Machine Translation, Exercise 03

Zishi Zhang

23-741-390

Task 1

1. The data was collected from the English subtitles of Rick and Morty Season 1-3. Most of the sentences are the conversation between characters in the show. There are various special attributes:

A. The data are all oral sentences inside conversations, the generated text will also look like talking to someone;

B. Names of characters appeared very often in the data, the generated text may include their names;

C. Weird synthetic words were made to fit in the alien style of the story, there will be some illogical use of certain words.

2. The generated text do looks like a conversation. And it contains some exotic words like “dinglebop” which is an alien substance in the show. Also the use of “eyeholes” seems illogical, due to the commercial lines of an alien cereal named “eyehole” in the show.

Github link: https://github.com/CeeeeeDZ/mt-exercise-03.git

Task 2

1. Test perplexity is always higher than validation perplexity by 5. Based on my result, dropout at 0.4 is the best. Because it has the lowest perplexity.

2. The quality didn’t improve a lot. Due to the perplexity at dropouts 0.4 and 0.5 are not too different. But it still reserves some special features of original text.

3. The text with the highest perplexity contains much less <eos> label compared to the lowest one. It is more random and illogical. Full of sentences that doesn’t make any sense.