Abstractive Summarization System

Amit Kumar[1], Manoj Kumar Gupta[2]

[1]Student, Shri Mata Vaishno Devi University, Katra, India
[2]Associate Professor, Shri Mata Vaishno Devi University, Katra, India

E-mail: [1]20mms002@smvdu.ac.in

Abstract

The World Wide Web has evolved into one of the world's most extensive information and knowledge repositories. Despite their ease of access, the great majority of such individual publications are extremely difficult to analyse or evaluate. Text summaries assist users in achieving such information-seeking goals by providing rapid access to the highlights or important features of a document collection. Abstractive summarization attempts to reduce a given text to its core components based on the user's preference for brevity. To summarise, there are two approaches: extraction and abstraction. Statistical techniques are used for extracting most important sentences from a corpus. Abstraction entails reformulating material based on the type of summary. This approach makes use of more adaptive language processing technology. Despite the fact that abstraction yields better summaries, extraction remains the favoured strategy and is widely employed in research. A number of approaches, including cosine, can be used to calculate the measure of resemblance between articles. Sentences' statistical & linguistic features are utilised to determine their importance. An abstractive summary is used to absorb the fundamental concepts of a material and then summarise them into plain English.

Keywords: Text summarising, Natural Language Processing (NLP), Extractive Summary, and Abstractive Summary

1. Introduction

Despite the fact that computational techniques for NLP have been explored for decades, achievements in natural language processing (NLP) are relatively recent. Textual data processing demands a slew of diverse building blocks from which complicated models can be built. Some of these building components may be challenging and complex on their own. Text summarising is the practise of creating a reduced version of a part while retaining

DOI: https://doi.org/10.36548/jei.2021.4.006
Received: 09.03.2022, received in revised form: 02.04.2022, accepted: 18.04.2022, published: 27.04.2022

its value. Removing this impediment is a vital step toward NLP. A concise and well-written summary may also benefit in the quicker absorption of text material.

Summarization systems are classified into two categories. Extractive models create summaries by cutting out key sections of the source text and combining it into a single entity. Abstractive types of models create summaries from scratch rather than reusing terms from the original text. Abstractive Text Summarization (ATS), the idea that summary articles may be created by gathering information from several sources and condensing it into a concise presentation while retaining the content component and overall tone. Algorithms for top-down information extraction look for a predefined number of information categories in the summary.

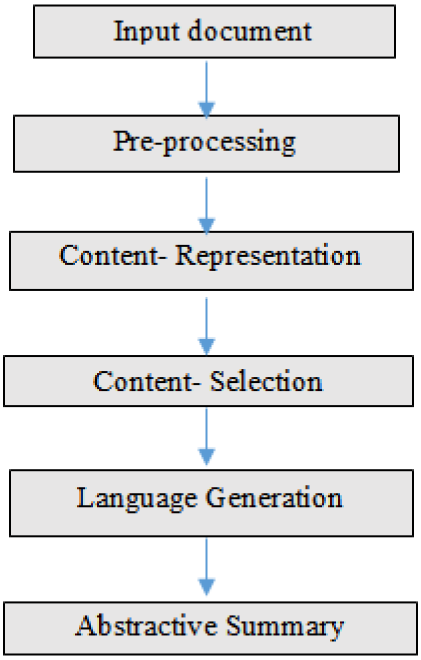

Figure 1. Workflow

Pre-processing: Text preparation encompasses all of the standard preprocessing required in Natural Language Processing (NLP). Tokenization, stop-word deletion, lemmatization, stemming, and other methods are commonly utilized during preprocessing. We can focus on crucial information rather than superfluous word fragments since the text has been preprocessed.

Content Representation (CR): Content representation semantically analyzes the input text using various statistical analytic methodologies. To improve the semantic analysis contained in the methodologies given below, Content Representation (CR) includes structuring the contents in various formats such as tree and Ontology (often referred to as intermediate form).

Content Selection (CS): The intermediate form developed in content representation is used in the Content Selection process to determine the most significant phrases as well as semantically equivalent terms using semantic analysis.

Language Generation (LG): Natural Language Generators (NLG) then examine the content selection's picked or paraphrased bits to provide the required abstractive summary for the given text. To achieve the desired outcome, NLGs employ statistical estimators and decision procedures such as the Markov model.

2. Literature Review

Song, S. et al. [1]. Abstractive text summarization using LSTM-CNN based deep learning. In the paper, the authors presented an ATS framework (ATSDL) based on LSTM-CNN that may create new sentences by looking at finer-grained parts than sentences, particularly semantic phrases. The ATSDL model extracts relevant phrases from the original text using a phrase extraction method known as MOSP, and then learns phrase grouping. After learning, the model will generate word sequences with the necessary linguistic structure. Furthermore, they employ word location information to address the issue of odd words, which practically all ATS methods experience. Finally, they do extensive testing on several datasets and found that the model beats province approaches of both semantic information and syntactic structure.

Chopra et al. [2]. Attentional recurrent neural networks were used to summarise abstractive sentences. The authors suggested a conditional (RNN) that provides a description of an input document in this work. During development, an attention-based convolutional encoder guarantees that the decoder is dependent on the correct input words. The article's most significant contribution is a unique convolutional attention-based conditional (RNN) model for abstractive phrase summarization.

Rush et al [4] employ a feed-forward nlp model for generation, while the author uses a RN network. They use an RNN architecture to improve the provience abstractive phrase summarization model (Rush et al. [4]). It is the reduced version of the machine translation encoder-decoder architecture (Bahdanau et al., 2014). Paulus et al. [3]. A sophisticated reinforcement model was employed for abstractive summarization. In this paper, the author presents a novel training technique that blends traditional supervised word selection with recurrent neural networks, providing a neural network model with innovative intraattention that adjusts for input independently and continually generated output. The unique model and training approach that produces reducing text summarization outcomes for CNN/Daily Mail, makes it a lot easier for generated summaries, and is best suited for long output sequences.

Rush et al. [4]. A neural attention(NN) model was created for abstractive phrase summarization. To construct each letter of the summary from the input text, the author of this study used a model depending on local attention. Zhang et al. [5] proposed a comprehensive framework for comprehending raw characters that is applied to a unique rnn integrated neural network model. It contains an updated language model that filters out superfluous information as part of the attention model. The optimal translation model considers the complete learning process. This model successively predicts each word in the output phrase and returns a fixed-length vector comprising all encoded data.

For translation prediction, statistical machine translation used aligned source terms and source contextual information as per Chen et al. [6]. Learning context representation, which is primarily focused through the neural network, may be utilized to enhance translation outputs. The authors next provide a unique neural technique for translation prediction based on Input contextual representation based on dependency. The approach presented is validated by incorporating it into any hierarchical or phrase-based translation paradigm. On large-scale Interpretation projects from english to german and from chinese to english, the proposed model surpasses several current context-enhanced approaches, according to experimental data.

A hierarchical sequence attentional NMT model was used to address the translational task in the way as suggested by Su et al. [7]. This structure stores phrases, words, and sentences at many levels, including clause and word level, using a RNN model with two hidden layers. The decoder translates sentences while learning inter clause information and anticipating intra clause translation using two types of attention models. To finish the difficult inquiry response, Yllias et al. [8] used a reinforcement learning problem for extractive multi-document summarization. The training data set includes a list of difficult questions, relevant papers for the specified subjects, and summaries generated by humans. The learning phase studies the characteristics of the variables as well as the weights associated with them, which is required for proper summary creation. The reward function compares human-generated and automated summaries for commonalities. The feature weight will be modified based on the reward function throughout the training phase. Throughout the testing process, the altered feature weights are employed to respond to more complicated queries. The evaluation findings indicated that reinforcement learning significantly improved.

Hsu et al. [9] created a unified framework with a novel attention mechanism that utilises both abstractive and extractive summarization. It also includes word-level consideration (based on the Pointer-Generator organisation See et al. [11] and sentence-level consideration (fixed on extractive outline Nallapati et al [10], with the expectation keywords in sentences should be forced to reduce consideration results and less is done to demonstrate that. Lloret et al. [12] produced extremely brief concept-level summaries. Following lexical analysis, it turns the input document into its syntactic representation. This software constructs language by collecting and summarising lexical elements. This strategy is founded on the premise that text lacks a meaningful representation.

SentiCircles, a lexicon-based technique, is employed in the system, which does sentimental analysis at the entity and tweet levels (Saif et al. [13]). The Twitter data was used for sentiment analysis. SentiCircles look for patterns of co-occurrence in tweets. SentiCircles are now more effective for fragmented sentences, but not for the entire context, which is what the approach is intended for. Barros et al. [14] created the Narrative Abstractive Summarization (NATSUM) approach for constructing a narrative and sequentially ordered summary of a target item from a collection of news items on the same topic. As a result, a narrative abstractive summary, rather than a chronological(sequential) summary, gives more relevant facts drawn from several sources.

Fang et al. [15] developed a word to sentence co-ranking technique which improves the sentence score. They contend that words and sentences interact in such a way that words appearing in high-scoring sentences and phrases containing high-scoring words receive high scores. The word-sentence connection is used in the specified activity. As a result of reciprocal influence, the underlying state of words or sentences may be revealed more precisely to graph-based rating algorithms. This co-ranking approach is intended for summarising a single document.

Tayal et al. [16] gave text summarization by utilising soft computing principles such as Subject, Verb, and Object. It uses a Pos tagging, an NLP parsing, semantic organization, and sentence reductions and combinations to handle data. In the field of machine learning, data reconstruction, along with feature selection and dimension reduction, has attracted a lot of attention. Chowanda et al. [17] developed a method for obtaining summary information from an online discussion forum. They used point-based methods, in which the verb as well as its syntactic parameters are included within the point. This technique is broken down into three sections feature extraction, feature curation, and summarize output. Feature extraction refers to the process of extracting subjects from literature; point curation refers to the process of strengthening the point; and summary creation refers to the process of constructing the summary.

Moradi and Ghadiri et al. [18] improved the effectiveness of the biomedical summarising technique by introducing domain knowledge into the Bayesian summary process using the UMLS concept. Using a six-criteria approach, they identified important themes in biomedical literature. The Unified Medical Language System has shown to be an effective tool for biomedical summarization. The vocabulary of the condensed text must be mapped onto the ideas it includes when using the UMLS. This is exacerbated by lexical ambiguity, which means that words can have many meanings depending on context. Biological materials have been found to be highly ambiguous, defying the widely held belief that technical domains are less ambiguous than general domains.

Hamza Shabbir et al. [19] compared several summarization methods. Text summarization, according to the author, is a major concern in the field of NLP. To create extractive and abstractive summaries, methods such as sentence extraction, paragraph extraction, deep understanding, and machine learning algorithms are utilised. As a result of merging one or more of these approaches with NLP techniques, several ways have been created. Even if these tactics are used to build the summary, the author claims that it is a challenging undertaking to produce a good summary.

Johan Hasselqvist et al. [20], proposed query-based summarising, which is a methodology for constructing text document summaries based on the user's demand, which is delivered in the form of a question. The author used a previously existing dataset of news stories. A similarity computation is used to compare the prepared summaries to the reference summaries. The author presented a sequence-to-sequence model. The summarizer receives both a document and a query as input. These are the word sequences that are sent to the document and query encoders. The encoder's output is sent into the decoder, which generates a summary. The issue with this strategy is that the reference queries returned for the two separate inquiries may be the same.

Recent research has concentrated on extracting and analysing entire phrases with data-driven extractive summarization based on neural networks (Dlikman & Last et al [21]) (Lapata and Cheng et al [22]). (Nallapati et al [23]) employed an over-extracting classifier technique, but word-level extraction necessitates linguistically acceptable output (Bui et al. [24]), which might be difficult to identify if it consists of a phrase and a filter that scores the items that are positively categorised. In contrast, human-generated summaries are frequently connected with ungrammatical key phrase extraction.

Table 1. Comparison of various datasets using different Machine Learning Algorithms

Author & Year	Dataset	Method	Approach	Accuracy
[1] (2016)	Gigaword corpus, DUC-2004	Conditional Recurrent neural network (RNN) RNN Encoder-Decoder	ABS+ Abstractive Sentence Summarization	ROUGE-1 33.10
[2] (2019)	CNN and DailyMail datasets	LSTM-CNN based Abstractive Summarization System.	In both semantic and syntactic structure, state-of-the-art techniques are used.	ROUGE-1(%) - 34.9 ROUGE-2(%) -17.7
[3] (2017)	CNN and DailyMail datasets.	RNN-based encoder-decoder	Word prediction and reinforcement learning .	ROUGE-1(%) - 42.94 ROUGE-2(%) -26.02
[4] (2015)	DUC-2004	For abstractive summarization, a neural attention encoder-decoder is used.	Abstractive phrase summarization using a data-driven technique.	ROUGE1-29.21
[5] (2016)	Subtitles of movies or television series from the Internet.	The encoder-decoder for a Recurrent Neural Network (RNN).	RNNembed learning that use sequence-to-sequence method.	NIST BLEU-25.29
[6] (2017)	LDC dataset (1.423 million sentence pairs)	NN based methods, for learning target mono lingual and discrete bilingual context representations.	Statistical + semantic representation of context.	NMT - 19.78

[7] (2018)	NIST 2005 (MT05).	A bi-attention based decoder is created using a hierarchical RNN encoder.	Machine learning technique is used to generate summary Neural network is used to enhance the sentence score.	BLEU - 20.36
[8] (2015)	DUC-2006 and DUC-2007	The technique of reinforcement learning Watkins' Q(λ) approach is used in recurrent neural networks to simulate the challenging question answering task.	The semantic and syntactic resemblances between a chosen summary and an abstract summary.	ROUGE 2 - 0.0863

3. Conclusion

Because of the vast quantity of information available on the internet, there is a growing demand for compressed, meaningful, abstract subject summarization as a subfield of NLP. Exact data boosts search productivity and efficiency, and to develop an abstractive summarising system that can provide output that is almost equivalent to human-framed summaries. The fundamental purpose of creating such a system is to suit the needs of readers who read books, articles, and other forms of media. This paper delves into the complexities of both extractive and abstractive strategies, as well as the methodology used, the results achieved, and the advantages and disadvantages of each strategy. Text summarising is valuable in both business and academia. Abstractive summarising is more challenging than extractive summarization since it involves more learning and reasoning. In contrast to extractive summarising, abstractive summarization generates more meaningful and relevant summaries. The suggested approaches outperformed others in testing utilising a corpus of words obtained from commercial newspaper websites and the CNN-daily corpora. Future researchers on low-resource languages can focus on basic research, such as building a corpus for further study and developing novel algorithms to handle specific languages.

References

[1] Chopra, S., Auli, M. and Rush, A.M., 2016, June. 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 (pp. 93-98).

[2] Song, S., Huang, H. and Ruan, T., 2019. Abstractive text summarization using LSTM-CNN based deep learning. Multimedia Tools and Applications, 78(1), pp.857-875.

[3] Paulus, R., Xiong, C. and Socher, R., 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

[4] Rush, A.M., Chopra, S. and Weston, J., 2015. A neural attention model for abstractive sentence summarization. arXiv preprint arXiv:1509.00685.

[5] Zhang, H., Li, J., Ji, Y. and Yue, H., 2016. 'Understanding subtitles by character-level sequence-to-sequence learning'. IEEE Transactions on Industrial Informatics, 13(2), pp.616-624.

[6] Chen, K., Zhao, T., Yang, M., Liu, L., Tamura, A., Wang, R., Utiyama, M. and Sumita, E., 2017. A neural approach to source dependence based context model for statistical machine translation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(2), pp.266-280.

[7] Su, J., Zeng, J., Xiong, D., Liu, Y., Wang, M. and Xie, J., 2018. 'A hierarchy-to-sequence attentional neural machine translation model'. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(3), pp.623-632.

[8] Yllias Chali, Sadid A Hasan&Mustapha Mojahid 2015, 'A reinforcement learning formulation to the complex question answering problem', Information Processing & Management, vol. 51, no. 3, pp. 252-272.

[9] Hsu, W.T., Lin, C.K., Lee, M.Y., Min, K., Tang, J. and Sun, M., 2018. A unified model for extractive and abstractive summarization using inconsistency loss. arXiv preprint arXiv:1805.06266.

[10] Nallapati, R., Zhai, F. and Zhou, B., 2017, February. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In Thirty-first AAAI conference on artificial intelligence.

[11] See, A., Liu, P.J. and Manning, C.D., 2017. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

[12] Lloret, E., Boldrini, E., Vodolazova, T., Martínez-Barco, P., Muñoz, R. and Palomar, M., 2015. A novel concept-level approach for ultra-concise opinion summarization. Expert Systems with Applications, 42(20), pp.7148-7156.

[13] Saif, H., He, Y., Fernandez, M. and Alani, H., 2016. Contextual semantics for sentiment analysis of Twitter. Information Processing & Management, 52(1), pp.5-19.

[14] Barros, C., Lloret, E., Saquete, E. and Navarro-Colorado, B., 2019. NATSUM: Narrative abstractive summarization through cross-document timeline generation. Information Processing & Management, 56(5), pp.1775-1793.

[15] Fang, C., Mu, D., Deng, Z. and Wu, Z., 2017. Word-sentence co-ranking for automatic extractive text summarization. Expert Systems with Applications, 72, pp.189-195.

[16] Tayal, M.A., Raghuwanshi, M.M. and Malik, L.G., 2017. ATSSC: Development of an approach based on soft computing for text summarization. Computer Speech & Language, 41, pp.214-235.

[17] Chowanda, A.D., Sanyoto, A.R., Suhartono, D. and Setiadi, C.J., 2017. Automatic debate text summarization in online debate forum. Procedia computer science, 116, pp.11-19.

[18] Moradi, M. and Ghadiri, N., 2018. Different approaches for identifying important concepts in probabilistic biomedical text summarization. Artificial intelligence in medicine, 84, pp.101-116.

[19] Moiyadi, H.S., Desai, H., Pawar, D., Agrawal, G. and Patil, N.M., 2016. NLP based text summarization using semantic analysis. International Journal of Advanced Engineering, Management and Science, 2(10), p.239678.

[20] Hasselqvist, J., Helmertz, N. and Kågebäck, M., 2017. Query-based abstractive summarization using neural networks. arXiv preprint arXiv:1712.06100.

[21] Dlikman, A. and Last, M., 2016, January. Using Machine Learning Methods and Linguistic Features in Single-Document Extractive Summarization. In DMNLP@ PKDD/ECML (pp. 1-8).

[22] Cheng, J. and Lapata, M., 2016. Neural summarization by extracting sentences and words. arXiv preprint arXiv:1603.07252.

[23] Nallapati, R., Zhou, B. and Ma, M., 2016(b). Classify or select: Neural architectures for extractive document summarization. arXiv preprint arXiv:1611.04244.

[24] Bui, D.D.A., Del Fiol, G., Hurdle, J.F. and Jonnalagadda, S., 2016. Extractive text summarization system to aid data extraction from full text in systematic review development. Journal of biomedical informatics, 64, pp.265-272.

[25] Shi, T., Keneshloo, Y., Ramakrishnan, N. and Reddy, C.K., 2021. Neural abstractive text summarization with sequence-to-sequence models. ACM Transactions on Data Science, 2(1), pp.1-37.

[26] Yu, T., Liu, Z. and Fung, P., 2021. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. arXiv preprint arXiv:2103.11332.

[27] Gunel, B., Zhu, C., Zeng, M. and Huang, X., 2020. Mind the facts: Knowledge-boosted coherent abstractive text summarization. arXiv preprint arXiv:2006.15435.

[28] Zhang, J., Zhao, Y., Saleh, M. and Liu, P., 2020, November. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In International Conference on Machine Learning (pp. 11328-11339). PMLR.

[29] El-Kassas, W.S., Salama, C.R., Rafea, A.A. and Mohamed, H.K., 2021. Automatic text summarization: A comprehensive survey. Expert Systems with Applications, 165, p.113679.

[30] Rothe, S., Narayan, S. and Severyn, A., 2020. Leveraging pre-trained checkpoints for sequence generation tasks. Transactions of the Association for Computational Linguistics, 8, pp.264-280.

[31] Beltagy, I., Peters, M.E. and Cohan, A., 2020. Longformer: The long-document transformer. arXiv preprint arXiv:2004.05150.

[32] Otter, D.W., Medina, J.R. and Kalita, J.K., 2020. A survey of the usages of deep learning for natural language processing. IEEE transactions on neural networks and learning systems, 32(2), pp.604-624.

[33] Fabbri, A.R., Kryściński, W., McCann, B., Xiong, C., Socher, R. and Radev, D., 2021. Summeval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics, 9, pp.391-409.

[34] Liang, Z., Du, J. and Li, C., 2020. Abstractive social media text summarization using selective reinforced Seq2Seq attention model. Neurocomputing, 410, pp.432-440.