10-2022

# ESSM: An Extractive Summarization Model with Enhanced Spatial-Temporal Information and Span Mask Encoding

Ran Li

Fengbo Zheng

Gongbo Liang

Lifen Jiang

Panpan Wu

*See next page for additional authors*

## Authors

Ran Li, Fengbo Zheng, Gongbo Liang, Lifen Jiang, Panpan Wu, and Bowei Chen

# ESSM: an extractive summarization model with enhanced spatial-temporal information and span mask encoding

Ran Li, Fengbo Zheng, Gongbo Liang, Lifen Jiang, Panpan Wu, et al.

Ran Li, Fengbo Zheng, Gongbo Liang, Lifen Jiang, Panpan Wu, Bowei Chen, "ESSM: an extractive summarization model with enhanced spatial-temporal information and span mask encoding," Proc. SPIE 12451, 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), 124514A (20 October 2022); doi: 10.1117/12.2656799

Event: 5th International Conference on Computer Information Science and Application Technology (CISAT 2022), 2022, Chongqing, China

# ESSM: An extractive summarization model with enhanced spatial-temporal information and span mask encoding

Ran Li[#a], Fengbo Zheng[#a], Gongbo Liang[b], Lifen Jiang[*a], Panpan Wu[a], Bowei Chen[a]
[a]College of Computer and Information Engineering, Tianjin Normal University, Tianjin, China;
[b]Department of Computing & Cyber Security, Texas A&M University-San Antonio, TX 78224, USA
[#] These authors contributed equally to this work.
[*] Corresponding author: Lifen Jiang, E-mail: wxxjlf@sina.com.

## ABSTRACT

Extractive reading comprehension is to extract consecutive subsequences from a given article to answer the given question. Previous work often adopted Byte Pair Encoding (BPE) that could cause semantically correlated words to be separated. Also, previous features extraction strategy cannot effectively capture the global semantic information. In this paper, an extractive summarization model is proposed with enhanced spatial-temporal information and span mask encoding (ESSM) to promote global semantic information. ESSM utilizes Embedding Layer to reduce semantic segmentation of correlated words, and adopts TemporalConvNet Layer to relief the loss of feature information. The model can also deal with unanswerable questions. To verify the effectiveness of the model, experiments on datasets SQuAD1.1 and SQuAD2.0 are conducted. Our model achieved an EM of 86.31% and a F1 score of 92.49% on SQuAD1.1 and the numbers are 80.54% and 83.27% for SQuAD2.0. It was proved that the model is effective for extractive QA task.

**Key words**: extractive reading comprehension; spatial-temporal information; span mask

## 1. INTRODUCTION

Machine reading comprehension[1] is a valuable study in the field of natural language processing (NLP). Extractive reading comprehension[2] is to extract consecutive subsequences from the given article to answer question. Existing machine reading comprehension has some problems in solving practical problems. First, previous work adopted Byte Pair Encoding[3] (BPE)[2], which randomly selected token as mask unit that could cause semantically highly correlated words to be separated. Then, during feature extraction[4] the global semantic information cannot be effectively captured that leads to many semantically informational loss. Also, most of the existing models do not consider how to deal with the unanswered question, therefore the practicality of the question and answer (QA) task has been substantially reduced.

In this paper, we propose an extractive summarization model with enhanced spatial-temporal information and span mask encoding (ESSM). Firstly, in the Embedding Layer, the model adopts a mask method that based on span of geometric distribution[5] to maintain semantically correlated sequences. Secondly, in the TemporalConvNet Layer[6], the model enhances spatial-temporal information. During the feature extraction, the global semantic information from high-level features[7] is captured, which can reduce the loss of extractive feature information. More, we comprehensively consider how to deal with answerable and unanswerable questions. And we conduct experiments on datasets SQuAD1.1 and SQuAD2.0, which achieve an EM of 86.31% and a F1 score of 92.49% on SQuAD1.1 and the numbers are 80.54% and 83.27% for SQuAD2.0. Experimental results demonstrate that the model is effective for extractive QA tasks and greatly improves its performance.

## 2. RELATED WORK

With the popularity of deep learning, more and more researchers adopt neural networks to build models[8], such as BiDAF[9], Match LSTM[10], QANet[11] etc. In terms of mask methods[12], Google team proposed the Bert[13] model in 2018. It adopts Byte Pair Encoding (BPE) and randomly selects token as the mask unit. However, it could cause semantically highly correlated words to be separated. In 2019, Cui Y[14] et al. designed Whole Word Masking (WWM) utilizing

5th International Conference on Computer Information Science and Application Technology
(CISAT 2022), edited by Fuming Zhao, Proc. of SPIE Vol. 12451, 124514A
 · doi: 10.1117/12.2656799

word-based mask, to deal with the full-words. But this mask method is more suitable for combination of units in Chinese rather than independent units in English. Sunday Y[15] et al. developed ERNIE, which masked the complete named entities. However, before model training, it needs to label these words or phrases.

In terms of datasets, Rajpurkar[16] et al. constructed the extractive MRC dataset SQuAD1.1 utilizing the crowdsourcing service model (in order to distinguish the SQuADRUn dataset proposed by the author in 2018, the former is called SQuAD1.1, The latter is called SQuAD2.0[17)]. The two datasets are widely applied in natural language research related to question answering. In this paper, we also conduct experiments with these two datasets to verify the effectiveness.

## 3. MODEL

To maintain semantically correlated sequences, we propose a mask method that based on span of geometric distribution. At each iteration, a span will be sampled from a geometric distribution, then the starting position of this segment will be selected. This mask method could avoid semantically highly correlated words to be separated. To relief information loss, we enhances spatial-temporal information by capturing the global semantic information from high-level features. And we consider both answerable and unanswerable questions to increase practicality.

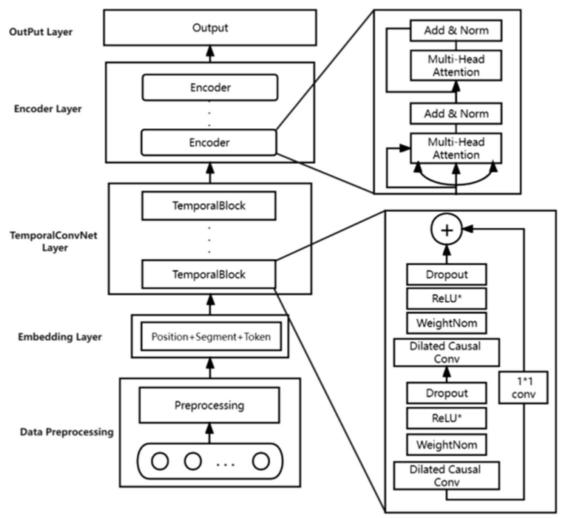


Figure 1. ESSM model structure

We develop an extractive summarization model with enhanced spatial-temporal information and span mask encoding (ESSM) to reduce semantic segmentation of correlated words and relief the loss of feature information. Our model consists of five layers, the model structure is shown in Figure 1. The data preprocessing gets the text information from datasets, and add the "is_impossible" parameter to determine whether the question can be answered. The default value is false that mean the question can be answered. When the value is true, it means the questions are unanswerable. The Embedding Layer implements random adjacency word segmentation span mask. The temporalConvNet Layer achieves high-level feature extraction. The Encoder Layer is composed of several identical encoder modules, and each encoder is a transformer encoder structure. It utilizes numerous multi-head attention mechanisms to connect each other. The Output Layer is adopted to predict the start position and end position.

### 3.1 Embedding Layer

Input the Paragraph and the Question to the model, then map it to a high-dimensional feature vector, its output dimension is [batch, seq_len, d_model]. The model structure is shown in Figure 2.

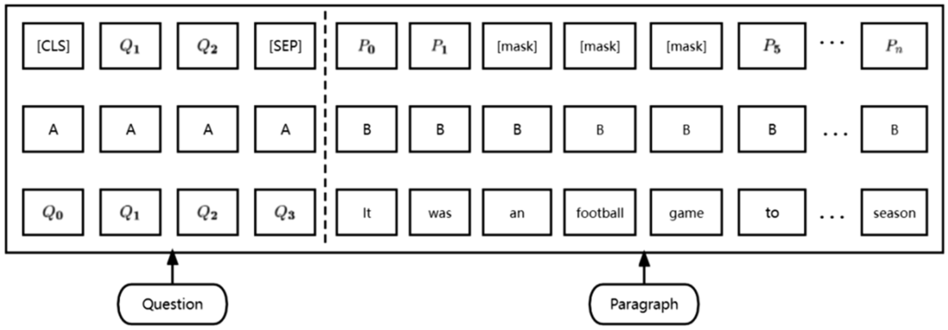


Figure 2. Embedding structure

Previous work masked 15% input tokens, but this method randomly selects tokens as its mask unit, resulting in semantically highly correlated words to be separated. This paper innovates on this basis, adopting a mask method based on span rather than token. In each iterative, it selects spans until reaching the budget 15%. First sample a span length from geometric distribution. Then select a starting point with uniform distribution as the first mask span.

### 3.2 TemporalConvNet Layer

After the Embedding Layer, in order to obtain the global information of the sequence, we captured the global semantic information from high-level features. It is helpful to improve the answer accuracy. The value at time t only depends on the value of the next layer and before layer. But the length of model is limited by the size of the convolution kernel. To capture longer dependencies, lots of linear layers must be stacked to obtain larger receptive field. However, after the pooling layer, it may cause information loss. In order to solve this problem, our model adopts 1D Convolutional Network. Finally, we introduce residual block, which is applied to replace one Convolutional layer. Each residual block contains two dilated convolution and two nonlinear mappings. At the same time, in each layer it adds WeightNorm and Dropout[18] for regularization.

### 3.3 Output Layer and Loss Calculation

The Output Layer is utilized to predict the start position and end position of the answer, and we adopt the maximum scores as output prediction. As for unanswerable questions we set the range of answer from start to [CLS].

The loss function adopts entropy loss, which consists of two parts. One part is the loss of the model mask. Other part is the loss of average answer prediction. The calculation method is shown in formula (1).

$$\ell = \ell_{\mathrm{MLM}} + \ell_{\mathrm{pred}} \tag{1}$$

## 4. EXPERIMENT

### 4.1 DataSet

This paper adopts extractive reading comprehension public dataset SQuAD, consisting of questions propose by crowdworkers on Wikipedia articles, and the answers of those questions are continuous text from each article. ESSM comprehensively considers how to deal with answerable and unanswerable questions. By testing these questions, the model enhances the practicality of extractive QA task. An example of SQuAD2.0 is shown in Figure 3.

Passage

At the time of its release, Twilight Princess was considered the greatest entry in the Zelda series by many critics, including writers for 1UP.com, Computer and Video Games, Electronic Gaming Monthly, Game Informer, GamesRadar, IGN, and The Washington Post. It received several Game of the Year awards, and was the most critically acclaimed game of 2006. In 2011, the Wii version was released under the Nintendo Selects label. A high-definition port for the Wii U, The Legend of Zelda: Twilight Princess HD, will be released in March 2016.

Question： What year will the game release a high-definition port for the Wii U console?
is_impossible： False
Answer： 2016

Question： What accolade did Radar Princess receive after its release?
is_impossible： True
Answer： [ ]

Question： How many Game of the Year awards did Twilight Princess receive?
is_impossible： False
Answer： several

Figure 3. An example of SQuAD2.0

### 4.2 results and analysis

In order to evaluate the performance of the ESSM, we conduct comparative experiments. Our model and other similar models such as QANet, Bert, TCN+Attention are tested on the dataset SQuAD1.1. The comparison results of above different models in dataset SQuAD1.1 experiment are shown in Table1. Through the experimental comparison results, it can be seen that ESSM has significantly improves the accuracy. Especially compared with bert, the EM value is improved by 5.07%, and the F1 value is improved by 0.42%.

Table1.SQuAD1.1 Model performance comparison

| Model | EM | F1 |
|---|---|---|
| QANet | 69.2 | 78.76 |
| TCN+Attention | 70.71 | 79.94 |
| Bert | 81.24 | 92.07 |
| AE-ESFS | 86.31 | 92.49 |

On the dataset SQuAD2.0 with unanswerable questions, we conduct a comparative experiment with the Bert model, the result is shown in Table2. According to the result, it proved that ESSM increases the value both on EM and F1 obviously. Compared with Bert, the EM value is improved by 2.73%, and the F1 value is improved by 7.13%.

Table2.SQuAD2.0 Model performance

| Model | EM | F1 |
|---|---|---|
| Bert | 72.99 | 76.14 |
| AE-ESFS | 80.54 | 83.27 |

Through the above experimental comparison results, we give an example on dataset SQuAD2.0 that is wrong prediction on Bert, but correct on ESSM shown in Figure 4.

Passage

One of the first Norman mercenaries to serve as a Byzantine general was Herv00e9 in the 1050s. By then however, there were already Norman mercenaries serving as far away as Trebizond and Georgia. They were based at Malatya and Edessa, under the Byzantine duke of Antioch, Isaac Komnenos. In the 1060s, Robert Crispin led the Normans of Edessa against the Turks. Roussel de Bailleul even tried to carve out an independent state in Asia Minor with support from the local population, but he was stopped by the Byzantine general Alexius Komnenos.

Question： When did Herve serve as a Byzantine general?
Answer of Bert： [ ]
Answer of ESSM： 1050s

Figure 4. An example of comparison between Bert and ESSM

## 5. CONCLUSION

We proposed an extractive summarization model with enhanced spatial-temporal information and span mask encoding (ESSM) to promote global semantic information. Firstly, in the Embedding Layer, the model adopts a mask method that based on span of geometric distribution to maintain semantically correlated sequences. Secondly, in the TemporalConvNet Layer, the model capture the global semantic information from high-level features to reduce the loss of feature information extraction. Our model can also deal with unanswerable questions. On datasets SQuAD1.1 and SQuAD2.0, our model achieved substantial improvements compared with Bert.

## ACKNOWLEDGENTS

This work was supported by the National Natural Science Foundation of China (Grant No. 61902282), the Tianjin Municipal Education Commission Project for Scientific Research Plan (Grant No. 2018KJ155) and the Doctoral Foundation of Tianjin Normal University (Grant No. 043135202-XB1707).

## REFERENCES

[1] Baradaran, Razieh, Razieh Ghiasi, and Hossein Amirkhani. "A survey on machine reading comprehension systems." Natural Language Engineering 1-50 (2020).
[2] Zeng, Changchang, et al. "A survey on machine reading comprehension—tasks, evaluation metrics and benchmark datasets." Applied Sciences 10.21 7640 (2020) .
[3] Tacorda, Alfred John, et al. "Controlling byte pair encoding for neural machine translation." 2017 International Conference on Asian Language Processing (IALP). IEEE, 168-171 (2017).
[4] Nevatia, Ramakant, and K. Ramesh Babu. "Linear feature extraction and description." Computer Graphics and Image Processing 13.3 257-269 (1980).
[5] Altun, Emrah. "A new generalization of geometric distribution with properties and applications." Communications in Statistics-Simulation and Computation 49.3 793-807 (2020).
[6] Bai, Shaojie, J. Zico Kolter, and Vladlen Koltun. "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling." arXiv preprint arXiv:1803.01271 (2018).
[7] Zhang, Zhenli, et al. "Exfuse: Enhancing feature fusion for semantic segmentation." Proceedings of the European conference on computer vision (ECCV). 269-284 (2018).
[8] Sun, Yuan, et al. "Teaching Machines to Read and Comprehend Tibetan Text." Journal of Computer and Communications 9.9 143-152 (2021).
[9] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).
[10] Wang, Shuohang, and Jing Jiang. "Machine comprehension using match-lstm and answer pointer." arXiv preprint arXiv:1608.07905 (2016).
[11] Yu, Adams Wei, et al. "Qanet: Combining local convolution with global self-attention for reading comprehension." arXiv preprint arXiv:1804.09541 (2018).
[12] Dondo, Diego Gonzalez, et al. "Application of deep-learning methods to real time face mask detection." IEEE Latin America Transactions 19.6 994-1001 (2021).
[13] Lee, J. Devlin M. Chang K., and K. Toutanova. "Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[14] Cui, Yiming, et al. "Pre-training with whole word masking for chinese bert." IEEE/ACM Transactions on Audio, Speech, and Language Processing 29 3504-3514 (2021).
[15] Sun, Yu, et al. "Ernie: Enhanced representation through knowledge integration." arXiv preprint arXiv:1904.09223 (2019).
[16] Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." arXiv preprint arXiv:1606.05250 (2016).
[17] Rajpurkar, Pranav, Robin Jia, and Percy Liang. "Know what you don't know: Unanswerable questions for SQuAD." arXiv preprint arXiv:1806.03822 (2018).
[18] Lappan, Sara N., Andrew W. Brown, and Peter S. Hendricks. "Dropout rates of in-person psychosocial substance use disorder treatments: a systematic review and meta-analysis." Addiction 115.2 201-217 (2020).