STM JOURNALS

Journal of Artificial Intelligence Research & Advances
ISSN: 2395-6720
Volume 7, Issue 3
www.stmjournals.com

Automatic Text Summarization: A Review

Behjat Reyaz[1,]*, Falak Jan[1], Riyaz-ul-Haque[2]
[1]M-Tech Information Technology, Central University of Kashmir, Srinagar, Jammu and Kashmir, India
[2]Department of Information Technology, Central University of Kashmir, Srinagar, Jammu and Kashmir, India

Abstract

We present Encoder/Decoder a unique methodology for summarization of text documents. The summarization of text is to be done in two forms: one is abstractive and the other one is extractive. Summarization of text is the main job of natural language processing. With the help of new idea we make a unique amalgam model of abstractive-extractive to combine BERT (Bidirectional Encoder Representations from Transformers). Firstly we transform the summaries which are abstractive to their respective embeddings using BERT. Secondly, pre-train two models separately by implementing BERT pre-trained model. Then the extraction network and abstraction network are combined to form a unified model by reinforcement learning. As we know that automatic text summarization fundamentally compresses a long document into a shorter format while securing its intelligence content and overall meaning. A number of automatic summarizers exist which is able of producing good content summaries, but they do not focus on securing the underlying meaning and semantics of the text. So we also catch and preserve the semantics of text as the fundamental feature for summarizing a document using distribution semantic model. Also clustering and ranking of sentences is done by using various algorithms. In this study, automatic text summarization is performed on single documents and is generic based. We evaluate our summarizer with the help of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) on DUC-2007 dataset and compare with other state-of-the-art summarizers.

Keywords*: Encoder, decoder, BERT, summarizers, ROUGE*

Author for Correspondence E-mail: reyazbehjat@gmail.com

INTRODUCTION

As we all know that we have a textual data which is present in plentiful form, the substantial provocation is evolve those devices which constitute applicable information having precise shape. As it is rather hectic to perform manual summarization method, automatic summarization becomes more important. Text summarization is a method to get the data in squeeze form; it refers to the shortening of text; its intention is to create a logical and fluent summary. In fact the data is spreading tremendously at a high rate globally and is expected to increase from 4.4 to 180 ZB by 2025; therefore as huge data is spread in the digital space, there is need to develop Machine Learning algorithms that can automatically condense the text and deliver summaries so that it can easily pass the meaningful messages. Summarization of text is to be done with abstractive and extractive ways.

In extractive method, certain key-phrases are chosen from original document; then these key-phrases are merged in order to generate summary. Sometimes summary can be grammatically strange and simple. In abstractive method, the concatenation is done in such a way that it creates new phrases by taking original document into consideration and then merge those new phrases, it can overcome grammar inconsistencies and is more concise. While summarizing extractive approach is better but readability is poor; on the other hand in abstractive approach, there is a fixed length, also some neural networks are combined to get the advantages of both extractive and abstractive approach. One more prime thing is that there is a focus on summarization of text which is questionnaire based or common, but here on task of common based, a on single document. Various long-established approaches are used based on

statistical, linguistic, graph-based, or optimization based. Automatic summarization summarizes using several approaches: of supervised learning which requires an enormous data but it is too high-priced; on the other hand, in another approach which is unsupervised learning it is not too high-priced and it focuses on various features but here is the disadvantage that automatic summarization ignores the semantics of summary which means meaning of the input document. Deep auto encoders and variational auto encoder are applied on generic based text summarization and this is the first time it is to be done, otherwise these encoders are implemented on query-based single document text summary.

A novel strategy has been used in summary generation, where it utilizes a pre-trained language model BERT (Bi-Directional Encoder Representation), also for representation of text, BERT embeddings are used even though we have alternatives for representation like Word2vec which can create Word embeddings but BERT is trending and is the latest approach. For the whole system, DUC 2007 dataset and for evaluation the metrics, ROUGE is used. It works as the contrast between the original summary and the summary which is generated at the end.

LITERATURE REVIEW

Depending on task it is based on presenting novel outlook for summarization of text which employs CNNS and BERT. Due to the new motivation, we shortly discuss some different approaches for summarization of text as below:

Abuobieda *et al.* [1]

This paper describes that in the article initial sentence is main also even powerful for making of summary [1]. The proposal used was Feature Selection model; the features that are explored are by using a Pseudo Genetic Concept. Parameters are considered and encrypted with the help of genes, which are binary structured; these aspects can be controlled by implementing probabilistic approach.

Joshi *et al.* [2]

SummCoder has been introduced majoring in the extractive form of summarization of documented text. Skip thought models are highly recommended and give rise to embeddings for sentences [2]. These have certain rules about sentences having three major groups to fall out in a documented text form; these are:

1. Relevance,
2. Novelty, and
3. Position.

Whereas relevance is depending on an auto-encoder network, the novelty parameter is depending on the data that has a linkage with the data we provide, lastly but not the least, sentence positioning parameter is the most important of all because it forms the ground bases of the document and the sentences being used in the document to make it look extravagant. The determination of the score for a handcrafted document is based on these above given three sentence parameters.

Alami *et al.* [3]

They introduced another way for summarization but it was different to what we proposed due to two major differences, one, that they had encoders for representation of sentences in the latent forms but not the relationship in the middle of the sentences used in that document [3]. Another one being their application not on the general datasets of summarization but only on the Arabic text.

Kaikhah *et al.* [4]

They submitted the idea of understanding the salient features describing the sentences in the summary by using the neural networks that include machines by which we can summarize the document [4]. Not only that but also make it highly organized and effective and most rated in a scene of modifications of sentences. This technique might as well describe each of the silent characteristics of the sentences used in the particular article or a document that will do justice with the meaning and central idea of the summary. They also proposed that it is human friendly and can easily be used while doing your digital documentations and writing articles on different aspects.

Muzaffar Shah *et al.* [5]
Text analytics assist in taking understanding from the text [5]. Text processing and automatic text summarization are the lead tasks in this area. This idea will help making up the summary by taking the lead sentences in the document or article so that one can get the proper idea without reading the whole document. Moreover text summaries are used for compressing the length of the input documents without negotiating with its overall meaning and information content. For distribution similarity, they used Word2vec which is a Distribution Semantic model. Hence, text summarization is the data trimming process for quick utilization by the user.

Nallapati *et al.* [6]
By using an artificial neural-network approach for summarization which is extractive in nature so that the worthiness of the document can be defined properly [6]. In this approach, using generator pointer model can be eligible for creation of words in the document. This approach also helps marking the issues in the summary and making it effective by getting rid of unwanted words and unnecessary sentences that are just causing the summary or a document to be prolonged.

Wang *et al.* [7]
They described fundamental responsibility for extraction of info and summarization of text document which is widely applied in many fields such as natural language processing; their approach is worked on one of the types of machine learning, that is, reinforcement [7].

Sinha *et al.* [8]
They introduced an approach that is data-driven using feed-forward neural networks for summarization of text and focused with regard to summarization of text [8]. Submitted design is climbable and capable of generating the summary which is absolute sorted document. This paper also focuses on objects directly from the entire collection without modifying the objects themselves. To acquire the summary, which is best at the final, is selected as per the summary length. It also reveals that a straightforward and a relatively simpler approach generates outcome which is equivalent toward complex deep networks.

Yin and Pei [9]
For the generation of embeddings, they used CNN as a platform so that the important and valuable information can be extracted from the document [9]. For good representation of sentences we can use skip-thoughts methods for the evaluation of useful information from the documents on huge dataset. Book Corpus (2015) can also be used for the same agenda. Yin and Pei did not employ the importance of positioning of sentences whereas we use it as auto-encoder to know the relationship between the sentences of the documents and extract the valuable information.

Yousefi-Azar and Hamey [10]
Their approach includes deep auto-encoders (AE**)** for summarization of text, which is questionnaire based [10]. Noisy auto-encoders that enable us to use effective and worth fully information be extracted from the documents and make it an extravagant summarized form. Their experiments inspect general as well as local vocabularies. ENAE (Ensemble Noisy Auto-Encoder) is a random type of an auto encoder; most important thing is that it chooses those sentences which are highly knowledgeable. Encoders of deep learning understand characteristics instead of hand operated on it. The technique is quite simple and does not need any examination. Techniques generated the summaries which were acquired through grading the sentence. The summaries which were formed at the end were very simpler; their preparation and evaluation price was slighter; the ensemble can select a summary which is quite similar with the reference summary. On the basis of sentence ranking, they extract human friendly summaries that are globally acknowledged.

METHODOLOGY

This segment describes the methodology for summarization of text. Figure 1 shows architecture of our model. Precisely, our proposed approach consists of various modules which are described as below:

Preprocessing

Firstly, it recognizes the information of the text on which it does text normalization and removes inconsistencies, in the input like URLs, it is to be stripped off by the preprocessing module, the case of text in input document is modified to lowercase; also stops word removal, means those words which are meaningless for text summarization task, which

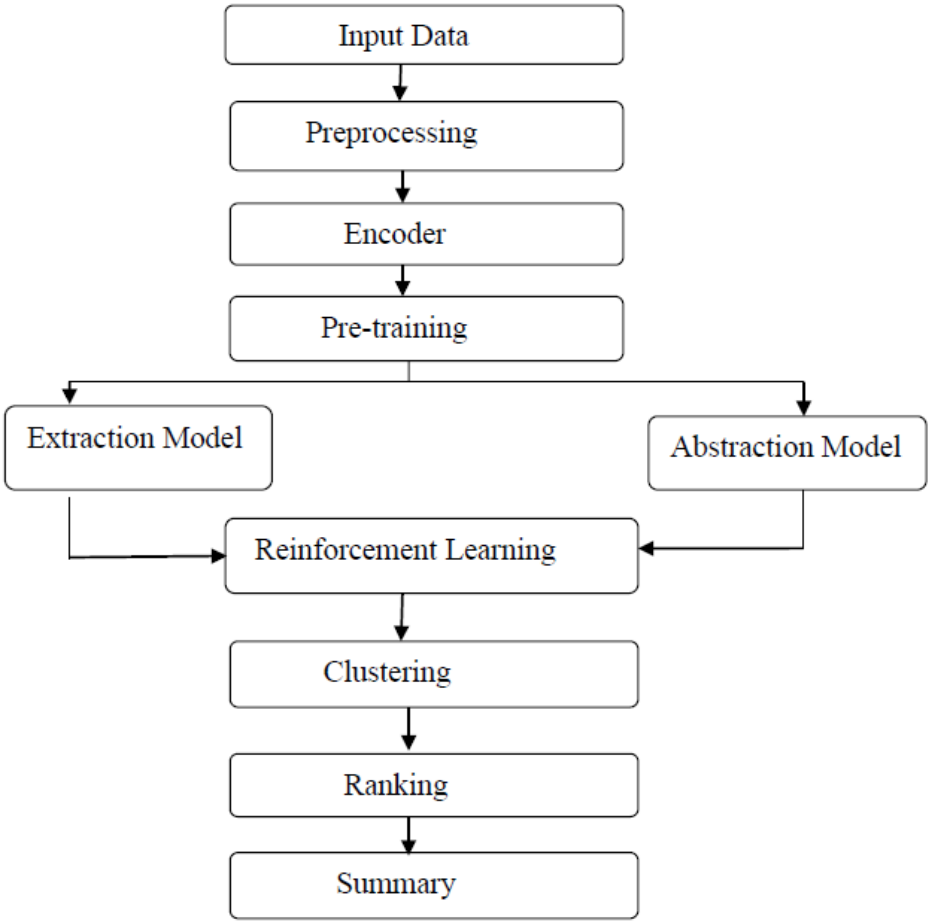

Fig. 1: *Text Summarization Proposed System.*

is done by Stanford core NLP language; after that, tokenization is to be put on means; document is tokenized into words; next, words are compressed to their stems, that is, lemmatization.

Encoder

Next module is to instruct a common encoder that represents text in to their respective vectors. Here the approach is dependent on the neural-network which is recurrent based. The neural network propels the approach to choose and extract some main attributes and then generate link between origin and a goal. So comes embedding which depends on supposition of representation of word. Embedding shows words of natural language as in vector representations which is low-dimensional. Nowadays, embeddings is used in various NLP jobs which include BERT, Word2Vec, Glove etc. BERT is mostly used in those functions of NLP which are in fine-tuning style. Here in our scenario we employ BERT as a feature-based style. It can indicate words which are tokenized to their equivalent embeddings. Embeddings of BERT are now as to load in first abstraction model and then in extraction model.

Pre-Training

Now we implement pre-trained model of BERT:

Extraction Model

Using BERT extraction design comprises of several parts, for encoding the document encoder, same as for sentence encoding again an encoder and for selecting the sentences extractor. The encoder which is used for sentence acquire BERT as for transforming, a direction of two-way gated recurrent unit which is used as for encryption of a text while as a direction which is one-way, used to determine the description of summary. Hidden states of GRU and document vector are mixed up in the calculation of the result of sentence. For encapsulating the knowledge, we can do it with the different directions comprises in onward direction $GRUh_t$ and also in reverse direction $GRUh_t$ GRU is uncomplicated, capable, possesses less criteria and is easy to execute. We use a GRU which is directional in two-way and is used to encrypt the documents and on other side GRU which is directional in one side helps to get the description of summary. Gated recurrent unit is a network, which is recurrent, consists of two gates: one for updating and the other is for resetting. The description of the entire document is designed as a sum of the sequence hidden stated H_D, and vector of weights, a_D, as return. The function for normalization of attention weights, which adds up to one is softmax(). To visualize the sentences as it belongs to the end summary, the extractor checked subsequently for this logistic layer form a decision. Also our extraction design works as a thinker in dual form, which distinguishes either the sentences in the original order are essential or no more.

Abstraction Model

This model reworks on earlier essential sentences and gives rise to the point and summary which is understandable. In the network model, we use pointer-generator which eases the duplicating of words from the original document through highlighting which upgrades correctness of out-of-vocabulary words, compared with Vanilla network, the novel network introduces words which are updated; secondly, we replace the LSTM with Gated Recurrent Unit. At last, the embedding of abstraction and extraction models are dissimilar. In our abstractive system, BERT is used; word2vec is used for vanilla network. The neural network is observed as a stabilizer in the middle of extraction as well as abstraction process.

Reinforcement Learning

It is for combining both abstraction and extraction models so as to make a fused model. We consider the data which is available for preparation as the simulation; and the extractive and abstractive models as decision making of reinforcement learning. The simulation perceives the condition and then performs the abstraction and extraction. The summary which is generated from abstraction is more alike to the recommended summary so as we get the highest result. The summary which is generated at the end; if it is not alike to the recommended summary a low result is what we achieve. Then we capture semantics of text by employing Distributional Semantic Models, it captures the coherence between two textual elements. This model states that words occurring in the same context have same meaning. To catch the semantics, we use Word2Vec. Various scientists use this Word2Vec for boot strapping; this is also used to solve the sense of words sense (Text Coherence problem).

Clustering

We have enormous sentences; for that purpose, we use clustering algorithm, which assembles those sentences which possess same meanings into familiar bunches as K-means Clustering algorithm to form the big clusters; after clustering is performed, various bunches are found on the basis of similarity of meaning in sentences.

Ranking

As the last component, we use the ranking algorithm which uses different characteristics like position of sentence, length of sentence, frequency (TF-IDF), phrases in sentence which can be noun or verb, cue-phrases etc. Extract top sentences from each bunch; first obtain the sentence with ranking scores and the final score is the sum of these scores. The total score is generated as aggregating outcome of each sentence; the sentences with

the good rank are selected to be candidate for the extractive summary, finest outcome of sentences from each bunch is selected and then normalize the scores for effective extraction. We use sentence position as a metric to choose a sentence among the two semantically similar sentences. Whichever sentence has highest position scores will be selected in summary.

CONCLUSION

A novel scheme is introduced for summarization of text. This study presents a unique summarization technique for producing better summaries using text summarization process. According to comparative analysis, our proposed approach performs better than the baselines. The main conclusions are:

1. All the sentences are represented in their embeddings using BERT.
2. We proposed a new strategy to merge extractive and abstractive methods implementing BERT pre-trained model on it.
3. Capturing the semantics, which is ignored by most of times in text-summarization process here considered as main characteristic for summarization of text also enhance precision in our summaries' usage.
4. Clustering is to be done so as for removing redundancy and semantically similar sentences.
5. Ranking algorithms rank sentences in each cluster to produce summaries by picking top sentences from each cluster.

REFERENCES

1. Abuobieda A, Salim N, Albaham AT, Osman AH, Kumar YJ. Text Summarization Features Selection Method Using Pseudo Genetic-Based Model. In *2012 IEEE International Conference on Information Retrieval & Knowledge Management*. Mar 13, 2012; 193–197p.
2. Joshi A, Fidalgo E, Alegre E, Fernández-Robles L. SummCoder: An Unsupervised Framework for Extractive Text Summarization Based on Deep Auto-Encoders. *Expert Syst Appl*. Sep 1, 2019; 129: 200–15p.
3. Alami N, En-nahnahi N, Ouatik SA, Meknassi M. Using Unsupervised Deep Learning for Automatic Summarization of Arabic Documents. *Arab J Sci Eng*. Dec 1, 2018; 43(12): 7803–15p.
4. Kaikhah K. Text Summarization Using Neural Networks. *WSEAS Trans Syst.* 2004; 3(2): 960–963p.
5. Mohd M, Jan R, Shah M. Text Document Summarization Using Word Embedding. *Expert Syst Appl*. Apr 1, 2020; 143: 112958p.
6. Nallapati R, Zhou B, Gulcehre C, Xiang B. Abstractive Text Summarization Using Sequence-to-Sequence rnns and Beyond. *arXiv preprint arXiv:1602.06023*. Feb 19, 2016.
7. Wang Q, Liu P, Zhu Z, Yin H, Zhang Q, Zhang L. A Text Abstraction Summary Model Based on BERT Word Embedding and Reinforcement Learning. *Appl Sci*. Jan 2019; 9(21): 4701p.
8. Sinha A, Yadav A, Gahlot A. Extractive Text Summarization Using Neural Networks. *arXiv preprint arXiv:1802.10137*. Feb 27, 2018.
9. Yin W, Pei Y. Optimizing Sentence Modeling and Selection for Document Summarization. In *24th International Joint Conference on Artificial Intelligence*. Jun 23, 2015.
10. Yousefi-Azar M, Hamey L. Text Summarization using Unsupervised Deep Learning. *Expert Syst Appl*. Feb 1, 2017; 68: 93–105p.

Cite this Article
Behjat Reyaz, Falak Jan, Riyaz-ul-Haque. Automatic Text Summarization: A Review. *Journal of Artificial Intelligence Research & Advances.* 2020; 7(3): 22–27p.