: This method assumes that sentences appearing in the beginning of document as well as the beginning of individual paragraphs have a higher probability of being relevant.

In general, there are two different approaches for automatic summarization: extraction and abstraction. Extractive summarization methods work by identifying important sections of the text and generating them verbatim; thus, they depend only on extraction of sentences from the original text. In contrast, abstractive summarization methods aim at producing important material in a new way. In other words, they interpret and examine the text using advanced natural language techniques in order to generate a new shorter text that conveys the most critical information from the original text. Even though summaries created by humans are usually not extractive, most of the summarization research today has focused on extractive summarization. Purely extractive summaries often times give better results compared to automatic abstractive summaries. This is because of the fact that abstractive summarization methods cope with problems such as semantic representation, inference and natural language generation which are relatively harder than data-driven

approaches such as sentence extraction. As a matter of fact, there is no completely abstractive summarization system today. Existing abstractive summarizers often rely on an extractive preprocessing component to produce the abstract of the text .Consequently, in this paper we focus on extractive summarization methods and provide an overview of some of the most dominant approaches in this category. There are a number of papers that provide extensive overviews of text summarization techniques and systems.

II. EXTRACTIVE SUMMARIZATION

As mentioned before, extractive summarization techniques produce summaries by choosing a subset of the sentences in the original text. These summaries contain the most important sentences of the input. Input can be a single document or multiple documents. In order to better understand how summarization systems work, we describe three fairly independent tasks which all summarizers perform:

1) Construct an intermediate representation of the input text which expresses the main aspects of the text.
2) Score the sentences based on the representation.
3) select a summary comprising of a number of sentences.

Intermediate Representation - Every summarization system creates some intermediate representation of the text it intends to summarize and finds salient content based on this representation. There are two types of approaches based on the representation: topic representation and indicator representation. Topic representation approaches transform the text into an intermediate representation and interpret the topic(s) discussed in the text. Topic representation-based summarization techniques differ in terms of their complexity and representation model, and are divided into frequency-driven approaches, topic word approaches, latent semantic analysis and Bayesian topic models. We elaborate topic representation approaches in the following sections. Indicator representation approaches describe every sentence as a list of features (indicators) of importance such as sentence length, position in the document, having certain phrases, etc.

Sentence Score - When the intermediate representation is generated, we assign an importance score to each sentence. In topic representation approaches, the score of a sentence represents how well the sentence explains some of the most important topics of the text. In most of the indicator representation methods, the score is computed by aggregating the evidence from different indicators. Machine learning techniques are often used to find indicator weights.

Summary Sentences Selection - Eventually, the summarizer system selects the top k most important sentences to produce a summary. Some approaches use greedy algorithms to select the important sentences and some approaches may convert the selection of sentences into an optimization problem where a collection of sentences is chosen, considering the constraint that it should maximize overall importance and coherency and minimize the redundancy. There are other factors that should be taken into consideration while selecting the important sentences. For example, context in which the summary is created may be helpful in deciding the importance. Type of the document (e.g. news article, email, scientific paper) is another factor which may impact selecting the sentences.

III. TOPIC REPRESENTATION APPROACHES

In this section we describe some of the most widely used topic Representation Approaches.

Topic Words - The topic words technique is one of the common topic representation approaches which aims to identify words that describe the topic of the input document. was one the earliest works that leveraged this method by using frequency thresholds to locate the descriptive words in the document and represent the topic of the document. A more advanced version of Luhns idea was presented in which they used log-likelihood ratio test to identify explanatory words which in summarization literature are called the “topic signature. Utilizing topic signature words as topic representation was very effective and increased the accuracy of multi document summarization in the news domain . For more information about log-likelihood ratio test. There are two ways to compute the importance of a sentence: asa function of the number of topic signatures it contains, or as the proportion of the topic signatures in the sentence. Both sentence scoring functions relate to the same topic representation, however, they might assign different scores to sentences. The first method may assign higher scores to longer sentences, because they have more words. The second approach measures the density of the topic words.

IV. KNOWLEDGE BASES AND AUTOMATIC SUMMARIZATION

The goal of automatic text summarization is to create summaries that are similar to human-created summaries. However, in many cases, the soundness and readability of created summaries are not satisfactory, because the summaries do not cover all the semantically relevant aspects of data in an effective way. This is because many of the existing text summarization techniques do not consider the semantics of words. A step towards building more accurate summarization systems is to combine summarization techniques with knowledge bases (semantic-based or ontology-based summarizers).The advent of human-generated knowledge bases and various ontologies in many different domains (e.g. Wikipedia, YAGO, DB pedia,etc) has opened further possibilities in text summarization , and reached increasing attention recently. For example, Henning etal. present an approach to sentence extraction that maps sentences to concepts of an ontology. By considering the ontology features, they can improve the semantic representation of sentences which is beneficial in selection of sentences for summaries. They experimentally showed that ontology-based extraction of sentences outperforms baseline summarizers. Chen et al. introduce a user query-based text summarizer that uses the UMLS medical ontology to make a summary for medical text. Baralis et al. propose a Yago-based summarizer that leverages YAGO ontology to identify key concepts in the documents. The concepts are evaluated and then used to select the most representative document sentences. Sankarasubramaniam et al. introduce an approach Text Summarization Techniques: A Brief Survey that employs Wikipedia in conjunction with a graph-based ranking technique. First, they create a bipartite sentence-concept graph, and then use an iterative ranking algorithm for selecting summary sentences.

V. THE IMPACT OF CONTEXT IN SUMMARIZATION

Summarization systems often have additional evidence they can utilize in order to specify the most important topics of document(s).For example when summarizing blogs, there are discussions or comments coming after the blog post that are good sources of information to determine which parts of the blog are critical and interesting. In scientific paper summarization, there is a considerable amount of information such as cited papers and conference information which can be leveraged to identify important sentences in the original paper. In the following, we describe some the contexts in more details.

Web Summarization - Web pages contains lots of elements which cannot be summarized such as pictures. The textual information they have is often scarce, which makes applying text summarization techniques limited. Nonetheless, we can consider the context of a web page, i.e. pieces of information extracted from content of all the pages linking to it, as additional material to improve summarization. The earliest research in this regard is where they query web search engines and fetch the pages having links to the specified web page. Then they analyze the candidate pages and select the best sentences containing links to the web page heuristically. Delort et al. extended and improved this approach by using an algorithm trying to select a sentence about the same topic that covers as many aspects of the web page as possible. For blog summarization, proposed a method that first derives representative words from comments and then selects important sentences from the blog post containing representative words. For more related works.

Scientific Articles Summarization - A useful source of information when summarizing a scientific paper (i.e. citation-based summarization) is to find other papers that cite the target paper and extract the sentences in which the references take place in order to identify the important aspects of the target paper. Mei et al. propose a language model that gives a probability to each word in the citation context sentences. They then score the importance of sentences in the original paper using the KL divergence method (i.e. finding the similarity between a sentence and the language model). For more information.

Email Summarization - Email has some distinct characteristics that indicates the aspects of both spoken conversation and written text. For example, summarization techniques must consider the interactive nature of the dialog as in spoken conversations. Nenkova et al. presented early research in this regard, by proposing a method to generate a summary for the first two levels of the thread discussion. A thread consists of one or more conversations between two or more participants over time. They select a message from the root message and from each response to the root, considering the overlap with root context. Rambow et al. used a machine learning technique and included features related to the thread as well as features of the email structure such as position of the sentence in the tread, number of recipients, etc. Newman et al. describe a system to summarize a full mailbox rather than a single thread by clustering messages into topical groups and then extracting summaries for each cluster.

CONCLUSION:

The increasing growth of the Internet has made a huge amount of information available. It is difficult for humans to summarize large amounts of text. Thus, there is an immense need for automatic summarization tools in this age of information overload. In this paper, we emphasized various extractive approaches for single and multi-document summarization. We described some of the most extensively used methods such as topic representation approaches, frequency-driven methods, graph-based and machine learning techniques. Although it is not feasible to explain all diverse algorithms and approaches comprehensively in this paper, we think it provides a good insight into recent trends and progresses in automatic summarization methods and describes the state-of-the-art in this research area.

REFERENCES :

[1] Amjad Abu-Jbara and Dragomir Radev. 2011. Coherent citation-based summarization of scientific papers. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.Association for Computational Linguistics, 500–509.

[2] RasimM Alguliev, RamizMAliguliyev,Makrufa S Hajirahimova, and Chingiz A Mehdiyev. 2011. MCMR:Maximumcoverage andminimumredundant text summarization model. Expert Systems with Applications 38, 12 (2011), 14514–14522.

[3] RasimM Alguliev, Ramiz M Aliguliyev, and Nijat R Isazade. 2013. Multiple documents summarization based on evolutionary optimization algorithm. Expert Systems with Applications 40, 5 (2013), 1675–1689.

[4] Mehdi Allahyari and Krys Kochut. 2015. Automatic topic labeling using ontology-based topic models. In Machine Learning and Applications (ICMLA), 2015 IEEE 14th International Conference on. IEEE, 259–264.

[5] Mehdi Allahyari and Krys Kochut. 2016. Discovering Coherent Topics with Entity TopicModels. InWeb Intelligence (WI), 2016 IEEE/WIC/ACMInternational Conference on. IEEE, 26–33.

[6] MehdiAllahyari and Krys Kochut. 2016. Semantic Context-Aware Recommendation via TopicModels Leveraging Linked Open Data. In International Conference on Web Information Systems Engineering. Springer, 263–277.

[7] Mehdi Allahyari and Krys Kochut. 2016. Semantic Tagging Using Topic Models Exploiting Wikipedia Category Network. In Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on. IEEE, 63–70.

[8] M. Allahyari, S. Pouriyeh, M. Assefi, S. Safaei, E. D. Trippe, J. B. Gutierrez, and K. Kochut. 2017. A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques. ArXiv e-prints (2017). arXiv:1707.02919

[9] Einat Amitay and Cecile Paris. 2000. Automatically summarising web sites: is there a way around it?. In Proceedings of the ninth international conference on Information and knowledge management. ACM, 173–179.

[10] Elena Baralis, Luca Cagliero, Saima Jabeen, Alessandro Fiori, and Sajid Shah. 2013. Multi-document summarization based on the Yago ontology. Expert System with Applications 40, 17 (2013), 6976–6984.

[11] Taylor Berg-Kirkpatrick, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1.Association for Computational Linguistics, 481–490.

[12] DavidM Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. the Journal of machine Learning research 3 (2003), 993–1022.

[13] Asli Celikyilmaz and Dilek Hakkani-Tur. 2010. A hybrid hierarchical model for multi-document summarization. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 815–824.