УДК 004.912
МРНТИ 50.05.19

OVERVIEW OF THE DIFFERENT TEXT SUMMARIZATION METHODS

DAUIT D., KEMALOV M., JAXYLYKOVA A.

Kazakh-British Technical University

Abstract: *Text summarization is one of the major problems because it has a high range of usage in various fields, it is most important to have an improved mechanism for the fastest and most effective extraction of the information. The extraction of the summary from all that available source of text data by hand is very difficult. In order to show the ways for solving the text summarization, this paper presents a brief survey of various text summarization methods like MatchSum (Zhong et al., 2020), BertSumExt (Liu and Lapata 2019) and SemSim (Yoon et al., 2020) which has shown the leading results in extractive and abstractive text summarization. This paper reviews those models and shows their advantages and disadvantages, makes a guess how text summarization can be improved.*

Key words: *text summarization methods, natural language processing (NLP), BertSumExt, MatchSum, SemSim)*

ОБЗОР РАЗЛИЧНЫХ МЕТОДОВ ОБОБЩЕНИЯ ТЕКСТА

Аннотация: *Суммаризация текста является одной из основных проблем, поскольку имеет широкий диапазон использования в различных областях, и наиболее важно иметь улучшенный механизм для быстрого и эффективного извлечения информации. Извлечение резюме из всего этого доступного источника текстовых данных вручную очень сложно. Для того, чтобы показать способы решения проблемы суммирования текста, в данной статье представлен краткий обзор различных методов суммирования текста, таких как MatchSum (Zhong et al., 2020), BertSumExt (Liu и Lapata 2019) и SemSim (Yoon et al., 2020).), которые показали наилучшие результаты в обобщении текста. В данной статье рассматриваются эти модели, показаны их преимущества и недостатки, и даются предположения, как можно улучшить суммаризацию текста.*

Ключевые слова: *методы суммирования текста, обработка на естественном языке (NLP), BertSumExt, MatchSum, SemSim*

МӘТІНДІ ҚОРЫТЫНДЫЛАУ ӘДІСТЕРІНЕ ШОЛУ

Аңдатпа: *Мәтінді қорытындылау негізгі мәселелердің бірі болып табылады, өйткені әртүрлі салаларда оны қолдану ауқымы жоғары, ақпаратты тез және тиімді алудың жетілдірілген механизмі болуы өте маңызды. Барлық қолжетімді мәтіндік дереккөздерден түйіндеме шығару өте қиын. Мәтіннің жалпылауын шешудің жолдарын көрсету үшін бұл жұмыста MatchSum (Zhong et al., 2020), BertSumExt (Liu and Lapata 2019) және SemSim (Yoon et al., 2020) сияқты сан түрлі мәтінді жинақтау әдістеріне қысқаша шолу ұсынылған) мәтінді экстрактивті және абстрактілі қорытындылау кезінде жетекші нәтижелер көрсетті. Бұл қағаз сол модельдерді қарастырады және олардың артықшылықтары мен кемшіліктерін анықтайды. Сонымен қатар мәтінді қорытындылаудың жолдарын жақсартуды болжайды.*

Түйінді сөздер: *мәтінді қорытындылау әдістері, табиғи тілде өңдеу (NLP), BertSumExt, MatchSum, SemSim*

Introduction

In the last fifty years, considerable work has been carried out in the area of text summarization. Novel methods that integrate linguistic elements into the summary have been established, and now the summary is not just the basic concatenation of sentences. This area of study is constantly growing, addressing new consumer demands and posing a range of challenges. Hence, in this section, emphasis is placed on the important issues that occur in this research area that the research community needs to tackle. Existing text summary methods are being updated with time as new machine learning algorithms are being employed to construct text summary systems. But the features (term frequency, position, etc.) needed for extracting essential sentences are not much modified. Therefore, some new features need to be found for terms and sentences which can remove essential semantic sentences from the text. The form of summaries is changing to match changing consumer requirements. Initially standardized single document summaries were produced but now they have gained prominence due to the availability of vast volumes of data in various formats and languages and due to the increasing growth of technology, multi-document, multi-lingual, multimedia summaries. This is also apparent from evaluation systems that are now focusing on different forms of overview channels. Summaries with defined emphasis are also being created, such as sentiment-based, customized summaries etc. But, another important issue is how such information can be presented. Currently most systems handle textual input and output. New approaches can be proposed in which input, other than text, can be in the form of meetings, videos, etc. and output in a format. Some other frameworks can be created in which input is in the form of text and output can be expressed by means of charts, tables, graphics, visual rating scales, etc. that allow visualization of the results, and users can access the necessary content in less time.

Numerous new approaches have been suggested dealing with linguistic characteristics and enhancing the consistency of summaries. But linguistic approach-based summary systems require more processor and memory space, as they need more linguistic knowledge and difficult linguistic techniques. Moreover, there is an additional complexity in employing linguistic resources (Context Vector Space, Lexical Chain, WordNet, etc) and linguistic analysis tools (discourse parser) of good quality as there is a scarcity of different language resources. Therefore, it is important to build statistical-based, effective synthesis systems that can summarize texts in all languages and produce a summary whose output matches that of a human summary. In addition to concatenating the sentences, the summary material has to be accurate. Therefore further needs to be done on an abstractive or mixed approach. Essential information can be picked, combined, compressed with hybrid approaches or any information can be omitted in order to provide new description information. To generate a high quality summary a hybrid approach can be created by integrating extractive and abstractive techniques together. Research is also generating abstracts so that the summaries produced by the machine fit closely with the human-written ones. The appraisal process is also another huge obstacle. This paper discussed both intrinsic as well as extrinsic types of assessment methods. Part of the assessment is fundamental in nature, and is further divided into informativity which consistency assessment, which is performed using modern techniques and instruments. Most of the latest instruments analyze the details contained in the summary, and very few approaches attempt to determine the consistency of the summary. New methods are being established to simplify the method of quality control which is a largely manual task carried out by professional judges. In general, accessible intrinsic evaluation approaches

rely on the can language between a machine-generated summary and reference summary. Analysis should be performed in an inherent assessment, thereby devising new approaches to assess the description based on the knowledge found therein and its delivery. The method of assessment is inherently subjective. Firstly, a reasonable criterion must be established so that what is relevant and what is not is transparent to the method. It is still unclear if this method can be streamlined enough. Similarly, summary quality assessment is also highly subjective, since expert judges perform it manually. For consistency assurance there are also certain criteria as grammaticality, coherence, etc. But when two experts evaluate the same summary different results are obtained. Text summarization is more than fifty years old and the science community is still active in this area, so that they can try to enhance current text summarization methods or create new summarization strategies to produce better quality summaries. But the output of summarizing text is still moderate, and summaries produced are not so good. This program can then be made smarter by integrating it with other systems so the new system can work better.

The purpose of text summarization is to simplify the original text to a version that would have the key substance and general meaning. Text summarization approaches may be categorized as extractive and abstractive summarization. Here we display the best findings tested for ROUGE-1, ROUGE-2, and ROUGE-L using full-length F1-scores using CNN / Daily Mail and Gigaword datasets.

Extractive text summarization

An extractive summarization is the process of selecting the main part of the document and concatenating it into a shorter version.

MatchSum (Zhong et al., 2020) model has shown 44.41 in ROUGE-1, 20.86 in ROUGE-2, 40.55 in ROUGE-L. MatchSum conceptualizes extractive summary as a problem that matches the semantic text. The paradigm is based on the premise that a strong description would be more semantically analogous to the source text as a whole. Semantic similarity matching is a key research concern in recognizing the resemblance that can be found in many ways between a source and a target text fragment. One of the most approaches for each text fragment is to learn a vector representation and then apply typical similarity metrics to calculate the matching scores. The model suggests a Siamese-BERT framework for measuring the similarities between the source text and the list of candidates. Siamese BERT leverages the pre-trained BERT in a Siamese structure to determine semantically important text embeddings that can be analyzed with cosine-similarity. Siamese-BERT consists of two BERTs with tied-weights and a cosine-similarity layer during the inference phase.

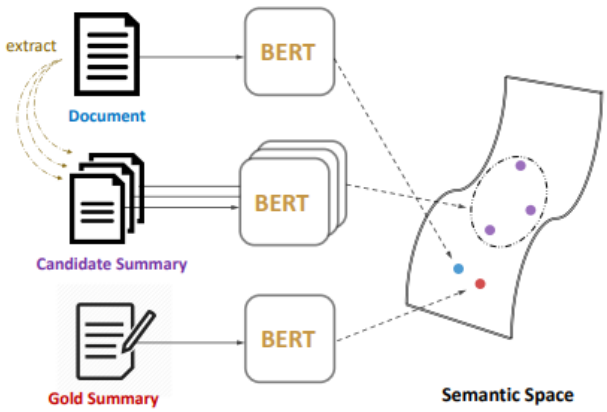

Figure 1. MATCHSUM model. Contextual representations of the document matched with aimed summary and possible summaries. Better possible summaries should be semantically closer to the document

MatchSum formulates extractive summarization as a semantic text matching problem and proposes a novel summary-level instead of scoring and extracting sentences, by this model overwhelms the problematic part of summary-level optimization by contrastive learning. This approach conducts an analysis to investigate whether extractive models must do summary-level extraction based on the property of the dataset. This model has shown the best performance on CNN/Daily Mail (44.41 in ROUGE-1) by only using the base version of BERT and seeks to observe where the performance gain comes from.

BertSumExt (Liu and Lapata 2019) model has demonstrated 43.85 in ROUGE-1, 20.34 in ROUGE-2, 39.90 in ROUGE-L and implements a novel text-level encoder based on BERT that can represent the meaning of a text and get representations for its sentences. The new version of pre-trained language models is Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2019). In a single very large converter, BERT blends word and phrase representations. This extractive model is built on top of this encoder by piling different transformer layers for phrasing text level functionality.

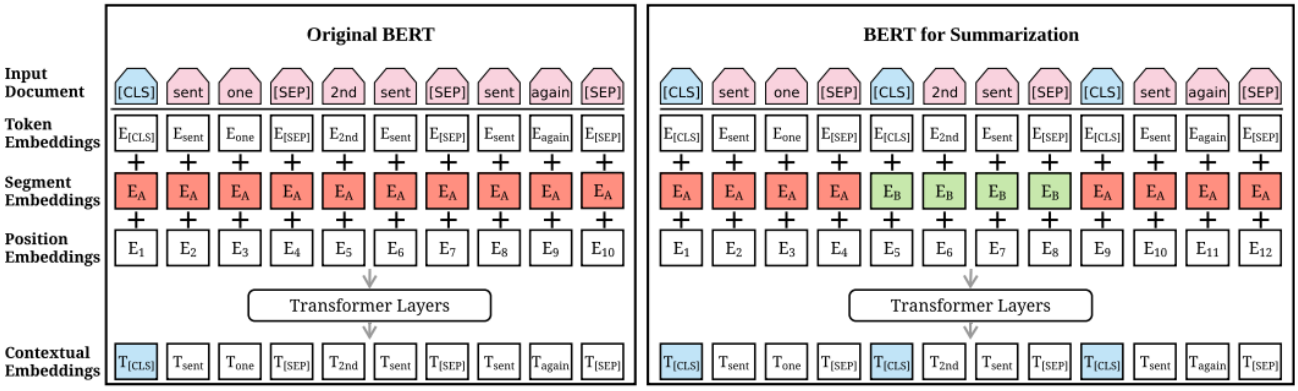

Figure 2. Initial architecture of BERT (left) and of BERTSUM (right). The series at the top is the text entry, followed by a summation of three forms of embedding for each token. The unified vectors are used to embed several bidirectional Transformer layers as inputs, creating contextual vectors for each token. BERTSUM extends BERT by inserting multiple [CLS] symbols to learn sentence representations and by using interval segmentation embedding to distinguish multiple sentences (illustrated in red and green colors)

BERT uses a bidirectional language model to retrieve masked tokens/spans for a given sentence, brings significant improvements to NLU tasks but are not suitable for generation tasks, proposes a masked language modeling (MLM) objective where some of the input sequence tokens are randomly masked and the goal is to predict those masked positions taking the corrupted sequence as input. BERT designed MLM to take advantage of bi-directional information during pre-training. It remains unclear whether there are pre-training objectives that are simultaneously more efficient and effective. The BertSumExt model produces a description by defining the main phrases in a text. Neural models consider extractive summary as a question of the classification of sentences: a neural encoder generates representations of sentences and a classifier determines which sentences will be chosen as summaries. Experimental findings through three datasets demonstrate that, under automated and human-based assessment procedures, our model produces cutting-edge performance around the board.

Abstractive text summarization

Abstractive summarizing is the method of recognizing and then explaining the key ideas in a text in a simple natural language. The supervised learning model and reinforcement learning (RL) algorithm were commonly used for abstractive summarization. Supervised learning counts to replace tokens with a reference synonym as incorrect, but RL-based models have shown remarkable performance but optimization is slow and requires significant computational effort to converge.

SemSim (Yoon et al., 2020) model has shown 44.72 in ROUGE-1, 21.46 in ROUGE-2, and 41.53 in ROUGE-L. The semantic similarity strategy uses the semantic distance between as a loss in the text summarization task. Maximizing the semantic similarity between the summary produced and the summary of reference is important in order to obtain a good model that is acceptable and adaptable. To calculate the semantic similarity between generated summaries and reference summaries – the computation

of semantic similarity score is needed. This model takes more training time compared to the approach of maximum probability since it learns by sequence level.

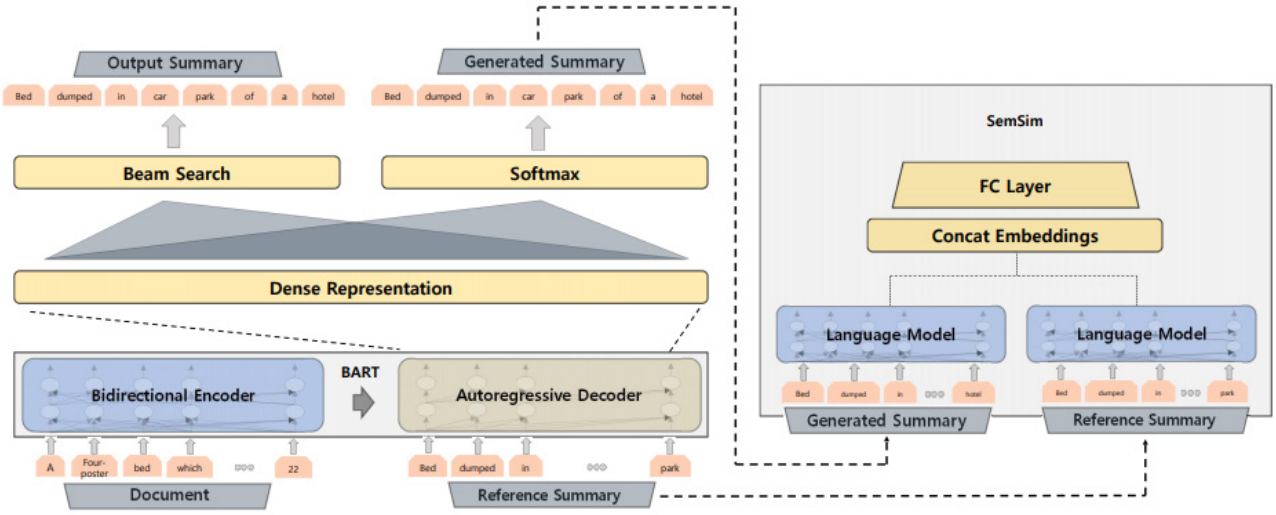

Figure 3. SemSim Overall Architecture, the BART structure was used to represent the generated summary. In the SemSim layer, Language Model, which is encoding the generated summary and the reference summary, is not updating the weights. SemSim layer calculates gradient

The model produces a series of word tokens of the generated description by the use of the BART algorithm. BART is an autoencoder that uses technology from sequence to sequence transformers. BART consists of two parts: one is a bidirectional encoder and the other part is a decoder of auto-regression.

Algorithm of semantic similarity strategy:

1) Define set of word tokens as a sequence of the original document – $S_{doc} = \{s_1^d, s_2^d, \dots s_n^d\}$

2) Define set of word tokens as a sequence of the reference – $S_{ref} = \{s_1^r, s_2^r, \dots s_n^r\}$

3) Generate a set of word tokens of generated summary – $S_{gen} = \{s_1^g, s_2^g, \dots s_n^g\}$ by auto-regressive process of BART model.

- The encoder part of BART encodes set of word tokens of original document (S_{doc})
- The decoder part computes probability distribution of token s_t^g at a step t – $P(s_t^g \vee s_1^g, s_2^g, \dots s_{t-1}^g, S_{doc})$ by previous word tokens and a sequence of original document.

4) Maximum-likelihood loss can be defined as a sum of logarithm of probabilities :

$$L_{ml} = -\sum_{t=1}^{m} logP(s_t^g \vee s_1^g, s_2^g, \dots s_{t-1}^g, S_{doc})$$

5) Calculation of the semantic similarity score:

- Generated summary S_{gen} and reference summary S_{ref} are encoded by pre-trained language model (LM). Model as a BERT, encodes each word as a dense vector and then computes the embeddings of whole sequence. Embedding of reference can be computed by the next equation $e_{ref} = LM(S_{ref})$, the same for generated summary $e_{gen} = LM(S_{gen})$;
- Semantic similarity score can be defined by the next simple linear equation :

$$Score_{semsim} = We + b$$

where – $eisaconcatenationofe_{gen} \wedge e_{ref}$; $e_{gen} \epsilon R^d$, $e_{ref} \epsilon R^d$; where d is a number of hidden layers of language model (LM); $W \epsilon R^{1*2d}$ are trainable parameters.

6) Semantic similarity loss defines as:

$$L_{semsim} = -Score_{semsim}$$

7) Training objective is to minimize Loss function which defines as: $Loss = L_{semsim} + L_{ml}$

Conclusion

In this paper, we compared several state of the art methods of text summarization in natural language processing task, such as MatchSum (Zhong et al., 2020), BertSumExt (Liu and

Lapata 2019) and SemSim (Yoon et al., 2020). We use CNN / Daily Mail and Gigaword datasets to compare these methods. The best results showed an abstractive summarization method SemSim (Yoon et al., 2020), the model has shown 44.72 in ROUGE-1, 21.46 in ROUGE-2, and a state-of-the-art result in ROUGE-L -41.53. SemSim model is more flexible than traditional maximum likelihood methods. This helps SemSim to create a summary using its vocabulary and structure of sentences and to give better effects.

Acknowledgment

We thank the Institute of Information and Computing Technologies for financial support, and we are also grateful to Alexander Alexandrovich Pak for useful advice and interest in the work. (Grant AP05132760, Kazakhstan).

REFERENCES

1. Lewis M. et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension //arXiv preprint arXiv: 1910.13461. – 2019.
2. Hermann K. M. et al. Teaching Machines to Read and Comprehend. arXiv. – 2015.
3. Zhong M. et al. Extractive Summarization as Text Matching //arXiv preprint arXiv:2004.08795. – 2020.
4. Liu Y., Lapata M. Text summarization with pretrained encoders //arXiv preprint arXiv:1908.08345. – 2019.
5. Yoon, Wonjin, et al. Learning by Semantic Similarity Makes Abstractive Summarization Better. 2020, http://arxiv.org/abs/2002.07767.
6. NLP progress, summarization. URL: http://nlpprogress.com/english/summarization.html