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 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.

在这个模型中，我们维护了一个新颖的多主题覆盖向量，它学习每个主题的权重，并在解码过程中顺序更新。
然后这个向量被输入到一个注意力模型来引导生成器

此外，我们自动构建了两个段落级别的中文散文语料库，305,000 个散文段落和 55,000 个问答对

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..

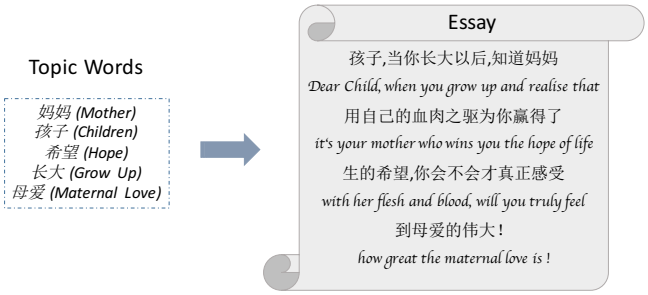

Figure 1: Chinese example for essay generation with five topics.

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.[1] 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

[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.

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 $T = \{topic_1, ..., topic_i, ..., topic_k\}$ 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' 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.

[2]Zhihu is a Chinese question-and-answer website, like Quora, where questions are created answered, edited and organized by the community of its users.

[3]Our code and data are available at: https://github.com/hit-computer/MTA-LSTM.

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:

$$T = \frac{\sum_{i=1}^{k} topic_i}{k}$$

where T is the topic representation, k is the number of input topic words and $topic_i$ is the word embedding of topic word i. After obtaining topic representation, we use LSTM to predict a probability distribution of the vocabulary set y to generate new words. At each step t of the generation, the prediction for the word y_t is based on the "current" hidden representation h_t of the LSTM. This can be formulated as follows:

$$\mathtt{P}(y_t|y_{t-1}, T) = \mathtt{softmax}(g(h_t))$$

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

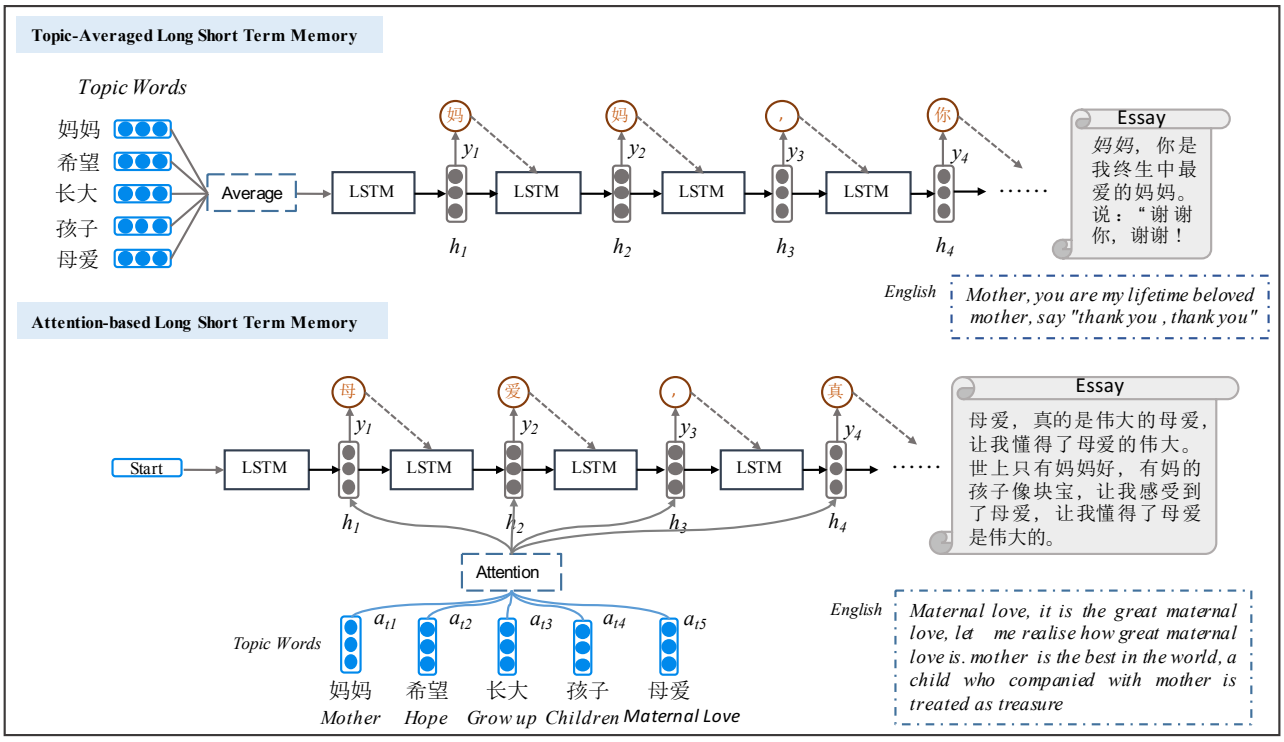

Figure 2: The basic long short-term memory approach with averaged topic words embedding and its topic-attention extension for essay generation.

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 α_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:

$$T_t = \sum_{j=1}^{k} \alpha_{tj} topic_j$$

where $topic_j$ is the word embedding of topic word j and α_{tj} is derived by:

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

and

$$g_{tj} = {v_a}^T \tanh(W_a h_{t-1} + U_a topic_j)$$

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 $topic_j$ at time step t. Therefore, the probability of the next word y_t can be defined as:

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

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

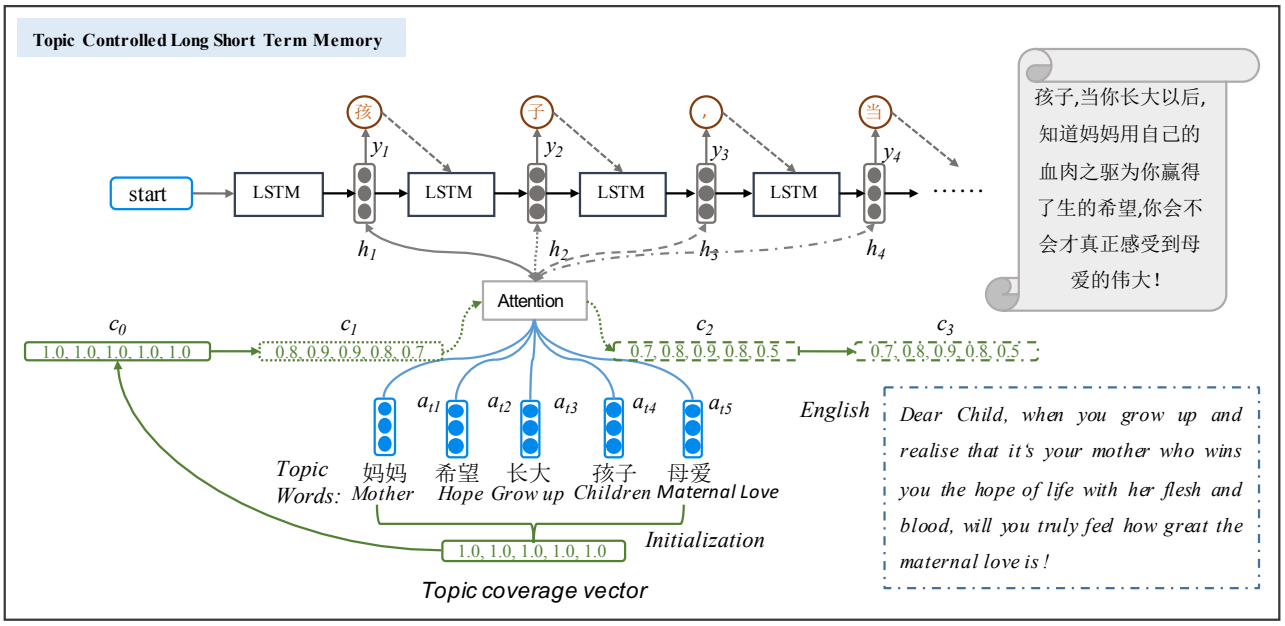

Figure 3: The multi-topic-aware long short-term memory approach for essay generation, where c are topic coverage vector.

topic coverage vector is updated by a parameter ϕ_j, which can be regarded as a discourse-level importance weight for $topic_j$.

An illustration of the model is shown in Figure 3. The input of semantic controlled unit is a topic coverage vector C_t, 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:

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

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

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

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

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

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., θ for decoding LSTM, and attention model) but also the parameters for coverage modeling (i.e., η 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:

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

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.

For implementation, we choose the 50,000 words which are frequently used as the vocabulary for both training and testing. The word embedding dimensionality is 300 and initialized by Word2vec [Mikolov *et al.*, 2013]. In detail, we pre-train the values of word vectors from 3.2 million junior or high school essays, which are crawled from the Internet. Moreover, we employ Language Technology Platform [Che *et al.*, 2010] to do Chinese word segmentation. The recurrent hidden layers of the two-layer LSTM model contains 800 hidden units. Parameters of our model were randomly initialized over a uniform distribution with support [-0.04,0.04]. The model was trained with the AdaDelta algorithm [Zeiler, 2012], where the minibatch was set to be 32. Specifically, in the testing phase, we use the beam search (beam=2) to generate diverse text.

Dataset	T-words	Training	Test
ESSAY	5	300,000	5,000
ZhiHu	5	50,000	5,000

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

Model	Score					Average Length
	Topic-Integrity	Topic-Relevance	Fluency	Coherence	Average	
PNN	2.46	2.77	3.67	2.25	2.79	97.17
SC-LSTM	3.20	3.36	3.99	3.15	3.43	95.31
TAV-LSTM	2.18	2.89	3.63	2.74	2.86	62.25
TAT-LSTM	2.34	3.14	3.71	2.88	3.02	77.82
MTA-LSTM	**3.28**	**3.92**	**4.18**	**3.37**	**3.69**	96.43

Table 2: Averaged ratings for Chinese essay generation with different methods on ESSAY datasets. Best scores in each group are in bold.

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 scoring 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.

Method 1	Method 2	P-Value
MTA-LSTM	PNN	0.000492
MTA-LSTM	SC-LSTM	0.026427
MTA-LSTM	TAV-LSTM	0.000023
MTA-LSTM	TAT-LSTM	0.000849

Table 3: Statistical Significance Test.

Model	ESSAY	ZhiHu
PNN	1.39	0.89
SC-LSTM	4.15	1.42
TAV-LSTM	3.49	1.18
TAT-LSTM	3.93	1.52
MTA-LSTM	**4.43**	**1.73**

Table 4: BLEU-2 scores with different generation methods on ESSAY and ZhiHu datasets.

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-

现在	未来	梦想	科学	文化
Now	Future	Dream	Science	Culture

在不断成长的过程中，我们更应该为自己的
During the process of growth, we should strive for our
梦想而奋斗，为理想而奋斗，为祖国的未来
dreams, struggle for our ideals and devote ourselves to
奉献出自己的一份力量。**未来**的我们也要像
the future of our motherland. Our future selves should
现在一样，要从现在开始，在这里，我们要
do as we do today. We should study hard from here and
好好学习，向着**科学**的方向前进。我们要努
here and now in the scientific direction to
力学习，长大后为祖国做贡献。作为一名学
make contributions to our motherland. As a student,
生，我一定要好好学习**科学文化**知识，学好
I will try my best to learn the scientific and cultural
科学知识，长大后成为国家的栋梁之才，让
knowledge with the hope to become the backbone of
祖国变得更加美好。
country and bring it a bright future.

Table 5: An example of Chinese essay with five topic words generated by the MTA-LSTM model.

inal articles. In addition, the semantically controlled mechanism of SC-LSTM can force the topic words to appear in the generated article. Therefore, SC-LSTM may generate some articles that are relevant to the topics but expressed inconsistently with original articles. Finally, compared with other text generation task (e.g. machine translation), the overall results of our methods are not very high, which indicates that the essay generation is definitely a more challenging task.

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 or/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

[Bahdanau *et al.*, 2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[Che *et al.*, 2010] Wanxiang Che, Zhenghua Li, and Ting Liu. Ltp: A chinese language technology platform. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, pages 13–16. Association for Computational Linguistics, 2010.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*, 2014.

[He *et al.*, 2012] Jing He, Ming Zhou, and Long Jiang. Generating chinese classical poems with statistical machine translation models. In *AAAI*, 2012.

[Jurafsky, 2000] Dan Jurafsky. *Speech & language processing*. Pearson Education India, 2000.

[Kiddon *et al.*, 2016] Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. Globally coherent text generation with neural checklist models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 329–339, 2016.

[Manning *et al.*, 1999] Christopher D Manning, Hinrich Schütze, et al. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999.

[McKeown, 1992] Kathleen McKeown. *Text generation*. Cambridge University Press, 1992.

[Mihalcea and Tarau, 2004] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into texts. Association for Computational Linguistics, 2004.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[Qin *et al.*, 2015] Bing Qin, Duyu Tang, Xinwei Geng, Dandan Ning, Jiahao Liu, and Ting Liu. A planning based framework for essay generation. *arXiv preprint arXiv:1512.05919*, 2015.

[Reiter *et al.*, 2000] Ehud Reiter, Robert Dale, and Zhiwei Feng. *Building natural language generation systems*, volume 33. MIT Press, 2000.

[Sauper and Barzilay, 2009] Christina Sauper and Regina Barzilay. Automatically generating wikipedia articles: A structure-aware approach. In *ACL*, pages 208–216. Association for Computational Linguistics, 2009.

[Sutskever *et al.*, 2011] Ilya Sutskever, James Martens, and Geoffrey E Hinton. Generating text with recurrent neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1017–1024, 2011.

[Tu *et al.*, 2016] Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. Modeling coverage for neural machine translation. *arXiv preprint arXiv:1601.04811*, 2016.

[Wang *et al.*, 2016a] Qixin Wang, Tianyi Luo, Dong Wang, and Chao Xing. Chinese song iambics generation with neural attention-based model. *arXiv preprint arXiv:1604.06274*, 2016.

[Wang *et al.*, 2016b] Zhe Wang, Wei He, Hua Wu, Haiyang Wu, Wei Li, Haifeng Wang, and Enhong Chen. Chinese poetry generation with planning based neural network. *arXiv preprint arXiv:1610.09889*, 2016.

[Wen *et al.*, 2015] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*, 2015.

[Xing *et al.*, 2016] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. *arXiv preprint arXiv:1606.08340*, 2016.

[Yan and Wan, 2015] Su Yan and Xiaojun Wan. Deep dependency substructure-based learning for multidocument summarization. *ACM Transactions on Information Systems (TOIS)*, 34(1):3, 2015.

[Yin *et al.*, 2015] Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. Neural generative question answering. *arXiv preprint arXiv:1512.01337*, 2015.

[Zeiler, 2012] Matthew D Zeiler. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*, 2012.

[Zeng *et al.*, 2016] Wenyuan Zeng, Wenjie Luo, Sanja Fidler, and Raquel Urtasun. Efficient summarization with read-again and copy mechanism. *arXiv preprint arXiv:1611.03382*, 2016.

[Zhang and Lapata, 2014] Xingxing Zhang and Mirella Lapata. Chinese poetry generation with recurrent neural networks. In *EMNLP*, pages 670–680, 2014.