Chinese Text Summarization Generation Algorithm based on ERNIE-GEN with Few-shot Learning

Yongqing ZHU*, Peng ZHAO, Xiaodong MU, Kun BAI, Xuanang YOU
College of Operational Support, Rocket Force University of Engineering, Xi’an, Shaanxi 710025, China
*Corresponding author’s e-mail: zyq593350@gmail.com

Abstract. In view of the problems of exposure bias and insufficient annotation data in the existing Chinese short text summary algorithms, a new pre-training short text summary algorithm named ERNIE-GEN-CTS based on ERNIE-GEN is proposed. This paper uses pre-training language model to realize the double attention of word level and word level for short text, and fills the sub block with semantic information to optimize the exposure bias problem. After the pre-training model of large-scale Chinese corpus, the results are improved in LCSTS data set after the task tuning of the downstream summary, and the validity and stability of the model are verified.

1. Introduction
Different from the extraction and combination of important sentences in extractive text summarization, abstractive text summarization needs to analyse the structure and semantics of the original text, and regenerate the abstract text. With the evolution of deep learning and end-to-end generation model, sequence to sequence model has become the core model of generative summarization in recent years, and the effect is more prominent than that of extractive summarization. Since the development of deep learning, with the proposal of transformer framework, the original sequence to sequence model based on RNN, LSTM, GRU and other models as encoders and decoders has been developed into the sequence-to-sequence model based on transformer, and its framework has gradually formed the main framework of deep pre-training model. The demand for training set is also transformed from supervised learning to unsupervised learning. In the case of lack of enough labelled data, the effect has been greatly improved. Aiming at the problem of insufficient training set for Chinese text summarization and combining with the more and more popular pre training model, this paper analysed the effect of Chinese abstractive text summarization algorithm based on ERNIE-GEN\cite{1} with few-shot learning.

2. Related work
At present, the research on the abstractive text summarization is studied by domestic and foreign scholars to different extent, and different improvement methods are put forward for the common core problems such as the out of vocabulary, generation repetition, long-range dependence and evaluation standard of the abstractive text summary. See et al.\cite{2} proposed the pointer generator network (PGN), which is the replication mechanism. In view of the problem that the basic model of sequence to sequence often does not accurately reproduce the details of facts, the OOV problem is relieved by copying words from the source text targeted by pointer, which allows summary words to be generated by pointer copying the words of the source document. It also allows sampling and generating from fixed word base under certain probability. This paper proposes a coverage mechanism, which uses attention distribution
to track the words that should be selected at present. When attention points to the same content as the previous step, it will be punished to solve the problem that there are frequent repetitive fragments in sentences generated from sequence models based on attention. Celikyilmaz et al. [3] used LSTM to extract the semantic representation of sentences, and uses DCA to solve the problem of how to better gather information in long distance. The maximum likelihood estimation, semantic cohesion, sentence by sentence reinforcement learning strategy are used to improve the accuracy, consistency and abstraction of abstracts, it also used the self-critical training method of intensive learning to calculate the non-differentiable Rouge function. Yang et al. [4] aimed at the problem of limited input length of Bert, the paper proposes that the problem can be solved by applying reasoning to the sentence alone, and then aggregate the sentence scores to generate document scores. This idea can alleviate the long-range dependence of Bert. Li et al. [5] proposed that the SCST technology was used to optimize the SCST technology directly in convolution sequence to sequence frame, which alleviated the exposure deviation problem and realized the calculation of the irremediable summary measurement rouge. Xu et al. [6] combined graph convolutional network model, uses graph to connect the analytic tree of sentences in documents, and uses stack graph convolution network to learn the syntax representation of documents. Through selective attention mechanism, it is used to extract significant semantic and structural information and optimize the generation of summary results. Zou et al. [7] had achieved comparable supervisory effect by integrating the unsupervised training method with self-built large-scale corpus.

According the deep learning of the generated text summary model, we can find that the special model is specialized in solving specific problems, so it has its own algorithm core, codec, application scope, advantages and limitations, etc. As a result, it needs to be used after research according to the actual situation, and then integrate a better automatic text summary model. In view of the situation that the pre-training work of natural language generation seldom pays attention to the exposure bias [8] of downstream tasks, which is that the basic fact words are used in the training process, and the words generated are used for the reasoning that is easy to accumulate errors, whether the prediction is correct or not. ERNIE-GEN is an enhanced multi-flow sequence to sequence pre-training and fine-tuning framework, it proposes a solution, which is very helpful for generating tasks such as summary.

3. Chinese Text Summarization Generation Model

3.1. Data preprocessing
In order to unify the input of the data set, the data are processed uniformly, the same amount of data is intercepted for the experiment, and the training input data is converted into a more suitable format for ERNIE-GEN model before training, as follows: < serial number + ‘n’ + source texts + ‘n’ summary >.

3.2. ERNIE-GEN pre-training model

3.2.1. Model architecture
ERNIE-GEN model mainly includes noise module, attention module and objective function, and its framework is shown in figure 1. In the noise module, super parameters $\rho$ are set to control the text replacement ratio in the two stages of pre training and fine-tuning. The attention module is divided into word level and word level attention based on semantic information filling sub block. The attention mode of word level is consistent with that of pre training model such as Bert. Span level, as its innovation, improves the degree of semantic association within and between words. The proposition of semantic information filling sub block [ATTN], promotes the training of reasoning stage to generate summary based on the overall semantics. In order to avoid the influence of the previous output error on the overall summary generation, the paranoid exposure problem of pre-training model such as Bert is improved. The loss function combines the loss of word level attention module and word level attention module by using super parameter $\lambda$, and reduces the loss caused by the two models.
3.2.2. Contextual multi-flow attention mechanism

Contextual multi-flow attention mechanism is the attention module in ERNIE-GEN model, shown in figure 2. Multi-flow is actually a dual flow, that is, word level and span level generation flow. The word level generation stream adds semantic information to fill in the sub block [ATTN] after each input word processed by noise, and each [ATTN] is calculated separately as follows:

\[
a_{l1}^{i+1} \leftarrow MH - Attn(Q = a_i^j, KV = [S^i, t_i^j, a_i^j])
\] (1)

where \(a_i^j\) indicates the \(i\)-th vector representation of the \(l\)-th layer for the artificial symbol sequence \(A_W\). The continuous selection of span is realized by two steps. In the first step, the number of n-grams is determined by t-test. Based on the number of n-grams, 200000 binaries, 50000 triples and all words are filtered to form a span vocabulary; The second step is to label all triples, triples and words in order based on the span vocabulary. Then, the semantic information is added to fill in the sub block [ATTN] after each input word processed by noise. Each [ATTN] is calculated separately, but the [ATTN] of the words belonging to the same span are the same. It is computed as follows:

\[
a_{lj}^i \leftarrow MH - Attn(Q = a_i^j, KV = [S^i, t_i^j, a_i^j])
\]

where \(j \in [b_i, b_{i+1}]\), and \(a_i^j\) denotes the \((j-b_i)\)-th vector representation of the \(i\)-th span.

Finally, by sharing the context stream parallel computing, the fusion of the two generated streams is computed as follows:

\[
\begin{align*}
X_{l1}^{i+1} & \leftarrow MH - Attn(Q = X_i^1, KV = X_i^1, M_c) \\
A_{W}^{l1} & \leftarrow MH - Attn(Q = A_{W}^{l1}, KV = [X_i^1, A_{W}^{l1}], M_W) \\
A_{S}^{l1} & \leftarrow MH - Attn(Q = A_{S}^{l1}, KV = [X_i^1, A_{S}^{l1}], M_S)
\end{align*}
\] (3)

where \(X\) denotes the concatenation of \(S\) and \(T\), \(X_i^1\) is the vector sequence of the \(l\)-th layer for the contextual flow. \(A_{W}^{l1}, A_{S}^{l1}\) are vector sequences of the \(l\)-th layer for the word-by-word and span-by-span generation flow respectively.
3.3. Loss function
The objective function of ERNIE-GEN model is generated based on contextual multi-flow attention mechanism, and the formula is as follows:

\[
L(T) = \lambda L_{\text{word}}(T) + (1 - \lambda) L_{\text{span}}(T)
\]
\[
= -\lambda \log P(T | A^{t-1}_w) - (1 - \lambda) \log P(T | A^{t-1}_s)
\]

Among them, the pre training stage is set up \( \lambda = 5 \), control to reduce the loss of word level and word level generated streams, while in the fine-tuning phase, set \( \lambda = 1.0 \), which can make the abstracts generated by reasoning output in terms of words, and improve the effect of abstracts generation.

4. Experiments
In this section, we will show and analyse the data set, evaluation index, experimental environment and experimental results.

4.1. Experimental dataset and evaluation index

4.1.1. Experimental dataset
LCSTS (large scale Chinese short text summary data set) is a short text news summary Chinese data set proposed by Harbin Institute of technology from Sina Weibo. The corpus is composed of more than 2.4 million real Chinese short articles. The author of each article gives a short abstract, in which 10666 short abstracts are manually marked with the corresponding short texts. After Paddle treatment, the training set has 2400591 pairs of Chinese abstracts and the test set has 725 pairs of abstracts.

4.1.2. Evaluation index
The evaluation criteria of abstractive automatic text summarization can be divided into manual evaluation method and automatic evaluation method. Artificial evaluation means that experts make artificial evaluation, and comprehensively consider the fluency, central idea relevance, interpretability and other aspects of the abstract. The following mainly introduces the evaluation criteria of automatic evaluation, which are divided into internal evaluation and external evaluation criteria. The internal evaluation criteria include information quantity, coherence, readability, length, redundancy, etc., while the external evaluation criteria are indirect evaluation, including retrieval accuracy, classification accuracy, etc. Among the abstractive automatic text summarization tasks, the most commonly used evaluation criteria is ROUGE (Recall-Oriented Understudy for Gisting Evaluation)\(^9\). Rouge focuses on the recall rate, compares the automatic summary generated by the system with the standard summary generated manually, and evaluates the quality of the summary by counting the number of overlapping basic units between the two. In the current environment, rouge is the most commonly used automatic text summarization evaluation standard, while ROUGE-1, ROUGE -2 and ROUGE -L are the most commonly used standards in the series to evaluate the effect of automatic text summarization. The formula is:

\[
\text{ROUGE-N} = \sum_{S \in \text{Reference}} \frac{\sum_{g \in S} \text{Count}_{\text{word}}(g)}{\sum_{g \in S} \text{Count}(g)}
\]

For the reason that the rouge module itself cannot evaluate the quality of Chinese abstract, we need to transform Chinese into vector representation, and then calculate the vector by rouge. The vector representation can be word embedded vector or ID. in this paper, the ID method is used to cut the character level of each output, and based on the established thesaurus, the result is transformed into the vector in the form of ID.

4.2. Experimental environment
In order to verify the performance of Chinese text summarization method based on ERNIE-GEN model, this paper uses labelled text data set to verify and analyse it. The experimental environment and
configuration are shown in Table 1.

| Experimental environment | Environment configuration |
|--------------------------|--------------------------|
| CPU Processor            | 4 cores, 32G             |
| Graphics processor       | Tesla V100               |
| Memory                   | 32GB                     |
| Programming language     | Python 3.7               |
| Deep learning framework  | Paddle 2.1.0             |

### 4.3. Experimental results and analysis

In the experiment, the pre-trained ERNIE-GEN model is used to fine tune the Chinese abstract data set, in which the consistent parameter is set to the following values: input batch size is 32 and the maximum encoder input is 256 tokens, and the maximum decoder input is 64 tokens, the learning rate is 5e-5, warmup proportion is 0.1, the weight-decay is 0.1, the beam search size is 5, and its length penalty is 1.0.

| Model                  | RG-1 | RG-2 | RG-L |
|------------------------|------|------|------|
| RG-1                   | 38.71| 26.39| 36.01|
| RG-2                   | 38.71| 26.39| 36.01|
| RG-L                   | 38.71| 26.39| 36.01|

The fifth group’s model ERNIE-GEN-CTS is the model of this paper. It uses the context attention mechanism of combining word level and span level generation flow to construct attention mechanism, and trains encoder and decoder, and introduces weight attenuation L2 regularization and beam search method. From table 2, it can be seen that under the ROUGE evaluation system, the rouge-1, rouge-2 and rouge-l values of ERNIE-GEN-CTS model are higher than those of the first four groups. ERNIE-GEN-CTS model introduces the semantic substructure of [ATTN], which effectively compensates the problem of exposure bias in reasoning stage, improves the amount of information entering decoder, and greatly improves the summary effect. At the same time, the model of this paper uses cluster search method at decoder side, and improves the accuracy of the abstract to some extent.

As shown in Table 3, this paper first uses all the training set data in LCSTS for fine-tuning, and then takes 3000 iters as the standard to summarize the results under the same training times environment under different data volume. The data are intercepted as 10000, 50000, 100000, 1000000 and not intercepted respectively. The training loss are show in figure 3–8. The analysis shows that when the data volume is 50000, the model effect is equivalent or even better than most models, It shows that the model still has good reasoning ability in the environment of small-scale data.
Figure 3: Data volume=2400000
Iters=20000

Figure 4: Data volume=10000
Iters=3000

Figure 5: Data volume=50000
Iters=3000

Figure 6: Data volume=100000
Iters=3000

Figure 7: Data volume=1000000
Iters=3000

Figure 8: Data volume=2400000
Iters=3000

Table 4. Ablation Experiment of ernie-gen model in the same scale and different data training sets.

| Group | RG-1 | RG-2 | RG-L |
|-------|------|------|------|
| 1     | 32.80| 19.16| 30.55|
| 2     | 32.69| 18.72| 30.46|
| 3     | 33.01| 19.17| 30.50|
| 4     | **33.30**| **19.57**| **30.97**|
| 5     | 32.23| 18.63| 29.93|
| 6     | 32.30| 18.67| 30.16|

As shown in Table 4, in order to verify the stability of small-scale data in the model, six Chinese summary data sets with a scale of 10000 are obtained by randomly intercepting continuous small-scale
Chinese summary data pairs. The six small data sets are used for fine-tuning test, and it is found that the training effect of the model is different under the same scale. The difference between the best and the worst of rouge-1, rouge-2, rouge-l is 1.07, 0.94 and 1.04 respectively, but the results are similar on the whole, which shows that the generalization ability of ERNIE-GEN is good.

5. Conclusion
This paper proposes a Chinese short text summarization model ERNIE-GEN-CTS based on ERNIE-GEN pre-training language model. The attention module of context fusion in the model effectively optimizes the exposure bias problem in the past summarization model, and achieves good and stable results in small sample Chinese data sets. It has a good enlightenment for the future Chinese summarization model. However, there is still a lot of room for further research, including the short text summary data set outside LCSTS, the applicability of medium and long text summary data set, the supplement of attention flow such as longer word level and longer exclusive nouns.

References:
[1] Xiao D, Zhang H, Li Y, et al. (2020). ERNIE-GEN: An Enhanced Multi-Flow Pre-training and Fine-tuning Framework for Natural Language Generation. https://arxiv.org/pdf/2001.11314v3.pdf.
[2] See A, Liu P J, CD Manning. (2017). Get To The Point: Summarization with Pointer-Generator Networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada. pp. 1073-1083.
[3] Celikyilmaz A, Bosselut A, He X, et al. (2018). Deep Communicating Agents for Abstractive Summarization. In: Proc of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. New Orleans, Louisiana. pp. 1662-1675.
[4] Wei Y, Zhang H, Lin J. Simple Applications of BERT for Ad Hoc Document Retrieval [EB/OL]. [2021-03-10]. https://arxiv.org/pdf/1903.10972.pdf.
[5] Li W, Yao J, Tao Y, et al. (2018). A Reinforced Topic-Aware Convolutional Sequence-to-Model for Abstractive Text Summarization. In: Proc of the 27th International Joint Conference on Artificial Intelligence (IJCAI). Stockholm, Sweden. pp. 4453–4460.
[6] Xu H, Wang Y, Han K, et al. (2020). Selective Attention Encoders by Syntactic Graph Convolutional Networks for Document Summarization. In: 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Virtual. pp. 8219–8223.
[7] Zou Y, Zhang X, Lu W, et al. (2020). Pre-training for Abstractive Document Summarization by Reinstating Source Tex. In: Proc of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Punta Cana, Dominican Republic. pp. 3646--3660.
[8] Ranzato M, Chopra S, Auli M, et al. (2016). SEQUENCE LEVEL TRAINING WITH RECURRENT NEURAL NETWORKS. In: ICLR. Caribe Hilton, San Juan, Puerto Rico.
[9] Lin C. (2004). ROUGE: A package for automatic evaluation of summaries. In: Proc of the Workshop of ACL 2004. Stroudsburg: ACL. pp. 74-81.
[10] Hu Baotian, Chen Qingcai, Zhu Fangze, et al. (2015). LCSTS: A large scale Chinese short text summarization dataset. In: Proc of the 2015 Conf on Empirical Methodsin Natural Language Processing. Stroudsburg: ACL. Beijing. pp. 1967-1972.
[11] Gu J, Lu Z, Li H, et al. (2016). Incorporating Copying Mechanism in Sequence-to-Sequence Learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (EMNLP). Berlin, Germany. pp. 1631-1640
[12] Li P, Lam W, Bing L, et al. (2017). Deep Recurrent Generative Decoder for Abstractive Text Summarization. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP). Copenhagen, Denmark. pp. 2091-2100.
[13] Ayana, Shen S, Zhao Y, et al. (2016). Neural Headline Generation with Sentence-wise Optimization. https://arxiv.org/pdf/1604.01904. pdf.