

3.2 Word Embedding

Training Corpora

We use 109 popular boards in PTT⁵ and Chinese Wiki as our corpora.

| Corpora | Words | Vocabulary size (frequency > 29) | Capacity |
|----------|-------------|----------------------------------|----------|
| PTT,Wiki | 858,118,783 | 249571 | 5.38 GB |

Table 3.1: Word embedding training corpora.

Corpus Preprocessing

Remove tags, punctuations, numbers, non-Chinese, invalid and duplicate characters. There are some fixed format in different boards, such as opening hours of stores, the number of products or the reports written by which reporter. These words should be removed because they are not related to content. Because the fixed word length of one line (about 40 words), a sentence may be separated to multiple lines. Recombine these separated words back to original words if they are in vocabulary table.

Stop words removal Use stop words to remove some of content words and function words which have little semantic meaning. Content words include numerals and quantifiers. Numerals and quantifiers don't affect the main se-

⁵ Taiwanese Bulletin Board System(BBS) which consists of 13,309 boards in plain text.

semantic meaning of the concepts even though they do have meanings. Although the number of objects has slightly different in context, they are still the same thing. It doesn't change any of its physical properties. For example, someone can drive a car, but can't drive two cars simultaneously. The main semantic meaning of a car or two cars are both cars. The number of cars doesn't change any property of it. It still has four tires, headlights, car doors, engines, and so on.

Some adverbs like adverbs of degree or time can be filtered out. Adverbs of degree are used to strengthen the meaning of adverbs or adjectives. Adverbs of time are used to describe when or how often an action happened. A concept with or without these words are basically the same. For example, this dress looks elegant and this dress looks very elegant. I buy a cake today and I bought a cake yesterday. It's not important how intensity an action, an adverb or an adjective. It's also not important when did you buy something. The point is how concepts interact to each other. Remove these types of adverbs may lose some information within sentence or paragraph, but the relations between concepts are still retained.

Some adverbs like adverbs of frequency can't be filtered out. For example, kids are always late for school and kids are seldom late for school. It's totally different how often kids late for school. It will change concept semantic meaning if they are removed.

Function words include conjunctions, particles, prepositions, interjection and onomatopoeia. These function words have little concrete semantic

meaning and for grammar purposes such as 和、的、從、啊、砰. They are used to make sentence grammatically correct and relate to other words in grammatical relation. They don't change the relations between concepts. For example, the students study in a classroom. It doesn't matter which specific classroom they are studying in. The relation between students and classroom is still "AtLocation" when function word "a" is filtered out.

Function words may decrease the performance of word embeddings. If there are many function words in a sentence, it would become fragment after segmenting without removing them. The strength of relation between content words without removing function words may lower than the sentence segmented after removing. It's possible that most of the content words relate to function words because of their high frequency in paragraph. It's intuitive to know that collocated words have stronger relation than others. Function words may relate to each content word, and this make them meaningless in word embeddings. It's important to select stop words. Have to consider each situation whether stop words have semantic meaning or not.

Text segmentation We use Jieba⁶ API to segment words. Although it can find unknown words, which are not in self-defined dictionary, by a HMM-based (Hidden Markov Model) Viterbi algorithm, it also finds lots of unreasonable words, e.g., 裡你寫, 講你領, 將你傷. The distinct words contain "你" are found 9598 times in the result of discovering new words, and the other one are found 126 times. Therefore, we don't use HMM to combine unknown words. Exclude word frequency < 30 to reduce the impact of unreasonable

⁶ <https://github.com/fxsjy/jieba>

words and rare words. We recombine segmented words back to their original word, e.g., [喝了酒] \rightarrow [喝] [酒] \rightarrow [喝酒]. “了” is a stop words [喝] and [酒] are separated because of removing a stop word “了”. Recombine these separated words so that it can represent its original meaning.