A Systematic Survey of Automatic Text Summarization Using Deep Learning Techniques

Madhuri Yadav and Rahul Katarya

Abstract Today, with the advancement of web technology and the accessibility of the Internet, one can now share their opinion almost everywhere. The Internet user is now not limited to consume the information; instead, they are now information producer. This evolution has resulted in an enormous quantity of data produced over social media and other platforms. This overloaded information makes it complicated for the user to extract and summarize this content. Thus, text summarization is required to generate a shorter version of information to handle the complexity of data. This need leads to exposure to various summarization techniques. This paper gives an overview to the summarization process and studies recent advancements in deep learning-based summarization. Dataset and summary evaluation methodologies are discussed in the area of text summarization. Finally, we conclude the study with various observations, challenges, and future scope in automatic text summarization. This paper is a contribution in providing help to researchers to better understand summarization and find new opportunities in the field of summarization.

Keywords Abstractive summarization · Automatic text summarization · Extractive summarization · Deep learning · Natural language processing

1 Introduction

The web is becoming a repository of data as every user is generating data, resulting in a significant boost in the amount of information available over the Internet. Because of the ease of access to the Internet, there has been tremendous growth in Internet users in the past ten years. Another reason for data generation over the web is due to social media platforms. Due to this overflow of data, a lot of effort is wasted to extract useful information. Not everybody has sufficient time to read all the information available to find the desired information. One solution to this task is to summarize the text or data which is to be read. But manual summarization also becomes difficult and

M. Yadav (✉) · R. Katarya
Department of Computer Science Engineering, Delhi Technological University, New Delhi, India
e-mail: madhuriyadav.me@gmail.com

Fig. 1 Summary
classification based on type
of output summary
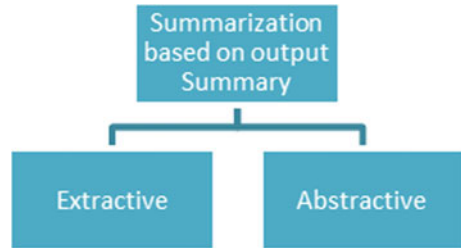
time-consuming when the text is large. Automatic summarization is the task which produces a distilled version of the text where important information is captured from the document. Automatic text summarization is an activity where extraction of significant information is done from one or more sources to produce a condensed version for particular tasks [1]. Basically, text summarization is obtaining the salient features of the input document and producing an output document. Automatic text summarization is divided on the basis of type of output summary: extractive and abstractive (Fig. 1).

The primary focus of extractive summarization is on finding out the salient paragraph and significant sentences that altogether represent the precise summary of the document [2]. The extractive summarization technique selects essential sentences, paragraphs, etc., and concatenates them to make a smaller and distilled version of the source document. The level of importance of sentences is determined on the basis of various statistical and linguistic features of sentences. The abstractive summarization approach, contrarily, understands the text and retells the text in lesser sentences as compared to the source text. For an abstractive summary generation, linguistics methods are used to understand and redefine the text [3].

The major contributions of this work are:

1. This work gives a clear understanding of the summarization process to the reader.
2. The latest work in deep learning is discussed which gives an insight into current progress in the summarization field.
3. Challenges of summarization are discussed for the modern web, which can give a direction to the reader to explore the area of summarization.

2 Automatic Text Summarization

The web is filled with information, and to handle and understand this gigantic data, we need some methods that generate a summary and tell us the important information. Generating a good summary needs a good understanding of the qualities of a good summary [4, 5]. The attributes of a good summary are given in Fig. 2.

Fig. 2 Qualities of a Good Summary
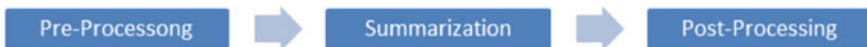
Fig. 3 Basic summarization process

The Summarization Process

The basic architecture of the summarization process can be given as follows in Fig. 3. It contains pre-processing, summarization, and post-processing [6].

i. Pre-processing: This is an important task that generates a structured representation of the source document by applying various linguistic techniques such as sentence segmentation, tokenization, stop word removal, part-of-speech tagging, and stemming [7].
ii. Summarization: In this step, the existing text summarization techniques are applied for generating the summary of the source document. The summarization approaches are discussed in 2.1.
iii. Post-processing: The generated summary may have some structural issues; to make the summary more fluent and structured, we must reorder the sentences. Such tasks are done post-processing [7].

Based on the type of output summary, the ATS approaches are divided as extractive and abstractive summarization.

2.1 Extractive Summarization

While summarizing a document, our main focus is to generate the summary that describes the overall conclusion of the document. For this task, extractive summarization selects those sentences or paragraphs that properly describe the document's importance in precise form. Because of the approach of directly choosing the informative sentences, the method of extractive summarization is easier in comparison with the other techniques. The importance of the sentences is dependent upon the linguistic features [2]. Extractive text summarization can be divided into the following independent tasks [8].

2.2 Abstractive Summarization

Produces a summary that includes the sentences and phrases that do not belong to the input text but holds a similar meaning as to the original text. Abstractive text summarization tries to locate the significant and relevant features of the text and produces a summary by use of phrases/words that may/may not be in the original text [9].

2.3 Evaluation Techniques

Despite the fact that data should be compressed from the huge data available, it is also essential to evaluate the generated summary. The summary evaluation before the automatic evaluation was typically manual. The summaries were manually judged on some specific measures.

Informativeness
It refers to the measure of information gain from the summary generated. To measure informativeness, various methods are proposed. The authors [10] divided these measures into questionnaire-based and overlap-based matrices: Questionnaire-based summary evaluation is based on a questionnaire set that is designed from the original documents. Overlap-based matrices for summary evaluation are based on the similarity of the generated summary to the reference summary.

There are three common criteria for proficiency evaluation of a summarization system [10]: Recall, Precision, given by Eqs. 1 and 2, respectively:

$$R = \frac{|S \cup C|}{|S|} \tag{1}$$

$$P = \frac{|S \cap C|}{|C|} \tag{2}$$

where S and C are the collection of sentences/terms in reference summary and candidate summary, respectively. Another metric that is used for summary evaluation is ROUGE(Recall-Oriented Understudy for Gisting Evaluation) was a summarization task in the Document Understanding Conference (DUC) [11]. ROUGE is based on overlap measures in comparison with the reference summary. The ROUGE matrices have several variants ROUGE-1, ROUGE-2, ROUGE-N, ROUGE-L, ROUGE-S, etc.

Table 1 List of the popular datasets

Dataset	Description	Research done
DUC[1]	The data available here consist of all the previous DUC containing documents and summaries (manually created, automatically created baselines, submitted summaries by participating groups), evaluation results, etc.	[12–14]
TAC[2]	Test data, summary, and evaluation results	[15–17]
Opinosis dataset[3]	This is a public dataset that contains sentences extracted from user reviews on 51 topics collected from Tripadvisor, Edmunds.com, and Amazon.com [18]	[18, 19]
Customer reviews[4]	A public dataset contains customer reviews of the product from Amazon.com, including information about the reviewer, review texts, ratings, etc. This dataset is best suited for aspect-based summarization [20]	[20, 21]
FIRE[5]	Previous year's news data for Indian languages such as Hindi. Bengali, Marathi, and Odia are available	[22]
CNN/Daily mail	It is a dataset collection of around 300 k English news article	[23]

2.4 Datasets

There are plenty of datasets available; some of the most popular are mentioned in Table 1.

3 Deep Learning-Based Automatic Text Summarization

Advancement in deep learning methods has been beneficial for artificial intelligence and NLP. Here, we study recent development in deep learning techniques in the summarization field. We divide this section into two subsections: extractive and abstractive text summarization using deep learning.

---

[1] https://duc.nist.gov/data.html.

[2] https://tac.nist.gov/data/index.html.

[3] http://kavita-ganesan.com/opinosis-opinion-dataset/#.XnCMuKgzZPY.

[4] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html.

[5] http://fire.irsi.res.in/fire/static/data.

3.1 Deep Learning-Based Extractive Summarization

Extractive summarization is basically a method where we try the sentence with some technique and then choose the most significant sentences to make the output sentence. In [24], the authors did a single document summarization using a hierarchical sentence encoder and attention-based extractor. The sentence extraction is done using supervised training. The authors used a word extractor to minimize the redundancy in the output. The dataset used was DUC2002 and Daily Mail. The sentence representation was done using a hierarchical document reader using convolutional neural network, and the selection of sentences was done by the attention-based extractor. The authors of [25] proposed a method called PriorSum, which is a multi-document summarizer which is a neural network-based technique that captures the important aspect with considering the context. Sentence ranking is based on improved CNN by obtaining context-independent features. For this work, DUC 2001, DUC 2002, and DUC 2004 were used. [23] Introduced extractive summarization as a problem of sequence classification. Here, the text is traversed sequentially to check whether the sentence should be included or not based on previous decisions made for sentences. Here, the authors use RNN-based classifier. The dataset used is DUC2002and Daily Mail.

Singh et al. [26] Introduced a bilingual (Hindi and English) multi-document summarization approach. Here, unsupervised deep learning was used. They used Restricted Boltzmann Machine. The pre-processing steps include segmentation, tokenization, stop word removal, part-of-speech tagging, and feature selection. These were done using TF-IST. [27] In this work, authors used unsupervised deep learning for single document query-focused summarization. Here, the authors used deep autoencoder for the summarization process. The model was divided into two tasks pre-training and fine-tuning. A noisy autoencoder was used for sentence representation and ranking. The sentence selection was done by applying cosine similarity on the concept vector which was generated using the encoder.

3.2 Deep Learning-Based Abstractive Summarization

In [28], the author developed an abstractive summarization approach. This work is about using the deep learning concept to generate the summary. Three encoders were used: a Bag-of-words encoder, CNN, and an attention network for decoding purposes. The author used a neural network language model as a decoder. For word generator, probability distribution was used. The experimentation was done on the Gigaword and DUC datasets. For the optimizer, the stochastic gradient descent was used.

Yao et al. [29] The author developed a deep learning-based summarization approach. Here, instead of using one encoder the authors tried to improve the process by adding a dual encoder. The secondary encoder is other than the regular encoder

models in sentence score based on the history. In the primary encoder, the bi-direction GRU-based RNN was used. For secondary, the unidirectional GRU-based RNN was used. This output is fed into an attention-based decoder. The experiments were performed on CNN/Daily Mail and DUC 2004. Authors of [30] also focused on CNN encoder and RNN decoder. For optimization purposes, SGT was used as used in [28]. In [31], introduced a framework which focused on encoding the keywords to the key information representation, used CNN-Daily Mail dataset encoder was a bidirectional LSTM and decoding was done using LSTM. Authors in [32] proposed a framework fact aware summarization using reinforcement learning to enhance the factual correctness of the summary. In work [33], authors developed an LSTM-based abstractive summarization that focused on finding the key features. For this, the authors divided the process into two phases: where the first one finds the important features and the second phase generated using DL techniques.

Here, we tried to discuss the most recent work done in summarization using deep learning. Despite being this, much research in the area of summarization this field still faces some challenges that are discussed in the next section.

4   Challenges in Text Summarization

Processing the text has been always a challenging task but with emerging times the level of challenges has also increased. Now, the people on social media are talking in their own languages other than English. Processing such text can be a difficult task. Also, the text over the web is not grammatically correct, spellings are wrong, and uses of slang are very common. Processing such text can be very challenging. Producing an abstractive summary is also very challenging as it focuses on conciseness and informativeness both at the same time. In the next section, we discuss some suggestive future work in this area so that these challenges can be tackled.

5   Conclusion and Future Work

Automatic text summarization is a very important task in natural language processing and for the current situation, i.e., web is overloaded with information, summarization becomes very important. In this survey, we discussed the entire process of summarization and approaches of summary generation. This has been concluded that the area of extractive summarization is explored and researched more than abstractive summarization. In this work, we have discussed deep learning-based research in the area of summarization. Supervised methods of extractive and abstractive summarization are explored well, but still challenges are being faced. In the future, we intend to develop a good unsupervised deep learning-based algorithm for text summarization, especially for abstractive summarization.

References

1. Belguith L, Ellouze M, Maaloul M, Jaoua M, Jaoua F, Blache P (2014) Automatic summarization. Natl Lang Process Semitic Lang, 371–408
2. Moratanch N, Chitrakala S (2017) A survey on extractive text summarization. In: 2017 international conference on computer, communication and signal processing (ICCCSP)
3. Dalal V, Malik L (2013) A survey of extractive and abstractive text summarization techniques. In: 2013 6th international conference on emerging trends in engineering and technology
4. Shareghi E, Hassanabadi L (2008) Text summarization with harmony search algorithm-based sentence extraction. In: Proceedings of the 5th international conference on soft computing as transdisciplinary science and technology—CSTST '08
5. Goldstien J, Mittal V, Carbonell J, Kantrowitz M (2000) Multi-document summarization by sentence extraction. In: Proceedings of the 2000 NAACL-ANLPWorkshop on automatic summarization
6. El-Kassas W, Salama C, Rafea A, Mohamed H (2021) automatic text summarization: a comprehensive survey. Expert Syst Appl 165:113679
7. Gupta V, Lehal G (2010) A survey of text summarization extractive techniques. J Emerg Technol Web Intell 2
8. Nenkova A, McKeown K (2012) A survey of text summarization techniques. Mining Text Data 43–76
9. Raphal N, Duwarah H, Daniel P (2018) Survey on abstractive text summarization. In: 2018 international conference on communication and signal processing (ICCSP)
10. Ermakova L, Cossu J, Mothe J (2019) A survey on evaluation of summarization methods. Inf Process Manage 56:1794–1814
11. Lin C (2004) ROUGE: a package for automatic evaluation of summaries. Text Summarization Branches Out 74–81
12. Luo W, Zhuang F, He Q, Shi Z (2013) Exploiting relevance, coverage, and novelty for query-focused multi-document summarization. Knowl-Based Syst 46:33–42
13. Canhasi E, Kononenko I (2014) Weighted archetypal analysis of the multi-element graph for query-focused multi-document summarization. Expert Syst Appl 41:535–543
14. Al-Sabahi K, Zhang Z, Long J, Alwesabi K (2018) An enhanced latent semantic analysis approach for arabic document summarization. Arab J Sci Eng 43:8079–8094
15. Jin F, Huang M, Zhu X (2009) A query-specific opinion summarization system. In: 2009 8th IEEE international conference on cognitive informatics
16. Campr M, Jezek K (2012) Comparative summarization via latent semantic analysis. Latest Trends Inf Technol
17. Jezek K, Campr M (2013) Comparative summarization via latent dirichlet allocation. Dateso 2013:80–86
18. Ganesan K, Zhai C, Han J (2010) Opinosis: graph-based approach to abstractive summarization of highly redundant opinions. In: Proceedings of the 23rd international conference on computational linguistics
19. Bhargava R, Sharma Y, Sharma G (2016) ATSSI: abstractive text summarization using sentiment infusion. Procedia Comput Sci 89:404–411
20. Hu M, Liu B (2004) Mining and summarizing customer reviews. In: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '04
21. Hu M, Liu B (2004) Mining opinion features in customer reviews. In: Proceedings of the 19th national conference on artificial intelligence (AAAI'04)
22. Verma P, Pal S, Om H (2019) A comparative analysis on Hindi and English extractive text summarization. ACM Trans Asian Low-Resource Lang Inf Process 18:1–39
23. Nallapati R, Zhai F, Zhou B (2017) SummaRuNNer: a recurrent neural network based sequence model for extractive summarization of documents. In: AAAI'17: proceedings of the thirty-first AAAI conference on artificial intelligence, 3075–3081

24. Cheng J, Lapata M (2016) Neural summarization by extracting sentences and words. In: Proceedings of the 54th annual meeting of the association for computational linguistics, vol 1: Long Papers
25. Cao Z, Wei F, Li S, Li W, Zhou M, Wang H (2015) Learning summary prior representation for extractive summarization. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, vol 2: Short Papers
26. Singh S, Kumar A, Mangal A, Singhal S (2016) Bilingual automatic text summarization using unsupervised deep learning. In: 2016 international conference on electrical, electronics, and optimization techniques (ICEEOT)
27. Yousefi-Azar M, Hamey L (2017) Text summarization using unsupervised deep learning. Expert Syst Appl 68:93–105
28. Rush A, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: Proceedings of the 2015 conference on empirical methods in natural language processing
29. Yao K, Zhang L, Du D, Luo T, Tao L, Wu Y (2020) Dual encoding for abstractive text summarization. IEEE Trans Cybernetics 50:985–996
30. Chopra S, Auli M, Rush A (2016) Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies
31. Li C, Xu W, Li S, Gao S (2018) Guiding generation for abstractive text summarization based on key information guide network. In: Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: human language technologies
32. Zhang M, Zhou G, Yu W, Liu W (2021) FAR-ASS: fact-aware reinforced abstractive sentence summarization. Inf Process Manage 58:102478
33. Song S, Huang H, Ruan T (2018) Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools Appl 78:857–875