A Transformer-based Pretrained Language Model for Automatic Document Summarization

Anonymous ACL submission

Abstract

Document summarization is an essential task in the information retrieval. User intention for specific documents is hard to satisfy for the information overload, thus there are rarely golden standards for matching a certain document. I propose a multi-grained summarization model for documents that uses attention on paragraphs, supervised learning on the sentence importance, and sequence-to-sequence learning on characters and words. Tested on the 2020 Chinese Law Summarization Dataset, the proposed model achieved better performance than other modern summarization models.

1 Introduction

Document matching is an essential task in daily work and study. Because the matching algorithm is not capable of choosing important words from the massive documents, to use the content of a document as the matching criteria often fails in returning unintended documents. Document summarization condenses long documents into short paragraphs by retaining core information. Thus, a much wiser way for the document matching is using the summarization instead of the content.

2 Related Work

Bhargava and Sharma (2020) use Convolutional neural network and Long-Short Term Memory network in the paraphrase detection of documents and summaries, then using multi-grained Convolutional neural networks as the sentence extraction model, as a result, the F1 scores on the English and Malayalam summarization corpus are 31.57% and 53.57%.

Erera et al. (2019) consider the summarization problem as a query problem, use query saliency, entities coverage, diversity, text coverage and sentence length as the objectives.

Altmami and Menai (2020) concluded the special pattern for the summarization of scientific articles, and found that abstract generation-based and citation-based are two notice-worth solutions.

Polsley et al. (2016) used a combined word frequency and domain knowledge summary method for the legal text abstraction, and achieved an average of 10.9% ROUGE score between automatically-generated summaries, and the expert-generated summaries.

Iqbal and Qureshi (2020) showed that the deep learning approaches have achieved many satisfactory results in several domains, Word embeddings, such as Word2Vec, Glove, and FastText are widely used in the Natural Language Processing. Variational Auto-Encoders and Generative Adversarial Networks have been used for the text generations.

Litvak et al. (2019) studied the evaluation system for text summarization, and showed that ROUGE, MeMoG for quality evaluation, Latent Dirichlet allocation for topic evaluation, and grammar analysis for the readability evaluation.

Deng et al. (2020) used a Sequence-to-Sequence generation model and an adversarial model in the text summarization for Chinese texts.

Zhang et al. (2019) used the Transformer network in encoder-decoder framework for the text summarization task, and achieved a Rouge score of 33.48% in the CNN/DailyMail dataset for news summarization.

Xiao and Carenini (2019) used the extractive LSTM-Minus model in the text summarization task, and by combining global document level, sentence level and topic segment representation in the network, achieved a Rouge score of 31.99% in the Pubmed summarization dataset.

Liakata et al. (2013) used the CRF and SVM models for the sentence extractive summarization for the scientific articles. Cohan and Goharian (2015) use the important sections in scientific pa-

pers as criteria for selecting important sentence.

Sutskever et al. (2013) started to use sequence-to-sequence neural networks in abstractive summarization. Vinyals et al. (2015) introduced copy mechanism from source to summary in sequence-to-sequence models.

3 Hierachical Multi-grained Summarization

Documents have their own organizations by different purposes, take the civil legal documents as an example, There are 123 different kinds of arguing reasons in the Chinese Court Sentence Documents. Largely list as follows:

- Disputes over personality rights,
- Marriage, family, inheritance disputes,
- Property rights disputes,
- Contract, management without cause, disputes over unjust enrichment,
- Labor disputes, personnel disputes,
- Maritime disputes,
- Civil disputes related to companies, securities, insurance, bills, etc.
- Inciting disputes

However, each of these documents follow the same organization of segments, for example, the purpose, the facts, the court opinions and the trail results. Thus, for a meaningful text summarization includes all there segments of importance and ignore other segments.

3.1 Segment Extraction

The document of the court sentence usually states the arguing point of the two entities. In this paper, a dictionary based arguing reason extractor was designed. A total of 123 arguing reasons were classified by the legal experts out of 70 million civil documents collected from the Chinese Court Sentence website.

Then I use several rule-based filters to extract meaningful segments. As Figure-1 shows, the document has been split into several paragraphs, the paragraph with important words are kept and other paragraphs are discarded.

3.2 Extractive Sentence Summarization

For the sentence importance scoring model, the experts wrote the summarization for each legal document. Base on the written summarization, each sentence importance can be easily computed by checking the overlap words. I use the pre-trained Chinese language model, which has multiple layers of transformers, for encoding sentences. As Figure-2 shows, the model layers are organized as follows: Take the last two layers of the RoBERTa encoding, then do the concatenation and take the Maximum element from the character sequence as representation for each sentence, then concatenating to a dropout layer, a softmax layer, and the output layer.

The learning rate is 10^{-4}, the max sequence length is 50, the dropout rate is 5%, and the batch size is 1. The epoch number is 3. The experiment was carried on a Google Cloud TPU v3, with 32GB of RAM, and 8 chips with 16GB of the high speed of memory each, which can provide 420 tera-flops of computation capability.

The classification dataset is unbalanced, thus I copy the positive samples multiple times to balance the dataset. In total, there are 782,895 samples for training and 42 samples for test in the Legal Sentence Importance dataset. The test accuracy is 92.06%.

3.3 Sequence To Sequence Summarization

For the token generation in the summarization step, I use a sequence-to-sequence model. By tagging each token existence in the corresponding summary, a sentence has its unique sequence of code for the token importance. For the 14,908 samples of important sentences for training and 108 samples for test in the Legal Token Summarization Dataset, I design a token summarization model as follows: First, encode the sentence by RoBERTa language model, and concatenate the first and last layers as the sentence representation. Then concatenate to a CapsuleNet layer, then concatenates the output layer.

The learning rate is 10^{-3}, and the bach size is 2, and the accumulation step is 8. The training epoch is 10, and the accuracy of the best model in the test set is 42.05%. From Table(1) we can see that different number of transformer layers has no significant performance improvement for accuracy in the test set. The RBT3, BRT3L and RoBERTa vary in number of neurons and number of layers, which was popular pre-trained language models

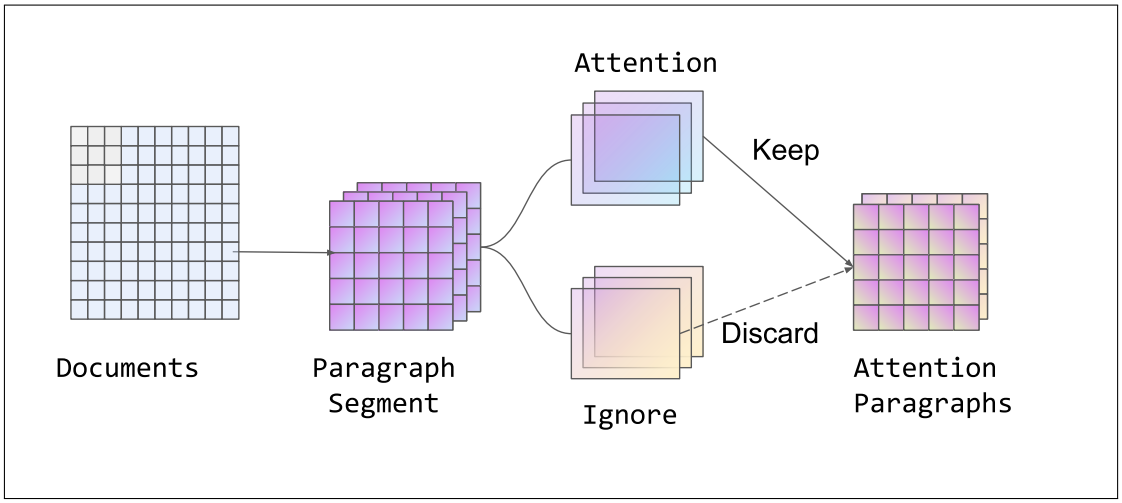

Figure 1: Rule based segment and importance based summarization.

Model	Neurons	Layers	Accuracy
RoBERTa	768	12	42.05%
RBT3	768	3	40.18%
RBT3L	1024	3	41.22%

Table 1: Different Pre-trained Language models accuracy in the sequence to sequece training steps.

R-1	R-2	R-L	R-W
32.62%	15.63%	31.79%	25.49%

Table 2: F-score for text summarization in the dataset.

trained on 4.5 billion of Chinese tokens by Cui et al. (2020).

For balancing the readability and quality of the summary, I did not use the sequence predicted by the token summarization model directly. Instead, a bigger threshold for the output of the model can improve the readability and the quality. As a practical result, I choose 87.5% as the threshold for keeping the important token. That is to say, keeping the most important 87.5% tokens for each inference shall get the best summarization result.

3.4 Summarization Evaluation

Lin (2004) represents a set of similar metrics such as ROUGE-N, ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. In this paper, I use ROUGE-1, ROUGE-2, ROUGE-L and ROUGE-W as the summarization evaluation metrics. The ROUGE-W score is computed using Formula (1).

$$R_W = R_1 * 0.2 + R_2 * 0.4 + R_L * 0.4 \quad (1)$$

4 Conclusion

In this paper, I introduced a novel summarization method, which leveraging hierarchical features of the document. By extracting multi-grained level from the document, the summary contains the paragraph-level, sentence-level and token-level key information. The evaluation results showed that the multiple layers of transformers have improved the Rouge evaluation performance and achieved a Rouge-W score of 25.49% on the Chinese Legal Case Abstraction dataset.

References

Nouf Ibrahim Altmami and Mohamed El Bachir Menai. 2020. Automatic summarization of scientific articles: A survey. *Journal of King Saud University-Computer and Information Sciences*.

Rupal Bhargava and Yashvardhan Sharma. 2020. Deep extractive text summarization. *Procedia Computer Science*, 167:138–146.

Arman Cohan and Nazli Goharian. 2015. Scientific article summarization using citation-context and article's discourse structure. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 390–400.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. In *Findings of EMNLP*. Association for Computational Linguistics.

Zhenrong Deng, Fuxin Ma, Rushi Lan, Wenmimg Huang, and Xiaonan Luo. 2020. A two-stage chinese text summarization algorithm using keyword

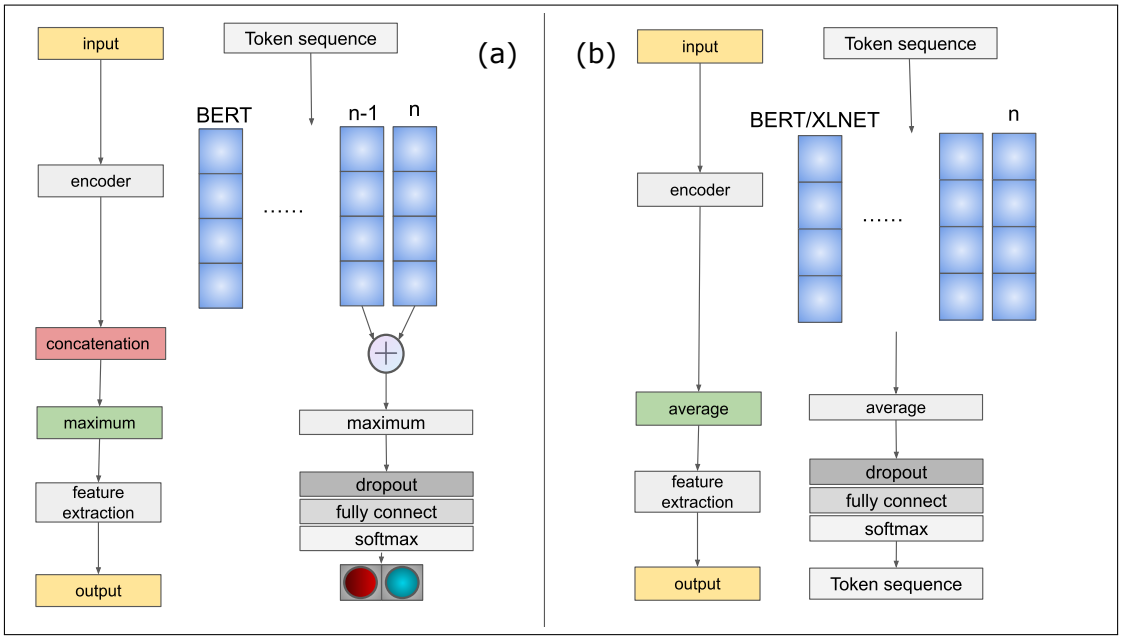

Figure 2: Different architectures for multi-grained summarization. (a) Extractive Sentence Summarization, (b) Sequence To Sequence Summarization.

information and adversarial learning. *Neurocomputing*.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, et al. 2019. A summarization system for scientific documents. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 211–216.

Touseef Iqbal and Shaima Qureshi. 2020. The survey: Text generation models in deep learning. *Journal of King Saud University-Computer and Information Sciences*.

Maria Liakata, Simon Dobnik, Shyamasree Saha, Colin Batchelor, and Dietrich Rebholz Schuhmann. 2013. A discourse-driven content model for summarising scientific articles evaluated in a complex question answering task. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 747–757.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Marina Litvak, Natalia Vanetik, and Yael Veksler. 2019. Easy-m: Evaluation system for multilingual summarizers. *SUMMARIZATION ACROSS LANGUAGES, GENRES AND SOURCES*, page 53.

Seth Polsley, Pooja Jhunjhunwala, and Ruihong Huang. 2016. Casesummarizer: a system for automated summarization of legal texts. In *Proceedings of COLING 2016, the 26th international conference on Computational Linguistics: System Demonstrations*, pages 258–262.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. 2013. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

Wen Xiao and Giuseppe Carenini. 2019. Extractive summarization of long documents by combining global and local context. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3002–3012.

Haoyu Zhang, Jingjing Cai, Jianjun Xu, and Ji Wang. 2019. Pretraining-based natural language generation for text summarization. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 789–797.