Reading Assignment 2

Jingyuan Wang, 260860682

In the paper *Hierarchical Neural Story Generation*, the authors proposed a hierarchical story generation model with a fusion as well as gated self-attention structure and claimed to achieve notable enhancement to produce fluent, topic-centered stories of good quality with several evaluation approaches.

As indicated in the paper, three mechanism were created or utilised for the model. The first one is using a hierarchical model structure, where short story premises were initially created by a language model and stories were generated based on those prompts. It is suggested that such model will generate stories with more topicality and variety compared to single-level language model. Another improvement is model fusion, which is to combine a seq2seq model with a pre-trained seq2seq model in order to complement the previous learning process in a sense of boosting. Moreover, since the length of stories is rather long, it is convincing to use convolutional seq2seq model for the task and meanwhile a multi-scale self-attention gate is designed so that at each time step the learning can be associated with any previous generated words. During the training, the attention heads can be selected with different time-scale so that different heads can learn different information. In terms of model evaluation procedure, they also made contributions regarding collecting and organizing a new dataset with prompt story pairs from WRITINGPROMPTS and raising new evaluation metrics to measure text validity, topic relevance and overall quality.

Four methods were utilised in the experiment, with two human evaluations and two automatic ones. First of all, perplexity can easily measures how close the result of a language model is with the training language, that is, how well the generated sentence is likely to exists. By perplexity, the authors illustrated that the proposed model has notably improved the fluency of the generated story. However, it only proves a good quality for language models, which is not a necessity for novel stories. Subsequently, they used prompt ranking accuracy, which is to examine if a story is mostly likely to be produced by the language model of its true corresponding prompt, compared to 9 random prompts. It can reflects the language model relation of prompts and stories, but several doubts can be raised during this process. Since the prompt is extremely short texts, it might only capture which prompt has most common keywords of the story, let alone the high frequency of stopwords. When checking the example prompt and story, we can tell that it is not crucial for a story to follow the prompt structure and a top-1 ranking may not be the best selection. The researchers should have present more conclusive results instead of giving an accuracy of merely 16% in a task baseline of 10%.

Following that, a triple pairing task was held where human judges were asked to pair stories with their corresponding prompts, after which an accuracy for performing that is counted. It is a straightforward way to assess the relevance of a prompt and the generated story, and quantity of stories and quantity of judges were acceptable. However, a significant weakness here is that the assessment is only about the topicality, which is not a necessary feature for story generation. There would be cases when the story is slightly off-topic, yet it has fluent and vivid words with creativity. Finally, they gave the judges stories generated by the hierarchical model and the non-hierarchical one and asked them to record their preference for each pair of stories. By this means, they could compare the overall quality of two model, but such story pairing is not fairly reasonable and they could increase the number of pairs to achieve fairness. Again, they only prove an enhancement towards the hierarchical structure. Choosing one from two does not justify the chosen stories are of good quality.

For a language generation system, creativity indicates generating a variety use of language and conducting new knowledge apart from the pre-existing basis. With a standard language model, produced texts are the least creative, since it will always predict the results that will maximize the likelihood, and thus repetition will occur greatly often. Nevertheless, being creative does not break the fundamental syntactic and semantic language rules. It is possible for a generation system to achieve creativity with supervised learning. A good model should learn different combination of words, sentences and styles for communication[1] from the training process which is likely to conduct context-appropriate and informative knowledge as long as given enough data.

References

[1] Harmon et al. *Revisiting Computational Models of Creative Storytelling Based on Imaginative Recall.* 2015.