Final Project Report -- TV

隊名：NTU_r06725028_rectangle
隊員：資工碩一 r06922008 李彥霆　電子碩一 r06943122 吳中群
　　資管碩一　r06725028 黃于真　資管碩一 r06725053 韓宏光

1. Introduction & Motivation : (1%)

　　這個題目主要的任務是依據給定的一段對話，可能是一個人的自言自語，也可能是多個人之間的交談，從六個選項中選出最可能在這段對話後出現的句子為何。自然語言處理一直都是個很有趣但也很有挑戰的議題，而這個任務可說是結合了自然語言處理/深度學習的許多面向，首先我們要處理的是相當口語化的短句，且說話對象還可能會轉變，增加分析上的難度，此外，還需要做word embedding，以及訓練RNN模型，。

　　我們認為當模型可以正確預測出下一句對白時，表示模型確實理解到這段對話的文意，這是令人驚豔的結果，因此我們覺得這是很值得嘗試的任務，而且資料集是來自電視節目中的對白，句子內容都相當生活化，甚至有流行用語，也增添不少趣味。

2. Data Preprocessing/Feature Engineering : (2%)

　　不同的模型我們有做不同的前處理，以下列出我們有做的前處理，至於各模型採用的方法，會在模型描述時再說明。

- 對話切分

　　由於訓練資料集中全部是連續的句子，因此我們需要自己切分出對話以及接續的下一句對白來訓練，由於testing data中對話最多有三句，所以我們將訓練資料集中每三句作為對話，接續的下一句則為該對話的回答。

- 切詞\切字、標點符號處理

　　常用的jieba中文斷詞較適用簡體字，所以我們使用它的繁體擴充版本Jseg來做切詞，此外我們也嘗試不切詞，而是切字，即每一個字都單獨作為一個詞，但遇到非中文，如英文單字時，

則不切開每個字母，仍維持單詞形式，如此可以減少token數目，另外也嘗試移除標點符號後再去訓練Word embedding，最後依據單純使用相似度的模型結果，發現切字且包含標點符號的效果較好。

- Word embedding、sentence embedding

  我們使用word2vec來訓練Word embedding，也嘗試使用兩種訓練方法hierarchical softmax和negative sampling來搭配兩種模型CBOW、Skip-Gram訓練，並且不濾掉低頻詞，即出現的每一個詞皆採用，最後依據單純使用相似度的模型結果，發現使用negative sampling搭配Skip-Gram的訓練效果較好。另外，sentence embedding部分，我們採取的方法是把句子中每個詞的word embedding做average後，直接作為sentence embedding使用。

- Padding for RNN

  使用RNN的模型，其input、output須padding成等長的序列，由於對話中最多有71個字，而回答中最多有44個字，因此我們會把兩者不足長度的部分皆補上0，再進行訓練。

- 增加負樣本

  為了讓RNN模型學會預測在一段對話後，出現某句回答的機率，除了使用訓練資料集切分出的對話和回答作為正樣本外，我們還需要自己增加負樣本，方式是隨機選取五個電視劇資料集的其中一句，計算該句和正確答案的jaccard similarity，如果小於0.5即採用，最後得到總共1200萬筆的training data。

3. Model Description (At least two different models) : (4%)

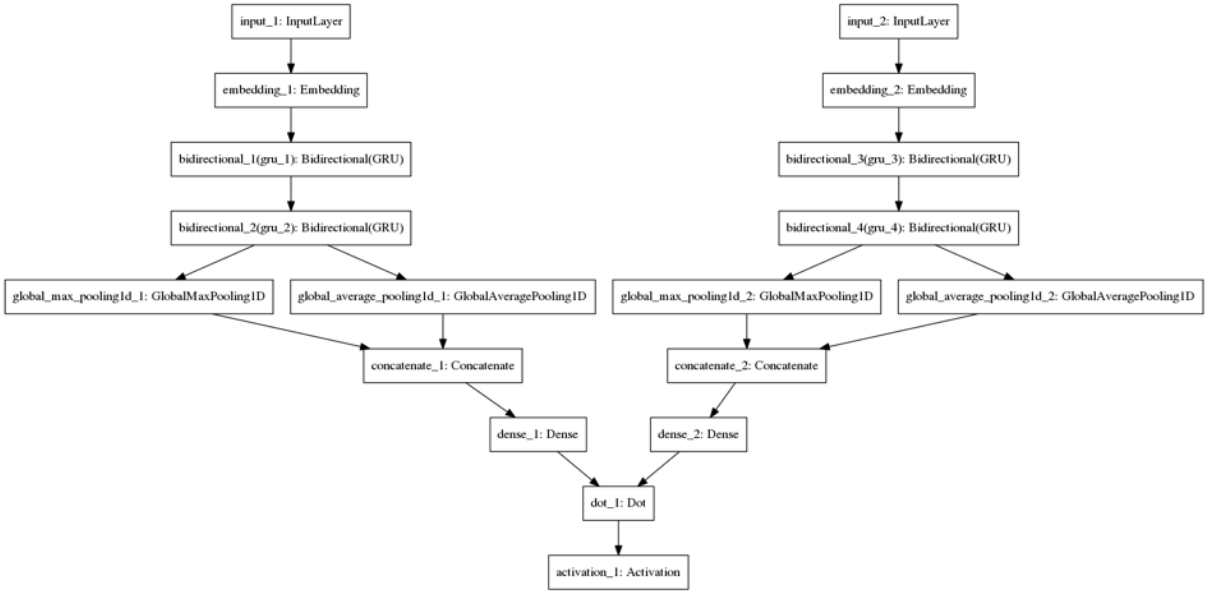
- only cosine similarity

  這是我們通過simple baseline的模型，在訓練部分，使用原本訓練資料集進行切字處理後，再用negative sampling搭配Skip-Gram的方式訓練word embedding，維度使用預設的100

維。而在預測部分，用average方式來計算testing data對話和選項各自的sentence embedding，接著計算對話和每個選項的cosine similarity，並作長度正規化，最高者即作為答案。

- RNN model

  這是我們通過strong baseline的模型，訓練部分使用training data切分出對話後，對話和答案皆需padding為等長，接著增加負樣本，再使用下圖中的模型訓練，也就是對話和答案都過相同的層後，再dot起來，最後透過sigmoid轉為機率形式。而預測部分，testing data的對話和選項同樣須先padding為等長，接著使用模型預測在此對話後出現該選項的機率，再選最高者作為答案。



4. Experiment and Discussion : (6%)

   - only cosine similarity

     在單純使用相似度的模型部分，我們比較了不同訓練word2vec的方法，首先是cbow和skip-gram這兩種模型，如下表格是使用切字前處理的結果，可以看到skip-gram的表現較好，我們認為這是因為這次的訓練資料集其實不算太大，這也是我們把min_count設為1的原因，希望涵括更多詞彙，但也

使得其中包含不少低頻詞，所以這樣的結果可能是由於skip-gram中每個詞作為中心詞時，都會依據周圍的詞做一次gradient desent，即每個詞都會進行K次的預測、調整，訓練次數會多於cbow，而當資料量較少、低頻詞多時， 這種多次調整使得詞向量相對的更加準確，讓模型的預測準確率也跟著上升。

	cbow	skip-grem
Public Score	0.36561	0.37747
Private Score	0.36758	0.38537

如下表格則是在切詞前處理並使用skip-gram下結果，可以發現有加入標點符號的表現就好，我們認為這可能是因為電視劇資料集中包含了許多口語，其中心情的表達多半會用標點符號來表示，像是驚嘆號！，還有...之類的，其實對於模型理解語義有一定的幫助，同時也能增加詞彙量。

	有標點符號	無標點符號
Public Score	0.36916	0.36166
Private Score	0.36877	0.36561

如下表格則是在使用skip-gram且包含標點符號下的結果，可以發現切字的效果比較好，我想是因為切字可以大幅減少字彙量，也就可以讓低頻詞降低，使得OOV的情況獲得改善，即使是沒出現過的多字組成的詞語，只要其中的字曾經出現，仍然可以表示，但切詞的模型就無法。

	切詞	切字
Public Score	0.36916	0.37747
Private Score	0.36877	0.38537

- 兩個模型比較

	cosine similarity	RNN
Public Score	0.37747	0.46324
Private Score	0.38537	0.47154

在模型間比較的部分，單純使用相似度的模型有一定的限制，在我們的嘗試下，大約到三成七左右就上不去了，而使用RNN則可以到接近五成，如果做ensembling還可以更高，實際查看預測出的結果有何差異，我們發現只使用相似度的模型偏好選擇字數較多的選項為答案，我們認為這可能和sentence embedding使用詞向量平均的方式有關，句子中越多字，各維度就有較多機會都有一定的值，在計算相似度上可能就有一些優勢，而RNN模型就可以避免這樣的情況，透過句子中每字的順序來學會口語化的短句用法，並且預測的是每個選項出現在該對話下一句的機率，這樣的訓練方式也較貼近我們最後要用來預測的方式，並且充分利用到訓練資料集包含的資訊，不像相似度模型只用訓練資料集來訓練詞向量。以下就是我們找出的例子，在這個對話下，相似度模型預測出的答案是5，但RNN模型就能預測出2，短短的「怎麼」兩字，但其實很符合文意，在我們看來應該是最適合這題的答案。

沒關係 闖嘴三郎應該有印象吧 警官，我想大家都誤會了

0:我知道今天易宏要結婚 1:我們叫計程車就回去了 2:怎麼 3:好，謝謝

4:老同學了，你說什麼 5:沒能力幫他處理公事就算了

- Ensembling結果

我們也有嘗試將多個RNN模型做ensembling，我們採用的做法是把每個模型預測出為正類的機率做累加，最後同樣是選擇機率最高的選項做為答案，發現預測準確率會有所提升，如下

表格結果。

	六個模型	單一模型
Public Score	0.50079	0.46324
Private Score	0.49920	0.47154

5. Conclusion : (1%)

這次的final project中，我們嘗試了不同方法，學習如何對於中文資料做前處理，原本的習慣都是做切詞，但這次發現切字會是比較好的做法，因為要考量到OOV的情況。此外，這次作業比較特別的一個地方是訓練資料和測試資料的格式並不一樣，因此還需要自己決定如何把訓練資料處理成和測試資料相同的格式，我們選擇的是隨機選取負樣本方法。另外一個特別的地方是，對話和選項各自都具有上下文這樣的順序關係，因此在訓練時是對話和選項都需要過RNN相關的模型訓練，並且使用pooling，即類似CNN的概念來做特徵的選取，最後再dot起來，因此這次作業的模型相對有些複雜。

6. Reference : (1%)

https://www.jianshu.com/p/c294e4cb4070
http://www.aclweb.org/anthology/C16-1063
https://github.com/amigcamel/Jseg
https://zhuanlan.zhihu.com/p/37477611