圖書資訊學刊　第19卷第1期　（2021.6）　頁43-65
doi: 10.6182/jlis.202106_19(1).043

電腦生成的新聞有多真？— 文字自動生成技術運用於經濟新聞的評估

How Genuine is Computer-generated News? — Evaluation of Automated Text Generation Applied to Economic News

曾元顯[1]　林郁綺[2]
Yuen-Hsien Tseng[1], Yu-Chi Lin[2]

摘　要

本研究以經濟新聞為範圍，探討GPT-2模型，在經過約30萬篇新聞訓練後產生15篇電腦生成之新聞（CGN），混合15篇人類撰寫之新聞（HCN），由12位受試者進行1到5分的可信度評價。結果在15篇HCN中，有1篇其平均可信度為2.92，不及3，原因為沒有邏輯、主觀性強等；在15篇CGN中，有2篇其平均可信度皆為3.33，大於3，原因為內容合理、細節符合邏輯。此2篇的部分內容與語料庫比對後，發現電腦移花接木再加潤飾的能力，已可欺騙專業人士。本研究也訓練BERT模型，以瞭解自動偵測電腦生成新聞之可能性。結果上述30篇新聞中，BERT只有2篇CGN預測錯誤，其餘皆正確，比受試者們集體的預測，有5篇錯誤，成效還要高。較大規模的實驗，顯示BERT的成效，可達0.96的F1分數。

關鍵字：電腦生成新聞、文字自動生成、新聞偵測、深度學習、人工智慧

Abstract

This research explores the GPT-2 deep learning model for economic news generation and evaluation. After training GPT-2 by about 300,000 pieces of news with a total of 150 million words, 15 news articles are generated by GPT-2. Together with 15 real news articles written by journalists, 12 subjects were invited to judge the credibility of the 30 news articles with 1 to 5 scales. As a result, 8 subjects who graduated from economic-related major were more capable of discriminating the human-composed news (HCN) from the computer-generated news (CGN); while 4 subjects who graduated from non-economic related major had poor discriminating ability, and one was even unable to tell the HCN from the CGN. Among the 15 HCN articles, 1 was rated as non-genuine news, with an average credibility of 2.92, which is less than 3, due to lack of logic and strong subjectivity. Among the 15 CGN articles, 2 were rated as genuine news, with average credibility of 3.33, which is greater than 3, because the content is reasonable and the details are logical. After comparing these two articles with the corpus, it is found that the computer's ability to substitute and retouch can deceive professionals. However, most of the CGN articles have been spotted, mainly because of obvious flaws in facts and incorrect digits such as dates and stock codes. The research also explores the possibility of automatically detecting computer-generated news using BERT-based neural network model. As a result, BERT had only 2 false predictions out of the above 30 news articles. Compared with the collective prediction by the 12 subjects with 5 errors, BERT performs better. Further large-scale experiments show that the effectiveness of BERT can reach an F-score of 0.96.

Keywords: Computer-generated News; Automated Text Generation; News Detection; Deep Learning; Artificial Intelligence

[1,2]國立臺灣師範大學圖書資訊學研究所
Graduate Institute of Library and Information Studies, National Taiwan Normal University, Taipei, Taiwan
* 通訊作者Corresponding Author: 曾元顯Yuen-Hsien Tseng, E-mail: samtseng@ntnu.edu.tw

Extended Abstract

1. Introduction

The rapid development of artificial intelligence (AI) technology has allowed computers to generate text that is indistinguishable from genuine articles. OpenAI announced the GPT (Generative Pre-trained Transformer) deep learning model in 2018, GPT-2 in 2019, and GPT-3 in 2020, which are effective techniques for generating such seemingly genuine texts. GPT-2 (GPT-2 Chinese, located at https://github.com/Morizeyao/GPT2-Chinese, was used for this study) is freely available for download, and those who install it have the ability to generate a sizeable quantity of texts, given enough training data. In the present study, we investigated the extent to which a computer can generate Chinese news articles that deceive professionals and whether such computer-generated news (CGN) could be automatically detected to prevent the spread of CGN in the future.

2. Research Questions

Having sufficiently sized corpus of satisfactory quality with which to train a computer results in notably realistic CGN. The present study investigated how people react to CGN, whether CGN can be filtered automatically, and compared responses to CGN with those to human-crafted news (HCN).

The present study used Chinese economic news to explore the following questions:

RQ1: How well do people with sufficient knowledge distinguish between HCN and CGN?

RQ2: How well can a computer automatically distinguish HCN from CGN?

Studies relevant to the preceding questions are presently scarce. Although OpenAI has not yet found evidence of the malicious application of text generation technology, we cannot rule out the possibility that some people may spread CGN widely to affect the economic market for their benefit. The aim of this research was to provide social science researchers with evidence that some CGN is as genuine as HCN and stimulate discussion of the impact on society of CGN.

3. Experiment Setups

The corpus that was used to train GPT-2 Chinese was composed of articles from the *Economic Daily News* from 2010 to 2013. The total number of news articles was 304,790 (423.2 MB or 150 million Chinese characters), and the average number of characters per news article was 519.

To answer RQ1, we trained GPT-2 Chinese for 10 epochs on a PC with Titan RTX GPU, which took a total of 75 hours; then, we let the

Note. To cite this article in APA format: Tseng, Y.-H., & Lin, Y.-C. (2021). How genuine is computer-generated news? — Evaluation of automated text generation applied to economic news. *Journal of Library and Information Studies, 19*(1), 43-65. doi: 10.6182/jlis.202106_19(1).043 [Text in Chinese].
To cite this article in Chicago format: Yuen-Hsien Tseng and Yu-Chi Lin. "How genuine is computer-generated news? — Evaluation of automated text generation applied to economic news." *Journal of Library and Information Studies* 19, no. 1 (2021): 43-65. doi: 10.6182/jlis.202106_19(1).043 [Text in Chinese].

computer generate 40,000 news articles (duration: approximately 10,000 minutes). Subsequently, we randomly sampled 90 articles from the 40,000 news articles by category and manually reviewed and selected 15 articles to be CGN articles. Similarly, 15 *Economic Daily News* articles were randomly sampled by category to serve as sample HCN articles. These 30 articles were each then trimmed to approximately 300 Chinese characters in length and were arranged in random order for later experiments.

We then recruited three groups of participants, who were divided into an expert group (economics major with master's or doctoral degree), middle group (major in economics or related fields), and general group (graduates from noneconomics departments in university). Each group included four participants for a total of 12 participants. We used the free SurveyCake service to create an online questionnaire, invited participants to read the 30 economic news articles, and asked them to judge and rate each article by using a 5-point Likert scale. The participants were not informed of the presence of CGN, but they were permitted to use the Internet to help judge the credibility of the news articles.

The questions included the following: "How do you understand this news?"; "How do you judge the credibility of this news? Why?"; "What do you think is the credibility of the news? (1–5, the higher the more credible)"; and "What is your willingness to repost/share this news?" The remaining questions regarded participant age, occupation, and time spent responding to the questionnaire.

For RQ2, 15,000 pieces of news were first generated by the GPT-2 Chinese trained on the aforementioned corpus for 50 epochs. Then, 1,500 pieces were randomly selected from the 15,000 CGN articles and 1,500 pieces were randomly selected from *Economic Daily News*, contributing a total of 3,000 articles. Among them, 900 were used as test data, 300 were used as validation data, and 1,800 were used to train machine classifiers. In each of these three data sets, HCN and CGN each accounted for 50% of the articles.

Three classifiers were trained to distinguish CGN from HCN; two were the traditional naïve Bayes and support vector machine (SVM) classifiers, and one was the Bidirectional Encoder Representations from Transformers (BERT), which is a deep neural network model that does not require manual extraction of features before classification. We trained each classifier for five epochs to classify 900 pieces of news.

In addition, to compare the performance of BERT with human's, we also applied BERT to predict the above 30 news articles investigated in RQ1.

4. Results and Discussion

4.1 Results and discussions for RQ1

Among the 12 participants, their ages ranged from 23 to 42 years, with an average of 30. The average time to read and evaluate 30 news articles, after deducting one outlier (who took approximately 33 hours), was 5,390 seconds, which is less than 1.5 hours.

In evaluating the credibility of the 30 news articles, the general group demonstrated the worst ability to discriminate; the average credibility ratings for HCN and CGN were 3.33 and 2.86, respectively, with a gap of only 0.47. The middle group had the best discriminative ability: the average HCN and CGN credibility ratings were 3.83 and 1.95, respectively, with a gap of 1.88. The

average credibility ratings among the expert group were 3.83 for HCN and 2.83 for CGN, with a gap of 1.0.

The poor discriminative ability of the general group may be attributed to their lack of background knowledge in economics, and this is consistent with our expectations. The expert group underperformed relative to the middle group. We contend that this may be due to the expert group's larger tolerance for imperfect news because its judgment of the credibility of HCN was at least the same as that of the middle group (3.83), whereas the rating for CGN was higher (2.83).

Of the 15 HCN articles, one article was rated as nongenuine news, and its average credibility rating was 2.92 due to its lack of logic and strong subjectivity. Among the 15 CGN articles, two articles were rated as genuine news, with an average credibility of 3.33, because the content was reasonable and the details were logical.

In one of the two most overestimated CGN articles, the text strings generated by the computer, such as "the current monetary policy of the global central bank," "mainly maintaining loose policies," and "Liu Lingjun believes that in a low interest rate environment," do not appear in the corpus of the 304,790 news articles at all. However, when they were reduced to shorter strings, they matched the text strings that appeared in the corpus: "the monetary policy of the global central bank" appeared in three news articles, "maintaining loose policies" appeared in 76 articles, "Liu Lingjun believes" appeared in 48 articles, and "in a low interest rate environment" appeared in 78 articles. Clearly, the computer's ability to substitute and retouch text strings to generate corpus-relevant articles can deceive professionals. However, most CGN articles were identified mainly due to their obvious flaws in facts and incorrect digits such as dates and stock codes.

4.2 Results and discussions for RQ2

Of the 900 test articles, the naïve Bayes classifier only achieved a 0.73 F1-score (a harmonic average of precision and recall) in predicting whether an article was CGN or HCN. Additionally, SVM achieved a 0.8 F1-score, but the effectiveness of BERT was as high as 0.96.

In predicting the 30 human-evaluated news articles, BERT had only two false predictions. By contrast, the collective predictions of the 12 participants contained five errors. BERT still performed better in this case. The classification experiments indicated that machines such as BERT may feasibly distinguish CGN from HCN.

5. Conclusion

News articles that can deceive professionals can be generated by using a sufficiently large text corpus and freely available software. Although most CGN have some flaws that can be identified by humans, some of the CGN articles were difficult to distinguish. If the CGN articles were to undergo human editing, whether humans or machines could distinguish between the edited CGN articles and HCN articles remains unknown.

How BERT predicted unedited CGN and HCN so accurately is also unknown. BERT works without extracting text features; this leaves us no clues for mechanically determining whether an article is CGN or not.

Future studies should address the aforementioned unknowns to manage daily news in the dawning AI age.

Acknowledgement

This work is supported by Ministry of Science and Technology with grant numbers: MOST 109-2410-H-003-123-MY3.

壹、前言

近年來，部落格、共享協作平台與社群網路（如：臉書、微博、Instagram等）的興起，網路媒體已經進入自媒體時代。閱聽眾不再只是資訊的被動接收者，透過社群平台進行訊息的獲取、分享、生產與傳播，每個人都具有訊息創作及傳播的能力，讓言論、知識快速流通。但訊息發佈門檻的降低，也導致不實消息的快速增長。不僅如此，近年來人工智慧技術的快速進展，自然語言生成技術可讓電腦產生擬真的文字、新聞等等，恐將更為加速假新聞的產製與傳播。身處資訊爆炸時代的現代人，具備區辨訊息真假、媒體識讀的能力，將越來越困難，卻也越來越重要。

以圖一為例，其以2010到2013年約30萬篇的經濟新聞語料庫，訓練電腦後，以圖中底線真實新聞的前導文字，讓電腦依前導文字產生出約300字的文稿。其中開頭的「據了解，金控」還出現在語料庫中，但「據了解，金控旗下」這連續八個字，以及最後的「參股對象不能只限一家」與「不能只有一家」，在30萬篇新聞中，都沒有出現過。顯示電腦能夠產生新的語句、文稿，且其擬真程度，以研究者的知識，無法判定其內容為真或假。

面對這樣的新聞，在不知其為電腦撰寫的情況下，多數人恐信以為真，進而參採，甚至分享或引用。這對於社會上的經濟行為、學術倫理等方面，恐造成傷害。

若吾人得知其為電腦撰寫之新聞，且認為電腦在讀過多篇新聞後這只是電腦寫的摘要，那麼要進一步求證其為真，幾乎需對每一句內容都求證，如：「包括台灣銀行、土地銀行、合作金庫銀行、第一銀行及華南銀行等，都是透過金控子公司，參股大陸銀行」，以及每個數字如：「參股對象不能只限一家」，都要進行求證，才能放行這則摘要。

現今，我們不僅面臨人為的假新聞，也面臨電腦產製的擬真新聞，而後者可以

根據金管會規劃，同一金控下的銀行，跟金控可以擇一參股大陸銀行，且參股總額度分開計算，金控「進可攻、退可守」，登陸投資彈性最大，成為兩岸金融開放後最大受惠者。據了解，金控旗下的銀行子公司，包括台灣銀行、土地銀行、合作金庫銀行、第一銀行及華南銀行等，都是透過金控子公司，參股大陸銀行。至於大陸銀行來台參股，銀行業者表示，參股對象是否有經營權，還是要看金管會規劃，但是參股上限是否放寬，還要看兩岸主管機關的規劃。金管會官員昨（7）日表示，目前尚未接到金管會的回應，但是，金管會已指示相關單位研議，將來可以研究。至於參股金控，銀行業者也說，目前還沒有具體的計畫，只能透過公司治理來解決。金管會官員表示，金控旗下子公司的投資架構，包括銀行、保險、證券、創投及證券等子公司，都是透過金控旗下證券子公司，投資大陸地區有價證券，因此，參股對象不能只限一家，不能只有一家。

圖一　電腦生成的經濟新聞範例

在極短時間內大量生產，光憑人力難以查證、過濾與阻絕。因此，探究自動化方式能否辨別電腦產生的新聞，有其實際上的需求。

本研究以經濟新聞為範圍，探討電腦在經過訓練後所產生的擬真新聞，受試者對其可信度的評價。另外，也以實驗檢驗自動偵測電腦生成新聞的成效，以瞭解自動化的方式是否可以幫忙過濾這類新聞。

下一節將介紹文字自動生成技術與工具，第三節說明本文的研究問題，第四節詳述實驗流程，第五節為實驗結果與討論，最後一節總結本研究，並提出未來進一步探討的方向。

貳、文獻探討

本節先介紹假新聞相關的概念，繼而簡介人工智慧的演進與文字自動生成技術，以做為本研究的背景知識。

一、假新聞相關研究

「假新聞」一詞目前並沒有明確一致的定義。Allcott與Gentzkow（2017）認為假新聞是刻意扭曲並可被證實錯誤的新聞文章。而更廣泛的定義是針對新聞內容的真實性或意圖，部分研究認為諷刺新聞屬於假新聞，因為儘管諷刺作品通常以娛樂為導向，並且對讀者顯示其娛樂性，但內容卻常是非真實或是過於誇大的。其他文獻則是直接將欺騙性新聞視為假新聞，其中包括嚴重的捏造、騙局和諷刺（Allcott & Gentzkow, 2017）。

假新聞本身並不是一個新問題。媒體生態隨著時間的推移，假新聞的媒介已從紙本、廣播、電視，轉變為網路新聞和社群媒體。Shu、Sliva、Wang、Tang與Liu（2017）認為假新聞在傳統新聞媒體上的樣態分為「心理基礎」與「社會基礎」。在心理基礎上，由於人性固有的認知偏見，假新聞一旦被讀者視為真實消息，便很難糾正它。而在社會基礎上，社會認同感對於個人的身份和自尊至關重要，即使來源是假新聞或者是證據不足的新聞，讀者也可能會為了社會認同感而選擇轉發訊息。

Shu等人（2017）也認為，現代社群媒體更多了「惡意帳戶」與「迴聲室效應」（Echo Chamber Effect）的影響。社群媒體上有些用戶可能含有惡意，在某些情況下甚至不是真正的人類，而創建社群媒體帳戶的低成本，也鼓勵了惡意使用者。另一方面，社群媒體的用戶傾向組成志同道合的群體，因社群媒體的推薦常導致意見兩極化，從而產生迴聲室效應，有時亦俗稱為「同溫層效應」。迴聲室效應透過下面兩項心理因素促使用戶相信假新聞：社會公信力與頻率啟發（Frequency Heuristic）。社會公信力意味著如果他人認為來源是可信的，那人們更有可能隨之將來源視為可信的，尤其是當沒有足夠的訊息可以確認真實性的時候；頻率啟發則意味讀者會自然偏愛他們經常聽到的訊息，即使這是假新聞。

在上述個人心理、社會制約以及社群影響下，假新聞的產製或散播比以往

任何時期，都更加快速氾濫。偵測假新聞的相關研究，也在近五年蓬勃發展，如ByteDance（2019）、Conroy、Rubin與Chen（2015）、Ruchansky、Seo與Liu（2017），以及Shu、Cui、Wang、Lee與Liu（2019）等。但目前多數以網路上蒐集到且經由人工標記的假新聞做為偵測對象，尚少針對電腦生成的假新聞，進行自動辨識。

二、人工智慧、深度學習之興起

人工智慧（Artificial Intelligence, AI）的研究目的，在建構具備智慧的機器，特別是建構具備智能的電腦程式（McCarthy, 2007）。1950年Turing提出了一種測試機器是否展示智慧表現而令人無法區別的操作型方法（Turing, 1950），現稱為Turing Test，協助形塑了機器智能的研究目標。1956年在Dartmouth學院舉辦為期二個月的工作坊，McCarthy首度使用了人工智慧這個詞彙，正式開啟了AI領域的時代（Russell & Norvig, 2009）。

近十年來由於電腦運算能力持續變強、可用的訓練資料變多，以及機器學習演算法的進步，使得AI的技術與應用有突破性進展。DeepMind的圍棋程式在2016與2017年陸續打敗日本、中國的圍棋高手，更促成各國紛紛投入AI的研究與投資（賴志遠，2018）。

前述AI的進展，幾可歸功於深度學習（deep learning）（LeCun, Bengio, & Hinton, 2015）的發展與運用。深度學習係指多層人工神經網路的學習與應用。以影像辨識為例，2012年Krizhevsky、Sutskever與Hinton（2012）採用8層的神經網路，達到比非使用多層神經網路的第二名影像辨識錯誤率26.2%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（deep neural network）或深度學習（deep learning）。2017年Vaswani等人（2017）提出了堆疊多層的Transformer編解碼器深度神經網路架構，在自然語言處理領域，也開啟了類似影像處理那樣突飛猛進的時代。

三、文字生成技術

2018年OpenAI公司提出GPT（Generative Pre-trained Transformer）模型（Radford, Narasimhan, Salimans, & Sutskever, 2018），其為12層Transformer解碼器疊加的深度神經網路，可以學習並準確估計下一個字詞的條件機率函數：

$$P\left(w^{(t)}\right) \mid w^{(t-1)}, \ldots, w^{(1)} \qquad (1)$$

亦即假設給予GPT前t-1個字詞 $w(1)$, $w(2)$, … , $w(t\text{-}2)$, $w(t\text{-}1)$，讓其預測下個 $w(t)$ 應該是那個字詞機率最高，則GPT比傳統方法最大似然估計（maximum likelihood estimation）或是較早出現的LSTM（Long-Short Term Memory）（Hochreiter & Schmidhuber, 1997）的準確度還高。簡言之，GPT的架構可學出非常優良的語言模型（language model）。

由於以神經網路為基礎的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層都有，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），效果好到開源釋出有所顧慮。所幸，其網站：https://openai.com/blog/gpt-2-1-5b-release/上公告，尚未發現有惡意應用GPT-2的強烈證據。

2020年5月28日OpenAI發表了GPT-3（Brown et al., 2020），其最大的模型有96層，可學習的參數有1,750億之多，使用的訓練資料將近45TB。其效果又比GPT-2好很多，可解決很多自然語言處理的問題。但由於模型太大，一般研究者無法有足夠的硬體設備可執行，GPT-3僅以申請其API使用的方式釋出。

本研究採用Du、Cheng、Chiu與Yida（2019）改版的GPT2-Chinese（https://github.com/Morizeyao/GPT2-Chinese），以用來學習中文的新聞語料，從而自動生成中文新聞。有關GPT的運作原理、學習方式以及執行範例，可參考楊德倫與曾元顯（2020）文章，在此不多贅述。

參、研究目的、問題與重要性

上述的相關研究顯示，有足夠大量、品質優良、主題一致性高的語料訓練下，電腦生成的文字或新聞（Computer Generated News, CGN）已相當擬真。未來，面對可能大量出現的CGN，能否自動過濾，以及人們反應如何，為本研究探討的主要議題。

具體而言，本研究擇定經濟新聞為主題範圍，對於人類撰寫之新聞（Human Composed News, HCN）與電腦生成之新聞（CGN）進行辨別，擬探討如下問題：

一、具備相當知識程度的民眾，辨別HCN與CGN的程度如何？

二、電腦能夠自動辨別的程度如何？

對研究問題一的探索，可以瞭解中文CGN的擬真情形，達到什麼程度，是否人為可控。對研究問題二的探究，可以得知自動化的作法，可否有效過濾未來大量的CGN。當然，這兩個問題都很大，牽涉的因素都很廣，包括控制GPT-2的訓練次數或是產生文章時的自由度：較貼近訓練資料，還是較不受控制而有創意。本研究僅就既有的計算資源，對某些參數以立意抽樣方式探究，希望所得結果對後續的研究能有所啟發，並提供部分例證。

雖然OpenAI尚未發現有惡意應用文字生成技術的強烈證據，但若誘惑極大，有心人士運用相關技術，從數據上造假或曲解數據，由此帶動市場情緒，將可能造成自由經濟市場的負面影響。此種最壞情況，仍不得不防範於未然。然而，目前相關研究非常稀少，本研究希望能喚起更多的探討，提供人文社會研究者瞭解人工智慧中文發展近況，並進而共同探究其可能的影響。

肆、實驗設計

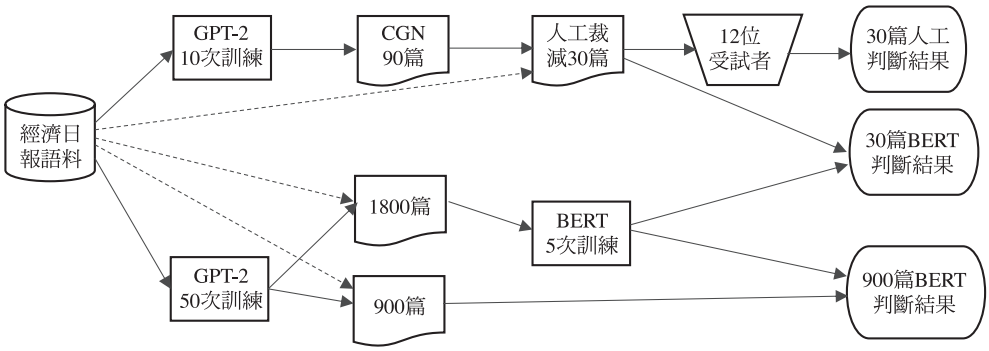
為探究上述兩問題，本研究進行如圖二的實驗步驟。圖二上面的流程為人工評估CGN、下面的流程為機器評估CGN，最左邊為用以訓練電腦生成新聞的語料庫，是2010到2013年共四年的經濟日報新聞抽樣版（每日新聞抽樣數量約70%），其總數計有304,790篇，共423.2 MB，約1.5億字，平均每篇新聞字數為519字。

本研究選擇經濟新聞為訓練語料，主因為其具有一定的知識量，亦即產生一篇經濟新聞，需具備相關的背景知識，才具有可信度。在多數人恐還不易接受電腦將可完全自動生成時，CGN適合跟記者寫的新聞，混合一起提供評估，用以對比出受試者的反應。另一方面，也由於需要具備一定程度的知識，才能看懂經濟新聞，因此辨別HCN與CGN，有一定的門檻，可做為電腦自動辨別HCN與CGN的好材料。

選定語料庫後，下一個步驟是訓練電腦，以便產生新聞。在具備Titan RTX的GPU（記憶體24GB）個人電腦上，採用GPT-2 Chinese語言建模程式，訓練423.2 MB語料一次所需的平均時間約7.5小時，訓練完後平均每分鐘可產生4筆新聞。

針對研究問題一，我們對電腦訓練了10次，總共花費了75小時，然後讓電腦產生四萬筆新聞（約一萬分鐘）。參考經濟日報官方網站之分類：金融、證券、期貨、商情、產業、理財、兩岸、國際共8類及其細類，從中按類隨機抽樣90篇，再以人工逐筆檢視、挑選、修剪出15篇，做為電腦生成之新聞（CGN）。同樣按類隨機抽樣經濟日報15篇出來（如圖二虛線），做為人類撰寫之新聞（HCN）。由於新聞文章長短不一，統一修剪成300字左右的文章後，將這30篇新聞隨機排列順序，做為實驗新聞語料。其內容主題與類別，如表一所示。

由於自動生成的文章，行文並非完美無瑕，中間偶有重複、疊句情形，這些明顯的現象，研究者視情況將其刪除。例如，題號13、原序號29：「包括六福



圖二　研究步驟流程

表一　用於人工評估的30篇新聞

題號	分類	內容主題	類別	序號	原序號
1	兩岸／兩岸焦點	阿里巴巴旗下淘寶商城	CGN	1	21
2	產業／政經大事	馬總統	CGN	2	30
3	兩岸／兩岸焦點	上海書展	HCN	1	10
4	理財／銀行保險	畢業生申辦信用卡	CGN	3	19
5	國際／國際焦點	印度女性銀行舉行開幕典禮	HCN	2	15
6	兩岸／陸港行情	香港股市局勢	HCN	3	5
7	產業／企業CEO	經濟部全球招商論壇	HCN	4	2
8	金融／金融脈動	壽險業投資不動產資金	CGN	4	27
9	商情／綠色產業	太陽能產業	CGN	5	23
10	金融／金融脈動	第一銀行亞太布局	HCN	5	12
11	國際／國際焦點	超級颶風重創紐約	HCN	6	11
12	國際／國際焦點	新興貨幣與債券市場趨勢	HCN	7	9
13	證券／營收快報	春節客房業務	CGN	6	29
14	兩岸／兩岸焦點	中國股市房地產	CGN	7	24
15	商情／綠色產業	國際綠色環保建材	HCN	8	4
16	產業／產業熱點	長榮航空貨運	CGN	8	20
17	金融／外匯市場	美元利率	CGN	9	28
18	金融／外匯市場	政治衝突影響貨幣寬鬆政策	CGN	10	18
19	商情／產學研訓	就業訓練	CGN	11	22
20	期貨／期貨市場	國際玉米小麥價格	CGN	12	26
21	商情／建材新訊	地板木材訊息	HCN	9	3
22	產業／產業熱點	台灣農林投資據點	CGN	13	16
23	證券／市場焦點	半導體產業股市趨勢	CGN	14	25
24	國際／國際焦點	日本311地震後經濟	HCN	10	6
25	商情／CSR.	高齡志工公益活動	HCN	11	1
26	產業／產業熱點	飛利浦全自動咖啡機租賃優惠方案	HCN	12	13
27	國際／國際焦點	希臘國債影響全球股債市	HCN	13	7
28	兩岸／兩岸焦點	中國房地產行情	HCN	14	8
29	商情／產學研訊	全國學生美展	CGN	15	17
30	商情／熱門亮點	台灣參與澳門發明獎	HCN	15	14

註：序號為HCN或CGN中第幾篇的序號，為後續圖表使用；原序號為人工檢視合併後的30篇新聞序號，供對照內文使用。

（2705）、晶華（2707）、王品（2707）、新天地（8940）、王品（2727）、夏都（2722）」，我們將第二個王品及其股票代碼刪除，再交給受試者評估（事後發現第二個王品其股票代碼才正確，第一個反而重複了晶華的正確股票代碼）。在15篇CGN中，此種微幅刪減的情形，只有4篇（原序號為：16、17、23、29），其餘11篇，都直接採用其前300字左右截至句號，而不再修剪。

本實驗於社交平台上招募三組受試者，分為：專家組（碩博士經濟相關系所畢業）、中等組（大學經濟相關系所畢業）以及一般組（大學非經濟相關系所畢業），每一組分別招募4人，共12人。實驗方式為利用免費的SurveyCake服務製作線上問卷，邀請受試者閱讀上述30篇經濟新聞，針對每篇新聞以五等尺度進行判斷，並註記相關判斷關鍵因素。受試者並未被告知有電腦產生之文章在其中，但可使用網路資源輔助判斷新聞的可信度。

受試者需回應的問題如下：「您對於這篇新聞的理解程度是如何的？（1為完全不理解；5為完全理解）」、「您判斷這篇新聞可信度的判斷關鍵字句為何？」、「您認為這篇新聞的可信度是如何的？（1為完全不可信；5為完全可信）」、「您轉發這篇新聞的意願度為何？（1為完全不願意；5為非常願意）」、「承上題，原因為何？」，最後調查受試者之年齡、職業與作答時間。

針對研究問題二，由於有先前15篇修剪其中4篇重複文句的經驗（比例4/15 = 26.7%），我們便將訓練次數，調高到50次，以期降低重複文句的比例、產出更順暢的文章。如此，先產生15,000篇新聞，再從中隨機篩選出1,500篇，加上從經濟日報隨機抽取的1,500篇（如圖二虛線），組成3,000篇語料。其中900篇拿來當測試資料（test set），2,100篇拿來當訓練資料（training set），而在訓練資料中，又拿其中的300篇當作訓練時的驗證資料（validation set），以瞭解訓練過程中，是否有過度擬合（overfitting）的現象。因此，實際的訓練資料為1,800篇。而不論是訓練資料、驗證資料或是測試資料，人類撰寫之新聞（HCN）與電腦生成之新聞（CGN）的比例都各佔50%。

我們除了應用傳統的Naïve Bayes、SVM分類器外，也使用近兩年來分類效果非常優良的BERT（Bidirectional Encoder Representations from Transformers）（Devlin, Chang, Lee, & Toutanova, 2018）。BERT跟GPT類似，是基於Transformer架構的深度神經網路，只是GPT使用Transformer的解碼器架構，BERT使用Transformer的編碼器架構，且BERT的訓練方式是基於遮罩式的語言模型（masked language model），也跟GPT不同。

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，我們以上述1,800篇訓練資料，再訓練5次，即用來對900篇新聞做分類。由於研究問題一中已有人工評估過的30篇新聞，我們也用BERT進行HCN與CGN的預測。為比較成效，Naïve

Bayes、SVM分類器也運用相同的資料集，進行訓練、預測。

伍、實驗結果與討論

一、研究問題一之實驗結果與討論

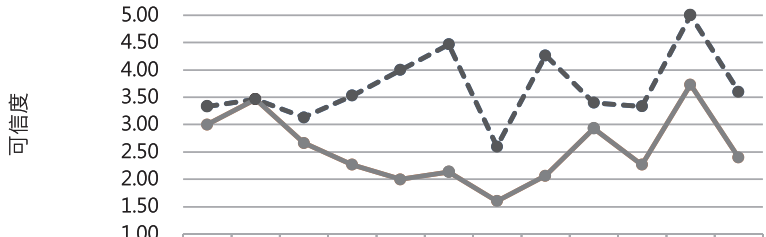
針對研究問題一的實驗，12位受試者年齡最低為23歲、最高是42歲，平均年齡為30歲。其閱讀、評估30篇新聞的平均時間，在扣除其中一位的異常值（花費118,474秒，約33小時）後，為5,390秒，一般組為4,992秒，中等組為6,531秒，專家組為4,647秒，亦即大都在2小時內完成。

圖三顯示12位受訪者評估HCN與CGN新聞的可信度統計。可看出中等組L、H與G三位受訪者對於HCN都能給予4以上的分數，CGN給予接近2的分數，明顯能辨別HCN與CGN。而一般組受訪者I則幾乎無法辨別，可信度平均分數都一致。專家組受訪者Y的給分偏高，但對於15篇HCN，都能正確判斷，都給予最高的5分，對CGN則平均給予3.73分。

圖四則為分組統計。其中一般組的鑑別度最差，因其HCN可信度評估為3.33，CGN為2.86，差距0.47最小；中等組的鑑別度最好，其HCN平均可信度評估為3.83，CGN為1.95，差距1.88最大；專家組的HCN平均可信度為3.83，CGN平均可信度為2.83，差距1.0。

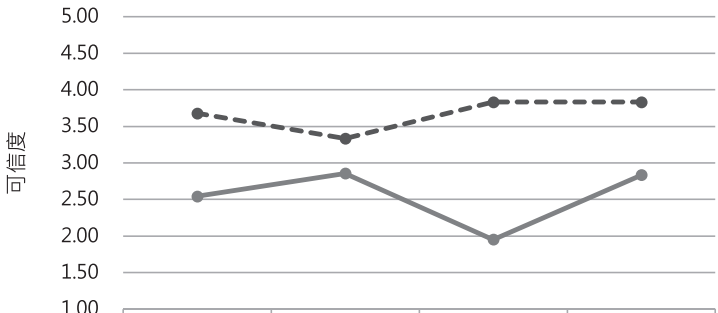
一般組可能因為較無經濟相關背景知識，致鑑別能力較差，這與我們的預期一致。而專家組在鑑別度的表現不如中等組，推測原因可能是專家組對於新聞誤謬的容許空間較大，因其對HCN的可信度判斷，至少跟中等組一致，都是3.83，對CGN則給予較寬容的高分。

圖五為各篇之理解程度、可信度與轉發意願的比較圖，可以看出三者有一定程度的關係：理解程度大都高於可信度、轉發意願；可信度也都高於轉發意願；轉發意願最保守，三者呈現高度連動關係。



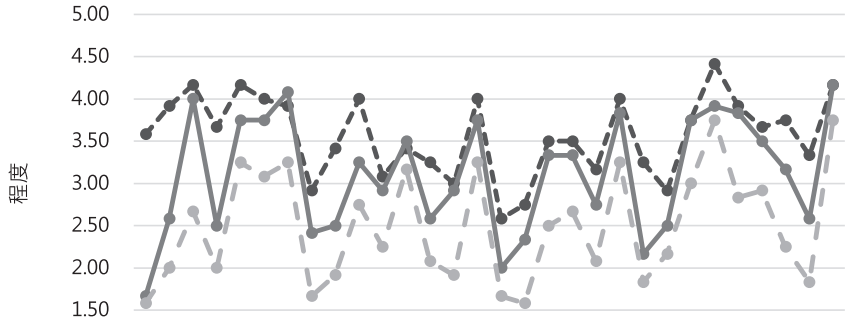
	U 一般	I 一般	W 一般	S 一般	L 中等	H 中等	M 中等	G 中等	D 專家	T 專家	Y 專家	Z 專家
HCN可信度	3.33	3.47	3.13	3.53	4.00	4.47	2.60	4.27	3.40	3.33	5.00	3.60
CGN可信度	3.00	3.47	2.67	2.27	2.00	2.13	1.60	2.07	2.93	2.27	3.73	2.40

圖三　12位受試者評估HCN與CGN的平均可信度



	平均	一般	中等	專家
HCN	3.68	3.33	3.83	3.83
CGN	2.54	2.86	1.95	2.83

圖四　12位受試者評估HCN與CGN的分組平均可信度



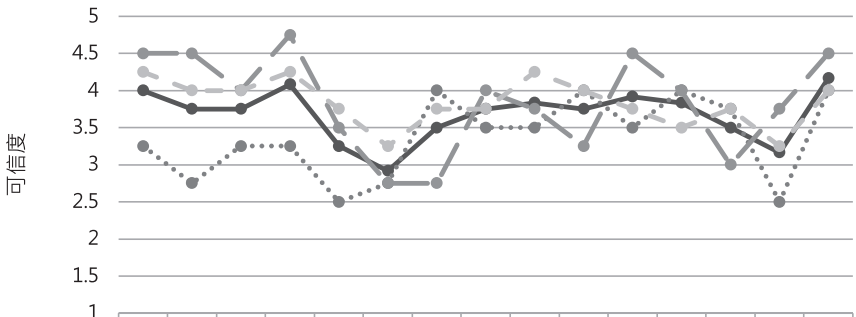
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
理解程度	3.5	3.9	4.1	3.6	4.1	4.0	3.9	2.9	3.4	4.0	3.0	3.4	3.2	3.0	4.0	2.5	2.7	3.5	3.5	3.1	4.0	3.2	2.9	3.7	4.4	3.9	3.6	3.7	3.3	4.1
可信度	1.6	2.5	4.0	2.5	3.7	3.7	4.0	2.4	2.5	3.2	2.9	3.5	2.5	2.9	3.7	2.0	2.3	3.3	3.3	2.7	3.8	2.1	2.5	3.7	3.9	3.8	3.5	3.1	2.5	4.1
轉發意願	1.5	2.0	2.6	2.0	3.2	3.0	3.2	1.6	1.9	2.7	2.2	3.1	2.0	1.9	3.2	1.6	1.5	2.5	2.6	2.0	3.2	1.8	2.1	3.0	3.7	2.8	2.9	2.2	1.8	3.7

圖五　各篇之理解程度、可信度、轉發意願比較表

圖六及圖七顯示各篇新聞經由12位受試者評估後的平均可信度。

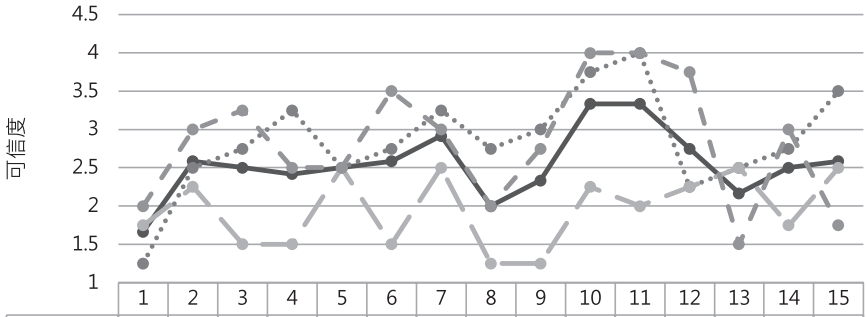
圖六中平均可信度最被低估的HCN，為序號第6篇（原序號11），平均可信度為2.92，其內容如圖八。

對於此篇HCN，12位受訪者中有1位給出了可信度為1的低分，認為這篇新聞完全不可信；認為可信度為2分的受訪者有4位；可信度為3分的有3位；4分的有3位；僅有1位受訪者認為這篇新聞非常可信，給出5分



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
平均	4	3.75	3.75	4.08	3.25	2.92	3.5	3.75	3.83	3.75	3.92	3.83	3.5	3.17	4.17
一般	3.25	2.75	3.25	3.25	2.5	2.75	4	3.5	3.5	4	3.5	4	3.75	2.5	4
中等	4.5	4.5	4	4.75	3.5	2.75	2.75	4	3.75	3.25	4.5	4	3	3.75	4.5
專家	4.25	4	4	4.25	3.75	3.25	3.75	3.75	4.25	4	3.75	3.5	3.75	3.25	4

圖六　12位受試者評估各篇HCN的平均可信度



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
平均	1.67	2.58	2.5	2.42	2.5	2.58	2.92	2	2.33	3.33	3.33	2.75	2.17	2.5	2.58
一般	1.25	2.5	2.75	3.25	2.5	2.75	3.25	2.75	3	3.75	4	2.25	2.5	2.75	3.5
中等	1.75	2.25	1.5	1.5	2.5	1.5	2.5	1.25	1.25	2.25	2	2.25	2.5	1.75	2.5
專家	2	3	3.25	2.5	2.5	3.5	3	2	2.75	4	4	3.75	1.5	3	1.75

圖七　12位受試者評估各篇CGN的平均可信度

滿分的可信度。給出最低分可信度的受訪者Z（專家組）認為：

「這篇新聞內容沒有邏輯，將交易量下降的原因完全歸究在紐約休市上並擷取分析師之片面之詞，意圖佐證記者自己的看法。」（受訪者Z／專家組）

超級颶風珊蒂不但重創紐約，也掃到歐股、原油市場、航空業與保險公司，對經濟的影響遠超過風暴本身的驚人威力。Global Equities證券交易員馬卡依斯說：「當紐約休市，巴黎的成交量約減少40%。真正的引擎仍是華爾街」。德國法蘭克福股市29日的交易量不到20億歐元（26億美元），幾乎是上周平均值的一半；巴黎股市交易量不到10億歐元，接近去年12月底創下的低點紀錄。歐股30日稍有起色，ETX資本公司分析師席迪基把它歸因於企業獲力強勁，而且「珊蒂對紐約市造成的最嚴重災害已經過了，華爾街31日恢復交易」。受航空企業財報佳音與亞股走揚激勵，歐股31日走揚，道瓊歐洲Stoxx 600指數盤中漲0.25%，為連續第2天上漲。席迪基表示，災情可能影響運輸、物流、工業與保險類股。

圖八　可信度最被低估的HCN內容

從理解程度給出了5分滿分來看，受訪者Z完全理解該篇新聞內容，但認為太過主觀的新聞內容可信度較低，且不願意轉發此新聞（轉發意願度僅為1分）。另外，給出可信度為2分的4位受訪者分別認為：

「這篇新聞內容跟之前看過的相關報導認知不同。」（受訪者U／一般組）

「內文『當紐約休市，巴黎的成交量約減少40%。真正的引擎仍是華爾街』看不懂紐約與巴黎的關聯，也不懂『引擎』的語意。」（受訪者W／一般組）

「新聞一開始說航空業受到颶風的負面影響，後面又說航空業財報佳音激勵，前後矛盾，不知所云。」（受訪者H／中等組）

「超級颶風珊蒂不但重創紐約，也掃到歐股、原油市場、航空業，但又受航空企業財報佳音與亞股走揚激勵，歐股31日走揚？無法理解文義。」（受訪者G／中等組）

給予最高可信度5分的受訪者則有不同看法，認為天氣災難確實可能影響股票，並且相當有意願轉發此新聞（轉發意願度為5分）：

「『超級颶風珊蒂不但重創紐約，也掃到歐股、原油市場、航空業與保險公司，對經濟的影響遠超過風暴本身的驚人威力。』——這篇新聞說明災難可能造成股票之損失，也說明財報公佈日對股票之影響。」（受訪者Y／專家組）

由受訪者的回饋可以看出，即便是HCN也可能被誤判成為CGN，其原因可能是寫作方式太過主觀或者前後文主題不同，而且專家之間彼此的看法，也會有極大的差異。

圖七顯示平均可信度最被高估的CGN，為序號第10與11篇（原序號18、22），可信度均為3.33。圖九顯示序號10的新聞內容（此篇原序號為18的內容，完全沒有人工刪減）。

對於此篇CGN，12位受訪者中有2位給出了可信度為5的高分，認為這篇新聞非常可信；可信度為4分的受訪者有5位；3分的

政治衝突事件不僅對全球股市造成衝擊，根據lipper統計，債券市場中偏向風險性資產的高收益債、新興市場債，在經歷政治事件衝突後一個月，平均表現亦難逃疲軟命運。摩根富林明債券產品策略長劉玲君指出，全球央行持續寬鬆政策，資金持續流入收益率較高的投資等級債券，加上企業獲利穩健且違約率維持低檔，將有助於債券價格續漲。劉玲君分析，目前全球央行的貨幣政策，主要是維持寬鬆政策，不會讓美國公債殖利率維持在目前水準，且在美國經濟溫和成長環境下，資金可望持續流向收益率較高的債券。劉玲君認為，在低利率環境下，投資級債券具備相對較佳的收益率，且利率風險相對較小，投資人若想在震盪環境中，尋求較佳的收益率，將是最佳選擇。

圖九　可信度最被高估的CGN內容（原序號18）

註：此篇的第一句前導文字，可在網路上查到：https://www.moneydj.com/KMDJ/Blog/BlogArticleViewer.aspx?a=09a03094-02b4-4eb1-8982-000000008934

有2位；2分的有1位；有2位受訪者認為這篇新聞完全不可信，給出1分最低分的可信度。給出可信度滿分的受訪者Y（專家組）在轉發意願的部分僅給出3分，受訪者Y認為該篇新聞可信度非常高，但投資人還是必須自行評估投資方針；另一位給出可信度5分的受訪者G（中等組）則認為該篇新聞不只可信度非常高，也相當願意轉發這篇新聞，轉發意願給出5分滿分：

> 「『目前全球央行的貨幣政策，主要是維持寬鬆政策，不會讓美國公債殖利率維持在目前水準，且在美國經濟溫和成長環境下，資金可望持續流向收益率較高的債券。』—這段話看起來相當合理，所以這篇新聞應該是真新聞，不過雖然說提供相關投資建議，但是投資人必須蒐集更多訊息，審慎評估投資意願。」（受訪者Y／專家組）

> 「『風險性資產的高收益債、新興市場債，在經歷政治事件衝突後一個月，平均表現亦難逃疲軟命運。…收益率較高的投資等級債券，加上企業獲利穩健且違約率維持低檔，將有助於債券價格續漲。…劉玲君認為，在低利率環境下，投資級債券具備相對較佳的收益率，且利率風險相對較小，投資人若想在震盪環境中，尋求較佳的收益率，將是最佳選擇。』—從這段話看起來，邏輯看起來好像沒有什麼問題，應該是可以相信的。」（受訪者G／中等組）

要注意的是，「目前全球央行的貨幣政策」、「主要是維持寬鬆政策」、「劉玲君認為，在低利率環境下」等這些語句，根本沒有出現在30萬篇新聞語料中。而若縮減成「全球央行的貨幣政策」則出現在3篇新聞中、「維持寬鬆政策」出現在76篇中、「劉玲君認為」出現48篇，「在低利率環境下」出現78篇。顯然，電腦移花接木，再加以潤飾的能力，已經可以欺騙大學程度且具專業領域知識的成年人了。

另外認為這篇新聞完全不可信的受訪者有兩位，分別有不同理由，且皆不願意轉發，轉發意願度僅有1分：

1. lipper統計，通常都是大寫Lipper
2. 摩根富林明債券沒聽過，只聽過摩根大通、富蘭克林
3. 全球央行的貨幣政策描述也有點奇怪（受訪者L／中等組）

「對債券市場不瞭解，所以也無法確定新聞可信度。」（受訪者M／中等組）

事實上，「摩根富林明債券」是存在的，且在123篇新聞中出現過，是受訪者L知識不足，亦未查證，且其第3點也與前面的受訪者Y（專家組）看法不同。

另一篇平均可信度同樣為3.33的CGN（序號第11篇，原序號22）其內容如圖十（此篇內容完全沒有人工刪減）。

對於此篇CGN，12位受訪者中有2位給出了可信度為5的高分，認為這篇新聞非常可信；可信度為4分的有6位；3分的有1位；有3位受訪者認為這篇新聞完全不可信。給出可信度滿分的受訪者Y（專家組）不太願意轉發此文（轉發意願度為2分），其認為內容中有主題不太關聯之處。另一位給出5分的受訪者H（中等組）認為該篇新聞不只可信度非常高，也相當願意轉發這篇新聞幫助需要求職的朋友，轉發意願給出5分滿分：

「『為提升勞工對產業及就業的了解，推動多元職訓課程，讓職訓中心有系統性、專業性與實務性的訓練課程，讓勞工能充分了解職業訓練內容、職業訓練及就業情形，並協助勞工職訓中心提供訓練資訊、職訓e網、勞委會職訓e網等，提供企業更多元的訓練資訊，同時也藉此提升職訓機制的正確性，提升就業市場競爭力。』—前段是提到如何加強勞工技能，後段是提到如何提升競爭力，兩者主題不同，看起來好像有點關聯，但又不太確定。」（受訪者Y／專家組）

為因應貿易自由化，加強輔導各產業從業人員參加相關訓練，來提升工作知識技能與就業能力，持續提升勞工職場能力。勞委會職訓局局長林三貴表示，為提升勞工對產業及就業的了解，推動多元職訓課程，讓職訓中心有系統性、專業性與實務性的訓練課程，讓勞工能充分了解職業訓練內容、職業訓練及就業情形，並協助勞工職訓中心提供訓練資訊、職訓e網、勞委會職訓e網等，提供企業更多元的訓練資訊，同時也藉此提升職訓機制的正確性，提升就業市場競爭力。職訓局長林三貴指出，職訓中心提供企業完善教育訓練及資源，提升就業力，同時也提升職場競爭力，對企業而言，可提供更多訓練機會，增加企業競爭力。（吳青常）為提升國內企業在國際市場競爭力，外貿協會將於11月24日於台北國際會議中心舉辦「2010年新興市場採購夥伴大會」，由外貿協會秘書長趙永全親自主持，邀請全球知名品牌及零組件供應商共襄盛舉，讓台灣品牌在新興市場的發展更具國際競爭力。

圖十　可信度最被高估的CGN內容（原序號22）

1.「勞委會職訓局局長林三貴表示」
2.「外貿協會將於11月24日於台北國際會議中心舉辦『2010年新興市場採購夥伴大會』，由外貿協會秘書長趙永全親自主持。」—內容合理且局長人名正確，有助於需要找工作的親友。（受訪者H／中等組）

另外認為這篇新聞完全不可信的3位受訪者皆無意願轉發此新聞（轉發意願度為1分），他們的看法是：

「『勞委會職訓局局長林三貴；（吳青常）為提升國內企業在國際市場競爭力；外貿協會將於11月24日於台北國際會議中心舉辦「2010年新興市場採購夥伴大會」』—首先，突然中間插入（吳青常）相當不合理，再來『2010年新興市場採購夥伴大會』是在3/31舉行。」（受訪者L／中等組）

「『（吳青常）為提升國內企業在國際市場競爭力，外貿協會將於11月24日於台北國際會議中心舉辦「2010年新興市場採購夥伴大會」，由外貿協會秘書長趙永全親自主持，邀請全球知名品牌及零組件供應商共襄盛舉，讓台灣品牌在新興市場的發展更具國際競爭力。』—這段新聞看起來不合理，並且對於提升勞工對產業及就業，推動多元職訓課程，讓職訓中心有系統性、專業性與實務性的訓練課程不感興趣。」（受訪者M／中等組）

「第三段新聞跳針，『（吳青常）』相當突兀，所以不會轉發。」（受訪者G／中等組）

這三位中等組都抓到了該篇新聞突兀、主題不太連貫之處。受訪者L甚至還查證了2010年會議的舉辦日期。但若內容中有提及公眾人物來背書，會提高可信度；而內容與生活相關，則會增加轉發意願度。

最後，圖七中平均可信度被評為最低1.67分的序號1（原序號21）新聞內容，有明顯的破綻，如：「淘寶商城是大陸電器龍頭，淘寶商城是台灣唯一的電器零售商」，因此可信度被打為1、2分，只有一位打3分。打3分的受試者，其判斷關鍵字句因為包括：「台灣設立官方旗艦店，在台灣已有11家店，去年營收達人民幣7,000萬元（約新台幣3.3億）」等具體數據，而傾向可信。

二、研究問題二之實驗結果與討論

表二顯示三種分類器，對900篇HCN與CGN測試資料的預測結果，由於資料的類別是平均分佈，Micro F1與Macro F1會很接近。表二顯示Naïve Bayes僅可達到0.73左右的成效，SVM則接近0.8，但BERT成效高達0.96。BERT只將36篇HCN預測成CGN，其餘預測皆正確，亦即450篇CGN預測為假、

表二　三種分類器預測900篇真假新聞的成效

分類器	Micro F1	Macro F1	詞彙特徵
Naïve Bayes	0.7211	0.7190	Count Vectors
	0.6833	0.6711	WordLevel TF-IDF
	0.7266	**0.7266**	Word N-Gram Vectors
	0.7211	0.7193	CharLevel Vectors
SVM	0.7722	0.7719	Count Vectors
	0.7977	**0.7977**	WordLevel TF-IDF
	0.7611	0.7610	N-Gram Vectors
	0.7944	0.7944	CharLevel Vectors
BERT	**0.9600**	**0.9599**	

HCN中其餘的414篇預測為真。顯示用機器來預測GPT產生的假新聞，是有效的。

由於BERT的分類成效明顯高於傳統分類器，我們針對受試者評估的30篇HCN與CGN，只用BERT預測，結果其Micro F1與Macro F1皆為0.933，只有2篇CGN預測錯誤（如表三中粗體底線部分），其餘13篇CGN皆預測為假，15篇HCN預測為真。

為了將BERT與12位受試者的評估結果作比較，表三列出12位受試者平均的預測結果。其算法是針對每一題，先將可信度打為3的受試者移除，假設剩下n位受試者（$n \leq 12$），將可信度評為1、2分的受試者個數算出，假設有x位，則x/n即為受試者們對該題預測為CGN的分數；同理，針對該題，將可信度評為4、5分的受試者個數算出，假設有y位（$x+y = n$），則y/n即為受試者們對該題預測為HCN的分數，結果如表三的第6、7欄。表三中粗體加底線的數字，為預測錯誤的情況。

如前述，BERT只錯了兩題，皆為CGN預測成HCN。而綜合12位受試者算出來的結果，則有5篇新聞預測錯誤，其中HCN有3篇、CGN有2篇，包括了題號11可信度最被低估的HCN，以及題號18、19可信度最被高估的CGN。但這5篇BERT都預測正確，而BERT預測錯誤的2篇，受試者卻不會搞錯。顯示人跟機器，有合作互惠之處。

使用Cohen Kappa係數計算BERT與受試者對於30篇新聞分類結果的一致性信度，Kappa係數為0.55，一致性落在0.4到0.6之間表示程度為一般，結果如表四。

陸、結論

本文針對現有的開源軟體工具，以品質優良的大量文字，訓練出仿真程度相當高的CGN。經由人工評估，多數人雖可區別，但少數人無法正確判斷；且多數CGN可被識破（多為文意邏輯矛盾或內容事實有破綻），但

表三　BERT與12位受試者的預測分數比較表

題號	原序號	解答	BERT分類結果		受試者分類結果		12人平均之可信度
			0-CGN	1-HCN	0-CGN	1-HCN	
1	21	CGN	0.99997	0.00003	1.00000	0.00000	1.67
2	30	CGN	0.99996	0.00004	0.83333	0.16667	2.58
3	10	HCN	0.03155	0.96845	0.00000	1.00000	4.00
4	19	CGN	0.99986	0.00014	0.72727	0.27273	2.50
5	15	HCN	0.00034	0.99966	0.14286	0.85714	3.75
6	5	HCN	0.00011	0.99989	0.20000	0.80000	3.75
7	2	HCN	0.00008	0.99992	0.10000	0.90000	4.08
8	27	CGN	**0.00008**	**0.99992**	0.75000	0.25000	2.42
9	23	CGN	**0.00848**	**0.99152**	0.77778	0.22222	2.50
10	12	HCN	0.02031	0.97969	**0.50000**	**0.50000**	3.25
11	11	HCN	0.00005	0.99995	**0.55556**	**0.44444**	2.92
12	9	HCN	0.00006	0.99994	0.30000	0.70000	3.50
13	29	CGN	0.99969	0.00031	0.70000	0.30000	2.58
14	24	CGN	0.99998	0.00002	0.55556	0.44444	2.92
15	4	HCN	0.00005	0.99995	0.20000	0.80000	3.75
16	20	CGN	0.99998	0.00002	0.75000	0.25000	2.00
17	28	CGN	0.99966	0.00034	0.77778	0.22222	2.33
18	18	CGN	0.99991	0.00009	**0.30000**	**0.70000**	3.33
19	22	CGN	0.99998	0.00002	**0.27273**	**0.72727**	3.33
20	26	CGN	0.99997	0.00003	0.66667	0.33333	2.75
21	3	HCN	0.00005	0.99995	0.18182	0.81818	3.83
22	16	CGN	0.99997	0.00003	0.83333	0.16667	2.17
23	25	CGN	0.99989	0.00011	0.72727	0.27273	2.50
24	6	HCN	0.00004	0.99996	0.18182	0.81818	3.75
25	1	HCN	0.00004	0.99996	0.11111	0.88889	3.92
26	13	HCN	0.00010	0.99990	0.18182	0.81818	3.83
27	7	HCN	0.00004	0.99996	0.25000	0.75000	3.50
28	8	HCN	0.00018	0.99982	**0.57143**	**0.42857**	3.17
29	17	CGN	0.99997	0.00003	0.75000	0.25000	2.58
30	14	HCN	0.00005	0.99995	0.00000	1.00000	4.17

表四　BERT與12位受試者分類結果的一致性信度

		受試者			SUM
		CGN	HCN	Both	
BERT	CGN	11	2	0	13
	HCN	4	12	1	17
	Both	0	0	0	0
	SUM	15	14	1	30

註：Cohen's Kappa Coefficient = 0.5503。

少數連專家等級的受訪者都會受騙。所幸，自動化的深度學習分類器，在獲取足夠的訓練資料後，可以高精準度地辨識出仿真的CGN。

可惜，GPT-2或BERT這種深度學習的技術，雖然可得出令人驚訝的輸出結果，但目前卻很難瞭解其為何會做出這樣的輸出，亦即其解釋性不足。例如，為何BERT會預測一篇新聞為CGN，依據哪些線索，而這才是我們真正想知道、要學習的部分。

目前實驗產生的文章，部分雖可騙過一些人，但到底只是綜合相關文章，模仿潤飾而成，還是做了極大創意的改寫，尚未完全清楚。若是後者，且內容資訊錯誤，其潛在的危害，勢必不小。這方面需要結合不同領域的專長，進行深入的研究，以評估其可能的影響範圍。

誌謝

本研究感謝科技部研究計畫補助，計畫編號MOST 109-2410-H-003-123-MY3。

參考文獻 References

楊德倫、曾元顯（2020）。建置與評估文字自動生成的情感對話系統。*教育資料與圖書館學，57*(3)，355-378。doi: 10.6120/JoEMLS.202011_57(3).0048.RS.CM【Yang, Te-Lun, & Tseng, Yuen-Hsien. (2020). Development and evaluation of emotional conversation system based on automated text generation. *Journal of Educational Media & Library Sciences, 57*(3), 355-378. doi: 10.6120/JoEMLS.202011_57(3).0048.RS.CM (in Chinese)】

賴志遠（2018）。*國際人工智慧政策推動現況*。檢自https://portal.stpi.narl.org.tw/index/article/10418【[Lai, Zhi-Yuan] (2018). *[Guo ji ren gong zhi hui zheng ce tui dong xian kuang]*. Retrieved from https://portal.stpi.narl.org.tw/index/article/10418 (in Chinese)】

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives, 31*(2), 211-236. doi: 10.1257/jep.31.2.211

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... Amodei, D. (2020). *Language models are few-shot learners*. Retrieved from https://arxiv.org/abs/2005.14165

ByteDance. (2019). *WSDM - Fake news classification*. Retrieved from https://www.kaggle.com/c/fake-news-pair-classification-challenge/

Conroy, N. K., Rubin, V. L., & Chen, Y. (2015). Automatic deception detection: Methods for finding fake news. In *Proceedings of the Association for Information Science and Technology* (pp. 1-4). St. Louis, MO: Association for Information Science & Technology. doi: 10.1002/pra2.2015.145052010082

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

Du, Z., Cheng, H., Chiu, H., & Yida (2019). *GPT2-Chinese: Tools for training GPT2 model in Chinese language*. Retrieved from https://github.com/Morizeyao/GPT2-Chinese

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *ImageNet classification with deep convolutional neural networks*. Paper presented at the International Conference on Neural Information Processing Systems, Lake Tahoe, NV.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436-444. doi: 10.1038/nature14539

McCarthy, J. (2007). *Artificial Intelligence*. Retrieved from http://jmc.stanford.edu/artificial-intelligence/index.html

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Retrieved from 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*. Retrieved from https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Ruchansky, N., Seo, S., & Liu, Y. (2017). CSI: A hybrid deep model for fake news detection. In E.-P. Lim & M. Winslett (Chairs), *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 797-806). New York, NY: Association for Computing Machinery. doi: 10.1145/3132847.3132877

Russell, S., & Norvig, P. (2009). *Artificial Intelligence: A modern approach* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Shu, K., Cui, L., Wang, S., Lee, D., & Liu, H. (2019). dEFEND: Explainable fake news detection. In A. Teredesai & V.

Kumar (Eds.), *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 395-405). New York, NY: Association for Computing Machinery. doi: 10.1145/3292500.3330935

Shu, K., Sliva, A., Wang, S., Tang, J., & Liu, H. (2017). Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter, 19*(1), 22-36. doi:10.1145/3137597.3137600

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433-460. doi:10.1093/mind/LIX.236.433

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need*. Paper presented at the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA. Retrieved from https://arxiv.org/abs/1706.03762

（投稿日期Received: 2020/8/4　接受日期Accepted: 2020/10/27）