Study of Automatic Essay Writing in Japanese

1910195 MA Congda

With the rapid development of computer science and technology, mobile devices make it becomes convenient to read electronic texts for relaxation and entertainment. Therefore, to satisfy the desire of people for reading, the demand for electronic texts is increased. Especially, short texts are suitable for reading since they can be read during small pockets of time. However, the number of authors (writers) tends to be decreased. It will cause imbalance between supply and demand of literary work. Text generation is one of the ways which can solve this gap. If texts such as stories and essays can be automatically generated by a computer, the number of texts that people could enjoy would be increased in the future. Text generation is also possible to support the creation of writers by utilizing generated texts as reference material.

Although research on text generation is popular, current Natural Language Generation (NLG) systems are still weak. Although stories can be generated by a neural language model, the lack of coherence harms the quality of them. It also influences the readability of generated texts. Therefore, how to maintain the coherence in the text generation is one of the urgent problems.

In this thesis, we propose a method that automatically generates an essay in Japanese. The essay is a special text with a little limitation in its writing style. The relations between events in an essay are not as close as those in a story or a novel. However, the coherence of sentences in a whole essay is still significant. Two subgoals are set in this thesis. One is to propose a model to automatically generate an essay from a given theme and keywords. The other is to improve the coherence in the automatically generated essays.

At the beginning, we define the essay generation task in this study. One theme and 4 keywords are given by a user to generate an essay. Both the theme and keywords are supposed to be nouns. The whole generated essay is required to describe something about the theme, while the main content of i-th sentence should be coincident with the i-th keyword. In this task, 4 sentences will be generated and be combined into an essay.

To realize the essay generation task, we propose the Essay Generation Model (EGM) that generates sentences of an essay one by one. In the EGM, the previously generated sentence is utilized as the input at the generation of the current sentence. It is inspired by the dialog system, in which an utterance is generated to reply to a user's previous utterance. In this way, the information in the previous sentence can be passed to the next generated

sentence to keep the relevance and coherence between sentences. To implement the EGM, the whole model is decomposed into 2 parts. The first part is used to generate the first sentence of an essay with a theme and a keyword. It is called the First Sentence Generation Model (FSGM). The second part is used to generate the rest of sentences in an essay. It is called the Content Sentence Generation Model (CSGM). In the CSGM, a theme, a keyword, and a previously generated sentence are used as the input, and a sentence is generated as the output. Both two models utilize the sequence-to-sequence model with the attention mechanism and coverage mechanism.

To improve the quality of automatically generated essays with respect to the coherence in them, we propose the Theme-Attention Essay Generation Model (TA-EGM) based on the EGM. The most important difference with the EGM is that the theme is given as the attention in the encoder, not in the input sequence. Through this method, the generated sentences are related to the theme so that the whole essay can be coherent on the theme. The TA-EGM is also decomposed into 2 parts. We call First Sentence Generation Model and Content Sentence Generation Model in the TA-EGM as FSGM-TA and CSGM-TA, respectively.

For training the EGM and TA-EGM, a new dataset is constructed by the essays downloaded from the web site "Aozora Bunko". Essays in "Essay, Review" category under the major category "Japanese Literature" are chosen. Some essays are rather old in "Aozora Bunko". To obtain relatively new essays, only the essays written by the new Japanese character system are downloaded. The number of retrieved essays is 2,140. Before the construction of the datasets for our models, several preprocessing are carried out. Firstly, the old character is replaced by the new character with the "Old and New Kanazukai comparison table". The sentences that do not meet our requirements are removed as well. Secondly, sentences in the essays are split into words by the morphological analyzer MeCab to obtain word sequences used as the input and output of our models. At the same time, the sentences containing more than 78 tokens are removed. Finally, the first noun other than a named entity in the title of an essay is extracted as the theme of the essay. The top 5 keywords extracted from an essay by TF-IDF based scoring are set to the keywords of the essay. From this corpus, two new datasets are constructed. One is a collection of triplets of (theme, keyword, sentence), which is used for training of the FSGM and FSGM-TA. The other is a collection of quadruplet of (theme, previous sentence, keyword, sentence), which is used for training of the CSGM and CSGM-TA.

In the experiment, automatically generated essays are evaluated by human subjects. A blind questionnaire with 4 questions is designed to evaluate the quality of essays. The first three (comprehensibility, relatedness to theme,

and relevance to keyword) evaluate individual sentences in an essay, while the last (coherence in essay) evaluates the overall essay. For every question, the subjects should give a score between 1 to 5. In addition to the EGM and TA-EGM, the baseline model is also evaluated and compared. It produces an essay by generating 4 sentences independently, where a theme and a keyword are given as the input.

From the results, our EGM can generate better coherent essays than the baseline. The human evaluation score on the coherence is increased by 0.12 point. Furthermore, comparing to the baseline model, the improvement by the TA-EGM is 0.31 point. It also outperforms the EGM by 0.19 point. Simultaneously, these three models are similar with respect to the grammatical correctness (comprehensibility). Although the EGM gets a lower score in the relevance between the generated sentences and the theme than the baseline, the TA-EGM where the theme is given as the attention can get comparable scores.

In the future, there are several explorations we should do to revise and improve our model. Firstly, we will search the better essays and explore the better way to extract the theme and keywords from the essay. Secondly, we reconsider what information should be used as the input to improve the quality of generated sentences. Finally, automatic evaluation of essays is essential for the optimization of hyperparameters in the training of the deep learning models.