Automatic Text Summarization of Extractive and Abstractive: A Survey

Abdulrahman Mohsen Zeyad, Arun Biradar

School of Computer Science and Engineering,
REVA University, Bengaluru.

Abstract.

Manual summary up is one of the old, time-consuming topics. It needs to have a good understanding of the content of the text and practice this task beforehand, with the rapid development of data and the profusion of its resources a data (such as social networking sites, News, Health, Economics, scientific articles, meetings, and books, etc.). It became clear the need to summarize these data and access information in a short time, considering the accuracy of the summary. To achieve the most significant benefit of ATS, the summary remains average in performance while striving to achieve this using modern methods, improving performance, and reaching a summary closer to the human summary. Among the modern methods used, the system of Extractive and Abstract will be discussed in this survey with the perception of ATS.

Keywords— Automatic Text Summarization, Extractive, Abstractive, ROUGE.

1. INTRODUCTION

The field of natural language processing (NLP) is one of the most important fields, given what it provides in several areas that are used daily in many aspects of life, for example, translation from one language to another, summarizing texts while preserving the meaning, reading texts from images and videos, extracting them and converting them to text, Converting speech to text, analyzing texts and giving an impression of the content, automatic text correction, and completion of missing words, generating text or writing a specific script based on the sentence or word and many more. With the rapid development of data and the multiplicity of its resources, whether in paper publications (such as magazines, newspapers, and books) or on the Internet (such as user reviews, blogs, communication networks, scientific papers, scientific websites, books, health and news in all its forms, other important documents) it is considered a source of The sources of vast data in the form of text are increasing rapidly daily. Hence, it consumes much time from users to find information that gives a summary of the search process amid a large amount of repetitive and unimportant data that does not carry any additional information for users, as well as wasting time, effort, and money, they cannot even read and understand all the textual content of the search results, so it is necessary to go to summarize the texts and present them to the user more acceptable to access the information. Manual summarization is one of the best ways to do that, and at the same time, it is expensive and consumes much time in terms of summarization and effort and productivity. Therefore, it has become difficult for humans to manually summarize vast amounts of text data. Therefore,

Automatic Text Summarization (ATS) is one of the tasks of NLP, and it can be the leading solution to this problem for several reasons:

- Given the significant information from a long text in a short time to reduce reading time
- Quick access to the most crucial information at the same time.
- Solve the problem of the correlation of the meaning of the summary with the evaluation of the summary previously.
- The summary is closely related to the original text Algorithm may be less biased than human summaries. Summarize texts in minutes using the software.
- Removes extra text while showing basic information and facts.
- Unlimited use of the summary with a reduction of the summary text.

However, the automatic summary is a task that is not easy because there are many issues, such as the repetition of words, the temporal dimension in the use of words, the standard reference to form the association of words with each other, the importance of arranging sentences with each other, the lexical ambiguity of the word many meanings Semantic ambiguity the sentence has multiple meanings, the syntactic ambiguity of the sentence has several parse trees. And so on, which needs particular attention when summarizing, which makes this task more complex and needs more research and study about Automatic Text Summarization of Extractive and Abstractive (ATS-ES).

2. ATS-ES SURVEY

They have applied BERT word embedding both to extraction, and abstraction with a new approach, using the Attention mechanism, Pointer mechanism, Copy mechanism, and WordPiece tokenizer, using deep learning, to use measurement the ROUGE they got on to (extractive ROUGE-1 42.54 and abstractive ROUGE-1 41.95), using the most popular databases such as “CNN/Daily Mail and DUC2002”[1].

They have applied transformer and seq2seq both to extraction, and abstraction with a new approach, using the Filtering TRANS-ext + filter +abs and Self-attention mechanism, using deep learning, to use measurement the ROUGE they got on to (extractive ROUGE-1 41.52 and abstractive ROUGE-1 41.89), using the most popular databases such as “CNN/Daily Mail and Newsroom”[2].

They have applied BERT-based BERTSUMEXTABS both to extraction, and abstraction with a new approach and New fine-tuning, using the Self-attention mechanism, and deep learning, to use measurement the ROUGE they got on to(extractive ROUGE-1 43.85 and abstractive ROUGE-1 42.13), using the most popular databases such as "CNN/DailyMail, XSum, NYT"[3].

They have applied Factorized Multimodal Transformer both to extraction, and abstraction with a new approach and New fine-tuning, using AVIATE, OCR, and Self-attention mechanism, and deep learning, using databases such as "How2 dataset and AVIATE dataset ", for text summaries of videos[4].

They Improved summary quality performance using a dual-attention pointer network (DAPT), and the LSTM decoder method uses a ‘‘teacher forcing’’ algorithm to use measurement the ROUGE they got on to(ROUGE-1 40.72)), using the most popular databases such as "CNN/DailyMail"[5].

3. ATS-ES APPROACHES

In general, ATS is a complex process that can take time and quality of summaries because computers do not understand natural human language; depending on the type of input documents, the type of databases and topics, the type and quality of the summary may vary.

ATS has two main methods: abstract and extractive. The abstract approach is more productive than the extractive approach to achieving summaries and is closer to a correct understanding of human language.

The advantages of good abstract production are repetition, keyword similarity, sentence position, sentence length, etc.

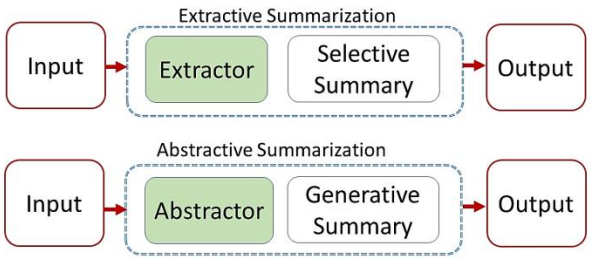

Fig.1 The architecture of an Extractive text summarization system and an Abstractive text summarization system

A. Extractive

It summarizes the extractive text: the method of identifying the essential words and sentences in the entered text to form a summary. This summary selects the most relevant group of words that have meaning. Extractive summarization seeks to arrange the sentences and words in order of importance to them with the exact contents of the words of the text, meaning that it uses the words in the original text. They are linked to form the final summary, with the style used being more robust to the use of existing phrases, but it lacks flexibility not to use new words that express better and not to reformulate the summary to simplify the summary problem.

B. Abstractive

It summarizes the abstractive text: It is an evolution of the traditional methods of text summarization, where the main sections and ideas in the text are identified by reformulating the text by creating new sentences and phrases, meaning representation and re-generation of a summary with sentences different from the original text and original sentences, to give the text greater flexibility and be coherent The meaning of the concept is closer to the human summarization, which depends on reformulating the summary and reducing the size of the summary text while preserving the general meaning. The abstract

summary is considered one of the most complex methods that attempt to solve problems related to the quality of the text and overcome the inconsistency of the text and link them, and supports identifying the most appropriate expressions Sentences based on the content that has been summarized.

4. ATS-ES PROCESSING TECHNIQUES

ATS is one of the most critical challenges of NLP and artificial intelligence in general and access to the use of deep learning methods. Hence, it is necessary to identify the most valuable parts of the text at different stages to be included in the final summary and document long, such as books that were created by the computer without reference to human summaries, so there are three phases of automated summaries:

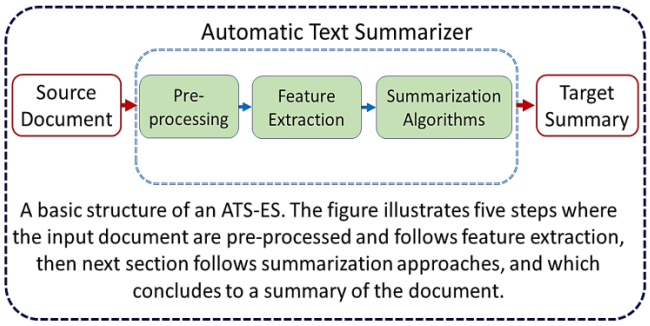

Fig.2 A basic structure of an ATS-ES.

A. Pre-Processing

Pre-processing: Techniques such as techniques of linguistics are used to pre-process the input documents, and linguistic techniques such as sentence segmentation and word encoding Remove stop words, part-of-speech marks, etc. The most commonly used procedures in pretreatment will also be mentioned:

1) Parts of Speech (POS): It is a technique for grouping the words of texts and organizing them according to the context of the speech, such as nouns, adverbs, adjectives, verbs, etc., and determining the type of this word in the grammar and linking it to the rest of the sentences.

2) Stop word filtering: It also determines the shape of the word, whether the word is letters or numbers, and whether it is one of the Stop words or not, and it is used to block words from appearing in the text, such as A, an and by.

3) Stemming: Returning words to an initial form for a group of words and returning them to their roots after removing the inflexions for them, meaning to return the word to its origin, all the words plays, playing, played, player refer to the word play.

4) Named Entity Recognition (NER): The words are recognized in the entered text as noun elements, which are related to the identification and classification of important words, such as the names of people, institutions, names of countries or cities, time, money, percentages, as well as identifying the names of the media from the names

People, companies, cities, currencies, and so on.

5) Tokenization: It is the process of dividing the sentence into a number of parts (words), and it is called one token, or it means separating each word separately to deal with it and know its type, etc. It is based on separating words from sentences so that Each word is alone, and there are two types of it: the word tokenizer, which means separating words, and the sentence tokenizer, which means separating sentences, each sentence separately and examining the words in the text.[9]

6) Lemmatization: It is similar to stemming in the idea, but it is more powerful and effective. It is not satisfied with removing the extra letters in words, but by searching for their meaning and basis, so the words, been, were, whose root is, be, and so on, as it takes into account the meaning in the sentence, the word meeting Its origin maybe meet if it is a present verb, and its origin may be the same as meeting in case it is a noun and not a verb (meaning to meet).

7) Matchers: word association: a tool that allows words to be linked together to make NLP realize that we mean that they have the same meaning, and it is imperative when we search for a specific word that is written in more than one way (Bangalore, Bengaluru) or words with different letters, but we want to We unite them in a specific word (artificial intelligence, artificial intelligence, machine learning, machine learning, deep learning)[6].

B. Feature of Extraction

Feature of extraction processing: It is a method for look the topic, first sentences, or features in the underlying data or from the source text, and the feature extraction method. Associated with them are the number of appearances of words and sentences or text and their importance. In this section, the most frequently used features are mentioned to represent the text and to generate the summary:

1) Term frequency (TF): It is the product of dividing the number of times the required word is repeated by the total number of words in one file to determine the importance and impact and to represent the weight of the word.

2) Term Frequency-Inverse Document Frequency TF-IDF: Two important values, first, Term Frequency, which means the frequency of the required word, and secondly, Inverse Document Frequency, which indicates how rare or common this word is. We have now mentioned the first value, TF; what about the second? The second value indicates how common the word is, and it is inversely proportional to the strength of the word. Note that word itself, if it is widespread and common among all texts, and its use increases in all

other documents, makes it have a weak effect on itself, meaning that any word spreads in many texts and Documents with different meanings. This word often has less value and influence in the meaning and in determining the type of text.

3) Position Feature: Considering that the first and last sentences will give more information about the text, giving an advantage in the summary.

4) Length feature: The length of the sentence may indicate whether it is to be added to the summary. This may cause an error assuming that the lengthy sentence is worth mentioning, compared to the length of the other sentences in the text, so neither relatively short nor long sentences are included in the summary.

5) Sentence-Sentence Similarity: It is partly about measuring the percentage of similarity between two words, whether in letters or the meaning in the text, and it may help summarize.

6) Title feature (Tif): Sentences containing terms from the title included in the summary may be presented because they contain part of the title.

7) Phrasal Information (PI): Attributable phrases are always valuable for summarizing. A group of P phrases also), verbal phrases (VP), nominative phrases (NP), and includes adjective clauses (ADJP)[6].

C. ATS-ES Algorithms

Based on the algorithm used to generate abstracts, ATS is based on two types shown below:

To summarize texts, two machine learning methods are applied - supervised and unsupervised machine learning:

a) Supervised Learning Methods:

Supervised summary, naming documents, the first step is how to learn by training on documents to identify the summarized and non-summarized ones with their labels, which requires a set of data that comes labelled for use in the learning process and also needs to train a sample Data by naming of human-assisted input text. Supervised learning, the learning method, comes at the sentence level to learn sentence discrimination and the characteristics of sentences embedded within the abstract. It also has significant drawbacks in making manual context summaries while requiring more training from previously classified samples for classification; English Wikipedia articles can be dealt with to summarize or group sentences using open-source data.

b) Unsupervised Learning Methods:

Unsupervised summary There is no need to categorize or label the text. Summarization is performed without assistance, i.e., identification of introductory sentences of the text by the user using advanced algorithms and the use of latent semantics to be used as user input and automation. These methods are helpful for big data to give summaries of Sentences that are logical and meaningful.

5. PERFORMANCE OF ATS-ES

Evaluating an abstract is a difficult task because there is no perfect summary of a document or text. Although a good summary is a summary that gives meaning to the original text without neglecting the details and coherently and coherently the, meaning without bias during the summarization that may result from the human summarization and with the various human scales of the summaries and also for the lack of a standard evaluation to determine the main content and important phrases and locate relevant information, Therefore, an automatic assessment is required that determines the effective and reliable assessment of the summary measure. The assessment scale is presented in the field of ATS-ES as follows:

Recall Oriented Understudy for Gisting Evaluation (ROUGE) is a sequence of assessments and matches the automatically generated summary to a set of pre-summarized summaries, ideally like human-generated summaries. Four different scales are ROUGE's: ROUGE-S, ROUGE-N, ROUGE-W and ROUGE-L; three were used; thus, it is a widely used summative assessment sponsored by the National Institute of Standards and Technology (NIST)[7].

ROUGE -1 (Unigram), ROUGE -2 (bi-gram), and ROUGE -L (the longest common suffix) are the most commonly used single-docs.Presence statistic based on N-gram. It can range from being Unigram to Bigram to Trigram, depending on the n-Gram length. Post-common Longest Common Sequence (LCS)-based statistics take into account the similarity of sentence-level structure naturally and determine the longest common occurrence in the sequence n-grams automatically[8]

6. RESULT AND ANALYSIS

Model	ROUGE-1
Extractive Model	
BEAR (ext +large) [1]	42.54
TRANS-ext + filter [2]	42.8
BERTSUMEXT [3]	43.85
Abstractive Model	
BEAR (large + WordPiece) [1]	41.95
TRANS-ext + filter +abs [2]	41.89
BERTSUMEXTABS [3]	42.13

Table 1 Performance comparison of models with Extractive and Abstractive on CNN/Daily Mail

Table 1 shows the performance of models on CNN / Daily Mail datasets. The data in the tables show that the Extractive Model outperformed the Abstractive Model in terms of 1-ROUGE values, which indicates that Extractive exceeds Abstractive based on the results despite the variety of models applied to both of them.

It is clear from this that abstraction provides an understanding and meaning of the text, in contrast to extraction, which links basic sentences or paragraphs without understanding their meaning.

7. CONCLUSION

The ATS is one of the most important applications of NLP today, and with the attempt by researchers to find ways to help improve the quality of summarization by developing summarizing methods and raising the quality of assessment with more effective tools to measure the performance of the summary text and obtain a perfect summarization. The text is better than the abstract summarization of the text, which is less efficient, and the difficulty in its application is the abstract summarization. The field is still open to solving various such complications in the field of ATS.

In the future, hybrid summarization will be added to its study, and there were some experiments to suggest hybrid textual summarization systems with abstract and extractive.

8. REFERENCES

[1] Q. Wang, P. Liu, Z. Zhu, H. Yin, Q. Zhang, and L. Zhang, "A text abstraction summary model based on BERT word embedding and reinforcement learning," Appl. Sci., vol. 9, no. 21, 2019, doi: 10.3390/app9214701.

[2] E. Egonmwan and Y. Chali, "Transformer-based model for single documents neural summarization," EMNLP-IJCNLP 2019 - Proc. 3rd Work. Neural Gener. Transl., no. Wngt, pp. 70–79, 2019, doi: 10.18653/v1/d19-5607.

[3] Y. Liu and M. Lapata, "Text summarization with pretrained encoders," EMNLP-IJCNLP 2019 - 2019 Conf. Empir. Methods Nat. Lang. Process. 9th Int. Jt. Conf. Nat. Lang. Process. Proc. Conf., pp. 3730–3740, 2019, doi: 10.18653/v1/d19-1387.

[4] Y. K. Atri, S. Pramanick, V. Goyal, and T. Chakraborty, "See, hear, read: Leveraging multimodality with guided attention for abstractive text summarization," Knowledge-Based Syst., vol. 227, p. 107152, 2021, doi: 10.1016/j.knosys.2021.107152.

[5] Z. Li, Z. Peng, S. Tang, C. Zhang, and H. Ma, "Text Summarization Method Based on Double Attention Pointer Network," IEEE Access, vol. 8, pp. 11279–11288, 2020, doi: 10.1109/ACCESS.2020.2965575.

[6] M. F. Mridha, A. A. Lima, K. Nur, S. C. Das, M. Hasan, and M. M. Kabir, "A Survey of Automatic Text Summarization: Progress, Process and Challenges," IEEE Access, vol. 9, pp. 156043–156070, 2021, doi: 10.1109/ACCESS.2021.3129786.

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," Expert Syst. Appl., vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.

[8] W. S. El-kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, “Expert Systems with Applications Automatic text summarization : A comprehensive survey,” Expert Syst. Appl., vol. 165, p. 113679, 2021, doi: 10.1016/j.eswa.2020.113679.

[9]. Singh, Konjengbam Dollar, and Syed Thouheed Ahmed. "Systematic Linear Word String Recognition and Evaluation Technique." *2020 International Conference on Communication and Signal Processing (ICCSP)*. IEEE, 2020.