教育資料與圖書館學

Journal of Educational Media & Library Sciences

http://joemls.tku.edu.tw

Vol. 57 , no. 3 (2020) : 355-378

建置與評估文字自動生成的情感對話系統

Development and Evaluation of Emotional Conversation System Based on Automated Text Generation

楊 德 倫 Te-Lun Yang
Master Student

曾 元 顯* Yuen-Hsien Tseng*
Distinguished Professor and Associate Dean
E-mail：samtseng@ntnu.edu.tw

English Abstract & Summary see link at the end of this article

Journal of Educational Media & Library Sciences
JoEMLS
http://joemls.tku.edu.tw

教育資料與圖書館學　57 : 3 (2020) : 355-378
DOI:10.6120/JoEMLS.202011_57(3).0048.RS.CM

研究論文

建置與評估文字自動生成的情感對話系統ψ

楊德倫[a]　曾元顯[bc*]

摘要

本研究藉由2019年中文情緒對話生成（CECG）評比任務所提供約170萬則語料，運用深度學習GPT-2與BERT等技術與工具，實作了具備情感對話的系統，並以CECG提供的測試發文評估其成效。由三位人工判斷的結果，顯示本研究建置發展的系統，與2019年CECG評測最佳團隊的系統有類似的成效水準。而進一步的案例分析發現，對於訓練資料中較普遍的話題，GPT-2的語言建模技術，的確可以生成創新、有趣、完美的回應文句。本研究的主要貢獻為：㈠將情感融入發文字串中做為條件求機率，以便簡潔地依原方式訓練並使用GPT-2；㈡運用BERT來預測回應文句的連貫性以做為排序的依據。雖然這兩項技巧分別源自GPT與BERT的訓練機制，但本研究稍加修改應用於CECG的任務上，獲得了不錯的效果。

關鍵詞：對話系統，文字生成，文意理解，深度學習，人工智慧

前　言

自從蘋果公司於2011年推出Siri個人語音助理後，透過對話方式與電腦設備互動，逐漸被廣大的使用者採用與接受，進而促使各大廠商紛紛推出自己的對話系統、服務或機器人，如三星的S-Voice、Google Now、微軟的Cortana、亞馬遜的Alexa、華碩的Zenbo等。近年來深度神經網路（deep neural network，簡稱DNN）機器學習技術的進步、雲端服務提供大量的運算資源，以及網路上大數據的易於取得，則加速了產業各界關注人機對話系統的商業發展。

ψ 本文兩位作者貢獻相同，由楊德倫實作、曾元顯提出解決方案並主筆完成。

[a] 國立臺灣師範大學圖書資訊學研究所碩士生
[b] 國立臺灣師範大學圖書資訊學研究所優聘教授兼副所長
[c] 科技部人工智慧生技醫療創新研究中心協同主持人
[*] 本文通訊作者：samtseng@ntnu.edu.tw

本文作者同意本刊讀者採用CC創用4.0國際 CC BY-NC 4.0（姓名標示-非商業性）模式使用此篇論文

2020/08/18投稿；2020/11/18修訂；2020/11/20接受

雖然人機對話系統已達商業應用階段，但多數以文句模版比對知識庫方式回應使用者，使用起來仍有生硬、呆板、低於預期之憾。要讓對話系統進步到更人性、同理、具情感的階段，需要更多自然語言處理相關的研究。例如，幽默的對話常可消解使用者抱怨程度、贏得好感等（Binsted, 1995; Binsted et al., 2006）；令人歡喜的回應，可提升專注或降低焦慮；而具備其他適當情緒的對話，也可具有陪伴、安撫或激勵等作用。

本研究目的，在研發相關的技術與系統，以探討中文人機對話中，如何運用情感或情緒來回應使用者，以便在後續人機對話的應用中，加入愉悅、安撫、吸睛、激勵、創意、降低煩躁等相關效應，做為增進良好對話體驗、提升應用系統滿意度、輔助人際溝通、陪伴安撫的工具。甚至，本研究的成果也可以應用於華語學習、華語教學等情境。

具體而言，本研究藉由2019年日本NII Testbeds and Community for Information access Research（NTCIR）計畫舉辦的第3屆短文對話（short text conversation task，簡稱STC-3）其中一項中文情緒對話生成（Chinese emotional conversation generation，簡稱CECG）任務所提供的語料，運用深度學習GPT-2與BERT等技術與工具，實作了具備情感對話的系統，並以CECG提供的測試發文與評估方式，評估其成效。三位中文系畢業生判斷的結果，顯示本研究建置發展的系統，與2019年CECG評測最佳團隊的系統，有類似的成效水準。而進一步的案例分析發現，對於訓練資料中較普遍的話題，GPT-2的語言建模技術的確可以生成創新、有趣、完美的回應文句。

下一節將簡要回顧相關文獻、介紹CECG評測任務，第三節說明本研究採用的理論、技術、方法與工具，第四節說明評測實驗並展示系統結果案例，最後一節總結本研究，並提出未來進一步探討的方向。

二、文獻探討

本節先介紹情緒對話系統相關的研究，再簡述2019年CECG的評比過程與結果，做為本研究的背景知識。

㈠情緒對話系統

在人機互動中，自動辨識人類情緒並做出適當回應，可讓人機互動更順暢有效。相關的研究顯示，同理情感的表達可增強使用者的表現（Partala & Surakka, 2004），並提升使用者滿意度以及促進正向的互動（Prendinger & Ishizuka, 2005）。Skowron（2010）提出了一種對話系統，稱為情感聽眾（affect listener），可以在內容和情感相關的層面上回應使用者的話語。這些研究主要是基於心理學領域的發現，而相關的系統主要依賴人工歸納的規則或少量的數據，使其難以應用於大規模的對話生成。

http://joemls.tku.edu.tw

考慮情感因素以生成文句的研究，有：Cagan等（2017）結合了語法資訊，運用情感和主題生成文字以回應有觀點的文章（opinionated article）；Hu等（2017）運用變分自動編碼器（variational autoencoders）提出的生成模型，可根據語言的某些屬性（例如正面或負面的情緒或評價）生成文句；Ghosh等（2017）提出了情感語言模型，以長短期記憶（Long Short-Term Memory，簡稱LSTM）語言模型為基礎，生成具上下文和五種情緒之一的對話。

上述三項研究採用的技術，各自不同且各具代表性，其在各自提出的實驗中，也都有不錯的成效。但跟本文最相關之研究，主要見於黃民烈提出的情感聊天機器（emotional chatting machine，簡稱ECM；Zhou et al., 2018）。

ECM擬解決的問題為：給定一則發文與指定的情緒類別，產生符合該情緒類別的回應文字。其採用序列對序列（Sequence to Sequence，簡稱Seq2Seq）的深度學習架構（Sutskever et al., 2014），加上情感嵌入、內部情緒記憶及外部記憶等三項機制，以約436萬則（發文、指定的情緒類別、回應）資料做訓練。其中發文與回應為蒐集自中國大陸微博Weibo的社群對話；而情緒類別有六類，分別為：Angry、Disgust、Happy、Like、Sad與Other，以中國大陸NLPCC 2013/2014情感分類挑戰賽的數據訓練，運用Bi-LSTM對發文與回應進行情緒類別自動分類，正確率為62.3%。因此，其訓練資料雜訊頗高，但因為資料量大（約436萬則），仍可訓練出有用的系統。在Titan X的GPU上，系統訓練時間費時約一星期。

就ECM產生的回應，以三位人員分別就內容與情緒做判斷。內容定義為回應是否適當、自然，因而可能是人為做出的，分0、1、2三等級評分。情緒定義為回應的情緒表達是否與給定的情緒類別一致，分0、1等級評分。對200則發文共1,200則回應（每則發文皆須就6類情緒回應）的評估，內容有51.4%達2分、26.3%達1分、22.1%為0分；而情緒為1分者有42.4%，0分者58.6%。內容與情緒合併看，內容2分且情緒1分者有27.2%，內容1分情緒1分者有10.8%，兩者合計38.0%。其效果較差者，多為發文本身的情緒與指定回應的情緒，在訓練資料中較稀缺者，如<Happy, Disgust>與<Happy, Angry>等。亦即發文呈現快樂的情緒，但回應是噁心、生氣的情形，在Weibo中屬少數，致使系統訓練不足，而成效較差。

㈡CECG評比任務

第14屆NTCIR（2018至2019年）舉辦的Short Text Conversation Task（STC-3）中，有一項黃民烈等人主辦的中文情緒對話生成（CECG）的評比（Zhang & Huang, 2019）。其任務定義為：針對使用者的發文（post），系統需輸出帶有指定情感類別（情緒）的回應（reply/response），其中情緒共五類：Anger、Disgust、Happiness、Like、Sadness。任務範例如下（本研究擬出的例子）：

Journal of Educational Media & Library Sciences
tku.edu.tw

使用者：昨天我的貓死掉了

系統：

[1：喜歡] 牠常裝，惹人愛的啊

[2：悲傷] 哦，光聽就令人難受

[3：噁心] 也好，省了照顧花費

[4：憤怒] 被弄死的嗎？算帳去

[5：幸福] 回天國，變成天使了

注意，此例中使用者發文可能表達悲傷的情緒，但要做出喜歡的回應，是相當困難的。

主辦單位提供的語料來自中國大陸微博使用者的發文和回應，總數約170萬則（2019年語料約有110萬則，2017年語料約60萬則）。對於每則發文與回應，主辦單位以分類器標記發文與回應的情感類別，其準確性約為62%，詳如前述。

主辦單位提供200則發文做為測試，參與隊伍需就每則發文，分別根據五種情緒做出回應，因此共需回應1,000則。最後有來自11個團隊的21份結果，其中16份符合格式要求（每隊可送出多份結果）。

所有團隊提交的16,000則結果先匯總在一起，刪除各團隊重複的回應後，共15,263則。將這些回應與其團隊編號隨機排序後，透過百度數據眾籌服務（Baidu Data Crowdsourcing Service）進行人工評估。經由示例與評估標註方式的培訓後，三位標註者各自獨立進行每則回應的評估。評估準則如下：

```
IF   Coherence and Fluency
     IF    Emotion Consistency
           LABEL 2   （得 2 分）
     ELSE
           LABEL 1   （得 1 分）
ELSE
           LABEL 0   （得 0 分）
```

其中Coherence表示與發文話題連貫，Fluency為回應語句流暢、合於文法，而Emotion Consistency則與要求情感一致。

在總共15,236則回應中，三位標註者分數皆相同者有9,527則（62%），只有二者相同者有5,150則（34%），三者皆不同者只有626則（4%）。若依多數決，則一致性程度高達0.96。各隊的每份結果，將得分加總後（最高為2,000分），除以送出的回應總則數（1,000），即為該隊該份結果的平均分數（因此理論上最高平均總分為2.0）。其中平均分數最高的團隊達0.953，其後分別為0.821、0.814、0.738、0.726，再其後的結果都遠低於這些分數。而各隊表現較差的情

緒類別，多為生氣與噁心，此乃由於訓練資料相對較少所致（Zhang & Huang, 2019）。

最佳團隊的作法，運用具注意力的Seq2Seq模型和複製機制，針對每類情緒實現了一個生成模型，再整合起來，並做了相當多的資料清理工作，以準備品質良好的訓練資料。生成模型本身的得分為0.738，加入基於規則的模組後，其得分增加到0.953。此規則模組在偵測到發文裡的關鍵詞後，即套用人工模版產生流暢、情緒明確的回應。顯然此規則加分不少。

三、研究方法

本研究採用的方法，異於前述，不是採用Seq2Seq模型，而是更進階的Transformer。本節循序漸進介紹使用的理論、方法、工具、系統架構以及訓練資料，期能建構有效的情緒對話系統。

㈠深度學習（Deep Learning）

近十年來由於雲端計算的技術讓電腦運算能力變得更強、大數據時代提供了許多可用於訓練電腦的資料，以及機器學習演算法的精進與突破，使得人工智慧在影像辨識、自然語言處理等方面，有突破性的進展，而被廣泛應用於搜尋引擎、智慧型手機、自動駕駛等系統，逐漸融入到人們的日常生活中。而此波人工智慧的進展，幾可歸功於深度學習（LeCun et al., 2015）的發展與運用。

以往，只有輸入層與輸出層的人工神經網路，其功能有限，無法學習出像是互斥或運算（exclusive OR operation）這類函數（Minsky & Papert, 1969）。之後科學家證明出只要在輸入層與輸出層中間多加一層隱藏層，就可以學習出任意的函數（Hornik, 1991）。可惜針對某一函數的學習（例如依照前導文，生成後續相關的文字），隱藏層中需要多少非線性處理單元才能學出該函數的輸入輸出對應，在理論上卻無法提供答案。

近年來透過多層的隱藏層，逐層的學出函數的對應，多層神經網路終於展現出其廣泛的應用能力。以影像辨識函數為例，2012年Krizhevsky等（2012）採用八層的神經網路，達到比第二名26.2%的影像辨識錯誤率更低的15.3%。從此之後，影像辨識使用神經網路的層數，從數十層，到上百層都有，而且效果越來越好。由於這數量遠超過理論上的需要值（理論上只要一層隱藏層），因此被稱為深度神經網路或深度學習。

2017年Vaswani等（2017）提出了Transformer編解碼器（encoder-decoder）深度神經網路架構，運用可以平行運算且能處理長距離依賴（long-distant dependency）的自我注意力機制（self-attention），取代需要循序運算的遞歸神經網絡（recurrent neural network），而在自然語言處理領域開啟了類似影像處理那樣突飛猛進的時代。

Journal of Educational Media & Library Sciences
http://joemls.tku.edu.tw

㈡語言模型（Language Model）

許多自然語言處理的任務中，都需要建構一個準確的語言模型。亦即在獲知前t個字詞後，需要準確的估計下一個字詞的條件機率：

$$P(^{(t+1)} \mid w^{(t)}, w^{(t-1)}, \cdots, w^{(1)})$$

例如，從語音訊號辨識出「mao zhuo lao shu」四個音以及前三個字「貓捉老」之後，下面的條件機率應該要符合：

$$P(\text{鼠} \mid \text{老}, \text{捉}, \text{貓}) > P(\text{樹} \mid \text{老}, \text{捉}, \text{貓})$$

亦即「貓捉老鼠」應有較大的機率，其音不僅跟「貓捉老樹」一樣（音調不計），依照常識在自然語言的表達上也更為常見、更為合理。

建構此語言模型的過程稱為語言建模（language modeling）。然而語言建模過去很難做得好，亦即「估計下一個字詞的條件機率」不易做得準確。以語料庫C = (貓跳, 狗躍, 貓奔, 狗跑, 跑車) 為例，以此極小（五個字句）的語料庫擬訓練出P(跳|貓)等條件機率，應用傳統方法「最大似然估計」（maximum likelihood estimation）可得出：

$$P(\text{跳}|\text{貓}) = \text{P}(\text{貓跳}) / \text{P}(\text{貓}) = 1/2 = 0.5$$
$$P(\text{奔}|\text{貓}) = \text{P}(\text{貓奔}) / \text{P}(\text{貓}) = 1/2 = 0.5$$

等數據，因為在C中貓出現二次，而貓跳、貓奔各出現一次。但是條件機率：

$$P(\text{跳}|\text{狗}) = \text{P}(\text{狗跳}) / \text{P}(\text{狗}) = 0/2 = 0.0$$
$$P(\text{奔}|\text{狗}) = \text{P}(\text{狗奔}) / \text{P}(\text{狗}) = 0/2 = 0.0$$

亦即 (狗跳, 狗奔) 兩詞的機率都為0，因為他們都沒有在語料庫C出現過。換言之，以C訓練出來的語言模型，幾乎不會產生 (狗跳, 狗奔) 的詞句。但事實上，不僅C中有類似的概念 (狗躍, 狗跑)，於一般語言常識上也應該允許該詞彙的出現。

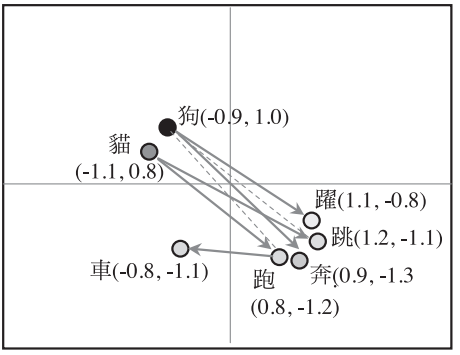
顯然傳統的方法不夠好，其源由為離散空間的字詞表達方式（discrete space word representation），亦即將貓、狗、跳、躍、奔、跑、車各字視為完全不同的概念，導致類似的概念無法類推，且訓練語料沒有的字句便難以估計其條件機率。

有鑒於此，神經機率語言模型於2000年左右被提出（Bengio et al., 2003），其將各個字詞w以n維實數向量V(w)表示，亦即每個字都是n維連續空間上的一個點，此種稱為連續空間的字詞表達方式（continuous space word representation）。當神經機率語言模型從語料庫學習完後，V(貓) 在連續空間上會近似於V(狗)、V(跳) 會近似於V(躍)、而V(奔) 會近似於V(跑)，因而可用於推論。

Journal of Educational Media & Library Sciences
http://joemls.tku.edu.tw

如圖1範例所示，每個字詞可以是二維空間中任意連續座標上的一點，當訓練完後貓的座標（-1.1, 0.8）接近於狗的座標（-0.9, 1.0），跳（1.2, -1.1）近似於躍（1.1, -0.8），依此類推。那麼語料庫原有的 (貓跳, 貓奔) 會有較高的機率值（例如0.45），而語料庫中沒有的 (狗跳, 狗奔) 也可在連續空間中推論出其機率值（例如0.35），而不再等於0了。而車則因為與貓、狗較不相似，難以類推，使 (車跳、車奔) 這些字詞的機率較低，但也並非完全不可能出現。

圖1　在二維連續空間表示貓、狗、跳、躍、跑、奔、車等字詞向量與關係之示意圖



註：實線為語料庫原有的詞句，
虛線為可推論計算出的詞句。

這種以連續空間表示字詞的向量表示法，解決了訓練語料無論大小，都無法窮舉所有語言現象的困擾；而且透過神經網路的學習，可以學出「類似概念有近似向量」（similar concepts having similar vectors）的表達方式。將類似「概念」轉換成近似「向量」的表達方式，被廣為運用，並以「嵌入」（embedding）一詞稱之。若此「概念」為詞，則稱為詞嵌入（word embedding; Mikolov et al., 2013），若為句，則稱為句嵌入向量（sentence embedding vector）表示法。將一群離散的物件進行嵌入轉換後，除了容易進行相似度計算、推論原先不存在的關係外，也可以從極大量資料中，訓練出品質較佳的嵌入向量（稱為預訓練模型pre-trained model），而可以分享給其他類似的任務進行微調（fine-tuning）運用。

使用大型的預訓練（語言）模型，以進行下游的自然語言處理任務，如：問答、摘要、分類等，目前已經越來越普遍。原因如下：1.從極大量資料中預先訓練出來的模型涵蓋較廣的知識、具備較佳的泛化能力（generalization），運用於特定任務時，可以利用這些特點而提昇效能；2.極大量資料通常不易取得與處理，訓練過程耗費許多時間與能源，在開源（open source）概念普及下，很多大公司紛紛釋出其耗資龐大的預訓練模型，以降低大家的總成本，並加快此領域研發與應用的進展。

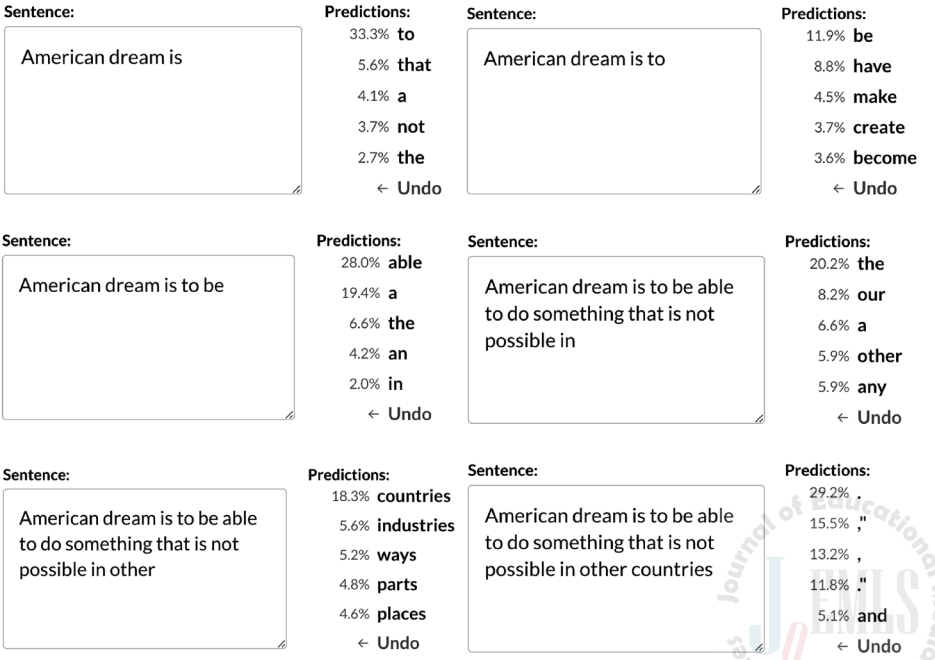
http://joemls.tku.edu.tw

㈢文字生成（Text Generation）

語言建模的方式與架構有很多種，效率與成效各異。OpenAI公司於2018年6月提出生成預訓練（generative pre-training，簡稱GPT）模型（Radford et al., 2018），其為12層Transformer疊加的深度神經網路，可用以學出估計下一個字詞的條件機率函數，且GPT比之前應用LSTM（Hochreiter & Schmidhuber, 1997）的準確度還高，且可預測更長的文句。亦即GPT的深度神經網路架構可學出非常優良的語言模型。由於以Transformers為基礎的GPT成效良好，OpenAI續於2019年推出GPT-2（Radford et al., 2019），其比GPT的模型架構更大，從12層到48層，最大的架構有15億個可學習參數，訓練資料量也更大（從GPT所用的5GB提升到40GB）。

GPT-2可用來準確地預測下個字，並且用預測出的下個字來預測下下個字，以致於能預測出整段文字，如圖2所示。在輸入「American dream is」之後，系統對所有下個字皆做出預測，但以「to」的機率最高，為33.3%，其次為「that」有5.6%。若選「to」則「American dream is to」的下個字為「be」的機率最高。即便依其預測選擇機率低者，如第4、5張圖的other，GPT-2依然能產生符合文法的下個字，直到選擇句點為止。

圖2 AllenAI網站（https://demo.allennlp.org/next-token-lm）使用GPT-2展示



2020年5月28日OpenAI更進一步發表了GPT-3（Brown et al., 2020），其最大的模型有96層Transformer，可學習的參數有1,750億之多，使用的訓練資

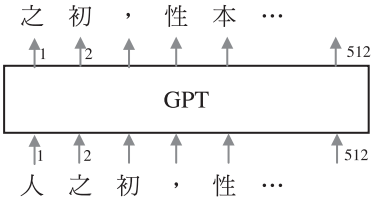
Journal of Educational Media & Library Sciences
JEMLS
http://joemls.tku.edu.tw

料將近45TB（模型訓練好的花費，網路上傳言約需460萬美金[1]）。其效果又比GPT-2好很多，可不用再訓練就足以解決很多自然語言處理的問題。

本研究採用Du（2019）改版的GPT2-Chinese（https://github.com/Morizeyao/GPT2-Chinese），用來學習中文的對話語料，從而自動生成中文回應。其有12層Transformer，在生成文字時，可接受的輸入長度最多512個字，輸出長度可指定，預設為1,024個字。

GPT系列的神經網路，是以自我監督方式訓練（self-supervised training）出來的。亦即，只要蒐集品質良好的語料，不必進行任何的人工標記與判斷，將語料中的每一句子當作輸入，如圖3之輸入：「人之初，性…」，並將該句子移走第一個字後的字串當作目標輸出，如圖3的「之初，性本善…」，然後要求GPT進行生成預測，若相對應位置的字詞有錯誤，就將誤差以倒傳遞（error backpropagation）方式，按梯度下降法（gradient descent）調整參數（Rumelhart & McClelland, 1986）。GPT使用內部遮罩機制，確保在預測第i個輸出字詞時，只用到輸入的第1到第i個字詞，不會用到（$i+1$）以後的字詞資訊。

圖3　訓練GPT學習三字經等文句示意圖



註：輸入字串最長允許512個字詞。

㈣文字理解（Text Understanding）

GPT只使用到Transformer的解碼器架構，且只看前面出現過的文字來預測下一個字。但在讀文句時，有時會需要看前後文，以便對文意有完整的理解。2018年10月Devlin等（2018）提出基於轉換器的雙向編碼器表示技術（Bidirectional Encoder Representations from Transformers，簡稱BERT），能夠接受整句或整段文字，進行如：主題分類、情感分析、自動問答、文意比對等需要某種程度文字理解的任務。

BERT的訓練方式是基於遮罩式的語言模型（masked language model，簡稱MLM），以及下一句的預測（next sentence prediction，簡稱NSP），如圖4所示。給予訓練資料「人之初，性本善」、「性相近，習相遠」等文句，在插入[CLS]、[SEP]特殊符號後，MLM的訓練是將輸入的15%字詞代換成遮罩符號（如[m]或[MASK]），要求BERT預測出正確的被遮罩字詞；而NSP則將下一句以50%

[1] https://www.reddit.com/r/MachineLearning/comments/h0jwoz/d_gpt3_the_4600000_language_model/

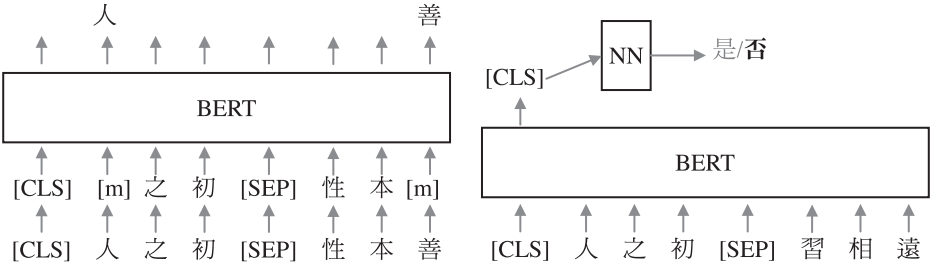
http://joeed.tku.edu.tw

的機率代換成語料中的其他句子，並將 [CLS] 位置的輸出送入一個簡單的神經網路（neural network，簡稱NN）學習，以預測用 [SEP] 分開的兩句話是否為上下句的關係，從而做出「是」或「否」的輸出預測。[CLS] 是特殊符號，大略代表整個輸入的句嵌入向量。以大量資料做這兩項任務的訓練，再針對下游任務（如分類、問答等）微調訓練後，BERT在11項自然語言處理評測任務上，達到當時最佳的效果。

由於GPT與BERT都是基於Transformer的深度神經網路，其輸入與輸出的每個字詞都經過嵌入轉換，使得語意相近的字詞有相近的向量，而每個字詞的嵌入向量維度為768維，亦即由768個實數值表示（GPT-3等大型的架構，維度更高）。

Google有釋出BERT的程式碼以及其訓練好的中文模型。在此基礎上，本研究運用BERT針對GPT-2產生的多項候選回應文字，進行下一句的線性回歸訓練，以便排序這些候選回應，選最佳者回應使用者的輸入。

圖4 BERT訓練方式示意圖



註：左圖為遮罩式語言模型，右圖為下一句的預測。

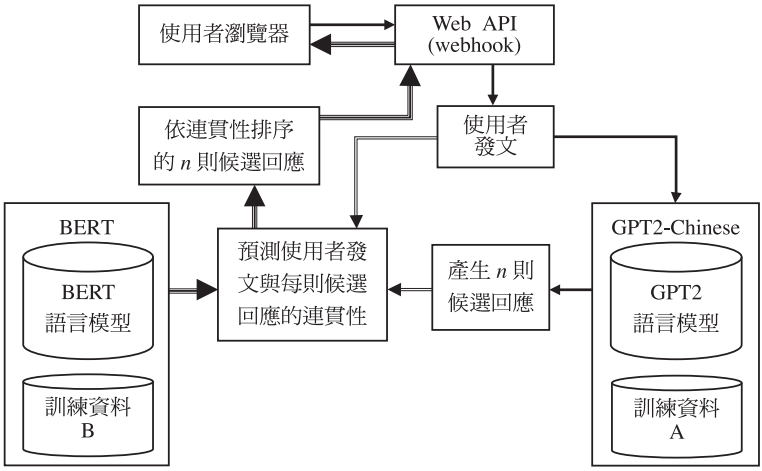
㈤情緒對話系統

本研究提出的情緒對話系統，其處理步驟流程如圖5所示。首先，系統從瀏覽器或以Web API方式獲得使用者的發文（post）輸入，以此發文加上情緒標記做為前導文字要求訓練好的GPT-2產生*n*則回應（reply）。為了評估哪一則回應較為適當，將使用者發文與每一則回應輸入到訓練好的BERT，以預測該則回應為使用者發文的下一句之連貫性（coherence）分數，依此分數排序這*n*則回應，最後將排序最高的回應傳回給使用者。

在上述的流程中，需事先對GPT-2與BERT做訓練。目前訓練的原始語料為STC-3主辦單位提供約110萬組的對話，其另提供2017年的語料約60萬組，兩者總計約170萬組對話。這些對話每則都以 [[post_*i*, post_*i*_label], [reply_*i*, reply_*i*_label]]的格式儲存，如下範例所示：

http://joemls.tku.edu.tw

圖5　情緒對話系統流程圖




```
[
  [
    ["最近 事 好多", "0"],
    ["忙 並 快樂 著", "1"]
  ],
  …
  [
    ["不管 你 是 小 獅子 還是 小 蟹子 。 我 都 一樣 的 喜歡 你", "1"],
    ["我 是 天蠍 ~ [ 哈哈 ]", "5"]
  ]
]
```

其中："最近 事 好多"為發文，"忙 並 快樂 著"為回應，而情緒標記為0, 1, 2, 3, 4, 5，分別代表：1是指喜歡（like），2為悲傷（sadness），3是噁心、厭惡（disgust），4是憤怒（anger），5是指開心、幸福、高興（happiness），而0是指其他（other），在本研究中不會用到。此外，STC-3 語料所使用的語系，原是簡體中文且斷詞過，本研究將其轉為正體中文後使用，且忽略其斷詞訊息。

為了引導GPT-2產生對應情緒的回覆，本研究將STC-3語料每則對話，轉換成這樣的格式：["發文[回應的情緒]回應"]。亦即CECG的任務需要兩個條件的語言模型：

$$P(\text{reply} \mid \text{post}, \text{emotion})$$

本研究透過文字串接將其合併成只用一個條件的語言模型（亦即原有的條件機率），如下：

http://joemls.tku.edu.tw

$$P(\text{reply} \mid \text{“post [emotion]”})$$

以上述範例第一則對話為例，將轉成：[“最近事好多[喜歡]忙並快樂著”] 做為GPT-2學習的一句話，其中「最近事好多」是原始發文，「[喜歡]」是其回應的中文情緒詞彙（情緒標記為1），「忙並快樂著」是原始回應。亦即，將原始語料轉成：

```
[
  [ “post_1 [reply_1_label_term] reply_1” ],
  [ “post_2 [reply_2_label_term] reply_2” ],
  …
]
```

得到總共約170萬句（圖5中的訓練資料A），用以訓練GPT-2共100回合（epoch），使用Titan RTX 24GB的GPU，費時約200小時（約8.3天）。

訓練完應用GPT-2時，使用者輸入類似的發文「最近好多事」，並指定回應情緒為「喜歡」的文句時，系統將兩字串串接成：「最近好多事[喜歡]」做為前導文後，送入GPT-2使其生成類似的回應：「忙得快樂啊」等文句，而不僅有「您好多事啊」這種較無情緒的規則式回應。亦即希望GPT-2估計的條件機率有如下的表現：

$$P(\text{“忙得快樂啊”} \mid \text{“最近好多事[喜歡]”}) >$$
$$P(\text{“您好多事啊”} \mid \text{“最近好多事[喜歡]”})$$

GPT-2可以生成多個回應文句，而首次生成的文句，不見得最好，但如何評估？前述BERT的訓練過程中，有NSP預測下一句的訓練，但其為「是」或「否」的輸出，不利於將多個回應做排序。本研究從原來的語料中，抽出一部分，做如下的安排：

```
[
  [ post_1, reply_1, 1.0 ],     <= 成對的對話
  [ post_3, reply_7, 0.0 ],     <= 不成對的對話
  [ post_8, reply_8, 1.0 ],     <= 成對的對話
  [ post_9, reply_2, 0.0 ],     <= 不成對的對話
…
]
```

亦即從原始語料中抽樣，取出成對的發文與回應，並給予連貫性分數1.0，同時抽樣不成對的發文與回應，並給予連貫性分數0.0。然後針對BERT之[CLS]位置的輸出向量（有768維實數），當作線性回歸（linear regression）的輸入，

Journal of Educational Media & Library Sciences
JEMLS
http://joemls.tku.edu.tw

其對應的連貫性分數當作線性回歸的輸出，以15,000則對BERT進行微調（fine-tuned）訓練1回合（費時約3分鐘），其中成對與不成對的對話各50%，亦即各7,500則（圖5中的訓練資料B）。

四、實驗結果與討論

本研究根據前一節描述的系統，就CECG提供的測試發文200則，產生了1,000句回應，然後找3位中文系畢業生，根據主辦單位的規則，分別判斷這1,000則回應，部分結果如表1所示。其中第一欄SID是發文回應的文句編號，第二欄使用者發文有兩則範例；第三欄是指定的回應情緒，第四欄是系統根據發文與指定情緒的輸出，第五、第六欄為依規則要求人工判斷的結果，第七欄為依據前兩欄計算出的分數，第八欄為本研究後續再人工判斷的結果。表1中第二則使用者發文的「神馬」應為「什麼」、「木」應為「沒」之意，為網路上部分人的慣用法，語法沒錯、語意清楚。

以SID為1的回應這句話「我們看完了,好看哈!」，其文句通順、合乎語法，因此第八欄給1，且其跟發文話題連貫，因此第五欄給1；但人工判斷認為此句話跟「喜歡」的情緒不符，因此第六欄給0；綜合第五欄與第六欄的結果，依據規則其得分在第七欄為1。而SID為54的回應「暖氣什麼時候變冬天啊有木有」其通順度為0，通順與連貫性也為0，但語意被判斷為有符合要求的「憤怒」情緒而給1，惟依據規則，分數只能給0。

表1　測試發文與系統回應之部分範例以及人工進行判斷結果

SID	使用者發文	回應情緒	系統回應	連貫且通順	符合情緒	分數	是否通順
1	為什麼 為什麼 為什麼 你們 都 不陪 我 看 電影 ! ! [怒罵]	1：喜歡	我們看完了,好看哈!	1	0	1	1
2		2：悲傷	我也不陪你	1	0	1	1
3		3：厭惡	你要是全陪你看電影都不好意思了	0	0	0	0
4		4：憤怒	什麼為什麼你不陪我看	1	1	2	1
5		5：快樂	看電影看的爽吧	0	1	0	1
51	為 神馬 合肥 木 有 暖氣 ! 凍死 我 也 …	1：喜歡	來北京吧~~~這裏很暖和的	1	1	2	1
52		2：悲傷	凍死我都沒覺得有暖氣~~	1	1	2	1
53		3：厭惡	暖氣沒有,這麼冷?	1	0	1	1
54		4：憤怒	暖氣什麼時候變冬天啊有木有	0	1	0	0
55		5：快樂	來北京啦!金山最近有暖氣	1	1	2	1

上述判斷皆由三人獨立完成，在1,000則回應中，分數三者皆相同者有490則，只有二者相同者有457則，三者皆不同者只有53則。若依多數決，則一致性程度高達0.947 (= (490 + 457) / 1000)。三者皆不同者，再由其中一位看過後，統整大家的判斷，做最後分數的裁決。

經上述統整後，第八欄的通順度達0.903 (= 903 / 1000)，第五欄的符合問句程度達0.591 (= 591 / 1000) 約60%，第六欄的符合情緒程度達0.508 (= 508 /

1000) 剛好超過5成，第七欄最後的總分為0.987 (= 987/ 1000) ，與STC-3成績最好的0.953相差無幾。嚴格來說，由於判斷的人與時間不同，不能跟STC-3的結果比較。但由於一致性程度0.947與STC-3的一致性0.96都很高，顯示不同的人做的判斷，不會差距太大，因此某種程度上，兩者還是能相提並論。表2呈現本系統與STC-3成績最好的RUCIR系統的差異，以供比較。

表2 本系統與STC-3成績最好的RUCIR系統比較表

	方法	模型架構	資料前處理	得分
本系統	生成法	GPT-2生成＋BERT排序	僅將訓練語料調整文句順序	987
RUCIR	規則法＋生成法	Seq2Seq（GRU＋注意力機制）＋拷貝機制＋排序	濾除非中文、太短的訓練文句；刪除出現次數超過100次的文句	953

圖6 本系統透過程式介面的輸出結果


```
"predictions": [
  {
    "coherence": 1.0328320264816284,
    "text_a": "我每天都被自己帥醒，壓力好大[喜歡]",
    "text_b": "你最近不錯，我壓力也很大啊！"
  },
  {
    "coherence": 1.0220566987991333,
    "text_a": "我每天都被自己帥醒，壓力好大[喜歡]",
    "text_b": "帥氣可愛的小夥兒"
  },
  {
    "coherence": 1.020853877067566,
    "text_a": "我每天都被自己帥醒，壓力好大[喜歡]",
    "text_b": "帥氣也是條不歸路啊"
  },
  {
    "coherence": 0.9385004639625549,
    "text_a": "我每天都被自己帥醒，壓力好大[喜歡]",
    "text_b": "是的，我就喜歡這樣的自己。"
  },
  {
    "coherence": 0.823341429233551,
    "text_a": "我每天都被自己帥醒，壓力好大[喜歡]",
    "text_b": "恩，你也一樣！"
  }
],
```


除了產生上述1,000則結果外，本研究也實作出系統介面，可供後續應用與持續評估。圖6是Restful API的輸出結果，其對應的輸入文句為：「我每天都被自

己帥醒，壓力好大」，選擇的情緒為「喜歡」。圖7介面最下端為輸入區，使用者可輸入如：「我想永遠跟你在一起」，然後選擇「喜歡」的情緒，按送出後，系統內部設定由GPT-2產生五句回應，並且由BERT評估其為下一句的連貫性分數，再按此分數排序列出。圖左為情緒選擇喜歡的結果，圖右為選擇噁心的結果。

圖7　系統輸入、輸出介面與展示範例

本研究運用Elasticsearch搜尋引擎，另外設計查詢系統，可用來比對GPT-2產生的文句是否出現在170萬則發文或170萬則回應的訓練資料中（共340萬則）。圖8中顯示圖7的「我也想永遠跟你在一起」不在340萬則文句中，是GPT-2自己生成的。其他像「我知道你說的是真話。。。真的希望你能在一起」、「為什麼一定要跟我在一起啊」、「我還是覺得這話說的有點深奧」，以及圖6中的「帥氣也是條不歸路啊」，這些有趣、完美的回應，也都沒有在340萬則的訓練資料裡，而是GPT-2自學而成的。

圖8　文句比對系統介面與查詢範例

id	es_score	text	type	label
1193558	28.35546	我們 在 一起 在 一起 永遠 永遠	request	1 喜歡 (Like)
1339408	27.021397	當然 。 我們 永遠 在 一起 。 也 永遠 在 一起 喝 茅臺 。	response	1 喜歡 (Like)
192515	25.858036	永遠 在 一起	response	5 開心 (Happiness)
1254701	25.327236	我們 永遠 在 一起 ！	request	1 喜歡 (Like)
1439628	25.327236	我們 永遠 在 一起 ！	request	1 喜歡 (Like)

經由後續的自我評估，本研究覺得若發文的話題，在訓練資料裡有豐富的語料，則GPT-2可生成具有創意的文句。若發文的話題，在訓練資料中相對稀缺，則產生的文句其通順度、話題連貫性，都會降低。此與之前的研究有類似的結論。

Journal of Educational Library & Information Science
http://joemls.tku.edu.tw

五、結　論

本文依據STC-3的CECG任務，實作了一套系統，雖然未能參加2019年的評比，但本研究自我評估後，其可媲美當時最佳團隊的成效，甚至猶有過之。本系統中有數個重要的變數，影響其成效。首先為GPT-2產生回應的參數，可設定為較具創意或較為保守（依照訓練資料）的文字生成方式。其次為候選句子的個數，個數越多，可供排序的候選者越多，若排序準確，效果越好。第三為運用BERT預測下一句連貫性的排序，需要決定要拿多少資料來訓練、訓練多久（幾回合）。由於人工評估的成本不菲，本研究僅依據少數的個人觀察，即決定相關的參數與作法，供人工進行成效評估。

特別是BERT的排序部分，GPT-2已經產生了相當優良的候選句，例如發文：「我每天都被自己帥醒，壓力好大」，回應候選句中有：「帥氣也是條不歸路啊」，這種在訓練資料裡面未曾出現又同時具備隱喻、文青、符合發文意境的回應，卻未能被排序在最前面（排在第3順位）。這是因為BERT的下一句預測，並未考慮文青、隱喻、意境等文學因素。另外，查詢網路上的回應為：「幹話我聽多了，就你的最精彩」，雖較具趣味、有力，但粗俗了些。因此，若這部分能做得更好，則此系統的文學造詣，當令一般人驚艷、讚嘆，甚至仰望。當然，從其原理可知，系統沒有智慧或文學造詣，它只出現有智慧的結果。

運用GPT-2進行語言建模的好處是，即使訓練資料中有雜訊，只要資料量大、雜訊不高，透過後續的候選句排序等處理，不用像先前的研究進行繁複的資料清理，即可自動過濾掉頻繁出現、簡短、萬用、不令人期待的制式回應。

本文的主要貢獻如下：㈠將情感融入發文字串中做為條件求機率，以便簡潔地依原方式訓練並使用GPT-2。㈡運用BERT來預測回應文句的連貫性以做為排序的依據。雖然這兩項技巧分別源自GPT與BERT的訓練機制，但本研究稍加修改應用於CECG的任務上，獲得了不錯的效果。

本系統需要的計算量，比先前研究採用的方法還要高。隨著機器算力的進步，此問題將逐漸解決，其可整合至其他系統一起應用，將指日可待。目前，本研究正將其應用於華語文教學領域，畢竟其產生的回應，具有可供比較、評價與學習的效果。希望由此提供未來更多的應用案例。

誌　謝

本研究感謝科技部研究計畫補助，計畫編號：MOST 107-2221-E-003-014-MY2與MOST 109-2410-H-003-123-MY3。

Journal of Educational Media & Library Science
JoEMLS
http://joemls.tku.edu.tw

參考文獻

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137-1155.

Binsted, K. (1995). *Using humour to make natural language interfaces more friendly* [Paper presentation]. Workshop on AI, ALife and Entertainment, International Joint Conference on Artificial Intelligence, Montreal, Canada.

Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., & O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, *21*(2), 59-69. https://doi.org/10.1109/MIS.2006.22

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …, Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165v4

Cagan, T., Frank, S. L., & Tsarfaty, R. (2017). Data-driven broad-coverage grammars for opinionated natural language generation (ONLG). In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1331-1341). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1122

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805v1

Du, Z. (2019). GPT2-Chinese: Tools for training GPT2 model in Chinese language. Retrieved January 12, 2020, from https://github.com/Morizeyao/GPT2-Chinese

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 634-642). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1059

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257. https://doi.org/10.1016/0893-6080(91)90009-T

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (PMLR 70, pp. 1587-1596). ML Research Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the 25th international conference on neural information processing systems* (pp. 1097-1105). Neural Information Processing Systems Foundation.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L.

Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 26th international conference on neural information processing systems* (Vol. 2, pp. 3111-3119). Neural Information Processing Systems Foundation.

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16*(2), 295-309. https://doi.org/10.1016/j.intcom.2003.12.001

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence: An International Journal*, *19*(3-4), 267-285. https://doi.org/10.1080/08839510590910174

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. MIT Press.

Skowron, M. (2010). Affect listeners: Acquisition of affective states by means of conversational systems. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Lecture notes in computer science: Vol. 5967. Development of multimodal interfaces: Active listening and synchrony* (pp. 19-181). Springer. https://doi.org/10.1007/978-3-642-12397-9_14

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 27th international conference on neural information processing systems* (Vol. 2, pp. 3104-3112). Neural Information Processing Systems Foundation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 30th international conference on neural information processing systems* (pp. 6000-6010). Neural Information Processing Systems Foundation.

Zhang, Y., & Huang, M. (2019). *Overview of the NTCIR-14 short text generation subtask: Emotion generation challenge* [Paper presentation]. 14th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan.

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *The thirty-second AAAI conference on artificial intelligence* (pp. 730-738). Association for the Advancement of Artificial Intelligence.

楊德倫 ORCID 0000-0002-3351-1785
曾元顯 ORCID 0000-0001-8904-7902

http://joemls.tku.edu.tw

Journal of Educational Media & Library Sciences 57 : 3 (2020) : 355-378
DOI:10.6120/JoEMLS.202011_57(3).0048.RS.CM

Development and Evaluation of Emotional Conversation System Based on Automated Text Generation[ψ]

Te-Lun Yang[a] Yuen-Hsien Tseng[bc*]

Research Article

Abstract

Based on the corpus provided by the 2019 Chinese Emotional Conversation Generation (CECG) evaluation task, an emotional conversation system is implemented in this paper using deep learning and other technologies such as GPT-2 and BERT. The effectiveness of the system is evaluated based on the test data and criteria provided by CECG. The results based on three human annotators show that the system has a similar effectiveness level with that of the best team participating in the 2019 CECG task. Further case studies reveal that the more post/reply pairs about a topic in the training data, the better the language model of GPT-2 to generate innovative, interesting, and perfect response sentences for that topic. The main contributions of this study are: 1. Integrating emotion into the post string as a condition for computing probability, so as to simply train GPT-2 and make GPT-2 predict in the original way; 2. Applying BERT to predict the coherence of response sentences as a basis for ranking. Although these two techniques are derived from the training mechanisms of GPT and BERT respectively, we have slightly modified them to fit the task of CECG and achieved good results.

Keywords: *Conversational system, Text generation, Text understanding, Deep learning, Artificial intelligence*

SUMMARY

Introduction

In human-computer interaction, automatic recognition of human emotions for appropriate response can make human-computer interaction smoother and more effective. Related research shows that the expression of empathy can increase user satisfaction and promote positive interaction.

[ψ] Both authors have the same contribution. Te-Lun Yang implemented the whole system, while Yuen-Hsien Tseng proposed the solution and complete the paper writing.

[a] Master Student, Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

[b] Distinguished Professor and Associate Dean, Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan

[c] Co-Principal Investigator, MOST AI Biomedical Research Center

[*] To whom all correspondence should be addressed. E-mail: samtseng@ntnu.edu.tw

The Author acknowledges that the Article is distributed under a Creative Commons CC BY-NC 4.0.

http://joemls.tku.edu.tw

In pursuit of the above delicate interaction, this paper presents the building of a Chinese dialogue system that emphasizes on emotional conversation using state-of-the-art AI techniques, namely Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT). This emotional conversation system (ECS) is expected to respond to a user's post with a fluent and coherent reply conforming to a specified or detected emotion.

Problem

Specifically, this paper adopts the datasets and evaluation criteria from the Chinese Emotional Conversation Generation (CECG) Shared Task held in Short Text Conversation Task (STC-3) in the 14th NTCIR Workshop (2018-2019) to train the proposed system and evaluate its performance.

The CECG task is defined as: for a user input post, the system needs to output a response (or reply) with a specified emotion category, of which there are 5 types of emotion: Anger, Disgust, Happiness, Like, and Sadness. A possible example of the post/reply is as follows:

User's post: My cat died yesterday.

System reply (given a specified emotion type in the squared bracket):

[1: like] Oh she likes to make believe. That's cute.

[2: Sadness] Oh, I'm so sorry for your loss.

[3: Disgust] That's fine. That would save you a lot of trouble.

[4: Anger] Was it killed? Let's find out who did it!

[5: Happiness] How fortunate! She's an angel now in heaven.

Note that in this imaginative example, the user's post may express sadness, it is rather difficult to make a like response, even for human.

The datasets provided by the CECG Shared Task are based on the pairs of posts and responses of Weibo users mainly from mainland China, with a total of about 1.7 millions of pairs (about 1.1 millions of pairs in 2019 and about 600,000 in 2017). For each post or replied text, a machine classifier was trained to label the emotion type of the text, and its accuracy is about 62%.

The CECG Shared Task evaluates each reply based on the post and the specified emotion, by human, according to the following criteria:

```
IF Coherence and Fluency
    IF Emotion Consistency
        LABEL 2
    ELSE
        LABEL 1
ELSE
        LABEL 0
```

Journal of Educational Media & Library Sciences
JoEMLS
http://joemls.tku.edu.tw

Note that Coherence means that the reply is consistent with the topic of the post, Fluency means that the reply text is smooth and grammatically correct, and Emotion Consistency denotes that the reply's emotion is consistent with the specified emotion.

Method

The developed ECS system consists of a user interface, a GPT-2 model for text generation, and a BERT model for text understanding and coherence prediction. The ECS takes input post from users through a Web API (or Web UI). An open source GPT-2 Chinese model was trained to output k candidate replies. These candidates were further ranked by a Chinese BERT model trained to predict the coherence of the reply based on the post. The highest ranked candidate was then chosen as the output reply.

The training data from CECG were in the form: [[post_*i*, post_*i*_emotion], [reply_*i*, reply_*i*_emotion]]. They were converted into the form: ["post_*i* [reply_*i*_emotion] reply_*i*"] so as to conform to the training data format of GPT-2. In other word, the CECG problem asks us to predict the reply based on two conditions:

$$P(\text{ reply} \mid \text{post, emotion})$$

By concatenating the two conditions into one string, we reformulated the problem into the original language model learnable and predictable by GPT-2:

$$P(\text{ reply} \mid \text{"post [emotion]"})$$

The GPT-2 was trained on a Titan RTX GPU with 24GB RAM. It took approximately 200 hours to train the 1.7 millions of post/reply pairs for 100 epochs.

The GPT-2 can be configured to output *k* candidate replies. To rank these candidates, a Chinese BERT pretrained model from Google was downloaded and fine-tuned on part of the original training data with the following format for coherence prediction:

```
[
  [ post_1, reply_1, 1.0 ],
  [ post_3, reply_7, 0.0 ],
  [ post_8, reply_8, 1.0 ],
  [ post_9, reply_2, 0.0 ],
…
]
```

Journal of Educational Media & Library Sciences JoEMLS
http://joemls.tku.edu.tw

In other words, the BERT model was trained to do linear regression prediction: if the input is the original post/reply pairs from the training data, the desired output has a score of 1.0; if the input is the scrambled post/reply pairs, the desired output is 0.0 to indicate that the post and reply are not coherent. The BERT model was fine-tuned on 15,000 pairs for one epoch (about 3 minutes), in which paired and scrambled post/reply are 50% each.

Findings

Based on the evaluation criteria of the CECG Shared Task in 2019, the evaluation of 1,000 ECS generated replies by three native speakers majored in Chinese linguistics indicated that 90.3% reply texts are grammatical correct, and 59.1% are coherent to the posted text, and about 88% reply sentences are novel (not in the 1.7M training texts). This result outperformed the top-ranking system in the 2019 task, where a hybrid method of using both text generation and rule-based mechanisms was applied. Further case studies revealed that the ECS could generate innovative, interesting, and perfect response sentences for popular topics in the training data.

As an example, for the post stated "I would like to be with you forever," example replies would look like: "I would also like to be with you forever." if the specified emotion is "Like"; and "Why do you have to stay with me? It's not fair!" if the specified emotion is "Disgust".

More exploration of the ECS showed that, If the topic of the post is rich in the training data, the GPT-2 can generate creative sentences; if the topic of the post is relatively scarce in the training data, the smoothness and topic coherence of the generated sentence will diminish. These results are similar with conclusions from previous studies.

Conclusions

The main contributions of this study are: 1. Integrating emotion type into the post text as a single condition for language modeling, so as to train and apply GPT-2 in the original way; 2. Applying BERT to predict the coherence of response text for ranking the generated replies. Although these two techniques are derived from the training mechanisms of GPT and BERT, respectively, we have slightly modified the techniques to fit the task of CECG and achieved good results.

This work sheds light on the pursuit of delicate human-computer interaction with emotion. Future work is needed to achieve the goal of better response texts (through better language modeling or larger training dataset) and to propose effective response strategies to yield proper emotional reply once a corresponding emotion was detected in the post.

http://joemls.tku.edu.tw

ROMANIZED & TRANSLATED REFERENCE FOR ORIGINAL TEXT

Bengio, Y., Ducharme, R., Vincent, P., & Janvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research*, *3*, 1137-1155.

Binsted, K. (1995). *Using humour to make natural language interfaces more friendly* [Paper presentation]. Workshop on AI, ALife and Entertainment, International Joint Conference on Artificial Intelligence, Montreal, Canada.

Binsted, K., Bergen, B., Coulson, S., Nijholt, A., Stock, O., Strapparava, C., Ritchie, G., Manurung, R., Pain, H., Waller, A., & O'Mara, D. (2006). Computational humor. *IEEE Intelligent Systems*, *21*(2), 59-69. https://doi.org/10.1109/MIS.2006.22

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …, Amodei, D. (2020). *Language models are few-shot learners*. arXiv. https://arxiv.org/abs/2005.14165v4

Cagan, T., Frank, S. L., & Tsarfaty, R. (2017). Data-driven broad-coverage grammars for opinionated natural language generation (ONLG). In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 1331-1341). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1122

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://arxiv.org/abs/1810.04805v1

Du, Z. (2019). GPT2-Chinese: Tools for training GPT2 model in Chinese language. Retrieved January 12, 2020, from https://github.com/Morizeyao/GPT2-Chinese

Ghosh, S., Chollet, M., Laksana, E., Morency, L.-P., & Scherer, S. (2017). Affect-LM: A neural language model for customizable affective text generation. In R. Barzilay & M.-Y. Kan (Eds.), *Proceedings of the 55th annual meeting of the Association for Computational Linguistics* (Vol. 1, pp. 634-642). Association for Computational Linguistics. https://doi.org/10.18653/v1/P17-1059

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251-257. https://doi.org/10.1016/0893-6080(91)90009-T

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (PMLR 70, pp. 1587-1596). ML Research Press.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Proceedings of the 25th international conference on neural information processing systems* (pp. 1097-1105). Neural Information Processing Systems Foundation.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. *Nature*, *521*, 436-444. https://doi.org/10.1038/nature14539

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L.

Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Proceedings of the 26th international conference on neural information processing systems* (Vol. 2, pp. 3111-3119). Neural Information Processing Systems Foundation.

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. MIT Press.

Partala, T., & Surakka, V. (2004). The effects of affective interventions in human-computer interaction. *Interacting with Computers*, *16*(2), 295-309. https://doi.org/10.1016/j.intcom.2003.12.001

Prendinger, H., & Ishizuka, M. (2005). The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence: An International Journal*, *19*(3-4), 267-285. https://doi.org/10.1080/08839510590910174

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language models are unsupervised multitask learners*. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. MIT Press.

Skowron, M. (2010). Affect listeners: Acquisition of affective states by means of conversational systems. In A. Esposito, N. Campbell, C. Vogel, A. Hussain, & A. Nijholt (Eds.), *Lecture notes in computer science: Vol. 5967. Development of multimodal interfaces: Active listening and synchrony* (pp. 19-181). Springer. https://doi.org/10.1007/978-3-642-12397-9_14

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, & K. Q. Weinberger (Eds.), *Proceedings of the 27th international conference on neural information processing systems* (Vol. 2, pp. 3104-3112). Neural Information Processing Systems Foundation.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Proceedings of the 30th international conference on neural information processing systems* (pp. 6000-6010). Neural Information Processing Systems Foundation.

Zhang, Y., & Huang, M. (2019). *Overview of the NTCIR-14 short text generation subtask: Emotion generation challenge* [Paper presentation]. 14th NTCIR Conference on Evaluation of Information Access Technologies, Tokyo Japan.

Zhou, H., Huang, M., Zhang, T., Zhu, X., & Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *The thirty-second AAAI conference on artificial intelligence* (pp. 730-738). Association for the Advancement of Artificial Intelligence.

Te-Lun Yang ORCID 0000-0002-3351-1785
Yuen-Hsien Tseng ORCID 0000-0001-8904-7902

Journal of Educational Media & Library Sciences
http://joemls.tku.edu.tw