3.1 ConceptNet Data Cleaning

Although ConceptNet is a large CKB, it contains many types of mistakes, such as typographical errors (typos), redundancy, relations in reverse order or meaningless data, etc. Input data is as important as model when training a neural network based system. Therefore, we refined ConceptNet to reduce its error rate. In the mean time, we increased the quality of ConceptNet. The quality here refers to correctness, coverage and number of concepts. Wider concept coverage isn't necessarily better, it may contain more errors because of crowd-sourcing limitations. We attempt to expand the coverage of each concept in ConceptNet and the number of concepts, and maintain its quality at the same time. We will describe different kinds of errors or downside in ConceptNet in section 3.1.1. In section 3.1.2, we will describe some methods that we adopt to clean data. In section 3.1.3,

we expanded ConceptNet by plesionyms (near-synonyms) in order to increase the coverage of ConceptNet.

3.1.1 Disadvantages of ConceptNet

Limitations of Crowd-sourcing

In order to reduce the cost of acquiring knowledge from experts or collecting manually, crowd-sourcing is an alternative method to acquire knowledge from the non-expert general public. It is more balanced compared to acquiring knowledge from experts or plain text corpus. The quality of translation datasets collected by the crowd-sourcing has been proved to be comparable to the professional translators given appropriate quality control [73]. English ConceptNet originated from OMCS, and Chinese ConceptNet used GWAP system to collect data. Both of them are crowd-sourcing projects. They collected common sense from voluntary web users all around the world.

However, the human crowd-sourcing techniques in collecting big data still exists some limitations. Semantic noise or errors of CKB in neural NLG is a problem which will affect the performance of the result [59]. The knowledge of expert systems is collected from domain experts. It is more specific, narrow and professional in some domains. Expert system KB contains in general fewer errors than the CKB collected from crowd-sourcing. Data Cleaning is required to replace or modify those incomplete, inconsistent and incorrect data in database which may lead to poor performance.

Knowledge acquired by crowds tend to be noisy, redundancy and meaningless especially for unguided projects without supervision and voluntary participants.

The coverage of uncurated crowd-sourcing may be wider than the curated ones, but the side effect is that it contains more errors. In order to increase the quality of answers, concise pre-statements are required. The situations of without supervision are similar to voluntary participants. The quality of data acquired from volunteer crowds may be lower than the paid crowds. This is because they usually answer the questions or fill in the answers based on their intuition. They don't ponder the questions carefully or pay much attention to it, and just want to finish the questions as soon as possible in spite of game-based questions. They usually ignore the pre-statements or instructions about that project. The pre-statements still can't guarantee the high quality of the results. Therefore, post-processing of evaluation and filtering is necessary after collecting the answers.

Missing Commonsense Links

Commonsense knowledge in real world actually is very large and versatile. It's hard to completely cover all. Therefore, lots of missing commonsense links between concepts are possible. For example, if the knowledge that Corgi is an animal not in the CKB, machine won't be able to infer that Corgi can breathe although it does know that animal can breathe. Without carefully carrying out the inheritance inference over the CKB can by itself commit errors. In [74–76], they discussed more about missing links in CKB.

Non-uniform Distribution

Figure 3.2 shows a non-uniform distribution of ConceptNet. The horizontal axis from left to right indicates the first N concept of frequency in descending order.

The red line represents the degree of concepts greater than 3 which may contribute to the paragraph generation (concepts with fewer connections to others may have lower contributions). The total degree of concept in the first 10% (12160) in ConceptNet accounts for 74.3% (517093/695854). The average degree is 5.7, and the first 10% is 42.5. Most of the data is in the range of the first 10%. Non-uniform distribution in ConceptNet will make generated paragraphs monotonous no matter how well the generated system is. Concepts need to connect with other concepts to generate paragraphs. The first concept connects to the second one, and the second one connects to next one, and so forth. If common concepts included in the path to generate, it will move back to the common concepts somehow after few steps. That or those common concepts will appear frequently in generated paragraph, and make it monotonous.

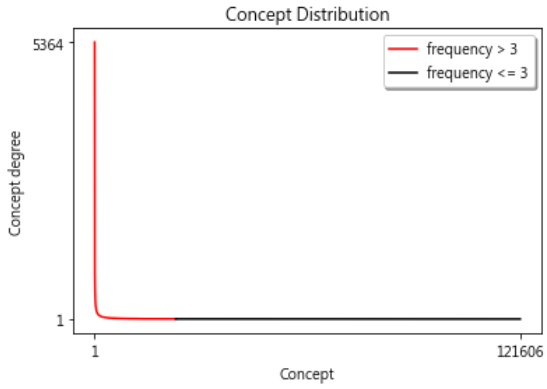

Figure 3.2: ConceptNet distribution.

Ambiguity of Concept

Common concepts tend to have ambiguity. An ambiguity in a sentence is a bit strange. For example, concept "squash player" relate to concept "play squash" and "play ball". If a sentence template is like "squash player ____ when play

squash/play ball". The blank can be "back to the T-position" or "hit the side wall" if related concept is "play squash". And the blank can be "slam dunk", "hit a home run" or "throw the ball" if related concept is "play ball". "play ball" is a more common concept than "play squash", but it contains ambiguities which are not suitable for this template. "play squash" is a rare concept compared to "play ball", but its connected concepts are more related to squash player. Therefore, higher degree of concepts is not always better than lower one.

3.1.2 Data Cleaning

Data cleaning methods (modification and deletion) which are adopted to clean data among all relations in ConceptNet will be listed in this subsection. Not all of the problems listed here are ConceptNet errors, some of them are modified to conform our system needs.

Modification

- Correct the typos by CKIP, a lab called Chinese Knowledge and Information Processing in Academia Sinica, Chinese Spelling Check System[1] [77]. Some typos have the same pronunciation, and some are similar words that usually be misunderstood, e.g., 綿、棉, 記、紀, 戴、帶. 戴 means wear something on body or face and 帶 means bring something. There are still lots of typos after automatic detection. The rest of them are corrected manually.

- Revert assertions which are in reverse order. A simple example is "[可食用] 是 [餅乾] 的"[2] ([edible] is [cookie]). Another example is "你可以在 [珠寶

[1] https://ckip.iis.sinica.edu.tw/service/typo/
[2] text inside square brackets [] refers to concept in ConceptNet

店] 找到 [珠寶]". The concept in Start or End field should be consistent. [珠寶店] (jewelry store) is a location which should be placed in End field of relation "AtLocation". However, the fields in ConceptNet is [珠寶店] AtLocation [珠寶]. The order in SurfaceText is not necessarily the same as the order in ConceptNet fields.

- Change the relations which are unreasonable or inappropriate between concepts.

- Concept in Start or End field in different relations has different parts-of-speech, such as verb in Start field of relation "CapableOf" is incorrect.

- Remove redundant words if they are already in SurfaceText, e.g., [上課的時候] 的時候，你會 [講悄悄話]。. Redundant words include 有點, 覺得, 時候, etc.

- Different SurfaceTexts in a relation have corresponding parts-of-speech and situations. For example, [公園] 的時候會 [慢跑] → 在 [公園] 會 [慢跑] ([jog] when you [the park] → [jog] when you in [the park]). The park is not an action that someone can take but a location. Another example is [鑽石] 的時候會想要 [咬] → [鑽石] 會令人想要 [咬].

- Remove subject like 你 (you) in SurfaceText to avoid specific pronoun. A generated paragraph has its own subject which is user's input. It's wrong to put subject in the SurfaceText again. It will conflict to original subject. Relations with pronoun in SurfaceText include AtLocation, Causes, CausesDesire, HasSubevent, MotivatedByGoal and UsedFor.

Concept Unification

Concept unification is a process of unifying similar concepts together. Unify Chinese variants which are words have the same pronunciation and semantic meaning but with different glyphs. Most of them are allographs (other writing), e.g., 濕、溼, 臺、台, 匯、滙. Unify segmented words in reverse order or have similar meaning, e.g., [很多作業] and [作業很多], [抹乳液] and [塗乳液] → [擦乳液]

Besides those unifications, ConceptNet is very non-uniform distributional. ConceptNet is incomplete due to limited resources and time. The number of concepts is imbalanced between common concepts and the others. The sum of first 10% concepts accounts for 74.3% in entire ConceptNet database. Balanced CKB is helpful when generating paragraphs. Decrease the number of common concepts to avoid paragraphs are always composed of them. It will make generated paragraphs more varied.

Because isolated assertions and low-degree concepts contribute less when generating paragraphs, they need to be decreased. Isolated assertions are totally isolated from the others in the semantic network. They can't be accessed by any other concepts. It would be a waste if they contain useful information. Low-degree concepts refer to concepts with fewer connections to others. Concepts with only one degree account for 65.3% (79398/121606) in ConceptNet. They have lower probability to generate paragraph successfully. Some potential nodes may not be simulated because the model spends time to explore low-degree concept which may be a dead end concept (can't be expanded any more). Most of them can't play an important role in paragraphs because they lack information to describe.

There are two methods to deal with low-degree problem. Remove these low-

degree concepts or increase their degree so that they are not low-degree. ConceptNet is large enough to include most of the daily life vocabularies, and isolated assertions or low-degree concepts are usually extensions or variants of existing concepts. Paraphrase these concepts to their most similar concepts (find similar concepts by average word embeddings) to decrease their number if they have. Increase low-degree concepts manually as more as possible if they don't have similar concepts to paraphrase. Each plesionym may have different connections with different concepts. Merge these plesionyms and share the information they have to increase the size of the group which have similar meanings. Decrease the number of low-degree concepts can construct a more complete concept with different descriptions, and avoid spending time to explore low-degree concepts. There are more choices when that concept is selected to generate paragraphs, and it will make paragraphs more varied.

Concept Abstraction

Concept abstraction is a process of abstracting core semantic of a concept, it is slightly different from concept unification. Remove prefix or suffix modifier is a simple example of concept abstraction, e.g., 很, 非常, 了, 的. Concept with detailed information make it more specific. KB is more rich and complicated if different kinds of individual characteristics concepts included in. However, constrained by the limited resources of KB. The more details of a concept, the less data it would be in ConceptNet. Therefore, we have to abstract specific concepts to its core concept, and removing detailed information to generalize. Concepts with specific number, name or location are unnecessary in KB sometimes, although

they provide more rich and detailed information. For example, [臺北市立民生國小] → [國小]. Taipei Minsheng Elementary School is still an elementary school no matter which city it located in. Another example, [全聯福利中心] → [超級市場]. Store name is also not important, though different supermarkets may have different products. Most of products in supermarket A could be found in supermarket B. Generally speaking, concept with or without those specific characteristics may not alter its core concept meaning.

Use ConceptNet relation "IsA" to find these specific concepts belong to which category. ConceptNet is a semantic network but not hierarchical structure like WordNet or E-HowNet[3] [78]. Select a random concept as a starting node, and find upper layer nodes by relation "IsA". Root node will diverge after few steps. Concept doesn't have an unique corresponding root node in ConceptNet. It won't become a formal hierarchical structure, even if concept has an unique root node. Because ConceptNet is constructed by voluntary web users, they fill anything which seem to reasonable in sentence template. They wouldn't consider overall structure. Figure 3.3 shows an example of root nodes of [學校] (school). Total number of root nodes is 1351. Some of them are related to [學校], but most of them are not. Web users fill concepts they can relate in template "[學校] 是一種 __" . It's less likely to build a hierarchical structure. Therefore, concept which is direct connected to starting node is better than root nodes for the use of abstraction.

[3] http://ehownet.iis.sinica.edu.tw/ehownet.php

[交友地方，好去處，教育單位，設施，學習的地方，知識的地方，變相的監獄，實習，教育場所，教育機構]
[教育設施設施，集體社會，學習地點，運動地點，無底深淵，奮鬥，旅途，滿足的過程，蜜或毒藥，無底洞]
[運勢，冒險，經驗，訓練，奧妙，深淵，巧合，愛的組合，生老病死，放鬆的方法]
[生理需求，釋放，必需，活力的來源，作夢，猶豫，每天做的事，平靜的方式，充電，必要的]
[補充體力，例行公事，靜宜，持續性，銀樓，實力，必須，技巧，災難，進修]

Figure 3.3: Root nodes of [學校].

Decrease the Number of Segmented Words

In order to translate concept to machine-readable representation, word embedding is needed. Because a concept may consist of words, it would be segmented to multiple words. However, average word embeddings of segmented words can't fully represent or even differ from original meaning, especially when the concept contains ambiguous segmented words or concept is a term that can't be separated.

We list two different levels of bias and how we deal with them. The first one is the concept meaning is completely different from the original one if it is separated (segmented).

Ambiguity

A word with fewer characters tend to have ambiguity, especially when a word is a single character. It may be an abbreviation of other words.

Take [賞 巴掌] (slap someone's face) as an example (blank between two words means they are segmented). "賞" means rewarding somebody for something, appreciating or admiring something. "賞" in [賞 巴掌] means giving something to someone. Word with ambiguities is hard to tell which one will be used in word embedding. Paraphrase [賞 巴掌] to its similar concept [打耳光] to disambiguate.

Metaphor

Some concepts can't represent their meaning if they are separated. They

imply a particular meaning that we can hardly know from each word's denotation. E.g., [腦袋 開花] → [腦袋 掛彩]. [腦袋 開花] doesn't mean flowers really blossom in someone's head. It means someone's head is injured. Idiom is similar to metaphor but with group of words having fixed combination and order, e.g., [輸到 脫褲子] → [輸光] (lose one's shirt at the track, English translation is not accurate.). Machine can't infer from [輸到] and [脫褲子] to [輸光] by averaging their word embeddings. A concept should be paraphrased to similar concepts if it is metaphor.

The second one is the concept meaning is biased from original semantic if it is separated though each word doesn't have ambiguity.

Segmented words combinations

Combine adjacent segmented words and rematch to vocabulary list to decrease the number of segmented words. For example, [聽 廣播] → [聽廣播] → [收聽廣播] (listen to the radio).

Remove prefix or suffix character of segmented word and combine with others to check whether it's in vocabulary list or not. For example, [蓋上 被子] → [蓋上被子](x) → [蓋上子](x) → [蓋被子] (tuck oneself in).

The meaning of combined segmented words is different from the original concept, even if the concept isn't metaphor and all of the segmented words don't have ambiguities. Concept semantic meaning is determined by its context. Combined segmented words and original concept usually don't have the same context, because they are separated when associating to context. Therefore, they don't have the same meaning as original concept.

For example, a segmented concept [返回 家鄉] (return to one's hometown) can be paraphrased to [返鄉]. Related concepts of [返回] (return) are [回到] (back), [離開] (leave), [前往] (go to), [抵達] (arrive). All of related concepts are about going to or back to somewhere. Related concepts of [家鄉] (a place that family has been living there for generations) are [故鄉] (someone's birthplace but no longer resident in there), [老家] (used to live in there), [南部](southern). Some of them are plesionyms of [家鄉] and some of them are about living in somewhere. All related concepts of [返回 家鄉] consist of related concepts of [返回] and [家鄉]. And related concepts of [返鄉] are [回老家] (return to the place someone used to live in), [探親] (visit relatives), [春節] (lunar New Year), [連假] (long weekend). It's obvious to know that combined of segmented words is different from the original concept.

Segmented words paraphrasing

The concept with more connections to others will make generated paragraph more varied. Paraphrase segmented words to short and common concept which are in vocabulary list to increase the connections to other concepts, e.g., [非常 好吃 的 食物] → [美食]. The less the segmented words, the more precise the word embedding can represent that concept.

In general, ConceptNet network after paraphrasing becomes more closely within a cluster than the original one, and word embeddings are also more precise.

Deletion

- Out-of-vocabulary (OOV), duplicate, some simplified Chinese concepts, offensive concepts, unreasonable and meaningless data, such as "[女人] 是一種 [錢包] ([woman] is a [purse])" or "[帥哥] 是為了 [生活]" ([handsome_guy] in order to [life]).
- Some data in ConceptNet are not precise, but not wrong technically. It could be rare conditions happen in special cases. However, these data are unhelpful when generating paragraphs. For example, it's unhelpful to know a table below universe, toilet paper is part of world or eating while jogging.

To summary this subsection, there are different levels of importance of modifications to make ConceptNet more suitable to generate paragraphs.

Number of connections between concepts The most important task is to generate paragraphs successfully, and it will be failed if there is not enough data to do it. The more connections between concepts, the more likely the paragraphs can be generated. If there is a region with concepts mutually connected in ConceptNet network, it tends to have similar topics or properties. A paragraph consists of similar topics would be more coherent.

Correctness of assertions and SurfaceText Incorrect concepts in a paragraph are obviously wrong and unreasonable. Different SurfaceTexts have different corresponding parts-of-speech.

Number of segmented words The semantic meaning of average word embeddings differs from original semantic when the concept is segmented. The

meaning will be severely biased when the number of segmented words increased.

Concept degree Provide more choices to generate more varied paragraphs.

Number of distinct concepts More different concepts can cover more different regions (topics) of ConceptNet network.

CKB size It's intuitive to know the quality of CKBs when they have considerable disparity in the size of database. Small CKB doesn't have enough data to generate paragraphs, even if the large one has much more errors than small one. It's hard to know which CKB is better when the size are closed.

Besides these conditions, the generated paragraphs need to be examined by human sometimes. Generally, size of these conditions are directly proportional to paragraph coherence and variation except for the number of distinct segmented words. The process of data cleaning in this subsection are among all relations, but each relation has different methods. The detailed information see appendix B.

3.1.3 ConceptNet Expansion

In order to make generated paragraphs more varied, we expand ConceptNet by different relations. User can also access data without precise concepts.

- Add new data to Start field of relation "CapableOf", "Desires", "HasA" and "NotDesires" manually. Increase the number of initial concepts, which is the root concept to simulate the MCTS, to avoid mismatching with user's input. Initial concept in those relations are usually a subject that can do actions.
- The paragraph will lack variation when the concepts are common and have

relatively less data in some relations compared to others. For example, the proportion of [考生] (examination candidates) in relation "NotDesires" and "HasProperty" is 126:1. There may be a template with "NotDesires" and "HasProperty" in the same time. If [考生] is the search node, a node connects with the other two nodes in a compound relation, of them. The generated paragraph may have different concepts in "NotDesires". However, no matter how many concepts in "NotDesires", there is always one fixed concept in "HasProperty". The bottleneck is controlled by the relation which has less data, even if the others have much more. Add new data to these concepts to balance the numbers in each relation.

Expanded by Antonyms

HasProperty

If A has property B, it may have property of antonym of B.

"[產品] 是 [昂貴] 的" → "[產品] 是 [便宜] 的".

[product] is [expensive] → [product] is [cheap].

Desires & NotDesires

If A doesn't desire B, A may desire antonym of B.

"[總統] 厭惡 [敗選]" → "[總統] 想要 [勝選]".

([president] doesn't want to [lose the election] → [president] want to [win the election])

If A desires B, A may not desire antonym of B.

"[女人] 想要 [減肥]" → "[女人] 懼怕 [發胖]".

([woman] want to [lose weight] → [woman] hate [gain weight])

Some of pairs are inappropriate.

“[人民] 懼怕 [旱災]” → “[人民] 喜歡 [水災]”.

([people] afraid [drought] → [people] like [flood])

Expanded by Synonyms

Synonyms are words which sound different but have the same meaning. In fact, most synonyms are plesionyms (near-synonyms) [79] rather than absolute synonyms which have exactly the same meaning. Absolute synonyms are fully inter-substitutable. They can be substituted for each other in any context situations without changing the original semantic meaning. They have the same denotation, connotation, implication, usage and POS. However, the pairs of absolute synonyms are rare. Therefore, we use plesionyms which are very close to absolute synonyms and mutually substitutable instead to expand ConceptNet.

Plesionyms usually don’t have the same meaning except denotation. Denotation and connotation are different aspects of a word. Denotation is a word’s explicit definition in dictionary. For example, denotation of table in cambridge dictionary is “a flat surface, usually supported by four legs, used for putting things on”. A word may have its connotative meaning not just denotation. Connotation of a word is its implicit meaning associate to emotion (positive, neutral or negative), past experience, environment or culture. A word can have different connotative meanings depending on different contextual situations.

Plesionyms have some central semantic features, i.e. semantic traits in [79], overlapping but differ in one or more peripheral features. For example, [苦痛, 苦

楚, 痛楚, 酸楚, 苦處, 苦水],[痛苦],[痛處],[苦頭, 苦難]. Although these plesionyms seem to be similar, they are not substitutable. Words in a sentence are closely related. Substitute words with plesionyms in a sentence or a paragraph will usually change its original meaning and style. The word would stick out like a sore thumb in a sentence if there exists slightly different in connotation. In order to find precise plesionyms that are very close to absolute synonyms, some conditions need to be excluded.

Ambiguity We exclude single character words and words which the number of categories in Cilin [4] more than one to disambiguate. We delete data of Cilin from 63,213 to 47,611. The goal is to increase the number of single category words as more as possible. Some of the ambiguities have relatively low frequency which can be ignored. If plesionyms consist of several groups (plesionyms in the same group are more similar than the others), we retain the group which is larger than the others.

Connotation

Formal/Informal Words in specific situations would be different. For example, [告訴] (tell) and [告知] (inform) both refer to say something to someone. [告訴] is informal and direct, and [告知] is more formal and usually used in business.

Archaic word Words are used in ancient times and still be used in book or movie, e.g., [爸爸] and [阿爹]. People barely used archaic words in everyday life. The meaning of some archaic words have been changed,

[4] HIT IR-Lab Tongyici Cilin (Extended), 哈工大信息检索研究室同义词词林扩展版

e.g., [大人] means father in ancient times, and now it means elders. Both of these two situations need to be excluded.

Emphasis Different plesionyms may have different emphases.

For example, [孤掌難鳴] and [無計可施]

[孤掌難鳴] emphasizes someone can do nothing by himself/herself.

[無計可施] emphasizes someone can't find any solution to solve the problem

Emotion [稚氣] (childlike) and [幼稚] (childish) both refer to an adult behave like a child. [稚氣] is a positive emotion to describe an adult is innocent, energetic, honest or curious like a child. [幼稚] carries a negative connotation which describes an adult behave badly like a child, such as stubborn, moody or immature. Emotion strength like [生氣](mad), [火冒三丈] (foam at the mouth) can be substituted though they are different. Different emotion strengths are acceptable.

Word frequency The word frequency < 130 are excluded.

Multiple segmented words The new concept replaced by plesionyms may not represent the original semantic meaning. Segmented words are closely related, e.g., [隱藏缺點] → [躲藏缺點]. Although [隱藏] and [躲藏] have similar meanings, they are not mutually substitutable.

Part-of-speech Exclude plesionyms which have different parts-of-speech. They can't be substituted for each other.

Besides excluding plesionyms from different sources, we also add new data manually. Find top 15 similar concepts by word embedding for each concept. Add

these data if they are in ConceptNet (avoid OOV) and not in synonym set, which consists of different sources. The total number of plesionyms is 16843, and 7774 match the concepts in ConceptNet will be used to expand.