ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter1058102020.pdf
https://doi.org/10.30534/ijeter/2020/1058102020

Top approaches to abstractive text summarization: A survey

Deepak Sahoo[1], Chayan Paul[2], Praveen Tumuluru[3]
[1]

ABSTRACT

We are in the age of information overload where a high volume of data and information is present over the internet. There are different types of text mining tasks are available to extract the appropriate data as per the user need. In text mining, summarization is a task where the gist of the source text is generated by the system. Extractive and abstractive are the two variants of text summarization based on context. Extractive approach requires less engineering and linguistic effort whereas abstractive text summarization is still a demanding task among natural language processing researchers. Abstractive text summarization system understands, interpret the original text and presents the text in new form therefore abstractive summarization require more engineering and linguistic efforts.

Key words: Abstractive summarization, Extractive summarization, Text mining, Text summarization.

I. INTRODUCTION TO TEXT MINING

In text mining [1], unstructured or semi-structured textual data is processed to find useful numerical indices that can be used by the various data mining algorithms to extract meaningfull informatin from the text. In general terms text mining will ”turn text into meaningful indices” and used in different analyses such as predictive data mining applicatons, clustering and classification. These techniques are discussed and explained in the work by Manning and Schütze [2]. The typical tasks of text mining shown in Figure 1 can be given as document clustering, organization, classification, and Information extraction.

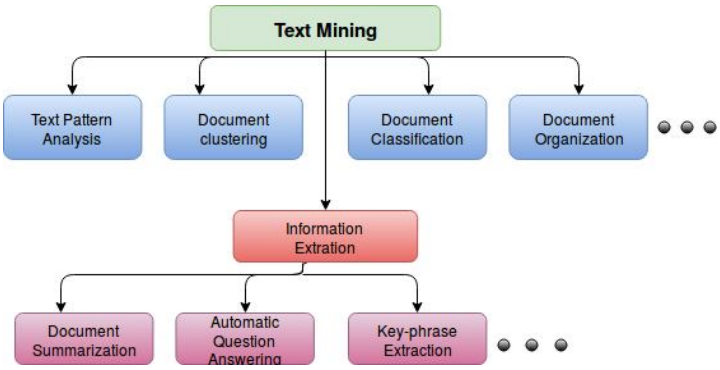

Figure 1: Text mining approaches

2. DOCUMENT SUMMARIZATION

Text summarization is an important tool to summarize large text because summarizing a large text manually is a tough task. The purpose of automatic text summarizaton is to present shorten and non-redundant verrsion of source text without negotiating the general meaning of the text.

This summarization task can be classified based on form, dimension and context given in Figure 2. An extractive summarization method [3, 4, 5, 6] has two parts, extracting most salient sentences from source text and fuseing them into shorter form. The most infromative sentences are extracted based on linguistic and statistical attributes of sentences.

In the abstractive approach, it is not only extracting the sentences but the system needs to understand the meaning, merge the sentences, add new words, remove unnecessary words and generate a new sentence. Therefore, abstractive summarization is challenging that require more linguistic effort compared to extractive summarization.

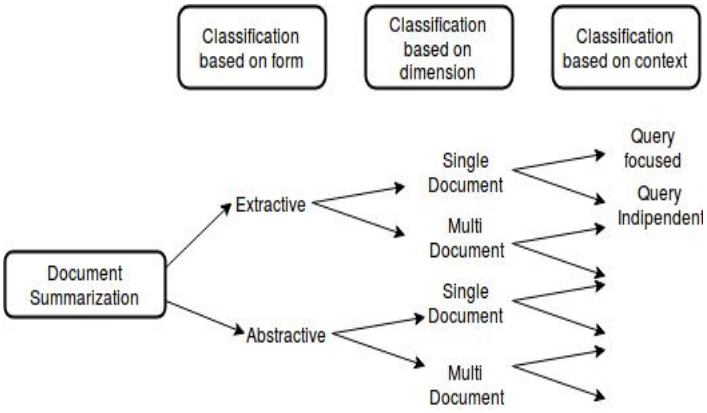

Figure 2: Classification of text summarization

Abstractive and extractive summarization can be multi document or single document summarization. if the system takes one document as input then it is a single document extractive/abstractive summarization. If the system takes more than one document as input then it is considered as multidocument extractive/abstractive summarization. Again the single/multi-document extractive/abstractive

summarization can be query focused or query independent. If the summary is generated based on a user query then its query-focused summarization or query independent summarization.

2.1 Back ground of extractive summary

A lot of work has been done and many approaches are used to extract best and informative sentences from the document since its inception in 1958 [7] where a sentence of a document can be mapped as a function of high-frequency words. In addition to standard keywords, the summarization system [8] used below three techniques to calculate weight of the sentence.

1. Cue technique: In this technique, the weight of a sentence is assessed by the occurance of certain cue words.

2. Title technique: In this technique, the weight of the sentence is computed taking the sum of all the content words present in the title.

3. Location technique: This technique assumes that, more relevant sentences are present in initial positions of the text.

Text extraction system ANES [9] is an domain independent news summarization sustem has following four components.

1. Analysis of corpus: tf*idf-weights of all terms are calculated in this component.

2. Selection of key words (Statistical approach): the top tf*id weighted words added with headline-words.

3. Sentence weight: summing the weight of all the words in the sentence and some other factors like relative position of sentence.

4. Sentence selection: Rank the sentences based on weight and choosing top ranked sentence.

The trainable document summarizer [10] extracts roughly about twenty percent of the original text, based on following weighting heuristics.

1. Length of sentence: Sentence length should be greater than five words.

2. Paragraph: sentence location in a paragraph.

3. Fixed phrase: sentences containing manually chosen phrases.

4. Thematic word feature: words that occurs most frequently (except stops) are called as thematic words. Sentence score is a function of frequencies of these thematic words.

Other most important techniques like inter-document link generation [11], rhetorical structures of texts [12], lexical chain and WordNet [13] are used to generate the summary of text. MEAD [14] a multi-document summarizer creates the clusters of related sentences based on topic detection and generates the summay from cluster centroids.

2.2 Drawbacks of extractive summary

In extractive summary key sentences or passages are extracted from a large text, based on statistical analysis of individual or group of features such as cue words, location and word/phrase frequency to locate and extract the best sentences. The “most important” content is considered as the “most frequent” or “most favorably positioned” content but the issues with extractive summary [15, 16] are:

1. Generally extracted sentences are longer than the average sentence length. Therefore, some unnecessary segments are also included in the summary, consuming space.

2. It is difficult to capture important and relevant informaition as it is spread across the sentences.

3. Difficult to present (sometimes may not be) inconsistent information correctly.

4. In extractive summary ”dangling” anaphora is a frequent and serious issue. Let two sentences have an anaphoric relation in a document and during extraction sentence containing the pronoun is selected but the sentence containing the proper-noun/common noun that refers to that pronoun is not selected then the overall summary becomes incoherent.

2.3 Abstractive summary and challenges

It is observed from the background study that there is less opportunity to explore in extractive approach which requires less linguistic effort but the abstractive summary is still a challenging area to work on that require a grate many linguistic effort. Abstractive summarization system tries to realize the theme of the document and express it in natural language. This requires paraphrasing, removal of unnecessary words, and addition of new words.

The biggest challehge for the abstractive summary is to understand the text and represent it in natural language. In a specific domain, it may be feasible to come up with an

appropriate structures, but it is difficult and challenging for open domain semantic analysis.

3. APPROACHES TO ABSTRACTIVE SUMMARY

In this section, we mostly discussed abstract summarization. Different researchers follow a different approach to generate an abstract summary of a text document. The important approaches to generate abstract summary are given below.

1. Template based approach
2. Graph Based approach
3. Discourse and rhetorical based approach
4. Structural approach
5. Statistical and structural approach
6. Hybrid approach
7. Optimization based approach
8. Word association approach
9. Machine learning approach
10. Deep learning approach

3.1 Template based approaches

Generally template based approach creates templates to generate summary of a document. In 1998, Radev et al.[17] proposed a summarization system named as SUMMONS which is a linguistic and conceptual summarization system. To generate summary SMMONS takes a set of templates.

The proposed work [18] is baes on fully abstractive approach. This is an improvement over the previously proposed work by the same authors Genest et al.[19, 20] based on abstraction schemes. The abstraction schemes are nothing but templates designed for each word in criminal domain.

Advantages: This approach is a good approach to generate summary of documents of a specific domain.

Disadvantage: Template based systems are not generic in nature and it is restricted to a specific domain. For each domain we have to define different schemes or templates for different domains to get the summary of a document. The issue with the system is that creating schemes or templates even for a domain is not scalable.

3.2 Graph based approaches

Graph is an important technique where many applications can be modeled as a graph based [21, 22, 23] problem. In this approach, each sentence or word represented as node in the graph and the edge represents the strength of relationship between two nodes. Then some technique is applied to generate summary.

The proposed graph based work [24, 25], first clustered the sentences then to find the difference and commonality between the clusters multiple-sequence alignment (MSA) technique is used. To calculate MSA score only word level similarity is considered but the semantics between the sentences has not been taken into account.

Opoinosis, one of the important graph-based model proposed by Ganesan et al. (2010) [26] generates concise, a non-redundant abstract summary of opinions or reviews given by the users regarding a product. The issue with this system is that it does not validate the gramatical correctness of the newly generated sentence.

Katja Filippova [27] proposed a simple, robust and graph-based model which is almost similar to Opinosis [25] they claim that this is the first technique that requires neither a parser, nor handcrafted rules, nor a language model to generate a grammatically correct sentence.

Other important abstractive summarization approach are [28, 29] use Rich Semantic Graph (RSG) and vertex constrained shortest path scheme to generate the summary,

Fei Liu et al. [30] conduct the research to explore the viability of an abstractive summarization system based on transformations of semantic representations such as the Abstract Meaning Representation (AMR; Banarescu et al.) [31]. In this work, the three steps of summarization are: (1)Creating AMR graphs for each sentence using parser. (2)Combining all the graths into a single summary AMR graph . (3) Generate the text from summary AMR graph.

Advantages: This is one of the popular approach to generate summary. Using graph based approach is used to identify the N-gram phrases. This approach is very useful to identify a valid start and end word of a sentence.

Disadvantages: Almost all the graph based techniques that are applied to generate a summary of a document or a group of reviews on a product try to find valid paths(Valid informative sentence) joining the connected nodes. It is required to validate the grammatical correctness of the generated sentence. Some systems do not validate the generated systems; some systems use syntactical analyzer to validate the sentence which is very costly in terms of execution time.

3.3 Discourse and rhetorical approach

In this approach, researchers use discourse parser to generate discourse tree of the sentences and use the discourse structure to find relationship between sentences. Then find strongly related sentences for summarization. Some researchers use discourse parser to generate aspect hierarchy tree of product reviews to generate generalized message about the product.

The System SEA [32] takes a hand-crafted feature set; then natural language summary is generated using the concept content stuucturing, lexical selection, sentence planing and realization given by Reiter et al. [33].

Rhetorical structure [34, 35, 36, 37, 38, 39], is an important concept used for abstractive summarization. However, it need to parse the text fully, which is a time consuming and complex process.

Advantages: To generate the summary this approach considers the context i.e it finds the relationship with its near by sentences in the document.

Disadvantages: This approach requires complete rhetorical parsing of sentences to understand the rhetorical structure of the sentences. If the corpus is huge then this system is not recommended as full parsing takes lots of time.

3.4 Structural approach

This approach relies mostly on the grammatical structure of a sentence. In the system [43], the sentence passed through a syntactic parser and the output of the parser taken by the system and regenerates the sentence using a NLG (Natural language generation) engine. Summary is generated from the selected regenerated sentences based on the document frequency of contained words.

The work presented in [44, 45] generates abstract summart from abstract representation of the source document not from the sentences. This abstract representation depends on the Information Items (INIT), the smallest coherenet information in the text.

Advantage: The grammatical correctness of the summary generated by this approach is better. This approach performs better in single document summarization compared to multi-document summarization.

Disadvantages: As discussed this approach based on syntactic analysis of sentence therefore these systems rely heavily on syntactic analyzer or syntactic parser. In this approach each sentence need to be passed through the syntactic analyzer in a document. For multi-document summarization its not a good idea to to use structural approach.

3.5 Statistical and structural approach

This is a very popular approach to generate summary. The basic idea of the approach [46] is to extract important sentences based on the words frequencies, position in the sentence and syntactic information of the words. The assumptions of this technique is as follows:

1. The sentences that are related closely to theme of the text occur frequently in the text.

2. The sentences that are related to the topic often occur in some particular structure.

Advantage: This is a very simple approach to generate the summary.

Disadvantages: This approach is more extractive than abstractive.

3.6 Hybrid approach

In this approach, bests of more than one technique or approach are merged to generate the summary. The proposed work Starlet-H [40] is a hybrid approach that takes bests of abstractive and extractive approach. The salient qutes are filtered using using extractive technique and makes abstract summay using Rhetorical Structure Theory (RST) [41].

The proposed method by Le, H. T. et al. [43] generates abstarct summary using discourse rules, syntactic constraints, and word graph. In this approach sentences are created from keywords using discourse rule and syntactic constraints. Word graph is used to combine several sentences into one.

Advantage: Advantage of this system is that it uses the best of the of the different approach and put it in a single pipeline therefore the quality of the summary is somehow better.

Disadvantages: Although the hybrid approach takes the best of other approach still it relies on same discourse parser, syntactic analyzer, sentence parser and word graph etc. therefore the system complexity and the execution time is a somehow increased.

3.7 Optimization based approach

In this approach, the summarization problem is framed as a maximization or minimization problem. The proposed work [47, 48] has two steps, in step one facts and concepts are

created using the phrases from the input document. In step two, new sentences are created by selecting and fuseing informative phrases while satisfy the sentence construction constraints.

Another important work of this approach AdaSum [49] that assumes topic representation and summary can be boosted manually. It aims to optimize the topic representation and extract summary simultaneously.

Advantage: In text summarization it is required to identify noun phrase, verb phrase and talked about topic in a document. Optimization technique frames the summarization problem as a maximization problem to identify noun phrase, verb phrase and topics.

Disadvantages: The problem with the optimization technique is that it uses external solver to solve the maximization problem.

3.8 Word association approach

The proposed method [55] for document summarizatioversion, aims to generate an abstract version of a news story. This models takes a document 'D' and a background corpus 'B' as input. It has two parts, computating document-specific word associations and selecting the sentences with strong word associations.

Disadvantages: Needs a big corpus in backend to get the word association to select the sentence.

3.9 Machine learning based approach

With advancement in technology (i.e Machine learning, Deep learning), there is an increase in active research in abstractive summarization among the researchers. Machine learning is applied in various area [50, 51, 52, 53, 54] and it is giving good results. Following are some important works in abstractive summarization using machine learning.

The proposed fully data-driven approach [56] to abstractive sentence summarization utilizes a local attention-based model that generates each word of the summary conditioned on the input sentence. The abstractive text summarization using Attentional Encoder-Decoder Recurrent Neural Networks [56, 57] and show that they achieve state-of-the-art performance on two different corpora.

3.10 Deep learning based approach

Now-a-days deep-learning techniques are used in many research area [58, 59, 60, 67, 68] to get more accuracy and improved result. Deep-learning technique is used to improve the quality of abstractive summary proposed in [65, 66]. The authors follow a pyramid structure given in Figure 3 to extract knowledge from a text document. Deep-learning technique is used in one or two stages in pyramid structure.

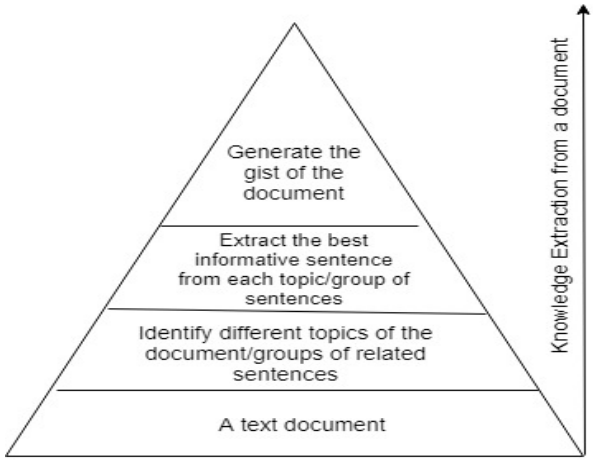

Figure 3. Pyramid approach to knowledge extraction

The summary of some relevant abstractive system elaborated in the table 1.

Table 1: Some selected works on abstractive summarization

Sl. No	**Author& Title**	**Approach /Technique & Input**	**Dataset**	**Evaluation Metric**	**Accuracy**	**Limitation and Future Work**
1	Zhang, J. et. al. (2008), "AdaSum: An Adaptive Model for Summarization"	Optimization Based Ap-proach & multi document	DUC2007	Rouge-one Rouge SU4	ROUGE-2 = 0.1172 ROUGE-SU4 = 0.1692	
2	Bing, L. Et al. (2015), "Abstractive Multi-document	Optimization Based Approach & multi docu-	TAC 2011	Pyramid score	Pyramid score = 0.905	

	Summarization via Phrase Selection and Merging"	men				
3	Gross O. Et al. (2014), "Document Summarization Based on Word Associations"	Word Association & Single document	DUC 2007	Rouge Score	Rouge-one = 0.424 Rouge-two = 0.104 Rouge-3 = 0.036 Rouge-L = 0.384	
4	Rush, A. M. Et al. (2015), "A Neural Attention Model for Abstractive Sentence Summarization"	Machine-learning Based Approach & Single sentence	DUC2004 and Gigaword	Rouge Score	(Using DUC) Rouge-one=28.18 Rouge-two = 8.49 Rouge-L= 23.81	
5	Chopra, S. Et al. (2016), "Abstractive Sentence Summarization with Attentive Recurrent Neural Networks"	Machine-learning Based Approach & Single Sentence	DUC2004 and Gigaword	Rouge Score	(Using gigaword) Rouge-one = 33.78 Rouge-two = 15.97 Rouge-L = 31.15 (Using DUC2004) Rouge-one= 28.97 Rouge-two = 8.26 Rouge-L = 24.06	
6	Nallapati, R. Et al. (2016), Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond	Machine-learning Based Approach & multi sentence	DUC2003	Rouge Score	Rouge-one= 28.97 Rouge-two = 9.46 Rouge-L = 25.24	
7	Fabbrizio et al. (2014), "A Hybrid Approach to Multi-document Summarization of Opinions in Reviews"	Hybrid Approach & Product Reviews	Two sets of labeled data: one for the restaurant domain and the other for the hotel domain.	Manual Readability Correctness Completeness Compactness	Readability 3.75 Correctness = 3.58 Completeness = 3.58 Compactness = 3.72	Takes less number of aspects but require huge amount of training data to learn the ordering of aspects.
8	Genest, P. E. Et al. (2010), "Text Gen-eration for Abstrac-tive Summarization"	Structural Approach & Single document	NA	pyramid and overall responsiveness scores	Pyramid score - 29, Overall response - 33	
9	Genest, P. E et al. (2011), "Framework for Abstractive Summarization using Text-to-Text Generation"	Structural Approach & Single document	NA	pyramid, linguistic quality and overall responsiveness scores	Pyramid score – 29, Overall response – 29, Linguistic quality - 33	
10	Khan, A. et al. (2015), "A framework	Structural Approach &	DUC - 2002	Pyramid score & Average	Pyramid score 0.50, Average	

	for multi-document abstractive sum marization based on semantic role labelling"	Multi document		precision	precision - 0.70	
11	Ren, F. J. (2005), "Automatic Abstracting Important Sentences"	Statistical and Structural Approach & Single document	NA	Manual	Extraction Percentage - 75%	
12	Katja Filippova (2010), "Multi-sentence compression: Finding shortest paths in word graphs"	Graph Based & Multiple Sentences	news articles presented in clusters on Google News	Manual evaluation by human judges.	grammaticality and informativity scores on three-point likert scale for English and Spanish (1.44/1.25 and 1.30/1.25)	
13	Ibrahim F. Moawad et al. (2012), "Semantic Graph Reduction Approach for Abstractive Text Summarization"	Graph Based & Single document	NA	Through case studies	NA	
14	Song et al. (2005), Toward Abstractive Summarization Using Semantic Representations	Discourse and Rhetorical Based Approach & Single document & Single document	Two test data set from Korea Institute of Science and Technology Information	Human judges	NA	
15	Carenin et al. (2013), "Multidocument summarization of evaluative text"	Discourse and Rhetorical Based Approach & Multi document & Multidocument	DUC	Manual mated and Automated	Grammatical, Non-redundancy, Recall Precision	This approach requires a set of hand-crafted features for each product which is not scalable.
16	Gerani et al. (2014) "Abstractive Summarization of Product Reviews Using Discourse Structure"	Discourse and Rhetorical Based Approach		Pairwise preference by crowd sourcing (Manual)	Preference user one 71% Preference user two 69%	
17	Genest et al. (2012), "Fully Abstractive Approach to guided summarization"	Template Based & multi-document	Guided summarizati on task at TAC	Pyramid, (pyramid is a content metric)	Pyramid score 0.54	You must define schemes and its extraction rules for different domains which is not scalable.
18	Opinosis by Gane-	Graph Based &	Document	Automatic	Rouge1-0.330,	This method is

	san et al. (2010),	Reviews of a product	containing reviews of a query	(ROUGE score) and Manual (readability test)	Average readability score = 0.67	more extractive than abstractive and does not validate the grammatical correctness of sentence.

4. EVALUATION AND RESEARCH ISSUES IN AUTOMATIC TEXT SUMMARIZATION

The manual evaluation of a summarization system is a challenging task. Manual evaluation of summary is costly, time taking and likely to suffer from human variability. Therefore, automatic evaluation of summarization system gained popularity among researchers.

Different researchers proposed different automatic evaluation techniques [61, 62, 63, 64] for summary evaluation. Pyramid score [61] and ROUGE score [62] are automatic summary evaluation techniques that considers N-gram lexical overlapping between system generated summary and human created summary for comparison.

Other than any summary evaluation method, ROUGE toolkit has gained popularity and become standard automatic machine generated summary evaluation technique.

5. CONCLUSION

The growth of data and information is very high due to the World-Wide-Web, therefore, there is a high demand to design and develop an efficient and accurate summarization system. Research on automatic text summarization started 50 years back but still a long way to go in this area. In initial days, research on text summarization started with summarizing the research and scientific articles then it shifted to news articles, advertisements, product reviews, electronic mail messages, and blogs. There are two approaches of text summarization that is Extractive and abstractive, and researchers tried both the approach based on the application.

The biggest challenge for text summarization is to extract gist of text from number textual sources for a user. The summarizer should produce a fruitful summary in less time and with least redundancy. This survey emphasizes the abstractive summarization approach. Usually, abstractive summarization requires a-lot-of engineering for language generation and is difficult to replicate or extend to broader domains.