4.3 Discourse Coherence Model

We compare different hyperparameters of DNN coherence model in this section. Each paragraph consists of 4 or 5 sentences, and the max number of words in a sentence is 10. The number of articles in training, validation and test data is 559537, 112279, 107002 (7:1.5:1.5). (The less important experiments are shown in appendix.)

We first evaluate whether the word embedding affects the training or not. Table 4.11 shows the comparison of joint trained and non-joint trained word embeddings. The accuracy of joint trained word embedding is lower than non-joint trained model, and the training is slower too. In table 4.12, the word embedding with higher Spearman score has higher accuracy than the other one.

WE joint trained	Train		Validation		Test		time/epoch
	loss	acc	loss	acc	loss	acc	
Yes	.277	.821	.410	.803	.448	.785	162 mins
No	.088	.972	.214	.929	.305	**.902**	**138** mins

Table 4.11: Word embedding joint trained.

Spearman score of WE	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
.604	.376	.749	.366	.763	.370	.760
.681	.263	.835	.348	.783	.327	**.800**

Table 4.12: Different scores of word embeddings.

The number of words in a sentence and how bidirectional model merges forward and backward hidden states doesn't seem to affect the performance. The results are shown in appendix table A.9

Table 4.13 shows the experiment of whether paragraphs with duplicate sentences will affect the performance or not. The size of moving sentence window is 1. The next sample is the original samples moving down by one sentence. The accuracy in both are pretty much the same in training samples, but the duplicate one perform worse in validation and test samples, which means training samples without duplicate sentences can handle more different kinds of data. Replaced rate is the probability of replacing context in negative samples. Each sentence in

Duplicate sentence	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
Yes	.216	.916	.291	.885	.366	.860
No	.189	.928	.151	.944	.142	**.947**

Table 4.13: Paragraphs with duplicate sentences or not.

a paragraph has x% to be replaced. In table 4.14, the more concepts replaced, the higher the accuracy. It's also intuitive to know that a paragraph is less coherent if it dissimilars more from the original one. From table 4.15 to 4.18, we can know that low replaced rate model can handle high replaced rate test data, but not vice versa.

Replaced rate	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
.35	.260	.828	.297	.813	.306	.806
.50	.251	.828	.279	.816	.288	.808
.65	.235	.836	.250	.826	.256	.822
.80	.247	.825	.245	.827	.239	.831

Table 4.14: Compare different replaced rate of negative samples.

Replaced rate	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
.25	.204	.925	.147	.946	.136	.950

Table 4.15: Training replaced rate 0.25.

Replaced rate	Test	
	loss	acc
.10	.536	.790
.20	.197	.925
.30	.112	.962
.40	.092	.971
.50	.086	.972
.80	.086	.973

Table 4.16: Replaced rate 0.25 in different test data.

In table 4.19, we can see the accuracy of replacing with arbitrary concepts is higher than replacing with connected concepts in ConceptNet. It means the

Replaced rate	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
.65	.041	.989	.012	.997	.012	.997

Table 4.17: Training replaced rate 0.65.

Replaced rate	Test	
	loss	acc
.10	2.19	.541
.20	1.29	.679
.30	.558	.840
.40	.173	.945
.50	.043	.986

Table 4.18: Replaced rate 0.65 in different test data.

model can distinguish the coherent or incoherent paragraphs easily if negative samples are replaced by arbitrary concepts. Intuitively we know that paragraphs being replaced by arbitrary concepts are obvious incoherent. Our MCTS-based model selects concepts from the connected ones. If the negative samples are replaced by arbitrary concepts, the performance isn't very well. All the scores of paragraphs generated by MCTS-based model are over 0.95. Namely, it can't distinguish between the good or bad ones. Consequently, we make negative samples being replaced with connected concepts. The negative samples in preceding experiments are replaced by arbitrary concepts, and in the following experiments they are replaced with connected concepts.

Negative samples	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
Arbitrary concepts	.329	.859	.268	.889	.271	.888
Connected concepts	.501	.757	.473	.779	.470	.777

Table 4.19: Negative samples replaced by arbitrary concepts against the connected ones.

In table 4.20, BiLSTM and BiGRU are almost the same, and both of them are a bit better than LSTM and GRU. The small batch size performs better than the large one, but it needs more time to train. We select 64 as our batch size. The model performs better when we use more hidden units in general.

NN architecture	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
BiRNN	.526	.737	.531	.735	.528	.739
BiLSTM	.286	.880	.308	.873	.309	**.873**
BiGRU	.285	.881	.316	.871	.315	**.871**
RNN	.541	.726	.526	.737	.529	.738
LSTM	.324	.861	.339	.857	.341	.857
GRU	.326	.860	.335	.857	.333	.859

Table 4.20: NN architectures.

Batch size	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
16	.491	.763	.446	.791	.447	**.790**
32	.501	.757	.473	.779	.470	.777
64	.508	.756	.454	.790	.456	**.788**
128	.519	.744	.477	.776	.465	.779
256	.534	.735	.491	.764	.482	.772

Table 4.21: Batch size.

Hidden units	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
8	.540	.730	.517	.750	.509	.753
16	.503	.756	.463	.781	.463	.782
32	.463	.782	.422	.807	.415	.810
64	.433	.800	.391	.824	.392	.825
128	.410	.813	.379	.836	.379	.834
256	.398	.820	.360	.842	.361	**.842**

Table 4.22: Hidden units.

Optimizer	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
SGD	.557	.712	.534	.736	.540	.733
RMSprop	.423	.811	.389	.834	.382	.835
Adam	.385	.828	.353	.846	.350	**.850**
Adagrad	.637	.639	.601	.683	.600	.683

Table 4.23: Optimizers.

Learning rate	Train		Validation		Test	
	loss	acc	loss	acc	loss	acc
.1	.696	.500	.693	.501	.695	.500
.01	.638	.641	.625	.663	.658	.614
.001	.503	.756	.463	.781	.463	**.782**
.0001	.569	.708	.520	.746	.521	.745
.00001	.674	.583	.654	.611	.654	.612

Table 4.24: learning rate.

Architecture	Epoch	Batch size	Batch numbers	Hidden units	Optimizer	Learning rate	Dropout rate
BiLSTM	10	64	16801	256	Adam	.001	.2

Table 4.25: Best settings of hyperparameters.

4.3.1 Examples

We show some examples of low and high coherence score in table 4.26. The paragraphs on the left side are the positive samples and the right side are the negative ones. The scores of positive samples are all higher than negative ones. Four or five words in the first two examples are substituted by other connected words in ConceptNet, and the coherence model can distinguish them easily. The rest of the examples only substitute two words, the model can still predict correctly.

In table 4.27, we replace the word manually. The paragraphs in the left side are replaced by plesionyms, and the ones in the right side are replaced by less related words. We find that the coherence score is the same or better than the original paragraph if replaced by plesionyms, and the score is lower than the original one if replaced by less related words. Although some of the scores are 0.7 and 0.8, they still lower than the original one. It proves that the coherence model can distinguish the coherent and incoherent paragraphs.

	High coherence	Low coherence
1	來到 角板山 行館 高處 **眺望** 溪口 台地 好 風光 美好 **景色** 一覽無遺 前往 **角板山** 樟腦 文化 特展 路上 經過 寧靜 **池子** 適合 爸媽 **走走** score:0.978	來到 角板山 行館 高處 **推下來** 溪口 台地 好 風光 美好 **人生** 一覽無遺 前往 **蟑螂** 樟腦 文化 特展 路上 經過 寧靜 **大海** 適合 爸媽 **拌嘴** score:0.004
2	父親節 家裡 **關係** 變 不好 父母 對 哥哥 **特別** 好 今天 **難得** 姐姐 休假 姐姐 提前 爸爸 約 **今天** 吃 父親節 大餐 哥哥 無法 排休 難 調班 score:0.994	父親節 家裡 **玩遊戲** 變 不好 父母 對 哥哥 **沮喪** 好 今天 **開心** 姐姐 休假 姐姐 提前 爸爸 約 **夜半** 吃 父親節 大餐 哥哥 無法 排休 難 調班 score:0.046
3	鄉民 推薦 三民區 三立 飯丸 該店 開業 招牌 **荷包蛋** 飯丸 口 咬下 流出 濃郁 蛋液 傳統 肉燥 菜脯 肉鬆 油條 佐料 加 豆皮 **香氣** 十足 包 滷蛋 大小 飯丸 score:0.787	鄉民 推薦 三民區 三立 飯丸 該店 開業 招牌 **店鋪** 飯丸 口 咬下 流出 濃郁 蛋液 傳統 肉燥 菜脯 肉鬆 油條 佐料 加 豆皮 **黃豆** 十足 包 滷蛋 大小 飯丸 score:0.543
4	物聯網 裝置 可貴 透過 終端 節點 蒐集到 資訊 經由 網路 回 雲端 進行 大數據 分析 **數據** 分析 有用 商業 政策 **發展** 寶貴 資訊 score:0.997	物聯網 裝置 可貴 透過 終端 節點 蒐集到 資訊 經由 網路 回 雲端 進行 大數據 分析 **名嘴** 分析 有用 商業 政策 **愛護** 寶貴 資訊 score:0.424
5	財團法人 高雄市 文武 聖殿 董事長 說 首度 試辦 愛心 餐券 清寒 小朋友 **平時** 在校 營養 午餐 寒假 期間 家長 全天 上班 無人 在家 怕 學生 **挨餓** score:0.721	財團法人 高雄市 文武 聖殿 董事長 說 首度 試辦 愛心 餐券 清寒 小朋友 **零分** 在校 營養 午餐 寒假 期間 家長 全天 上班 無人 在家 怕 學生 **上床睡覺** score:0.004

Table 4.26: Coherence model results on test dataset.

Original paragraph	
請 大家 幫忙 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 注意 score:0.904	
Replaced by plesionyms	Replaced by less related word
請 大家 **協助** 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 注意 score:0.969	請 大家 幫忙 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 **天晴** 發現 不見了 請 大家 幫忙 注意 score:0.094
請 大家 幫忙 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 **留意** score:0.906	請 大家 幫忙 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 **水災** score:0.03
請 大家 **協助** 朋友 照顧 貓咪 凌晨 開 **窗戶** 出去 早上 **起身** 發現 不見了 請 大家 幫忙 **留意** score:0.955	請 大家 幫忙 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 **超人** score:0.736
	請 大家 **吃飯** 朋友 照顧 貓咪 凌晨 開 紗窗 出去 早上 起床 發現 不見了 請 大家 幫忙 注意 score:0.806

Table 4.27: Coherence model results testing.