4.4 Paragraph Generation

We evaluate the generated paragraphs by human rating. We employee 30 web users to rate the paragraphs from 1 to 6 according to coherence, fluency and correctness. The paragraph with lower score is less coherent, and the one with higher score is more coherent.

We compare 4 different types of paragraphs, and each one has 7 different paragraphs based on different templates. The paragraphs are generated by human, our generation model with all sub-models and enhancements, our model without coherence evaluation and our model extracting commonsense knowledge from the original Chinese ConceptNet (without data cleaning). The results of human evaluation are shown in table 4.28.

	Human-written	Our model	Without coherence model	Original CN
Human rating	4.70	3.23	2.14	1.55
Human-written as gold standards	5	3.43	2.39	1.71

Table 4.28: The result of human rating.

We can see that the paragraphs generated by human get the highest score which is what we expected. The second one is paragraphs generated by our model. The third one is paragraphs generated by our model without coherence evaluation. The last one is extracting concepts from original ConceptNet. Most of the participants rate the last one as lowest score (1). There are two reasons why. First, original ConceptNet contains lots of errors and redundancy which we discussed in section 3.1. Secondly, the ConceptNet that we use to extract connective con-

cepts to replace the context words when training the coherence model is refined one, not the original one. Consequently, the coherence model doesn't work in the original ConceptNet. Also, we found that extracting concepts from the original ConceptNet are the most time-consuming method to generate paragraphs. The generated model can't find connective concepts, it has to change the initial concept or template over and over, and it still can't generate paragraphs successfully most of the time.

Because the human evaluation is subjective, everyone may have different opinions. Some participants even rate the human-written paragraphs as incoherent. Some people tend to rate higher scores, and some tend to rate lower scores. Therefore, the absolute scores are not precise. If we see the human-written paragraphs as gold standards, then we normalize the score from 1 to 5 (the second row of table 4.28). The score of our model is 3.43 which is over half of five. It means that the paragraphs generated by our model are coherent to some degree. The concepts in the paragraphs are combined closely way more than extracting concepts randomly in ConceptNet.

In table 4.29, we list some examples which are in the human evaluation dataset. The examples on left side are the highest score with that method, and the ones on the right side are the lowest score. The lowest score of our model (the second row in table 4.29) is 2.41. Participants may see the similar meaning of [優秀員工] and [好員工], hence they rate it as incoherent. We can see that the participants rate paragraphs to lower score even there are only one or two words that incoherent to other words.

A	學生上課時首先要拿筆作筆記 辛苦的讀書是為了將來 讀書時不能懶惰會令人落榜 可以邊讀邊聽音樂來沉澱心情 score: 5.11	朋友工作時懼怕出錯，會令人自責 他希望有錢時能買房子來過上快樂的生活 但沉重的房貸會帶來壓力 壓得他喘不過氣 score:4.59
B	黑道會為了耍帥而飆車 刺激的開快車時會躲避警察 耍酷時會抽菸來假裝成熟 危害健康的汙染會帶來損害 score:4.37	店裡的雇主是體恤員工的 老闆會照顧員工和賺大錢 想要優秀員工和好員工 是會令人勞累的職業 score:2.41
C	教師騎車時害怕摔車 摔車會帶來嘲諷 因為騎車而累時會喝保力達 是為了用心的精力 難過的睡不著時會想要打架 score:2.78	警方打架時要優先臭罵來動手 麻煩的拌嘴是為了確定答案 爭執時分手會令人生氣 離婚時移居是為了購屋 score:1.74
D	暖和的教師是有些很機車的 是會學習的職業 會令人想要就學、去上課 上學時會健康的吃早餐來出門買早餐 score:1.70	錢的老闆是討厭的的 老闆會給錢和性騷擾員工 想要開除員工和懶惰 是會難吃蛋餅的會罵你的人 score:1.19

A: human-written
B: our generation model
C: our generation model without coherence model as reward function in MCTS
D: extract concepts from original ConceptNet

Table 4.29: Some examples of generated paragraphs in human evaluation dataset.