Topic-to-Essay Generation with Corpus-Based Background Information

Dan Luo¹, Xinyi Ning², Chunhua Wu¹*, Maonan Wang¹ and Jing Wu³

¹Department of Cyberspace Security, Beijing University of Posts and Telecommunications, Beijing, 100876, China
²Department of International, Beijing University of Posts and Telecommunications, Beijing, 100876, China
³Department of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing, 100876, China
*E-mail address: wuchunhua@bupt.edu.cn

Abstract. This study aims to generate more topic-related and coherence essays based on user-defined topic words. Existing research generates essays without considering the semantic information from the corpus. However, the corpus contains statistical relationships of words which can be used to guide the model to generate more coherent and fluent essays. To fill this gap, we propose a corpus-based topic-to-essay generation model (C-TEG). We elaborately devise a background network based on the co-occurrence relationships of words from the corpus. The empirical results demonstrate that our approach has achieved 4.14 average score in subjective evaluation and a better BLEU-2 score, which shows that our model is able to generate more topic-related and coherent text than existing models.

1. Introduction

Natural language generation (NLG) has been an important part of natural language processing (NLP), which has been widely applied in machine translation, dialogue system and news generation. The research of NLG has undergone a long development. The traditional approach uses some simple syntax and grammar rules as patterns or templates, which guide the model to generate text like answering a fill-in-the-bland question. At present, with the development of neural networks, more researches have been done on incorporating NLG with machine learning technologies. Through training deep neural network model with large scale corpus to automatically generate text, for example, Siri of Apple and Tay of Microsoft. As the mainstream technology is evolving, NLG technology has been successfully applied in many business fields to minimize manual involvement in practice. The needs of users are constantly improving as well. According to the given conditions to generate text that meets expectations is a new challenge for NLG. In this paper, we focus on using user-defined topics as the conditions to generate text.

At present, there is little research on text generation with given topics. Chinese poetry generation is a similar task which uses topics as conditions. Considering that a poetry usually conveys a specific topic, Chinese poetry generation tasks take user-defined words as the theme and then generate related poetry line by line. Feng et al. [9] first research on topic-to-essay generation. They proposed MTA which models the topic-to-essay generation process the same way as Chinese poetry generation. However, after implementation the generation performance is not satisfactory. The main reason is that the semantic
information provided by the user-defined words is not adequate to guide the model to generate high quality text.

In this paper, we imitate the Chinese poetry generation tasks using the corpus to provide the model additional topic semantic information. We proposed a corpus-based topic-to-essay generation approach (C-TEG). We model the problem as a sequence-to-sequence problem based on the long short-term memory (LSTM). First, we build a background network utilizing co-occurrence relationships of words from the corpus. The additional topic information is obtained through mapping user-input topic words with the background network. We further concatenate user-defined words with additional information to provide the model with more topic-related semantic information.

2. Related Work
There are mainly three approaches in NLG. First, using templates to generate text has high controllability over the generated content. Second, planning based generation method has less control as the generation process usually implemented through connecting several modules. Last, using deep learning methods generates text freely.

Corpus has been playing a key role in NLG task using deep learning methods. Corpus stores large scale knowledge or background content of the generation task. Typically, corpus contains the generation target with little noise. Through utilizing the corpus, the hidden helpful semantic information can be captured by the model. For example, the co-occurrence distribution relationships of words can be derived from corpus-based approach [1]. There are a flurry application using Corpus-based approaches, such as semantic similarity [2], query completion [3]. The quality of the corpus can influence the model performance. Corpus has been working as a key information source in poetry generation. Jukka M et al. used two corpuses for poetry generation [4], one for content using background graph and one for grammar. Marjan Ghazvininejad proposed Hafez [5], a program which automatically generates poems based on related words computed from corpus. Tim Van de Cruys trained the poetry generation model exclusively on a generic model [6] to generate more poetic verse. Yi et al. uses maximum likelihood estimation (MLE) based model to capture the poetry patterns of the corpus [7]. Malte Loller-Andersen et al. used corpus-based methods and deep learning approaches to generate subjectively topic poetry [8]. Topic information is given to the poetry generation model through a user-defined words [4] and then poetry lines are generated based on the sub-topic words or combined with previous generated lines [10-11]. Topic-to-essay task is first proposed by Feng et al. [9] which models multiple topic information with attention mechanism. But the implemented result is not in good quality as the model is going to generate a long paragraph based on five input words. The input topic information is too insufficient to generate expected text.

3. Proposed Approach
The model we proposed takes inputs from users aiming to generate text which is both fluent and topic-related. The motivation is that people read related content before writing. Inspired by Jukka M et al. [4] using background network to plan poetry content, we maintain a background network to record the related semantic information. Figure 1. intuitively shows the model structure. We follow Feng et al. [9] adopt encoder-decoder LSTM architecture with attention mechanism to dynamically assign weight of each topic.

3.1. Corpus-based Background Network
Jukka M. et al. [4] used background graph in poetry generator for content. It uses a large scale of corpus as a knowledge source which is close to what we need. The first step of our approach is to build a background graph, which is a network of word relationships based on the corpus. The basic idea is to model the frequency (co-occurrence relationships) of words. More specifically, we follow Jukka M. et al. using log-likelihood ratio test (LLR) to measure the relationship. For any word pairs \( \{a, b\} \) from the corpus a frequency table is built as in Table 1.
Table 1. Frequency Table of word pairs.

|       | $\alpha$ | $\sim\alpha$ | Total          |
|-------|----------|--------------|----------------|
| $b$   | $p_{11}$ | $p_{12}$     | $p(b; \text{Corpus})$ |
| $\sim b$ | $P_{21}$ | $P_{22}$     | $1 - p(b; \text{Corpus})$ |
| Total | $p(\alpha; \text{Corpus})$ | $1 - p(\alpha; \text{Corpus})$ | 1 |

Figure 1. The sketch of our proposed corpus-based topic-to-essay generation model.

The final log-likelihood is defined as

$$LL(a, b) = -2 \sum_{i=1}^{2} \sum_{j=1}^{2} k_{ij} \log \left( \frac{p_{ij}^{\text{independent}}}{p_{ij}} \right)$$

where $k_{ij}$ is the number of co-occurrence, $p_{ij}^{\text{independent}}$ calculated assuming the word pairs are independent, $p_{ij}$ is calculated based on the observed frequency. LL stores the score list of word frequency with the largest LL values listed at the top. The corpus network is built based on the LL. In practice, first the model takes input from users which is a word set constructed with $k$ user-defined words $T = \{\text{topic}_1, \text{topic}_2, \text{topic}_3, \ldots, \text{topic}_k\}$. In order to find words that are most semantically related to the given topic words, every topic word of $T$ is used as a query, which will be searched in the corpus network. The retrieved result is the first-level related neighborhood words of query word, which is used as the additional topic information. The final topic information is obtained through simply concatenating the user-defined topic words and retrieved results as follows:

$$\text{Topic} = \{\text{topic}_1, LL_{\text{topic}_1}, \text{topic}_2, LL_{\text{topic}_1}, \text{topic}_3, LL_{\text{topic}_1}, \ldots, \text{topic}_k, LL_{\text{topic}_k}\}$$

where $LL_i$ is the corresponding retrieving results of $\text{topic}_i$. The additional information is appended adjacent with the topic word rather than at the end considering the coherence.

3.2. Topic-related Text Generation

Based on the obtained additional topic information, the attention mechanism is adopted to score the semantic weight of each input. The final topic semantic information at each generation step $t$ is:
\[ Topic_t = \sum_{j=1}^{K} \alpha_{tj} topic_j LL_{topic_j} \]

where \( topic_j \) and \( LL_{topic_j} \) are the embeddings, and \( \alpha_{tj} \) is the attention weight of topic calculated by:

\[ \alpha_{tj} = \frac{\exp (g_{tj})}{\sum_{i=1}^{K} \exp (g_{ti})} \]

\[ g_{tj} = C_{t-1,j} \nu^T \tanh (w_a h_{t-1} + U_a topic_j) \]

where \( \nu, w_a \) and \( U_a \) are parameters optimized during the training phase. \( C_t \) is a coverage vector initialized as a K dimensional vector which controls the topic semantic calculated as:

\[ C_t = C_{t-1,j} - \frac{1}{\phi_j} \alpha_{tj} \]

where \( \phi_j = N \cdot \sigma(U_j[topic_1, LL_{topic_1}, topic_2, LL_{topic_2}, \ldots, U_{k}]) \), \( U_j \in \mathbb{R}^{kdw} \). Typically, the next word \( y_t \) is generated as:

\[ P(y_t | y_{t-1}, Topic_t, C_t) = \text{softmax}(g(h_t)) \]

during the training phase, all the parameters are trained to maximize the log-likelihood of the training corpus:

\[ (\theta^*, \eta^*) = \underset{\theta, \eta}{\text{argmax}} \sum_{t=1}^{N} \log P(y_t | Topic_t; \theta, \eta) \]

where \( \theta \) parameters of decoding LSTM and \( \eta \) are parameters of attention guidance.

### 4. Experiment Results

#### 4.1. Datasets

We use the Essay dataset made by Feng et al. [9] through crawling Chinese college essay data from the Internet. Table 2 shows the statistics of how we utilizing the dataset data. A two-layer LSTM model with 512 hidden units is trained by the AdaDelta algorithm. Other basic settings are the same as Feng et al. set.

| Corpus Network | Training | Testing | Total  |
|----------------|----------|---------|--------|
| 4,94,944       | 300,000  | 5,000   | 4,94,944 |

#### 4.2. Evaluation Metrics

**4.2.1. Human Evaluation.** Human evaluation is adopted as the major evaluation method as text generation involves innovation. 5 Chinses experts are invited to score 200 randomly selected generated samples. Four dimensions including integrity, relevance, coherence and fluency are evaluated through scoring ranged from 1 to 5.

**4.2.2. Automatic Evaluation.** Bilingual Evaluation Understudy (BLEU) [12], typically BLEU-2 is adopted as the automatic evaluation metrics. It mainly compares the generated sentence with the testing samples.
4.2.3. **Comparison Model.** PNN [10] is a topic Chinese poetry generation model which generates all the lines based on the topic. The MTA model proposed by Feng et al. [9] is used as the baseline model as it simply inputs the topic words without additional information.

4.3. **Experimental Results**

Human evaluation results are shown in Table 3. Our model scored the highest on all four dimensions, especially in Relevance. The main reason is that additional background information helps providing more related topic information. In order to better utilize the background information instead of introducing noise, we expand the additional topic information adjacent with the topic words rather than adding at the end which helps machines to understand the topic.

BLEU evaluation results are shown in Table 4. It is obvious that the automatic evaluation results are consistent with the evaluation result of human experts. Our proposed model gains the highest BLEU-2 evaluation scores. The main reason is that using the corpus-based network provides semantic information which is more consistent with learned expression habits.

Finally, our proposed model gets the highest evaluation scores on both evolution metrics, which shows that our proposed method is effective.

| Model   | Integrity | Relevance | Fluency | Coherence | Average Score |
|---------|-----------|-----------|---------|-----------|---------------|
| PNN     | 2.46      | 2.77      | 3.67    | 2.25      | 2.79          |
| MTA-LSTM| 3.40      | 2.80      | 3.72    | 2.84      | 3.19          |
| C-TEG   | 4.13      | 4.32      | 4.12    | 4.10      | 4.14          |

| Model    | BLEU Score |
|----------|-------------|
| PNN      | 1.39        |
| MTA-LSTM | 3.19        |
| C-TEG    | 3.36        |

4.4. **Case study**

Table 5. shows two examples that generated by MTA-LSTM and our proposed model. These two models both show a poor understanding of “Science” but our model generates topic-related essays that contains less redundant sentences and more coherence.

| Topic: 现在 Present, 未来 Future, 科学 Science, 梦想 Dream, 文化 Knowledge |
|-------------------------------------------|
| MTA-LSTM                                 |
| 我的梦想是当一名科学家，发明一种科学的科学，我想，未来的未来一定会实现未来的未来。 |
| My dream is to be a scientist and to invent a scientific science. I think that the future of the future will surely be realized. |
| C-TEG                                    |
| 啊！现在的我已经长大了，现在已经是一名中学生了。我的梦想是当一名科学家，发明出一种新的科学，让我们的生活更加美好，让我们的祖国更加繁荣富强，更加富饶！ |
| Now I have grown up and I am a middle school student now. My dream is to be a Scientist and invent a new kind of science to make our lives better and to make our motherland more prosperous and more powerful! |
5. Conclusion and Future Work
In this paper, we propose using a corpus-based background network neural model (C-TEG) to generate essays based on user-defined topic words. The background network is built to model the co-occurrence relationships between words which further works as a background information source for the model. In this paper, we verified the advantages of the proposed model through automatic and subjective evaluation. Our model is able to generate topic-related and more readable essays. Although this paper shows that using the background network can improve the generation performance, the model poorly understands the common sense for example, “science”. In the future, we plan to provide the model more common-sense knowledge especially nouns. We will try to integrate our model with other tasks, such as news generation, dialogue systems.

References
[1] Jiang, J. J., & Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. arXiv preprint cmp-lg/9709008.
[2] Cardon, R., & Grabar, N. (2020, May). A french corpus for semantic similarity. In Proceedings of The 12th Language Resources and Evaluation Conference (pp. 6889-6894).
[3] Rossiello, G., Caputo, A., Basile, P., & Semeraro, G. (2019). Modeling concepts and their relationships for corpus-based query auto-completion. Open Computer Science, 9(1), 212-225.
[4] Toivanen, J., Toivonen, H., Valitutti, A., & Gross, O. (2012). Corpus-based generation of content and form in poetry. In Proceedings of the third international conference on computational creativity. University College Dublin.
[5] Ghazvininejad, M., Shi, X., Choi, Y., & Knight, K. (2016, November). Generating topical poetry. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1183-1191).
[6] Van de Cruys, T. (2020, July). Automatic poetry generation from prosaic text. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 2471-2480).
[7] Yi, X., Sun, M., Li, R., & Li, W. (2018). Automatic poetry generation with mutual reinforcement learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3143-3153).
[8] Loller-Andersen, M., & Gambäck, B. (2018). Deep Learning-based Poetry Generation Given Visual Input. In ICCC (pp. 240-247).
[9] Feng, X., Liu, M., Liu, J., Qin, B., Sun, Y., & Liu, T. (2018, July). Topic-to-Essay Generation with Neural Networks. In IJCAI (pp. 4078-4084).
[10] Wang, Z., He, W., Wu, H., Wu, H., Li, W., Wang, H., & Chen, E. (2016). Chinese poetry generation with planning based neural network. arXiv preprint arXiv:1610.09889.
[11] Ghazvininejad, M., Shi, X., Choi, Y., & Knight, K. (2016, November). Generating topical poetry. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (pp. 1183-1191).
[12] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).