Research and Implementation of Network Comment Automatic Generation Algorithm Based on Pre-trained Language Model

Zejun Wang, Xi'an Jiaotong University,
Department of Electronics and Information Science

***Abstract*—With the explosive growth of information, how to quickly generate high-quality comments on various news content has become a hot research topic. Unlike traditional machine learning models with automatic comment generation algorithms, emerging deep learning models with automatic comment generation algorithms have their own unique advantages. Starting from the pre trained language model, we trained a large number of corpora and obtained a model that performs well in the task of automatic news comment generation. This proves the feasibility and unique advantages of the pre trained language model based automatic comment generation algorithm, and provides a foundation for further exploration of more efficient pre trained language model based automatic comment generation algorithms in the future.**

***Index Terms*—Automatic news commentary, deep learning, pre trained language model， BART**

I. INTRODUCTION

Today's society has entered the era of Internet information. All kinds of information are exploding, and there are many kinds of news; However, human time is limited. In order to quickly find news that interests oneself from the sea of news for reading, in addition to reading news headlines, there is another method that is to read news comments. Moreover, the Chinese people's tendency to watch and comment together makes news comments a very popular feature. Online comments can not only provide users with rich supplementary information, reduce the difficulty of understanding articles, but also attract users to participate in interactions, enhance user stickiness, and increase website activity, thereby increasing website traffic. Sometimes, it can also enable relevant government agencies to timely understand the public's attitude towards a certain event or product. Good online comments can help promote exchange of views and knowledge sharing, and promote social progress.

The traditional automatic generation of news comments is mainly based on statistical machine learning methods of feature engineering, utilizing retrieval methods. Given a news article, using similarity calculation method, retrieve some news related to it from the news corpus, and then select the most relevant comment from the corresponding comments of the retrieved news, and use it as the generated comment. Although the quality of generated comments is ensured through retrieval, as it is a genuine comment, this retrieval based automatic comment generation method largely relies on the type of news corpus. When there is a significant difference between the given news and the searched news corpus, it is likely that suitable comments cannot be found, so this method has weak scalability. Against the backdrop of the rise of deep learning, researchers have begun to attempt to generate news comments based on deep learning, which is not limited by news corpora and can flexibly generate diverse comments. News articles and comments are not a pair of parallel texts. News articles are much longer than comments and contain a lot of information unrelated to comments. Therefore, it is not suitable to directly apply the Seq2Seq model, which has been proven effective in other NLG tasks (such as machine translation), to the task of automatic news comment generation. This may generate a lot of noise, leading to insufficient sentence fluency. However, compared to comment generation methods based on retrieval models, deep neural networks have strong representation learning capabilities, enabling them to represent vocabulary, sentences, and even articles in vector space. This method does not rely on news corpora and can generate diverse comments freely and flexibly, making it possible to build a true news comment automatic generation system.

II. RELATED WORK

At present, there are two main methods for implementing the automatic generation of news comments, namely machine learning based retrieval models and deep learning based Seq2Seq models.

For the research on machine learning based retrieval models, Ma et al.[1] used a retrieval based comment framework, introduced variant topic models to represent topics, matched articles and comments based on topic similarity, in order to bridge the semantic gap between articles and comments. Yuan et al. [2]used the simhash algorithm to retrieve relevant news from a massive news database, then sorted the relevance of news comments based on word vectors, and finally classified the sentiment of news comments based on convolutional neural networks to generate the final comments. The experimental results show that the introduction of simhash algorithm

This paper was submitted on March 31, 2024. This work was supported by Xi'an Jiaotong University, Department of Electronics and Information Science.

Zejun Wang is with the Xi'an Jiaotong University(e-mail:1298683025@qq.com)

effectively improves the speed of comment generation. This new method can incorporate topic and emotional constraints into the comment generation process and achieve high accuracy.

With the rapid development of deep learning, the Seq2Seq model is widely used in tasks such as machine translation, automatic summary generation, and dialogue generation. Therefore, more researchers have conducted research on the automatic generation of news comments based on the Seq2Seq model. In order to generate higher quality comments, researchers have improved the structure and algorithm of the SeqSeq model and conducted some research work. Zheng et al. [3]proposed a GANN model to generate news comments and designed a gate controlled attention mechanism to extract contextual information from input. To ensure the diversity of comments, the author uses random samples and correlation control to generate comments with different themes and relevance. Inspired by the idea of Generative Adversarial Network (GAN)[4], the author also applied additional discriminators to distinguish between real and fake comments, in order to improve the comment generator. Yang et al. [5]simulated the human habit of reading articles and designed a deep architecture for automatically generating news comments - DeepCom, which consists of a reading network and a generating network. Firstly, the information fragments in the news are extracted from the reading network, and then the generating network utilizes the information fragments to generate comments. The effectiveness of this architecture has been demonstrated through experiments. Due to the recent popularity of the transformer architecture[6], which has achieved good results in many generation tasks, Seong et al. [7]introduced this architecture for research on comment generation. The experimental results demonstrate the effectiveness of the architecture for comment generation tasks. Li et al. [8] were inspired by graph convolutional networks [9] and proposed using the Graph2Seq model to generate comments. This model uses graph convolutional networks as encoders and models news using topic interaction graphs to understand the internal structure of news. Experiments have shown that this model can generate comments with information content.

In addition, Lin et al. [10] proposed a method that combines retrieval models and generative models to address the low quality of existing training data. The author believes that the more likes a comment has, the better the quality of the comment. Therefore, the author first uses the number of likes from users to rate the quality of comments, and uses the TF-IDF algorithm to calculate the similarity between articles and comments, proposing a model for rating comments. Next, the author uses a search model to find a candidate set of comments related to the input article, and then uses a comment scoring model to select the most relevant comments to the article. Finally, the article and selected comments are used as inputs for the generation model to improve the quality of training data. Yang et al. [11] argue that in practical situations, online news typically contains multiple thematic content. For example, graphic news contains a large number of images in addition to text. Yang et al. believe that content beyond text is also important because it not only makes it more attractive to readers, but also provides key information. Therefore, Yang et al. proposed a collaborative attention model to capture the dependency relationship between text and visual information, aiming to publish comments by integrating multiple topic contents.

In order to generate personalized comments, some researchers have conducted research on the characteristics of news topics, emotional tendencies, and user information. Park et al. [12] constructed four emotional language models using existing Seq2Seq models and constructed a system for generating emotional comments from newspaper articles. Zeng et al. [13] proposed a personalized comment generation network called PCGN based on the real comments and corresponding personal information of users on Weibo. PCGN utilizes user functions embedded in gated memory to personalize modeling based on user information, thereby generating comments related to the user. Yuan et al. [14] proposed an HM BiLSTM model that introduces topic attention mechanism algorithms and syntactic structure information to generate Chinese comments with specific topics.

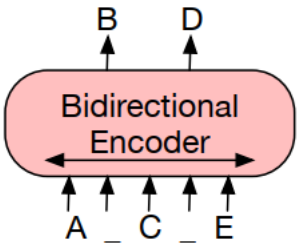
In addition, research has also been conducted on comment generation for specific fields. Choi et al. [15] constructed a system based on a recursive neural network model to automatically comment on published articles in a specific social network according to the plan. Tai et al. [16] first conducted topic classification on texts in specific fields of social networks to construct different recursive neural network models and generate initialized comment sets. Based on the similarity between the initial comments and the text topic words, text replacement is performed on the initial comments to correct the model and ultimately generate a set of comments. Yan et al. [17] learned a Joint Evolutionary Knowledge Graph (EKG) in a multitasking framework and designed a Graph to Sequence model to generate domain specific (novel) reviews in conjunction with EKG.

III. MODEL

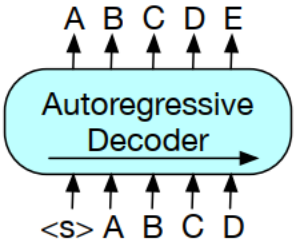
This experiment uses the Bart model as a deep learning model. BART is a denoising autoencoder that maps a corrupted document to the original document it was derived from. It is implemented as a sequence-to-sequence model with a bidirectional encoder over corrupted text and a left-to-right autoregressive decoder .

A. Architecture

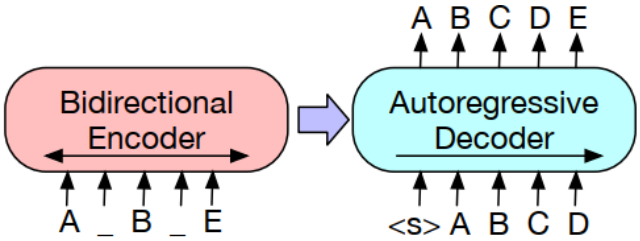
BART uses a standard Tranformer-based neural machine translation architecture which, despite its simplicity, can be seen as generalizing BERT (due to the bidirectional encoder), GPT (with the left-to-right decoder), and many other more recent pretraining schemes (see Figure 1). For the base model, BART use 6 layers in the encoder and decoder, and for the large model BART use 12 layers in each. The architecture is closely related to that used in BERT, with the following differences: (1) each layer of the decoder additionally performs cross-attention over the final hidden layer of the encoder (as in the transformer sequence-to-sequence model); and (2) BERT uses an additional



(a) BERT: Random tokens are replaced with masks, and the document is encoded bidirectionally. Missing tokens are predicted independently, so BERT cannot easily be used for generation.



(b) GPT: Tokens are predicted auto-regressively, meaning GPT can be used for generation. However words can only condition on leftward context, so it cannot learn bidirectional interactions.



(c) BART: Inputs to the encoder need not be aligned with decoder outputs, allowing arbitary noise transformations. Here, a document has been corrupted by replacing spans of text with mask symbols. The corrupted document (left) is encoded with a bidirectional model, and then the likelihood of the original document (right) is calculated with an autoregressive decoder. For fine-tuning, an uncorrupted document is input to both the encoder and decoder, and we use representations from the final hidden state of the decoder.

Figure 1: A schematic comparison of BART with BERT (Devlin et al., 2019) and GPT (Radford et al., 2018)

feed-forward network before wordprediction, which BART does not. In total, BART contains roughly 10% more parameters than the equivalently sized BERT model.

B. Pre-training BART

BART is trained by corrupting documents and then optimizing a reconstruction loss—the cross-entropy between the decoder's output and the original document. Unlike existing denoising autoencoders, which are tailored to specific noising schemes, BART allows us to apply any type of document corruption. In the extreme case, where all information about the source is lost, BART is equivalent to a language model.

BART uses several transformations to pre-train. The transformations are summarized below, and examples are shown in Figure 2.

Token Masking Following BERT (Devlin et al.,2019), random tokens are sampled and replaced with [MASK] elements.

Token Deletion Random tokens are deleted from the input. In contrast to token masking, the model must decide which positions are missing inputs.

Text Infilling A number of text spans are sampled, with span lengths drawn from a Poisson distribution ($\lambda = 3$). Each span is replaced with a single [MASK] token. 0-length spans correspond to the insertion of [MASK] tokens. Text infilling is inspired by SpanBERT (Joshi et al., 2019), but SpanBERT samples span lengths from a different (clamped geometric) distribution, and replaces each span with a sequence of [MASK] tokens of exactly the same length. Text infilling teaches the model to predict how many tokens are missing from a span.

Sentence Permutation A document is divided into sentences based on full stops, and these sentences are shuffled in a random order.

Document Rotation A token is chosen uniformly at random, and the document is rotated so that it begins with that token. This task trains the model to identify the start of the document.

IV. EXPERIMENTAL DESIGN

A. Dataset

The dataset is collected from Tencent News (news.qq.com), one of the most popular Chinese websites of news and opinion articles. Table 1 shows an example data instance in the dataset. Each instance has news content and corresponding comments.

We selected 100000 comments as the training set and 10000 comments as the validation set.

B. Experimental hyperparameter

We trained 3 epochs, with a batch size of 100 each time for both training set and validation set. We set warmup_steps to 500, and set weight_decay to 0.01.

C. Test and evaluating indicator

We selected 100000 news contents as the test set, and use the perplexity as the metric to evaluate the model's result on the test set.

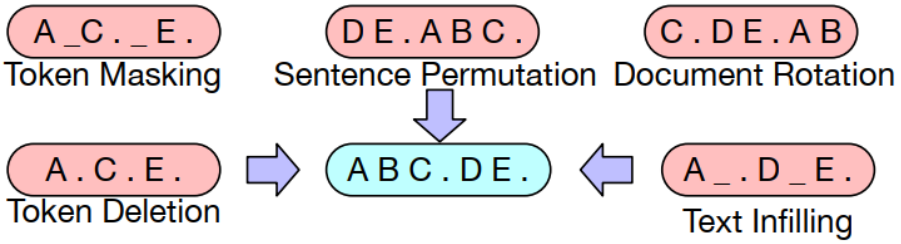


Figure 2: Transformations for noising the input that we experiment with. These transformations can be composed.

TABLE 1
A DATASET EXAMPLE

Number	NEWS CONTENT	News comments
1	【#张帅首进辛辛那提网球公开赛 16 强#】中国球员张帅 17 日在女子网球协会辛辛那提公开赛第二轮比赛中，以 6:3 和 6:4 击败俄罗斯球员亚历山德洛娃，职业生涯首次打进该赛事 16 强。接下来，中国“金花”将面对 2 号种子、爱沙尼亚球员康塔维特，争夺一张四分之一决赛门票。	不错，恭喜晋级
2	【#女孩被困 12 楼外平台消防员翻窗救援#】近日，山东聊城，一名 16 岁女孩为捡帽子，被困在 12 楼外的平台上，无法自行脱困。消防员赶到后，翻窗跳跃到平台上，利用板带固定住被困女孩后，将其成功转移到安全区域。	别的不说，这胆量我是服的
3	【习近平：党始终在人民身边】17 日下午，习近平总书记来到辽宁沈阳市皇姑区三台子街道牡丹社区，了解基层党建、为民服务等情况。习近平强调，老旧小区改造直接关系人民群众的获得感、幸福感、安全感，是提升人民生活品质的重要工作。改造老旧小区，既要改善居住环境和生活设施，也要加强社区服务、提高服务水平。我国已经进入老龄化社会，老年人越来越长寿，对老年人的服务要跟上。要抓好老龄事业、老龄产业，有条件的地方要加强养老服务设施建设。孩子们现在都是宝，对孩子们的养育和培养等工作要加强。教育“双减”工作开展后，社区要多开展公益性校外实践活动，让孩子们首先把身体锻炼好，确保身心健康。要让老百姓体会到我们党是全心全意为人民服务的，党始终在人民身边。	要让老百姓体会到我们党是全心全意为人民服务的，党始终在人民身边。
4	【#樊振东孙颖莎单打世界第一#】日前，国际乒联更新本年度第 31 周世界排名。男单方面，樊振东、马龙、梁靖崑排名前三。孙颖莎、陈梦、王曼昱、王艺迪包揽女单前四，日本名将伊藤美诚位列第五。此外，王曼昱孙颖莎女双排名世界第一。	两个超厉害的小可爱
5	【#豫剧演员高温天下乡演出挥汗如雨#】近日，安徽阜阳，豫剧演员在高温天下乡演出，穿着戏服卖力表演几个小时后挥汗如雨。观众纷纷送菜送肉，用最朴素的方式表达他们的喜爱和敬意。	这才是人民的文艺工作者

V. EXPERIMENT RESULT

Most comments can achieve semantic coherence and relevance to the news content, while a small number of comments may appear unrelated to the news content, and there are also a few comments that may have semantic inconsistencies. Table 2 shows some examples of test results.

The results of this experiment indicate that using the BART pre trained model as an automatic news comment generation model is feasible. This method saves more time and computational resources compared to training a model from scratch, and the trained model is not inferior to the latter in terms of semantic smoothness and content relevance. But, in terms of the length of comment content, most of the comment results tested by the model trained in this experiment are relatively short comment content.

However, in terms of perplexity evaluation indicators, the perplexity of model testing results is generally high, even to the point of incomprehensibility (most results have a perplexity of several hundred thousand or even several million). First of all, let me explain that the evaluation metric used in this experiment is the performance metric in the evaluate module, using a bart-base-Chinese model, while the huggingface official website uses gpt-2, and the input text for the prediction field is in English. At present, there are two possible issues that may arise. The first is that perplexity problem is caused by differences in Chinese and English, and the second is the problem caused by different model IDs. We use English text as the prediction input for the bart-base-Chinese model, but the perplexity is still high, so it can be basically determined that it is a problem with the model ID. We speculate that it may be due to the incomplete evaluation module of the Bart base Chinese model.

TABLE 2
TEST RESULTS EXAMPLE

Number	NEWS CONTENT	Comment
1	【夜读丨袅袅秋风起 最美人间秋】明日#处暑#，#秋天真的要来啦#！戳↓送你一份最美的秋色愿你我不负光阴，珍惜岁时，加倍努力！	你好，晚安
2	【#4 招学会跳绳正确姿势#】你真的会跳绳吗？怎么跳才能减少绊脚，跟着视频解锁跳绳的正确方法	学到了学到了
3	【现场视频！#海军登陆舰编队展开实战训练#】舰员奔赴各自战位，组织火力击退“敌”武装渔船，随后，编队继续向登陆海域航渡并做好对岸射击准备！近日，南部战区海军某登陆舰大队组织舰艇编队，全天候开展对海防御、对岸火力支援、抢滩登陆等多个课目的实战化训练。	了不起的祖国
4	【夜读丨哪些励志的话，一瞬间就击中了你？】#18 句提振心气的话#送给你↓要在每一次想要放弃的时候，再给坚持一点理由。苦熬心志，努力追上那个被赋予重望的自己。哪些励志的话，一瞬间就击中了你？	加油每一天！
5	【#5G 领域最新技术应用都有啥#一起探展！】今天，#2022 世界 5G 大会开幕#。本届大会以“线上+线下”形式举办，利用线下精品展览和云上展厅相结合的方式展示 5G 技术的各项应用场景。5G 领域的最新技术应用都有哪些？6G 技术开发如何？	哇哇哇，厉害了

VI. CONCLUSIONS AND FUTURE WORK

Starting from the issue of automatic generation of news comments, we explored the method of using pre trained models to train models for automatic generation of news comments. There has been good progress in the experimental results, and the trained model can be used to generate semantically coherent and news related comments. But most of them are comments with short content, and there are still a few comments with unclear semantics or unrelated to the news content, and they perform very poorly in terms of perplexity evaluation indicators. In future research, we will delve deeper into the fine-tuning methods of pre trained models and optimize some parameter layers to make them more suitable for the task of automatically generating news comments. At the same time, we will also conduct more detailed research on the evaluation indicators of the bart-based-Chinese model to solve the problem of high perplexity that occurred in this experiment.

REFERENCES

[1] MA S, CUI L, WEI F, et al. Unsupervised machine commenting with neural variational topic model[J]. arXiv preprint arXiv:1809.04960, 2018.

[2] YUAN J, SUN Y, CHEN G, et al. A Fast Chinese Review Generation Method Integrating Theme and Sentiment[C]//2019 IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). IEEE, 2019: 1652-1655.

[3] ZHENG H T, WANG W, CHEN W, et al. Automatic generation of news comments based on gated attention neural networks[J]. IEEE Access, 2018, 6: 702-710.

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//Advances in neural information processing systems. 2014: 2672-2680.

[5] YANG Z, XU C, WU W, et al. Read, attend and comment: a deep architecture for automatic news comment generation[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.2019: 5077–5089.

[6] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

[7] SEONG S, CHOI J, KIM K. A Study on Improved Comments Generation Using Transformer[J]. Journal of Korea Game Society, 2019, 19(5): 103-114.

[8] LI W, XU J, HE Y, et al. Coherent Comment Generation for Chinese Articles with a Graph-to-Sequence Model[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 4843-4852.

[9] YAO L, MAO C, LUO Y. Graph convolutional networks for text classification[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2019, 33: 7370-7377.

[10] LIN Z, WINATA G I, FUNG P. Learning comment generation by leveraging user-generated data[C]//ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2019: 7225-7229.

[11] YANG P, ZHANG Z, LUO F, et al. Cross-modal commentator: Automatic machine commenting based on cross-modal information[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 2680-2686.

[12] PARK C Y, PARK Y H, JEONG H J, et al. Automatic Generation of Emotional Comments on News-Articles using Sequence-to-Sequence Model[J]. 한국어정보학회: 학술대회논문집, 2017: 233-237.

[13] ZENG W, ABUDUWEILI A, LI L, et al. Automatic generation of personalized comment based on user profile[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.2019: 229–235.

[14] YUAN J, GUO Z, CHEN G, et al. Chinese Reviews Generation Based on HM-BiLSTM Model[C]//Journal of Physics: Conference Series. IOP Publishing, 2019, 1325(1): 012075.

[15] CHOI J, SUNG S, KIM K. A Study on Automatic Comment Generation Using Deep Learning[J]. Journal of Korea Game Society, 2018, 18(5): 83-92.

[16] TAI Y, HE H, ZHANG W Z, et al. Automatic generation of review content in specific domain of social network based on RNN[C]//2018 IEEE Third International Conference on Data Science in Cyberspace (DSC). IEEE, 2018: 601-608.

[17] YAN C, YAN J, XU Y, et al. Learning to Encode Evolutionary Knowledge for Automatic Commenting Long Novels[J]. arXiv preprint arXiv:2004.09974, 2020.