應用序列到序列生成模型於情感型聊天機器人之研究

**摘要**

受益於深度學習演算法的發展，商業聊天機器人中應用深度神經網路的方法變得愈發普及。在2018舉辦的NTCIR-14，CECG子任務內，傳統生成式聊天機器人中加入了情感分類的元素，旨在針對同一問句，為不同情感標籤生成不同的回應。

受此啟發，本文主要研究的生成式聊天機器人，透過深度學習的架構，使用seq2seq模型進行訓練，并以情感標籤為之分類。

最後，本文將透過便利抽樣的方式對30位使用簡體中文的微博用戶進行調查，以此來評測聊天機器人在加入情感標籤條件下與普通生成模型的表現差異。

\*研究發現

關鍵詞：聊天機器人、自然語言生成、深度學習、情感分析

2019年第三十屆國際資訊管理學術研討會論文本文

1. **導論**
2. 研究背景

近年來微博微信等新興SNS（Social Networking Services）飛速崛起，每天都有數以億計的對話在網路上傳播。在這些對話中又有著極大一部分來自於聊天機器人。

從回應答句的角度來看，可以將聊天機器人大致分成兩類：

（一）檢索式聊天機器人：回答是預設的，其中會使用到規則引擎、正則表達匹配或是使用深度學習預先訓練好的模型，从已有的知識庫中檢索一句最優答句進行匹配。

（二）生成式聊天機器人：不依賴也不使用已存在的回答句，可以自動產生新的答句。這種模型在訓練的過程中要求較大量的時間和語料，而語料通常包含post和response兩個部分。時下生成式聊天機器人大多會使用LSTM和RNN來訓練生成模型。這種方法早先在機器翻譯的領域中取得過非常好的成效。

目前而言，聊天機器人的商業化程度已有相當的規模，但受制于傳統聊天機器人大多是檢索式聊天機器人，預設的回答漸漸無法滿足企業與客戶溝通的需求。

1. 研究動機

美國雲端通訊公司Twilio在2016年的報告中指出，有接近89%的用戶更希望透過訊息（messaging）的方式跟企業或商家直接溝通，而不是電話或者郵件。但是只有48%的企業去實現了這種溝通的可能。

現在，生成式聊天機器人也許能為企業和客戶溝通建立一座新的橋樑。前文提到的傳統聊天機器人（檢索式聊天機器人）有一些缺陷，而生成式聊天機器人所倚賴的巨量資料和運算時間，在當今科技飛速的發展下也已然不是很大的問題。在Tensorflow、Keras、Pytorch等開源深度學習框架，以及強悍的GPU設備支持下，使用深度學習訓練生成式模型的時間，已經被壓縮在一個完全可以接受的範圍內。此外，在這個資訊爆炸的年代，想要在網路上收集到大量資料用於此種研究，也不再是一件難事了。於是如何才能夠提升生成答句的質量，成為了生成式聊天機器人研究的重點。

在NTCIR-14，CECG子任務中釋出的資料集里，除了問答對之外，還包含了句子（post & response）本身的客觀情感標籤，共分成6類（Other,Like,Sadness,Disgust,Anger，Happiness）。這個資料集啟發我，將情感標籤也加入生成模型的訓練，透過類神經網路，模型將會學習到類似自然人類高情商（Emotional Quotient）的概念，從而生成更合適的自然語言表達。

1. 研究目的

如上文所述，本研究將在序列到序列的生成中，加入情感標籤為參考要素，使用Pytorch的開源架構來實作生成式聊天機器人，以實現聊天機器人在大部分情況下，能夠透過問句及其情感，來生成更相關且更合適的回應。

為了讓實驗結果更有可信度，本研究會將部分的測試問答句，通過簡單抽樣的方式設計成問卷，交由簡體中文微博用戶，從語言流暢度，問答相關度和情感表達度三個方面進行評測，以全方位展示實驗的結果。

1. 論文架構

本論文共分五個章節，前述的部分為導論，闡述本研究的背景，動機及目的。其後第二章是文獻探討，將討論與本研究相關的各個技術，其在發展過程中產生的變化與影響。第三章將會對本研究的研究方法加以介紹。第四章將會說明本研究的實驗環境，實驗設計和結果。最後會在第五章總結本研究的實際成果與研究限制，並提出未來可能得以延伸的研究方向。

1. **文獻探討**

2.1 NTCIR-14

NTCIR是一個針對資訊架構（IA）技術的的評估比賽，其中包含資訊檢索，問題回答，文本摘要和資訊萃取等。最早由日本學術振興會（JSPS）和國立情報學研究所（NACSIS）聯合贊助，並於1998年開始籌備，並最終在1999年成功舉辦首屆workshop。經過二十年的發展NTCIR已然成為一項國際重要賽事，它設置了一系列基於中文，日文，英文三種語言的評估任務，目前舉辦至第十四屆。

在 2016 年，第十二屆 NTCIR 中首次設置了基於中文、日文兩種語料的評測新任務，即短文本對話任務（STC），這也是目前國際上唯一的“開放域”對話方面的評測比賽。

在今年的比賽中，依然有中文評測任務，而其中的短文本對話任務NTCIR14-STC3，在上一屆引入基於深度學習的生成模型對話的任務設置之後，又結合了情感標籤的元素，提出基於情感標籤的對話生成任務CECG。

我們團隊有幸參加了此次NTCIR 14-STC3的CECG子任務，也正是本次的比賽啟發我使用情感標籤分類，作為提升句子生成質量的一種管道。

2.2檢索式聊天機器人

2.2.1樣板式

樣板式聊天機器人的技術原理是經過人工在聊天庫中設定一些對話場景，然後根據不同場景塑造相對應的對話範本，並根據每一個問題設計其可能會出現的答案。基於這個技術的優點是精確性高，缺點是人工工作量大，可擴展性差，不同的場景要有不同的設定。大名鼎鼎的Siri對話機器人就是使用這樣的技術生成的，相比於其他技術的對話機器人其精確性非常高。

2.2.2規則式

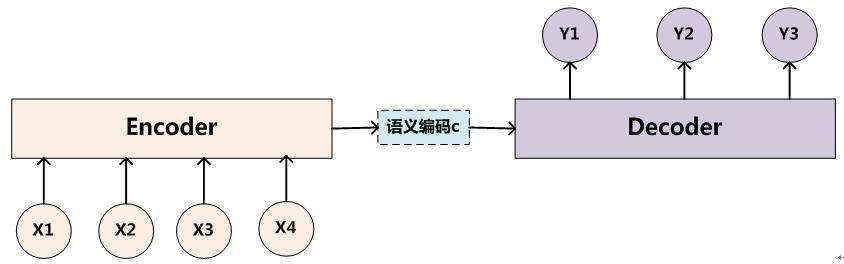
規則式聊天機器人的回應句也是提前設定好的。它的技術類似於搜索引擎，在製作一個這樣的聊天系統之前，需要創建一個聊天對話庫並建立索引，根據輸入的問題，在聊天對話庫中進行搜索查詢來進行模糊匹配，並根據我們預先設定的規則計算相對匹配度最高的問句，取得其索引值，將索引值相對應的答句返回給用戶。要提升這一類聊天機器人，很大程度上倚賴語料庫的擴增和搜尋匹配演算法的提升。

2.3生成式聊天機器人

近年在NLP領域中出現了很多新的模型，而在生成式聊天機器人中大部分都是在Seq2Seq模型框架下進行改進的。這種框架的好處主要有兩點，一是在未來的擴展上有很大空間，二是生成模型的泛用性也可根據訓練語料來調整。以下主要介紹本研究中所使用的生成式聊天機器人相關的幾個技術。在詞表示方面有word embedding，深度學習模型方面有RNN，GRU，Attention-Based-GRU模型等等。

2.4 sequence-to-sequence模型

seq2seq 可以簡單看做是兩個RNN組成的模型，一個Encoder–Decoder 結構的網路，它的輸入是一個序列，輸出也是一個序列， Encoder 中將一個可變長度的信號序列變為固定長度的向量表達，Decoder 將這個固定長度的向量變成可變長度的目標的信號序列。在實際應用中，seq2seq活躍在機器翻譯，對話生成，文本摘要等領域。因為要求輸出要與輸入相關，所以模型本身要能夠記憶輸入序列的訊息。下面將介紹seq2seq模型的運作原理。



2.4.1 深度學習

深度學習是經過構建大量訓練數據和大量隱藏層，學習更有用特徵，以達到提高分類或預測準確性的目的的機器學習模型。與傳統的淺層學習不同的是，深度學習模型結構的深度，通常有5層以上。此外，特徵學習的重要性被明確的強調，即經過一層層的特徵變換將原始空間中的特徵表示轉換到新的特徵空間，使得模型分類與預測更加的簡單。

學習特徵的方法有兩種：基於人工規則和使用大數據。前者適用於提升一些規則已知的學習，後者則可以更好地體現出數據本身所要表達的資訊。深度學習是一種端到端的結構，即可以從數據的訓練直接到最終生成結果，不需要額外的特徵提取的環節，也可以說所有的特徵都包含在隱藏層中。

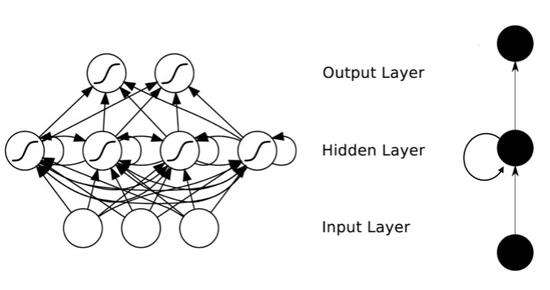
深度學習有許多經典的模型，像是深信网络模型（DBN），限制玻尔兹曼机(RBMs)，卷積神經網路（CNN），循環神經網路（RNN）等。

2.4.2 RNN（循環神經網路）

RNN是為了解決DNN（Deep neural network前饋人工神經網路）存在著無法對時間序列上的變化進行建模的問題（如自然語言處理、語音識別、手寫體識別），出現的另一種神經網路結構，迴圈神經網路。

RNN之所以稱為循環神經網路，即一個序列當前的輸出與前面的輸出也有關。具體的表現形式為網路會對前面的資訊進行記憶並應用於當前輸出的計算中，即隱藏層之間的節點不再無連接而是有連接的，並且隱藏層的輸入不僅包括輸入層的輸出，還包括上一時刻隱藏層的輸出。

一般來說，RNN可以處理任意長度的序列。



2.4.3 LSTM和GRU單元

由於梯度消失問題，S-RNN（Simple RNN） 很難高效訓練。在反向傳播（backpropagation）過程中，前面環節的誤差信號/梯度小時的很快，無法到達更前面的輸入信號，使得S-RNN 難以捕獲廣範圍的依賴。Long short-term memory (LSTM) 結構是 Hochreiter and Schmidhuber 2017 設計來解決梯度消失問題的，首創地引入了門限機制。

S-RNN 結構中每層都複用相同的矩陣 W，這將導致梯度計算包含多次的矩陣 W 的乘積運算，這很容易導致梯度值消失或者爆炸，門限機制解決這個問題的主要方法就是避免單個矩陣的重複乘積計算。

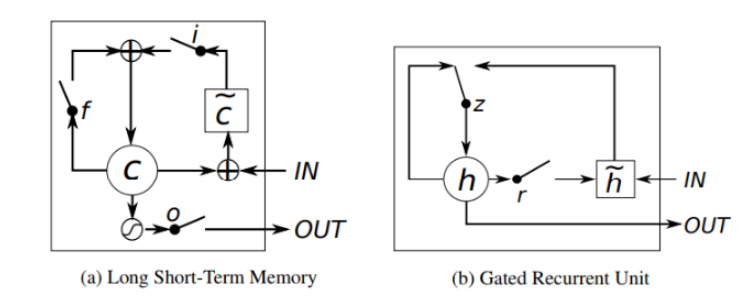
如果把 RNN 看做一個計算黑箱，其中狀態 s\_i 代表一個有限的記憶。函數 R 的作用就是讀取輸入 x\_i+1 和目前狀態 s\_i，然後將處理結果保存為新的狀態 s\_i+1。從這個角度看，S-RNN 顯然沒有提供控制記憶的方式，每次計算都是讀取整個記憶然後寫入整個記憶。門限結構提供的就是控制記憶存取的機制。

門限機制，即通過門限向量，控制對當前記憶狀態 s\_i 的讀取，且是動態的，即可以被當前記憶狀態和輸入所控制，且其行為是可以學習的。該機制保證了和記憶部分相關的梯度，即使經過很長的時間跨度，也保持較高的值，避免了梯度消失。

當訓練 LSTM 網路時，Jozefowics2015 強烈推薦總是將遺忘門的 bias 部分初始化接近1。

雖然 LSTM 結構效果很好，但由於結構過於複雜，導致難以分析，運算成本高。 Gated recurrent unit (GRU) 是 Cho2014 作為 LSTM 替代品提出的。 GRU 雖然基於門限機制，但減少了門限的數量，並取消了切分 memory 的思路。

目前 GRU 在語言模型和機器翻譯任務重表現很好，雖然相較LSTM孰優孰劣尚未有所定論，但可以肯定的是GRU在簡潔性上有優勢，可以有效降低訓練模型的時間成本。



LSTM和GRU模型示意圖

2.5 Word-embedding

Word Embedding是指透過特定算法將詞語轉換成向量編碼的表示法，表示方法主要有one-hot，N-gram，分布式表示(distributed representation)、共现矩阵等。

在早年很多機器翻譯的任務中，詞表示會使用One-Hot-Encoding。

One-Hot-Encoding是一種稀疏向量表示法，它將需要表示的元素總數作為向量的統一維度，每一個元素的對應向量，只有在特定的某一維表示為1，其餘均為0。這種詞表示法的優點是簡單，缺點也顯而易見。當語料的數量較大時，使用這種詞表示法容易產生維度災難。此外，這種詞表示法中每個詞都是獨立的，因此也無法計算詞之間的相似度。

本文中使用到的是另一種分佈式詞表示法（distributed representation）。該方法以深度學習為基礎，將文本中的每個詞訓練成不同的短向量，并將它們集中在一個向量空間中。在這個空間里有距離的概念，如餘弦相似度（Cosine similarity）。這樣一來，我們就可以用詞之間的距離來表示相關度。

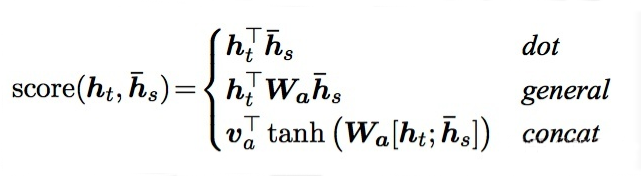
最常用的詞向量訓練法有兩種：CBOW和Skip-Gram。兩者的原理類似，實現的方式相反，在此不多贅述。

2.6 Attention Mechanism

固定長度向量承擔編碼輸入序列的整個“含義”的負擔，無論可能有多長。 由於語言的差異，這是一個非常難的問題。 想像一下兩個幾乎相同的句子，二十個單詞長，只有一個單詞不同。 編碼器和解碼器都必須細緻入微，以便將這種變化表示為空間中略微不同的點。

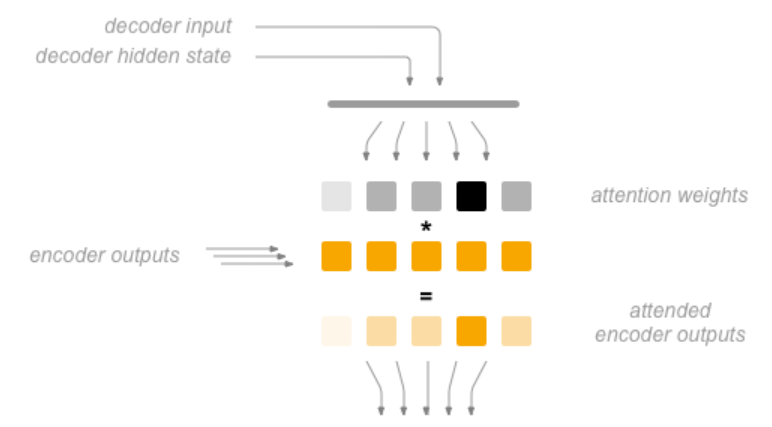
Bahdanau等人（2016）在論文中提出注意機制。 通過給Encoder提供一種“注意”部分輸入的方式來解決這個問題，而不是依賴於單獨的特定向量。Decoder可以根據輸入是句子的不同部分這一觀點，來考慮生成輸出。

在Multiplicative attention(Luong et al., 2015)中提到了三種attention score的計算方法，本文中會採用第一種。



三种attention score计算方式

注意力的計算與Decoder中的另一個前饋層有關。 該前饋層將使用當前的輸入和隱藏狀態來創建新的向量，該向量與輸入序列的大小相同（即是固定的最大長度）。 該向量通過softmax處理以創建注意力權重，該注意力權重乘以Encoder的輸出以創建新的上下文向量，然後用於預測下一個輸出。



Attention计算示意图

2.7 Sentiment Analysis

https://arxiv.org/pdf/1312.5419v3.pdf

1. **方法介紹**

3.1簡介

在早年商用聊天機器人的研究中，我們不難發現一個現象，即在問句的回答中，很容易出現讓人覺得尷尬的回應。所謂尷尬指的是該回應合理，卻不合情。比如當一位顧客在反應購買的商品有問題時，不排除客服機器人會發生以下這種情況。

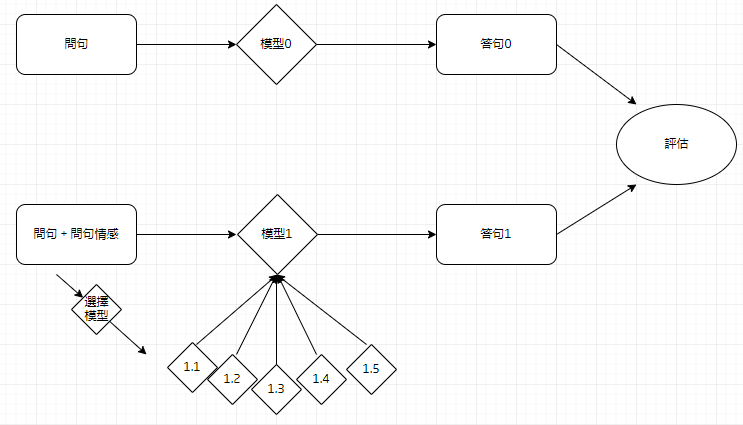
問：我买到商品为什么是坏的？

答：哈哈哈哈，我的也是。

這種回答合乎語法，也與問句相關，但顯然不是好的回答。為了解決這一問題，我們建置了一個基於seq2seq模型的聊天機器人系統，並加入情感標籤元素，讓系統可以在特定的情感標籤下生成對應的輸出，以提升聊天機器人對話的質量。

3.2系統架構

本研究的系統架構如下圖，可分成兩部分來看。模型1是對照組，模型2是實驗組，實驗組是五個由不同情感標籤的資料集分別訓練的。模型的本質都是Seq2Seq神經網路，透過模型1和模型2，會對同樣的問句產生兩種答句，最後再由人工對抽樣後的兩種答句進行評估。

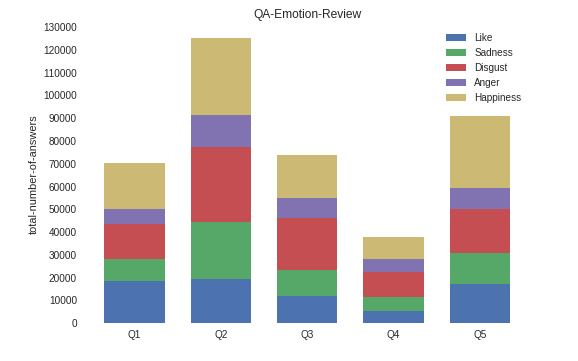


系統架構圖

3.3前處理

由於NTCIR官方提供的資料結構比較規範，在斷詞方面也已經處理完畢，因此本研究中沒有在進行重新斷詞。首先我們刪去了情感標籤的元素，為模型1的訓練做準備。

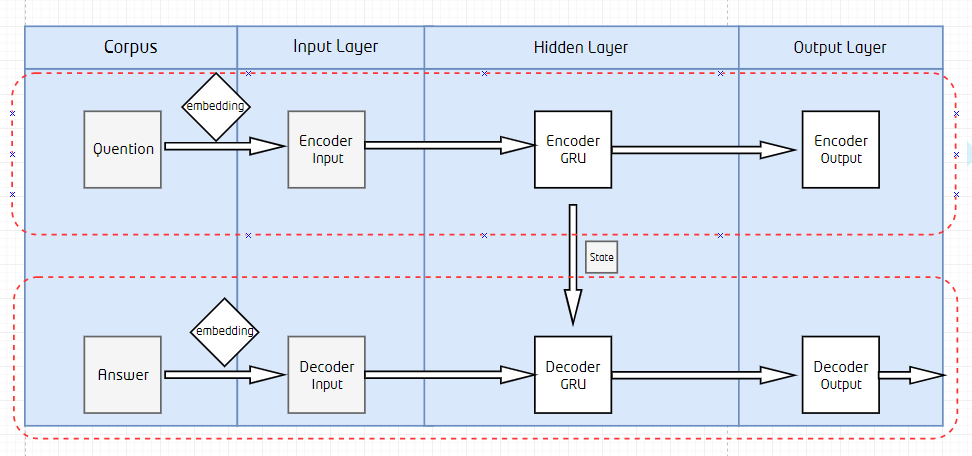
由於模型2包含5個子模型，所以我們根據有效情感標籤（不含標籤0），將資料集分成5份。下圖中橫軸的Q1到Q5分別表示情感標籤為Like,Sadness,Disgust,Anger，Happiness的問句，縱軸表示該情感問句對應不同情感答句的數量分佈。



資料集預覽圖

3.4 seq2seq模型

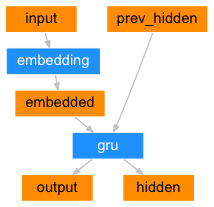
序列到序列網路或seq2seq網路或編碼器解碼器網路是由稱為編碼器和解碼器的兩個單獨的RNN組成的模型，下圖中的上半部分是編碼器（Encoder），下半部分是解碼器（Decoder）。Encoder將許多輸入編碼到一個向量中，並從一個向量由Decoder解碼為多個輸出，因此可以擺脫序列順序和長度的約束。 編碼序列由單個向量表示，在理想情況下，這一向量可以被理解為整個序列的“意義”。



Seq2seq系統架構圖

3.4.1 GRU-Encoder

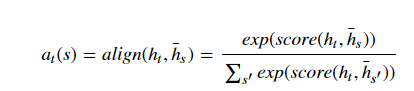
Seq2Seq網路的Encoder本質上是一個RNN，它為輸入句子中的每個詞輸出一個向量和隱藏狀態，並將隱藏狀態用於下一個輸入字。GRU結構可以用來解決梯度消失的問題。



Encoder示意圖

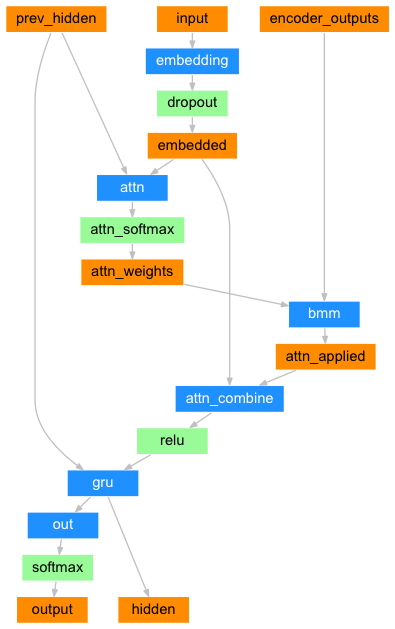
3.4.2 GRU-Decoder with Attention

一般的注意力計算來自Decoder的隱藏狀態（）和Encoder的狀態（），經由正規化後最終會得到一個總和為1的值。



Luong等人（2015）提出了一些Global Attention 的模型，來改進傳統注意力模型，他們之間的區別在計算注意力得分的方式（詳見2.6）。本研究中用到的得分計算方式是dot，顧名思義是兩個狀態之間簡單的乘積。

現在我們可以構建一個Decoder，在和3.4.1中提到的一樣的RNN之後插入這個Attn計算模組來計算注意力，並將這些權重應用到Decoder的輸出，以獲得上下文向量。

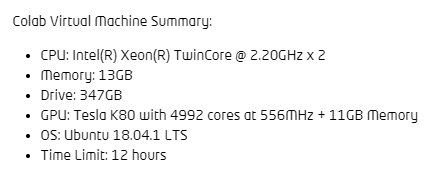


Decoder with Attention示意圖

1. **實驗設計與結果**

4.1實驗環境

以下將對本研究的環境進行介紹。基於成本和時間效益的考量，本研究的主要進程均在Google Colab（Colabortary）上進行。Google Colab是Google推出的免費GPU伺服器，預裝Jupyter Notebook環境，可以和Google Drive聯結，並且完全在雲端運行。下圖是Google Colab的環境配置圖。



Colab環境配置圖

除了上述的硬體環境，本研究還用到一些Python套件，這些套件也在一定程度上推動了這次實驗的進度。下圖是本研究中用到的套件列表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 深度學習框架 | 資料框 | 矩陣運算 | 詞向量 | 敘述統計 | 視覺化 |
| Pytorch | Pandas | Numpy | Gensim | NLTK | Pyplot |

使用套件表。

本研究中使用到的語料來自NTCIR大會官方。在第十四屆的比賽中，大會方提供了共計60萬筆簡體中文微博用戶的問答對（post & response），問答對中每一筆資料還有一個基於句子本身的客觀情感標籤。情感標籤分類器是一個簡單Bi-Lstm模型，它接受來自NLPCC情緒分類挑戰任務的資料集訓練。情感標籤共分六類（Other,Like,Sadness,Disgust,Anger，Happiness）。訓練資料集的格式如下所示：

资料集格式

以資料集中第一筆資料為例：

[['现在 刷 朋友 圈 最大 的 快乐 就是 看 代购 们 各种 直播 。 。 。 。 。', '5'],  
['卧 槽 我 也 是', '4']]

此外，我們還搜集了第十三屆比賽中用到的資料集，其格式與上述資料集相同，共計約112萬筆（1,119,207），斷詞後不重複的詞數有約9.9萬筆（98,627）。以下是本次試驗中用到的資料集對照表。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 資料集 | 問答數（筆） | 詞數（個） | 問句基本統計量（詞） | 答句基本統計量（詞） |
| 2017 train\_data | 1,119,207 | 98,627 | max:33.11  median:8.5  min:1 | max:33.14  median:10.7  min:2 |
| 2018 train\_data | 600,000 | 93,969 | max:83.17  median:13.9  min:0 | max:377.11  median:7.4  min:0 |

實驗資料集對照表

4.2實驗設計

4.2.1生成模型實驗

生成模型的實驗主要分成兩部分進行，第一部分是訓練不含情感標籤元素的傳統生成模型，以下簡稱模型1，第二部分是本研究提出的情感型生成模型，以下簡稱模型2。

模型2大致上可以看成5個小模型的整合，這些模型根據5種有效情感標籤區別（即不含Other）。首先需要從原始資料集中，按照情感標籤的不同將資料集分成5份，然後透過3.1本文所提出的模型，分別訓練5個情緒模型，按照Other,Like,Sadness,Disgust,Anger，Happiness的順序，記作模型2.1, 2.2, 2.3, 2.4, 2.5。

模型1和模型2在演算法的應用上並無區別，只是模型1中使用的語料會在透過前處理之後，將所有情感標籤去除。為了滿足模型1與模型2的對比實驗的科學性，兩個實驗使用到的訓練參數（RNN層數，隱藏層數，學習率，迭代次數）與訓練資料集的數量將完全一致，最終結果展示時用到的語料也將相同，以達到更好的參照效果。

在對模型1和模型2生成的回應句的評估上，將會採用問卷的形式，從生成句中分別隨機抽樣30句進行實驗。由於訓練語料源自微博，所以受測人員均是微博用戶簡體中文的使用者。受測人員將會從語言流暢度（s1），問答相關度（s2）和情感表達度（s3）三個角度對生成句進行評估，且三個角度的權重賦值相等，針對每一個角度的評估都是從兩種回應句中擇優選擇，優勝句計1分。最終得分計算方式如下：



4.2.2情感標籤分類實驗

4.3實驗結果

1. **結語與未來發展**