MLDS 2017 Spring HW1 - Language Model

B03901056 孫凡耕 B03901070 羅啓心
B03901032 郭子生 B03901003 許晉嘉

1 Environment

1. 硬體資訊：

OS	CPU	GPU	Memory
`Arch linux 4.10`	`i7 3.4 GHz`	`GTX 1070`	`32 GB`

2. 所使用的 `python` library:

`spacy 1.6.0`	`nltk 3.2.2`

2 Model description

3 Improvement

1. 僅使用 Training Data 中，長度在特定範圍之內的句子
特短的句子，大多屬於雜訊或是較無意義的短句（e.g. somebody said,）；特長的句子，也均爲雜訊居多，若納入特長的句子亦會使 training model 維度過高，造成記憶體空間不足。且特短及特長的句子，本身在 Training Data 中所佔的比例也不高（如 Figure 1.）。

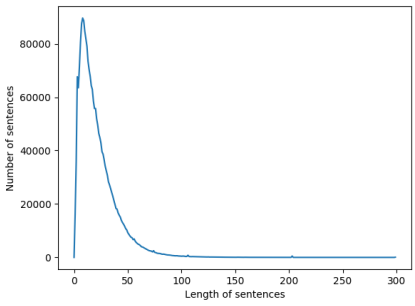

Figure 1: 句子長度分佈

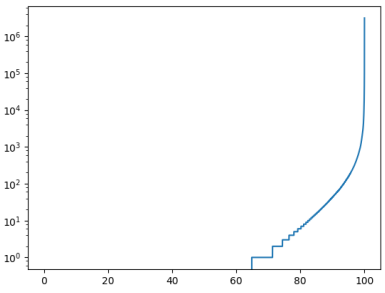

Figure 2: 單字出現次數分佈

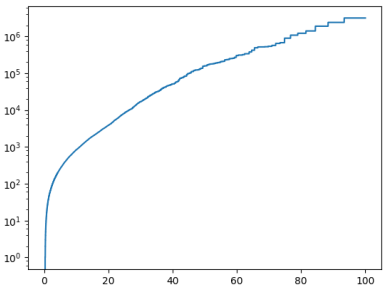

Figure 3: 單字出現次數比例分佈

2. 將出現次數過少的單字從 corpus 中移除
原本 corpus 的字量過大，會造成記憶體空間不足以及訓練困難，因此，將在 Training Data 中出現次數較少的單字移除（視爲 unknown word），可使訓練學習的過程加速。從上圖 Figure 2 可看出七成左右的單字不在 pretrained corpus 內或是出現次數少於兩次，而從 Figure 3 可看出這些單字又僅佔不到總單字量的一個百分點，因此，將其刪除對於訓練的過程有較大的助益。

3. 將 Training Data 中，開頭及結尾的部分刪去
由於 Training Data 中幾乎所有文章，開頭及結尾皆相當於目錄或版權資訊等，較不爲一般常用語的句子。有利訓練過程的進行。

4. Dependency Tree
Dependency tree 可表示句子當中各單字之間的關聯。因此，將資料轉爲 dependency tree 後，可更爲有效的分出每個句子中各個合法的語句，使訓練的資料更爲廣泛且一般。

4 Experiment

5 Team division

孫凡耕	
羅啓心	
郭子生	
許晉嘉	資料處理、撰寫報告