Topic-to-Essay Generation with Neural Networks

Xiaocheng Feng, Ming Liu, Jiahao Liu, Bing Qin*, Yibo Sun, Ting Liu
Harbin Institute of Technology, China
{xcfeng, mliu, jhliu, bqin, ybsun, tliu}@ir.hit.edu.cn

Abstract

We focus on essay generation, which is a challenging task that generates a paragraph-level text with multiple topics. Progress towards understanding different topics and expressing diversity in this task requires more powerful generators and richer training and evaluation resources. To address this, we develop a multi-topic-aware long short-term memory (MTA-LSTM) network. In this model, we maintain a novel multi-topic coverage vector, which learns the weight of each topic and is sequentially updated during the decoding process. Afterwards this vector is fed to an attention model to guide the generator. Moreover, we automatically construct two paragraph-level Chinese essay corpora. 305,000 essay paragraphs and 55,000 question-and-answer pairs. Empirical results show that our approach obtains much better BLEU-2 score compared to various baselines. Furthermore, human judgment shows that MTA-LSTM has the ability to generate essays that are not only coherent but also closely related to the input topics.

1 Introduction

Nature language generation (NLG), also known as text generation [McKeown, 1992; Sutskever et al., 2011], is a fundamental and challenging task in Natural Language Processing (NLP). NLG plays an important role in dialogue system [Wen et al., 2015; Xing et al., 2016], machine translation [Cho et al., 2014; Bahdanau et al., 2014] and summarization [Zeng et al., 2016]. In this paper, we focus on essay generation, which takes a set of topic words as input and outputs an essay (a paragraph) under the theme of the topics. Figure 1 shows a simple example of essay generation with multiple topic words. Automatic essay generation can be applied in many scenarios to reduce human workload, e.g. news compilation, mail generation, etc. Moreover, as writing is a skill which can only be mastered by human beings, we believe that teaching a computer program to automatically generate essays is a convincing way to test our progress towards artificial intelligence.

*Corresponding author.

1 For instance, quatrain is one of the most popular genres of poetry in China. The principles of a quatrain include: the poem consists of four lines and each line has five or seven characters.

Figure 1: Chinese example for essay generation with five topics.

We model the essay generation in the same way as the Chinese poetry generation. Both of them belong to the topic-to-sequence task. However, essay generation requires to consider multiple topics and outputs a long unstructured plain text, while poetry generation is always centering around one topic and predicting a strict structural output, like as quatrain. Therefore, essay generation is not only required to solve the common problems of NLG communities, like fluency and coherence, but also confronted with two extra challenges, which are topic-integrity and topic-relevance. The former stresses that the generated essay should involve the semantics of all input topic words and the latter one means that every generated sentence should closely surround one or more topics.

In this paper, to tackle aforementioned challenges, we develop a multi-topic-aware long short-term memory (MTA-LSTM) approach for essay generation. We represent the topic by mapping the topic words into an embedding space and employ LSTMs as generator. Afterwards, an attention model is used to construct the semantic relations between topic words and the generated words, which refers that the attention can guide the generator to produce topic-related output. Moreover, considering the fact that each individual essay is related to multiple topics with various relativities, we design a novel coverage mechanism to automatically assign a weight for each topic to denote its relativity to the target essay, and incorporate it into the coverage vector calculation formula to control the decay rate of the corresponding topic.

Furthermore, to the best of our knowledge, there is no
public large-scale dataset for essay generation yet. In order to verify the effectiveness of our approach, we construct two large-scale Chinese essay generation datasets by utilizing the naturally annotated web resources on Writing Website and ZhiHu^2. We employ TextRank [Mihalcea and Tarau, 2004] to extract topic words from the text of the former (305,000 essay paragraphs) and crawl the topic words of the latter one (55,000 question-and-answer pairs), which was annotated by the editor. Finally, we compare our model with a series of NLG models on these two datasets. Empirical results show that MTA-LSTM achieves the best performance in terms of BLEU metric. Besides, a comprehensive evaluation with human judgment demonstrates that integrating attention and coverage mechanism could improve the diversity and integrity of essays generated by a basic LSTM-based generator.

2 Task Definition and Data Collection

2.1 Task Definition

Given a set \( T = \{\text{topic}_1, \ldots, \text{topic}_i, \ldots, \text{topic}_k\} \) consisting of \( k \) topic words, essay generation aims at generating an article (a paragraph) under the theme of these topics.

2.2 Data Collection and Construction

In this part, we introduce how to collect data from the Internet and reorganize them to construct two standard corpora^3. ESSAY

In this paper, we target at essay generation, which can be seen as a variant version of Topic Composition. Topic Composition is a compulsory subject of Chinese College Entrance Examination and then there are a lot of excellent Topic Compositions for students to learn on the Internet. Therefore, a straightforward way is to collect these data as our essay candidates. In order to guarantee the quality of the crawled text, we only crawl the compositions which contain some reviews and scores. The process of the data collection is summarized as follows: a) We crawl 228,110 articles, which have high scores. b) We choose paragraphs composed of 50 to 120 words to be our corpus from these articles. c) We follow [Wang et al., 2016b] and also employ TextRank [Mihalcea and Tarau, 2004] to extract keywords as topic words. In the end, we obtain 305,000 paragraph-level essays and randomly select 300,000 as training set and 5,000 as test set. We name this dataset as ESSAY.

ZhiHu

In this paper, we also find some articles that conform to our requirements on ZhiHu, a Chinese question-and-answer website, where questions are created, answered, edited and organized by users in the community. In particular, users also give the topic words of each article. Based on the information mentioned above, we crawl a large number of Zhihu^2 articles and corresponding topic words. Referring to the standard of

ESSAY (five topic words and the length of the essay between 50 and 100), we select 50,000 articles as training data and 5,000 articles as test data.

3 Approach

We describe the proposed approach for essay generation in this section. We first present a topic averaged long short-term memory (TAV-LSTM) approach, which models the semantic representation of the topic as an average weighted summation of all topic words embeddings. Further, we extend LSTM with an attention mechanism (TAT-LSTM), where the semantic relatedness of each topic word with generator’s output are modeled. Lastly, we extend TAT-LSTM by considering multi-topic distribution, obtaining the multi-topic-aware long short-term memory (MTA-LSTM), which can continuously adjust the topic distribution along with the generating words.

3.1 Topic-Averaged LSTM (TAV-LSTM)

In this part, we describe a topic-averaged long short-term memory (TAV-LSTM) for essay generation. The topic semantics is represented as an average weighted summation of all topic words embeddings.

We use LSTM as it is a state-of-the-art performer in decoding process for nature language generation [Wang et al., 2016a; Wen et al., 2015; Bahdanau et al., 2014]. It is capable of computing the representation of a longer expression (e.g. a sentence) from the sequence of its input words one by one, which can be viewed as an encoding process. And the decoding phase can be seen as an inverse process of encoding. Therefore, an implicit topic representation can be naturally considered as the input, which is fed to the LSTM-based decoder to generate new sentences.

An illustration of the model is shown in Figure 2. We first learn the topic representation, which is calculated as the following equation:

\[
T = \sum_{i=1}^{k} \frac{\text{topic}_i}{k}
\]

where \( T \) is the topic representation, \( k \) is the number of input topic words and \( \text{topic}_i \) is the word embedding of topic word \( i \). After obtaining topic representation, we use LSTM to predict a probability distribution of the vocabulary based on the “current” hidden representation \( h_t \) of the LSTM. This can be formulated as follows:

\[
p(y_t|y_{t-1}, T) = \text{softmax}(g(h_t))
\]

Before each prediction, \( h_t \) is updated by:

\[h_t = f(h_{t-1}, y_{t-1})\]

where \( g(\cdot) \) is a linear function and \( f(\cdot) \) is an activation function that is determined by the LSTM structure.

3.2 Topic-Attention LSTM (TAT-LSTM)

The aforementioned TAV-LSTM model learns topic information through an average weighted summation of input topic words embeddings. That is to say, each topic word is
considered in a unified way. However, this topic representation is ambiguous and non-exclusive. For example, a topic representation is \([0,2,0.4]\). It can be obtained by two different sets of topic words, A: \(([0,1,0.3],[0,3,0.5])\) and B: \(([0,8,-0.3],[-0.4,1.1])\). Therefore, a reasonable approach is to directly model the generated word with the certain topic word.

Based on the consideration mentioned above, we go one step further and develop a topic-attention long short-term memory (TAT-LSTM). This model extends TAV-LSTM by introducing an attention mechanism, which scores the semantic relatedness of each topic word with the generating word and softly selects the relevant topic words to guide the model.

An overview of the attention model is illustrated in Figure 2. Compared with the previous model, the semantics of each topic word is transferred to the generated words by an attention component, which outputs a list of scalars: \(\alpha_1, \alpha_2, ..., \alpha_k\), where \(\alpha_i\) represents the score of topic word \(i\).

We follow [Bahdanau et al., 2014] and use neural networks as the calculators. Furthermore, the topic representation \(T_t\) of TAT-LSTM is sequentially updated. For each generation step \(t\), \(T_t\) can be formulated as follows:

\[
T_t = \sum_{j=1}^{k} \alpha_{tj} \text{topic}_j
\]

where \(\text{topic}_j\) is the word embedding of topic word \(j\) and \(\alpha_{tj}\) is derived by:

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

and

\[
g_{tj} = v_a^T \tanh(W_a h_{t-1} + U_a \text{topic}_j)
\]

where \(v_a, W_a\) and \(U_a\) are three matrices that need to be optimized during model training and \(g_{tj}\) is the attention score on \(\text{topic}_j\) at time step \(t\). Therefore, the probability of the next word \(y_t\) can be defined as:

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

and \(h_t\) is updated by:

\[
h_t = f(h_{t-1}, y_{t-1}, T_t)
\]

### 3.3 Multi-Topic-Aware LSTM (MTA-LSTM)

Although TAT-LSTM could make better use of the topic information, we argue that TAT-LSTM is still not good enough because it can not guarantee that the semantic of all the topic words are represented in the generated essay. Furthermore, conventional attention model, like TAT-LSTM, tends to ignore the past attentional historical information, which may lead to a situation where some topic words appear repeatedly while the others do not appear in the generated text.

To address both problems, we develop a topic-aware component on the aforementioned TAT-LSTM model and introduce a multi-topic-aware LSTM (MTA-LSTM) in this subsection. The basic idea is to maintain a topic coverage vector, each dimension of which represents the degree to which a topic word needs to be expressed in future generation, to adjust the attention policy, so that the model can consider more about unexpressed topic words. We believe that utilizing such topic distributed information could improve the thematic integrity and readability of the generated essay. Specifically, the
topic coverage vector is updated by a parameter $\phi_{t,j}$, which can be regarded as a discourse-level importance weight for $\text{topic}_j$.

An illustration of the model is shown in Figure 3. The input of semantic controlled unit is a topic coverage vector $C$, which will be initialized as a $k$ dimensional vector, $k$ is the number of input topic words, and each value is 1.0. For example, in Figure 3, we represent $C_0$ as [1.0, 1.0, 1.0, 1.0, 1.0]. When generating a new word at time step $t$, $c_{t,j}$ is calculated as follows:

$$C_{t,j} = C_{t-1,j} - \frac{1}{\phi_{t,j}} \alpha_{t,j}$$

where $\alpha_{t,j}$ is the attention weight of topic word $i$ at time step $t$ and $\phi_{t,j} = N \cdot \sigma(U_f[T_1, T_2, ..., T_k])$, $U_f \in \mathbb{R}^{k \times w}$. And $g_{t,j}$ is updated as follows.

$$g_{t,j} = C_{t-1,j} a^T \tanh(W_a h_{t-1} + U_a c_j)$$

Therefore, the probability of the next word $y_t$ can be defined as:

$$P(y_t | y_{t-1}, T_t, C_t) = \text{softmax}(g(h_t))$$

### 3.4 Training

In training phase, we follow [Tu et al., 2016] , and also take end-to-end learning for the MTA-LSTM model, which learns not only the parameters for the “original” attention-based generation model (i.e., $\theta$ for decoding LSTM, and attention model) but also the parameters for coverage modeling (i.e., $\eta$ for guidance of attention). And then all the parameters of the essay generation model are trained to maximize the log-likelihood of the training corpus:

$$(\theta^*, \eta^*) = \arg \max_{\theta, \eta} \sum_{t=1}^{N} \log P(y_t | T_t; \theta, \eta)$$

### 4 Experiment

We describe experimental settings and report empirical results in this section. Particularly, we compare the proposed model with two other popular methods: a simplified model in poetry generation [Wang et al., 2016b] and a semantically conditioned LSTM-based approach [Wen et al., 2015].

#### 4.1 Experimental Settings

We conduct experiments on these two datasets constructed previously, ESSAY and ZhiHu. Statistics of the datasets are given in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>T-words</th>
<th>Training</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESSAY</td>
<td>5</td>
<td>300,000</td>
<td>5,000</td>
</tr>
<tr>
<td>ZhiHu</td>
<td>5</td>
<td>50,000</td>
<td>5,000</td>
</tr>
</tbody>
</table>

Table 1: Statistics of the dataset, where T-words stands for input topic words.
4.2 Evaluation Metrics

Human Evaluation

The main evaluation is subjective and is conducted by 5 Chinese experts. Following [He et al., 2012; Zhang and Lapata, 2014; Wang et al., 2016b], these experts were asked to rate the essays by using a 1-5 scale on four dimensions: “Topic-Integrity”, “Topical-Relevance”, “Coherence” and “Fluency”. The score of each aspect ranges from 1 to 5 with the higher score the better and the rating scores are averaged as the final score. In addition, we evaluate the scorings consistency between multiple experts by computing the value of kappa. Finally, 200 randomly selected essays are evaluated by 5 experts and the rating scores are averaged as the final score. The value of kappa is 0.56 in 1-5 scale on four dimensions. Although we do not get very good consistency, we find their scoring results keep a consistent trend.

BLEU Evaluation

Bilingual Evaluation Understudy (BLEU) [Papineni et al., 2002] is widely used in automatic evaluation of machine translation systems. Considering that most words in traditional Chinese consist of one or two characters, we follow the Chinese Poetry Generation task[Zhang and Lapata, 2014] and take the BLEU-2 score as the automatic evaluation metric for essay generation.

4.3 Comparison to Other Methods

We compare our model with the following baseline methods on both datasets.

PNN. A Chinese poetry generation method with planning based neural network [Wang et al., 2016b]. Specifically, we remove all the special features of the poem, such as structural, rhythmical and tonal patterns.

SC-LSTM is a statistical language generator based on a semantically controlled Long Short-term Memory structure for response generation. Wen et al. incorporates a dialogue act 1-hot topic vector into the original LSTM model as coverage mechanism which enables the generator to output the topic-related information.

TAV-LSTM is a basic version of our approach, which models the semantic representation of the topic as an average weighted summation of all topic words embeddings.

TAT-LSTM is a simplistic version of our approach, which only uses attention mechanism to model the semantic relations between topic words and generating words.

4.4 Experimental Results

The results of the human evaluation study are shown in Table 2. Each column reports mean ratings of each dimension (e.g., Topic-integrity, Coherence). In the first, PNN performs very poor as it uses a kind of sequence to sequence framework, which makes its “Coherence” much lower than other models. From Table 2, we find that TAT-LSTM does yield obvious improvement over the TAV-LSTM, especially in terms of “Topic-Relevance”. The reason is that the attention mechanism can directly construct the semantic relation between topic words and generating words. Moreover, we also find that the proposed method and SC-LSTM get an overwhelming advantage in four evaluation metrics. The main reason is that both MTA-LSTM and SC-LSTM consider the multi-topic distribution. Afterwards we can see from Table 2 that in “Topic-Integrity”, “Fluency” and “Coherence”, MTA-LSTM has a slight advantage over SC-LSTM in performance, but the superiority is much obvious in “Topic-relevance”. It is because that SC-LSTM only uses the topic words as one-hot feature representation and then loses the semantics of different topics. Specifically, by comparing the length of the generated essay and the corresponding average score, we do not find the inherent relation between them. That is to say the length of the generated essay does not affect the quality of the essay. In addition, to demonstrate the difference between the results of these approaches are not random, we conduct paired t-test between each of these two approaches over these randomly selected essays to check whether the average difference in their performances is significantly different or not. Table 3 shows the two-tailed P values. The differences are all considered to be statistically significant while all p-values are less than 0.05.

To support the objective evaluation, we also show the BLEU results in Table 4. It can be seen that the BLEU results are highly consistent with the results of the subjective evaluation. A minor exception is that TAT-LSTM outperforms SC-LSTM on ZhiHu corpus. This may be caused by two reasons. In the first, the topic words of ZhiHu is labeled by the users but these annotated topic words may not appear in orig-
Table 5: An example of Chinese essay with five topic words generated by the MTA-LSTM model.

<table>
<thead>
<tr>
<th>现在 未来 梦想 科学 文化</th>
</tr>
</thead>
</table>

In the growing process, we should strive for our dreams, struggle for our ideals and devote ourselves to the future of our motherland. Our future selves should do as we do today. We should study hard from here and make contributions to our motherland. As a student, I will try my best to learn the scientific and cultural knowledge with the hope to become the backbone of the country and bring it a bright future.

Table 5 shows an example (about 150 words) produced by our model which received high scores with respect to topic-integrity and topic-relevance. After analyzing the generated results of our seq2seq model, we find that the model has the ability to generate topic-aware essay. We categorize the main issue into two groups: duplication and self-contradiction. We believe that this issues could be mitigated with a discourse-driven decoder, which takes the structure of the essay into consideration as the decoding processes. We leave this as a potential future work.

5 Related Work

We briefly introduce some related works from the task and method in this section.

Natural language generation (NLG) is a fundamental and challenging task in natural language processing and computational linguistics [Manning et al., 1999; Jurafsky, 2000; Reiter et al., 2000]. The task of essay generation could be viewed as a special kind of natural language generation [Qin et al., 2015]. In NLG communities, Chinese poetry generation, response generation and summarization, which are three similar tasks to essay generation. Chinese poetry generation [Zhang and Lapata, 2014; Wang et al., 2016b] is to generate a kind of structural text, which contains some specific structural, rhythmical and tonal patterns. Response generation [Yin et al., 2015; Xing et al., 2016] is a sequence-to-sequence task, the input and the output of which are both sentences and two sentences are always centering around one topic. As for summarization [Zeng et al., 2016], it condenses one long article to several sentences (usually less than 5) to highly summarize the article. In this paper, we regard essay generation as an extension of the poetry generation since they share the same input (both use words as input) and require the output to be a topic-related text. But essay generation is more difficult than poetry generation due to the following two reasons. In the first, some predefined structure patterns and regulations are utilized to train text generator in poetry generation task. In this situation, the generation model is easier to be constructed. In addition, poetry is short and implicit, which make the computer easier to imitate. Therefore, essay generation can be seen as a further exploration of poetry generation in artificial intelligence.

On the other hand, existing NLG approaches could be roughly divided into two categories: extractive-based methods and generative-based methods. Extractive-based methods focus on automatically learning some templates or patterns from the web and generating articles in an automatic way [Qin et al., 2015; Yan and Wan, 2015; Sauper and Barzilay, 2009]. For example, BABEL\(^4\) is a classical automated essay generation system, even though the system motivation is not quite the same. Generative-based methods rely on the ability of language model and sequential decoder to generate target sequences with long dependencies and the value of distributed representations [Cho et al., 2014]. In this work, we follow [Xing et al., 2016] and explore topic aware-based approach. [Bahdanau et al., 2014] proposed attention mechanism, which can jointly learn to align and translate in machine translation. However, attention mechanism tends to ignore past alignment information, which often leads to over-translation and under-translation. To address this problem, [Tu et al., 2016] proposed coverage-based neural model by maintaining a coverage vector to consider more about untranslated source words. [Kiddon et al., 2016] presented a neural checklist model, which models global coherence by storing and updating an agenda of text strings that should be mentioned somewhere in the output. Along the same direction, we develop a multi-topic-aware approach for essay generation to ensure the generated essay involves the semantics of all topic words.

6 Conclusion and Future Work

We develop a multi-topic-aware long short-term memory (MTA-LSTM) for essay generation. Compared with conventional nature language generator like attention-based sequence to sequence model, our approach takes into account the multi-topic distribution. We train the model in an end-to-end way on two automatically constructed large-scale Chinese essay generation datasets. Both automatic and subjective evaluation results verify that the proposed approach performs substantively better than several popular text generation methods. We have also demonstrated that our model has the ability to generate multi-topic related and expression-coherent essays by incorporating attention and coverage mechanism. In the future, we plan to integrate more

\(^4\)http://babel-generator.herokuapp.com
logical knowledge and common sense into existing model for generating discourse-level essays. We will also apply our approach to other forms of literary genres e.g. Wikipedia, news or essays in other languages.

Acknowledgments

This work was supported by the National High Technology Development 863 Program of China (No. 2015AA015407), the National Natural Science Foundation of China (NSFC) via grant 61632011 and 61772156.

References


