International Journal of Engineering & Technology, 7 (2.33) (2018) 456-460

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Review on automatic text summarization

Abirami Rajasekaran [1] *, Dr. R. Varalakshmi [2]

[1] *Research Scholar, Department of Computer Applications, School of Computing Sciences, VISTAS, Chennai - 600 117, India*
[2] *Dr. R. Varalakshmi, Associate Professor, Department of Computer Applications, VISTAS, Chennai -600 117, India*
**Corresponding author E-mail: abiraja2004@gmail.com*

Abstract

Due to the abundant information available in different forms of sources and genres, there is an immense need to summarize the data for humans. Text Summarization has gained more popularity in this fast-growing information age. Past few years have witnessed a rapid growth in the research of summarizing the text automatically using different approaches. This paper provides an in-depth review of the vari-ous approaches, techniques, methods involved in Automatic Text Summarization.

Keywords: *Automatic Text Summarization; Extractive; Abstractive; Summarization Approaches.*

1. Introduction

Automatic Text summarization is a subset of Natural Language Processing (NLP) and is the process of shortening the source text or set of text documents/paragraph while retaining the main in-formation content. The main aim of the Text Summarization is to create a reduced version of the text preserving its essential infor-mation. The main aspects which one should consider while sum-marization are:

- Generate short summaries.
- Less redundant information summaries.
- Preserve the important information of the source text.

Abstractive Summarization: Abstractive Text Summarization method is the process of using linguistic methods to examine and interpret the text in order to find the new concepts and expressions for generating a new shorter text that conveys the most important information from the original text. As the abstraction process uses linguistic methods and cannot be formulated logically or mathe-matically, this process is not as easy to implement as it requires a deep understanding of the linguistic skills and semantic under-standing of the text.
Extractive Summarization: Extractive Text Summarization selects important sentences, paragraphs etc. from the original text and concatenating them into shorter form without losing the meaning of the text. Majority of the research work done so far have gener-ated summarization systems which are extractive while some work has been done in abstractive summarization as the latter one is harder to develop and one should possess an in-depth knowledge of the linguistics. Automatic Text Summarization approaches involves redundancy elimination, significant sentence identifica-tion, coherent summary generation and evaluate the automatically generated summaries using evaluation metrics.

2. Earlier approaches in text summarization

The early works of literature categorizes three different approach-es to summarization:

Surface-level: This approach represents information in terms of its shallow features in order to identify the salient sentences for summarization such as thematic features, location, background, Cue words etc.
Entity-level: This approach builds an internal representation of the source text with its text entities and relationships between them to identify the patterns in the text which can help in determining the salient sentences for summarization. Some of the entity relation-ships referred in this approach are text similarity, proximity, co-occurrence, co-references, representation-based, logical and syn-tactic relations etc.
Discourse-level: This approach models the structure of the text and its relation based on the text format, threads of topics and rhetorical structure of the text.

3. Text summarization process

3.1. Source input

Identifying the type of source input source in advance can help us in planning for an effective summarization system.
Genre: Summaries can be in different genre such as web pages, scientific articles, legal cases, blogs, software bug reports, single email or email threads etc. Summarization system can process either single or multiple source inputs such as single document or multi-documents, single email or email threads etc.
Language: The source input can be in any type: Mono-lingual or Multi-lingual. In a mono-lingual system, the source input will be in one specific language and the expected output to be in the same language whereas for a multi-lingual system the source and output will be in vice versa format.
Subject Specificity: The source input can be restricted to a specific domain such as medical articles, legal documents etc. which re-quires a deep knowledge about the domain for identifying the concepts or topics of that domain to identify the relevant sentences.
Size: The size of the input document (short or long) can also be considered as an important factor as the performance of the system also relies on the time taken to process the input text.

Type: A source input file or document can be a plain text or can be provided in the form of a multimedia file such as video, audio, picture files etc. Recent years many of the works have been done in video summarization.

3.2. Text preprocessing

The raw text which comes from different sources and genre are generally unstructured, noisy in nature and not suited for summary processing. In order to summarize, these unstructured text needs to be preprocessed where the text goes through step by step process to achieve a structured representation for text processing. Different types of preprocessing techniques used in literature mainly Segmentation, Tokenization, Stemming, Stop Words Removal and Case folding. Other optional preprocessing steps such as Parts of Speech Tagging and Named Entity Recognition are also used.

3.3. Summarization methods

In general, summarization methods are classified into two main categories a) Extractive method selects the key important sentences from the document by using different statistical methods b) Abstractive method creates a semantic representation of the input text to generate a summary. Once the text is preprocessed, the important sentences are identified in the document which can be further considered for the summarization process. Text Summarization process uses several different features for determining the weights of each sentences in the document. The sentences scores are further computed based on the linear combination of these derived weights. Once the Sentence scores are calculated, the sentences are ranked in the descending order of their scores- starting from the highest score at the top to the lowest one at the bottom. The Sentences with the top scores are picked up for the summary generation. The most frequently used feature extraction methods are:

Statistical Methods: Below are the most commonly used statistical methods considered for deciding the importance of sentences in the literature.

Term Frequency-Inverse Document Frequency(tf-idf): Term Frequency (tf) refers to the number of times a term(t) appears in the provided input document(d) whereas the inverse document frequency refers to the number of times a word appears in the given text corpus through which it measures the salience of a word within the document.

Cue Phrases: This method assigns a weight to the sentences based on the presence of certain pragmatic words (classified as positive or negative) such as ‘develop’, 'significant', ‘purpose’, 'hardly', ‘aim’, 'impossible', ‘believe’ etc. as these words or phrases provide a rhetorical context for identifying important sentences [9].

Title: The words in the title, subtitle and headings for the input text are generally considered to be more important as it adds more significance in determining the weight of each of the sentences [9].

Location: Weights are assigned to sentences based on the location where it appears in the beginning or end of the document such as conclusion or summary or in the first and last sentences of a paragraph [10]

Thematic Word: The terms that occur frequently in a document can be more probably related to that topic and these thematic words in a sentence contribute more to compute the sentence scores which are further derived by the number of thematic words in the input sentence over its total length.

Sentence Score (S) = [Number of total Thematic words in the Sentence(S)] / (Total Length of the Sentence(S))

Sentence Centrality: Centrality of a sentence is derived based on the words which overlaps or occurs more frequently within the given sentence (Si) in a document with the other sentences in a document (Others).

Sentence Centrality (Si) = [W (Si) W (Others)] / [W (Si) W (Others)]

Other features such as length of the sentences, pronouns in the sentences, named entities (proper nouns), numerical data, fonts (bold, italic, underlined words) and biased words (domain specific words) etc. also considered to be more important in calculating the sentence scores.

Linguistic Methods: Sentence scoring with linguistic methods are bit more challenging than the statistical methods. Few of the Linguistic methods which are more broadly used by researchers are discussed below.

Graph Theory: The relationships and importance of sentences, can be illustrated effectively using Graphs where the sentences in the document are represented in the form of Nodes and the Edges represents the connection between those sentences. The connection between the sentences are associated based on the similarity relation. Further each sentence is scored and the sentences which has the top scores are selected for the text summarization [10].

WordNet [11]: Each individual word can have more than one sense and each sense can have one or more meaning. For example, a word ‘bank’ can be referred to a river bank or to a financial bank. Word Net, an online lexical English database organizes the English nouns, verbs etc. into set of synonyms for a sense (according to the meaning) called sys-nets and also provides a semantic relation between each sys-net. The Senses are further listed based on their occurrence i.e most frequently used sense in the document is considered to be more important.

Co-Reference chains [12]: Co-Reference refers to one word or phrase refers to the same real world entity. The ultimate aim is to resolve the anaphora in the corpus and to extract the sequences of references with its same referent. The longest co-reference chains are considered to be crucial in framing the sentence scoring.

Clustering methods [13]: This method groups the similar sentences or paragraphs into different clusters to detect a common theme or subtopic among them and then select the textual unit (a representative sentence) from these clusters one by one for summarization. The key factors one should consider in the clustering approach [14] are Sentence Score computation, clustering sentences, cluster Ordering and selecting representative sentences or the sentences with the highest scores from each clusters. The cluster is computed based on the number of important sentences present in it. Different clustering algorithms (K-means, Hierarchical Clustering, Expectation Maximization algorithms etc) have been used widely.

Machine Learning [15]: This approach can be applied when the set of input documents has their corresponding reference summaries for evaluation. A classical machine learning algorithm can be applied in a summarizing system and identify if the sentences generated after the preprocessing step belongs to the reference summaries or not based on the relevant set of features. The algorithm uses the so far learned pattern to classify if the given new sentences belongs to the reference summaries or not. Many supervised, unsupervised and semi supervised machine learning algorithms like Naïve Bayes(NB), Random Forests or Decision Trees, Hidden Markov Models(HMM), Conditional Random Fields(CRF), Support Vector Machines(SVMs) are applied in the automatic text summarization process.

Latent Semantic Analysis [16]: The input document is broken down into linearly independent base vectors or concepts. Singular Value Decomposition(SVD) method is applied to capture the most recurring and relevant word combination pattern from the input document and represents it as singular vectors along with the sentences containing this pattern. Each Singular vector implicitly represents the important topic or theme in the input document. The sentences which contains this word combination pattern will have the largest index value in the singular vector which are further ordered in descending based on the highest index value which is included in the summary.

Fuzzy Logic: This approach uses a Fuzzy Analyzers to calculate the rank of each sentence in the input source text form its statistical parameters. The relation between the statistical parameters are

described by using fuzzy rules (if-then rules). The sentences with the high rank are chosen for the final summary [17].
Neural Networks: The Neural networks are trained to learn the patterns in the given input sentences s to recognize the type of sentence required for summary generation. In the next step feature fusion is applied where the features that are uncommon are eliminated and the common features are collapsed using adaptive clustering technique. The sentences are further ranked based on the cluster. Several types of Neural networks such as Convolutional Neural Network (CNN), Recursive Neural Network (RNN), Recurrent Neural Network (RNN), Feedforward Neural Network etc has been applied in automatic summarization process in the past few decades.
Abstractive Methods:
Structure based: During the initial phase of summarization, the important sentences, phrases, paragraphs from the original text are collected in a predefined structured format without losing its meaning. The predefined structured format can be in any of the below structures:
Rule based: This method is based on abstraction schemes which consist of Information Extraction rules (IER), content selection methods and patterns for generating sentences. Before the extraction rules are created, the verbs and nouns sharing the similar meaning as well as the syntactical roles are identified. The IER translates the annotations into specific candidate answers based on the provided aspect. For each aspect, this IER find several candidates. The Content Selection module determines the best candidate to be included in the generated sentence for each aspect and sends them to further summarization [18].
Ontology based: This method mainly defines the relationship of domain-specific knowledge where the domain ontology is defined by experts of that domain. Mainly Ontology based approaches were successfully applied for extracting information from the input text by a specific domain of interest, Question and Answering (Q&A) systems, e-news summarization etc [19].
Tree based: For a tree based structured abstractive approach, a three step process (Dependency Tree Alignment, Sentence Fusion Computation and Generation) is followed for producing a grammatical and concise summary. The similar sentences are clustered and a dependency parse tree is generated for each sentences in it [20]. The trees are further aligned according to their structural similarity and similarity between their lexical items. The extraneous subtrees present are pruned in this fusion lattice computation process. In the final Generation process, the linearization of the fusion lattice is performed using the entropy-based scoring method.
Template based: In order to retrieve the relevant information from the source input, an Information Extraction system requires a template representation of topic. The Information Extraction system uses the linguistic patterns or other extraction patterns to identify the topic relevant information and populates the template with the text snippets which are the main indicators for the summary. Sentences are further ordered based on their reference resolution.
Lead and Body Phrase: This method was proposed to revise lead sentences in a news broadcast with the concept of insertion and substitution of phrase [21]. The same chunks or phrases in lead and the body sentences are identified in a news article and the identified chunks are aligned based on their metric of similarity. Substitution of a body phrase to a lead phrase occurs if the former has a corresponding phrase in the latter. Insertion of a body phrase to a lead phrase occurs vice versa.
Semantic based: The goal of this approach is to identify the noun-phrase and the verb-phrases by processing linguistic data by using different approaches described below:
Multimodal based: This approach focusses on generating abstractive summaries of a multimodal document [22]. The contents of multimodal documents are basically in the form of text and images. A semantic model is constructed to represent these contents (concepts) by Ontology (knowledge-based). Finally, the important concepts rated based on their relationships and completeness of its attributes are expressed as sentences and stored in a semantic model to form a summary.
Information Item based: This method focusses on the selection of content based on its abstract representation which depends on the Information Items (INIT) [23] which are the basic element of the input text. In order to identify and retrieve this Information Items, the semantic analysis of the original text has to be performed using Semantic Role Labelling, Word-sense disambiguation etc.
Semantic Graph based: In this method in order to represent the sentences in the input text semantically, a Rich Semantic Graph (RSG) [24] is created. Multiple RSGs are generated for each pre-processed input sentence in the input document. In each RSG, the graph nodes represent the verbs of the input text and the edges represent the semantic relations between sentences. Final RSG for the input document will be generated based on the sentences with the highest ranked graphs.
Use the same symbol into a definition over the entire article. Use correct symbols for physical or technical terms. (Example: ε0 and not ε0 for permittivity). Do not repeat definitions over the article. Refer to already defined symbols, equations, theorems by using the cross reference number (Example: As pointed in (1) the…).

4. Summarization techniques

Extractive Summaries: These summaries are generated by selecting the relevant sentences from the input document based on the different methods described in the above sections.
Abstractive Summaries: Unlike the extractive summaries, the abstractive summaries do not include the original sentences from the input text instead it reinterprets the original text in a different form.
The output of an automatic summarization system can also be further classified based on the below categories of summaries:

4.1. Usage

Informative Summaries: As the name suggests, the informative summaries provides detail about the main information or abstract of the text in few lines of summaries. These are summaries with a concise restatement of the main background or domain information of the text [25].
Indicative Summaries: These summaries contain the metadata of the text where it characterizes what the text is all about to the users and does not include any informative content about the text. It just provides an indication about the text and has only partial information about the text [25].
Critical Summaries: Critical Summaries are also called as Evaluative Summaries as it captures the author's summary based on a given subject.

4.2. Audience

Query based Summaries: In Query based summaries, the summaries are generated based on the user interests or it summarizes only the information relevant to the user queries.
Generic Summaries: Since the Query based summaries extracts only a part of information from the main content by satisfying the user queries, these type of summaries does not provide us an overall information of the input text. Whereas, the Generic summaries provides the overall summary of all the information of the input text maintaining minimum redundancy.
Sentence Fusion (or) Information Fusion: Sentence Fusion has been used in Multi-Document summarization and Question and Answering Systems where the common sentences or phrases among the documents are identified and combined to form one single document or a sentence without losing its meaning and maintaining the redundancy. Sentence Fusion has to generate new sentences focusing mainly on the coherence improvement and redundancy.

4.3. Evolution measures

The summaries generated by the summarizing system needs to be evaluated either manually (human) or automatically (system).
Manual Methods or Human Evaluation: The human judges or the assessors are requested to either create new summaries manually or rate through summaries and score them with respect to each of the linguistic factors [26] such as Grammar, Non-redundancy, Focus (Less irrelevant details), Sentence structure, Coherence and Referential Clarity and a five-point scale is adopted to assess the scores of each summary with 5(very good) ranging to 1(void) for each of the above indicators.
Automatic Evaluation: The sentences selected by the summarizing system ($S_{Summarized}$), after successfully applying the summarization techniques, are then compared with the sentences selected by human (S_{manual}) manually to evaluate the Precision (PC), Recall (RC) and F-Score (F). The generated summaries can either be evaluated or assessed by humans or through automatic evaluation methods – Precision (PC), Recall (RC) and using F-Score
Precision: Precision (PC) is the number of sentences which occurs in both summarized system and the manually selected sentences by the number of sentences generated by the summarizing system which in turn is the percentage of selected sentences which are correct.

$$PC = (S_{manual} \cap S_{Summarized}) / (S_{Summarized}) \quad (1)$$

Recall: Recall (RC) is the number of sentences which occurs in both the summarized system and the human summary by the number of sentences in the human summary which in turn is the percentage of correct sentences that are selected.

$$RC = (S_{manual} \cap S_{Summarized}) / (S_{manual}) \quad (2)$$

F-Score: The F-score is the mean of both the Precision and Recall computed above.

$$F\text{-Score} = (2 \text{ X } PC \text{ X } RC) / (PC + RC) \quad (3)$$

ROUGE: Recall-Oriented Understudy for Gisting Evaluation is a tool for evaluating the system generated summary (extraction based summarization) against the summaries produced by humans [27] with the set of metrics.
Pyramid: This pyramid model is used to calculate the quality of information content with the peer summary [28]. A set of span of words that expresses the similar meaning is called as Summarization Content Units (SCU). A weight is assigned to each SCU based on its occurrence in model summaries A Pyramid is formed with the SCU's at the top with greater weights and the SCU's with lower weights at the bottom. The final score of a pyramid for a peer summary is derived on the sum of the weights of SCUs in the summary and also by its maximum sum of SCU weights [29].

4.4. Summary generator

During the summarization process, all the information in the sentences (which has been identified as salient) in a document regardless of its relevance are included. This remains an issue when it comes during the summarization of large amount of text [30] as the final summaries will also hold unnecessary information in it. This problem can be addressed by two approaches [30]:
Sentence Compression: In Sentence Compression method, it "compresses" the sentences in the original text where it removes the unnecessary details or the irrelevant information from the identified key important sentences in order to produce concise summary.
Sentence Fusion (or) Information Fusion: Sentence Fusion has been used in Multi-Document summarization and Question and Answering Systems where the common sentences or phrases among the documents are identified and combined to form one single document or a sentence without losing its meaning and maintaining the redundancy. Sentence Fusion has to generate new sentences focusing mainly on the coherence improvement and redundancy.

5. Conclusion

This paper gives a brief insight about automatic text summarization [8], its techniques, approaches and evaluation measures etc. The main aim of an automatic summarization system should produce a summary with least redundancy and meaningful information within minimum amount of processing time. In future, latest computing techniques available in single or multi document extractive summarization tasks will be explored more in detail.

References

[1] Giuseppe Carenini, Raymond Ng and Gabriel Murray, "Methods for Mining and Summarizing Text Conversations," 2011.
[2] Dragomir R. Radev, Hongyan Jing and Malgorzata Budzikowska, "Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluatio," In ANLP/NAACL Workshop on Summarization, April 2000.
[3] Lin, C. Y. and Hovy, E., "Automated Multidocument Summarization in NeATS," In Proceedings of the Human Language Technology (HLT) Conference, 2001.
[4] Harabagiu, S. and Lacatusu, F, "Generating Single and Multi-Document Summaries with GISTEXTER," In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization), 2002.
[5] Radev, D, Weiguo, F and Zhang, Z, "Webinessence: A personalized webbased multi-document summarization and recommendation system," In NAACL Workshop on automatic Summarization, 2001.
[6] Svore, K., Vanderwende, L and Burges, C., "Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources," In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), p. 448–457, 2007.
[7] Hovy, E. and Lin, C. Y., "Automated Text Summarization in SUMMARIST," In Inderjeet Mani and Mark Marbury editors, Advances in Automatic Text Summarization, 1999.
[8] K. S. Jones, "Automatic summarising: The state of the art," Information Processing \& Management, vol. 43, pp. 1449-1481, 2007.
[9] H. Edmundson, "New methods in automatic extraction," ACM16 (2), p. 264–285, 1968.
[10] H. Luhn, "The automatic creation of literature abstracts," 1959, p. 159–165.
[11] Bellare, Kedar and Sarma, Anish Das and Sarma and Ati, "Generic Text Summarization Using WordNet."
[12] P. Lal, "Text summarisation," Unpublished M. Sc. thesis, Imperial College, 2002.
[13] K. Sarkar, "Sentence Clustering-based Summarization of Multiple Text Documents," TECHNIA – International Journal of Computing Science and Communication Technologies, vol. 2, 2009.
[14] Deshpande, Anjali R and Lobo and MRJ, "Text summarization using Clustering technique," International Journal of Engineering Trends and Technology, vol. 4, 2013.
[15] Neto, Joel and Freitas, Alex and Kaestner, and Celso, "Automatic text summarization using a machine learning approach," Advances in Artificial Intelligence, pp. 205-215, 2002.
[16] Y. Gong and X. Liu, "Generic Text Summarization Using Relevance Measure and Latent Semantic Analysis," Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 19-25, 2001.
[17] Kyoomarsi.F, Khosravi.h. and Eslami.E and Davoudi.M,, "EXTRACTION-BASED TEXT SUMMARIZATION USING FUZZY ANALYSIS," Iranian Journal of Fuzzy Systems, vol. 7, pp. 15-32, 2010.
[18] Genest, Pierre-Etienne and Lapalme and Guy, "Fully abstractive approach to guided summarization," Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, vol. 2, pp. 354-258, 2012.
[19] Embley, David W and Campbell and Douglas M and Smith, "Ontology-based extraction and structuring of information from data-rich unstructured documents," Proceedings of the seventh international conference on Information and knowledge management, pp. 52-59, 1998.

[20] Barzilay, Regina and McKeown and Kathleen R, "Sentence fusion for multidocument news summarization," Computational Linguistics, vol. 31, pp. 297-328, 2005.
[21] Tanaka, Hideki and Kinoshita and Akinori and Kobayaka, "Syntax-driven sentence revision for broadcast news summarization," Proceedings of the 2009 Workshop on Language Generation and Summarisation, pp. 39-47, 2009.
[22] Genest, Pierre-Etienne and Lapalme and Guy, "Framework for abstractive summarization using text-to-text generation," Proceedings of the Workshop on Monolingual Text-To-Text Generation, pp. 64-73, 2011.
[23] Genest, Pierre-Etienne and Lapalme and Guy, "Text Generation for Abstractive Summarization," 2010.
[24] I. F. Moawad and M. Aref, "Semantic graph reduction approach for abstractive Text Summarization," In Computer Engineering & Systems (ICCES), pp. 132-138, 2012.
[25] Wibisono and Yudi and Hendratmo Widyantoro and Dw, "Generating indicative and informative summaries for search engine results," Proceedings of the International Conference on Electrical Engineering and Informatics, 2007.
[26] Saggion and Horacio and Poibeau and hierry, "Automatic text summarization: Past, present and future," Springer, pp. 3-21.
[27] Hassel and Martin, "Evaluation of automatic text summarization," Licentiate Thesis, pp. 1-75, 2004.
[28] Steinberger, Josef and Je{\v{z}}ek and Karel, "Evaluation measures for text summarization," Computing and Informatics, vol. 28, pp. 251-275, 2012.
[29] https://tac.nist.gov//2011/Summarization/Guided-Summ.2011.guidelines.html. [Online].
[30] Dharmarajan, K., and M. A. Dorairangaswamy. "Discovering Student E-Learning Preferred Navigation Paths Using Selection Page and Time Preference Algorithm." International Journal of Emerging Technologies in Learning (iJET) 12.10 (2017): 202-211.
[31] R.V.V Muralikrishna, "A survey on sentence fusion techniques of abstractive text summarization," Int. Journal of Engineering Research and Applications, vol. 5, pp. 59-64, 2015.