4.1 ConceptNet Data Cleaning

In order to generate coherent paragraphs, the quality of input data is as important as the model. Although the coverage of original ConceptNet is wider than the modified one, it contains numerous errors. We refine the data, expand by different kinds of relations (mainly "Synonym") and ensure the quality of ConceptNet.

We add data of relation "Antonym" and "Synonym" from Chinese WordNet [1] [95], MOE revised dictionary [2] and manually. The number of each sources is shown in table 4.1 and 4.2. Use "Antonym" to expand "HasProperty", "Desires" and "NotDesires", and use "Synonym" to expand all relations.

	ConceptNet	Chinese WordNet	MOE revised dictionary
Size	31	100	9,146
Expanded size	HasProperty	Desires and NotDesires	
	2,650	830	

Table 4.1: Number of data from different sources and expanded size in Antonym.

[1] http://lope.linguistics.ntu.edu.tw/cwn/download/

[2] 中華民國教育部 (Ministry of Education, R.O.C.) « 重編國語辭典修訂本 » (版本編號：第五版) site:http://dict.revised.moe.edu.tw/

	ConceptNet	Chinese WordNet	MOE revised dictionary	Manually
size	1,018	1,700	9,756	2,511
Expanded size	844,448			

Table 4.2: Number of data from different sources and expanded size in Synonym.

Expanded ConceptNet is 3.21 times of the size of the original one. The actual number of each statistic in original ConceptNet should be smaller, because numerous incorrect concepts are included. The comparisons of data cleaning between original, modified and expanded ConceptNet are shown in table 4.3.

The words can be more precise if being represented by word embedding because the number of multiple segmented (> 1) concepts decreases up to 44%. We can see the significant difference of distribution in Figure 4.1. Though the number of distinct segmented concepts in the expanded ConceptNet is lower than the original one, we still have a much higher concept degree (Figure 4.2) which means more relations between concepts.

We unify low-degree concepts to other ones that is described in section 3.1.2 to increase the number of active concepts (degree > 3) and avoid spending time to explore dead end concepts.

CKB	CKB size	Distinct concept	Distinct segmented words
Original CN	352,411	121,606	1:30,692, 2:55,288, 3:10,720, 4:786
Modified CN	287,582	64,910	1:27,271, 2:33,801, 3:3,670, 4:60
Expanded CN	1,132,030	68,765	1:31,234, 2:33,801, 3:3,670, 4:60
	Active concepts	Average degree	
Original CN	23,174(19.1%)	5.7	
Modified CN	16,723(25.8%)	8.1	
Expanded CN	33,131(48.1%)	32.2	

CN: Chinese ConceptNet
Active concepts: concepts degree > 3

Table 4.3: Data cleaning comparison.

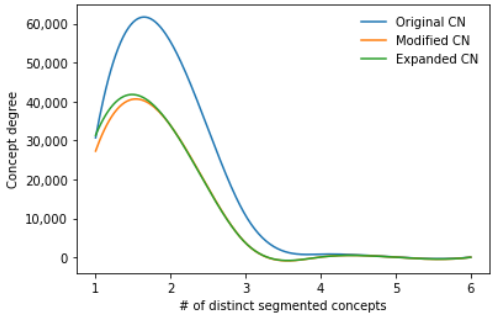

Figure 4.1: Distinct segmented concepts distribution.

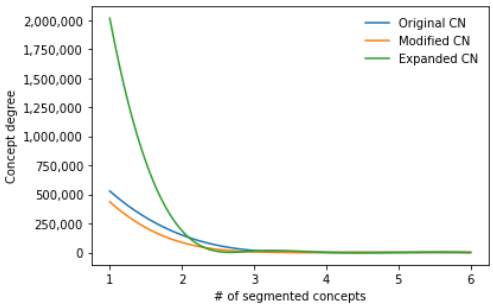

Figure 4.2: Segmented concepts distribution.

Table 4.4 shows the number of modifications, deletions and new data. The total number of modification and deletion is 325,417 which accounting for 92.3% in original ConceptNet.

Add			
Increased degrees	New concepts	Degree of new concepts	
88,927	29,519	66,191	
Modify			Delete
Concept pairs	Relations	SurfaceText	
142,939	31,463	51,959	99,056

Increased degrees: the number of increased degrees of concept if it exists in both original and modified ConceptNet.

Table 4.4: The number of data cleaning.