10.6924/TANET.201909.0041

聊天機器人之研製-以 PTT 八卦板文章為知識庫

陳志達[1] 吳雅卉[2]
[1、2]南臺科技大學資訊管理系
[1]andypony@stust.edu.tw
[2]MA690103@stust.edu.tw

摘要

近年來，人工智慧一直是熱門的議題，其中聊天機器人的應用最為廣泛，且因許多軟體公司推出聊天機器人開發平台，提供眾多強大的功能，例如：自然語言處理、語音辨識、關鍵字分析等技術，甚至有連接其他通訊平台的服務，讓開發者可以更輕鬆、快速客製化出聊天機器人。因此有不少企業開始使用聊天機器人來為消費者服務，如:點餐訂購、客服服務、金融服務等，但這些聊天機器人大多數都是屬於 Closed Domain 的框架，使用者只能詢問特定領域的問題，且得到的回覆普遍都是制式性的回答，缺乏靈活性，使用者較不能感受到互動聊天的感覺。本研究欲實作一個 Open-domain 的聊天機器人，使用 python 語言開發，首先運用爬蟲技術抓取 PTT 八卦板的文章，將這些資料進行自然語言處理，再將處理完畢後的語料用來訓練聊天機器人，並建構檢索式模型(Retrieval-based model)，藉此匹配問答，當使用者輸入語句時，可從模型當中找尋出與其相關之回答，並回覆給使用者。完成後的聊天機器人系統，將會發布在 line 平台上呈現，並達成問題檢索之有效性、快速性與趣味性之成果。

關鍵詞：聊天機器人、自然語言處理、Jieba、Word2vec、BM25。

Abstract

In recent years, artificial intelligence has always been a hot topic, among which application of chat robots(chatbot) are the most adopted. Many software companies have launched chatbot development platforms, providing many powerful functions, such as natural language processing, speech recognition, keyword analysis and other technologies. There are even services that connect to other communication platforms, making it easier and faster for developers to customize chatbots. Therefore, many companies have begun to use chatbots to serve consumers, such as ordering, customer service, financial services, etc. Not only save manpower and time, but also quickly analyze to understand the effectiveness of user behavior and marketing activities based on previous record of conversations, disclose information. However, most of these chatbots belong to the framework of Closed Domain. Users can only ask questions in specific areas, and the replies received are generally systematic answers and without flexibility and users can't feel the feeling of interactive chat. This study uses the python language to develop an Open-domain chatbot. First, we use the crawler technology to capture the PTT gossip articles, perform the natural language processing on those data, and then use the processed corpus to train the chatbot and construct it. The model is used to match the question and answer. When the user inputs query sentence, the relevant answer can be found from the model and replied to the user. The completed chatbot system will be published on the Line social media and will achieve the results of the validity, speed and fun of the problem retrieval.

Keywords: Chatbot, Natural Language Processing, Jieba, Word2vec, BM25

1. 前言

1.1 研究背景

人工智慧是指由人製造出來的機器所表現出來的智慧，如何讓電腦做到推理、知識、規劃、學習、交流、感知、移動和操作物體的能力等，一直是許多科學家追求的目標。早期 IBM 開發用來分析西洋棋的深藍超級電腦，隨後人工智慧發展遇到瓶頸，便沉寂一段時間，直到2016年，由 Google DeepMind 開發的人工智慧圍棋軟體 AlphaGo，擊敗世界冠軍韓國職業棋士李世乭，讓人工智慧再度成為發燒話題。因為科技的進步讓硬體設備能力大幅提升，GPU 運算能力及速度也比以前更強大、大數據資料分析技術的進步、深度學習技術的發展，才能讓人工智慧有如此大的突破。其中聊天機器人是現今最熱門的人工智慧應用之一，目前最早的 Chatbot 是在1966年由麻省理工學院實驗室推出的 ELIZA，是能夠與人簡單對話的機器人程式，隨著網際網路、智慧型手機的出現和即時通訊軟體的普及，聊天機器人便開始快速發展。在2010年由 Apple 推出的語音助理 Siri[9]，更是聊天機器人發展的重要里程碑，但如果要讓聊天機器人具備足以和人類對話的智慧，背後必須要有機器學習、深度學習等技術來支援，而近年來，這兩項技術日趨成熟以及陸續出現許多的開發平台，例如：Line Developer、Facebook Bot Platform、Google Dialogflow、Microsoft LUIS、HIGH5 Robin 等，造就了聊天機器人的崛起，在最近幾年內，各式各樣的聊天機器人更是如雨後春筍般不斷湧現。

10.6924/TANET.201909.0041

1.2 研究動機

聊天是人與人最基本的互動方式，隨著自然語言處理技術和語音辨識技術的發展，許多通訊軟體公司紛紛推出聊天機器人技術與開發介面，讓一般開發者可以在他們的通訊聊天平台上面開發不同的聊天機器人或應用程式，範圍涵蓋了食、衣、住、行、育、樂等。現今，聊天機器人不再是單純的聊天文字介面，它包含了各式各樣的應用及技術，在各個領域與產業都能看見聊天機器人的身影，但大多數都是屬於 Closed Domain 的框架，系統會引導使用者，將對話轉移至特定的主題並回答與預設主題相關之訊息，使用者得到的回覆普遍都是制式性的回答或罐頭訊息，較缺少人性化的回應，自然也少了與真人互動聊天的感受，且有些企業會將此類型的聊天機器人用來當作推播廣告訊息之工具，而缺乏互動功能，失去聊天機器人「聊天」的意義。為了使聊天機器人能更加聰明地回覆訊息，必需要有大量語句讓系統進行訓練和學習，本研究將採用 python 語言開發 Open-domain 之聊天機器人系統，選擇擁有大量日常生活用語的 PTT 八卦板文章作為訓練資料，使用爬蟲技術來進行蒐集，再搭配 Jeiba、Word2vec 兩項工具進行自然語言處理，系統根據使用者所輸入的語句比對詞向量，找出與該語句相似的文章標題，再利用 BM25演算法搜尋與該語句相對應之訊息，並回覆給使用者，讓使用者體驗到聊天互動的感覺。

1.3 研究目的

本研究以 PTT 八卦板文章作為訓練資料，使用 python 語言設計一個能與使用者進行聊天互動之聊天機器人系統，搭配 Jeiba、Word2vec 兩項工具來研究如何進行斷詞處理、比對字詞的相似度等自然語言處理，並建構出檢索式模型(Retrieval-based model)，藉由此模型分析使用者所輸入之語句，利用 BM25演算法檢索與該句子相關的訊息，排序出分數前3高的訊息並回覆給使用者。本研究之功能與特色共有四項，分別為：1. 資料蒐集 2. 斷詞處理 3. 訓練詞向量 4. 建構模型，詳細內容說明如下：

1. 資料選用與蒐集（Data selection and collection）

   為了讓聊天機器人更加貼近人類說話模式，資料必須具有大量日常生活用語，在台灣，目前較多人使用之社群平台為 Facebook、Dcard、PTT 等，因 Facebook 平台上的貼文大部分為照片或短文居多，語句字詞含量較少；Dcard論壇的使用者大多數是學生，文章內容以感情、課業相關居多；PTT 論壇有許多分類看板，文章內容涵蓋的範圍較多，因此，選用 PTT 論壇的文章，運用爬蟲技術蒐集八卦板的文章作為訓練資料。

2. 斷詞處理（Word segmentation Processing）

   本研究利用結巴(Jieba)工具將爬蟲程式蒐集到的文章進行斷詞處理，此步驟是為了將文章中的語句分成單一字詞並判斷該詞之詞性，使用演算法字典樹(Trie Tree)結構去生成這些句子中文字所有可能成詞的情形，從中觀察分出來的字詞是否正確，另外，Jieba 支援增加自定義字典，可透過增加字典來提高斷詞的準確度。

3. 訓練詞向量（Weight calculation of keyword）

   本研究將使用 Word2vec 工具來訓練詞向量，將這些字詞轉成向量來表示，並計算字詞間的距離，好的詞向量會使相似的詞在向量空間上的分布距離較為靠近，有助於訓練模型，藉由使用者所輸入的句子來進行詞向量比對，判斷該語句所表達的情境，找出相似的文章標題。

4. 建構模型（Word to vector calculation）

   文章內容經過上述2.3兩個步驟的處理後所產生的資料，將用來當作聊天機器人知識庫的語料，建構檢索式模型(Retrieval-based model)，當使用者輸入語句時，分析該語句欲想表達的意圖，再從模型當中找出相關的回答，並回覆給使用者，開發者可藉由訓練或調整模型讓機器人回覆的訊息更加準確。

2. 文獻探討

2.1 聊天機器人（ChatBot）

是一個可以對自然語言輸入做出回應的程式，並試圖以模仿真人的方式進行對話[12]。最早的聊天機器人 ELIZA 誕生於 1966 年,由麻省理工學院(MIT)的約瑟夫·魏澤鮑姆(Joseph Weizenbaum)開發,用於在臨床治療中模仿心理醫生，ELIZA 的實現技術僅為關鍵詞匹配及人工編寫的回覆規則[15]。聊天機器人是由相關研究人員開發的一種程式，用作與人交談的代理人，並試圖讓用戶覺得他們正在與真人對話[7]。有學者認為，聊天機器人指透過人工智慧 (ArtificialIntelligence; AI) 的方式，由機器學習程式模擬與使用者互動的對話，目的是幫一般民眾解決投資理財、瞭解服務項目內容及相關商品查詢等日常生活中的細節問題[5]。也有研究提出聊天機器人亦是一種特殊的自動問答系統，特點是模仿人的語言習慣，幾乎都是通過模式匹配的方式來尋找問題最合適的答案[6]。

10.6924/TANET.201909.0041

2.2 自然語言處理（Natural Language Processing，NLP）

是人工智慧和語言學領域的分支學科。此領域探討如何處理及運用自然語言；主要是讓電腦能妥善的處理中文、英文等自然語言，其最終目標是要讓電腦能『理解』自然語言，首先從單字開始，進而到片語(Phrase)、句子(Sentence)，加上文法(Syntax)、語意(Semantics)解析，才能理解一段自然語言的意義[4]。

2.3 詞向量表示法（Word Representation）

1986年時，Hinton提出了分散式表示法(Distributed Representation)做為詞的表示法,這種向量表示是將詞表示成一個較低維度的實數向量，每個詞彙之間的關係可以利用餘弦或是歐式距離計算找出兩個詞向量間的語意相似度[8]，我們將這些詞向量稱為詞表示法。而在2013 年 Google 所提出了 Word2Vec 方法[11]，透過類神經網路的學習方式，將語句中的字詞與字詞關係轉變成具有語法結構與語意關係的向量。Word2Vec 包含兩種模型，分別為連續型詞袋模型(Continue Bag-of Words，CBOW)與跳躍式模型(Skip-grams，SG)。

2.3.1 連續型詞袋模型（Continue Bag-of Words；CBOW）

該架構類似於前饋類神經網路 (Feed-Forward Neural Network)。不同之處在於連續型詞袋模型移除非線性隱藏層 (Non-Linear HiddenLayer)[1]。如此，該模型僅使用線性表示能力來計算詞的實數表示向量，仍然可以保持良好的性能。在連續型詞袋模型中，是透過一個詞的上下文 (Context) 來預測該詞的機率，如下圖1所示：

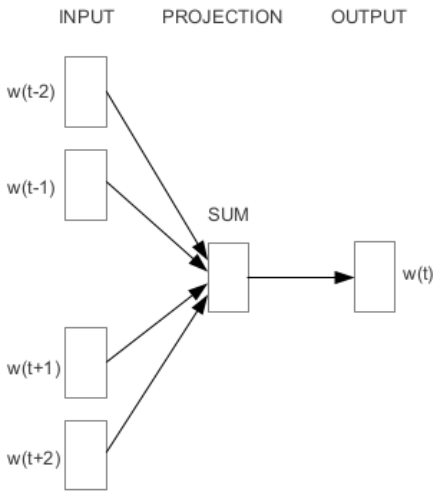

圖1：連續型詞袋模型

2.3.2 跳躍式模型(Skip-grams；SG)

該模型以簡化的前饋類神經網路透過逆向訓練目標來學習詞表示法[1]。該模型與連續型詞袋模型相反，是給定一個詞 w 來預測其上下文 Context(w)的機率，如下圖 2 所示：

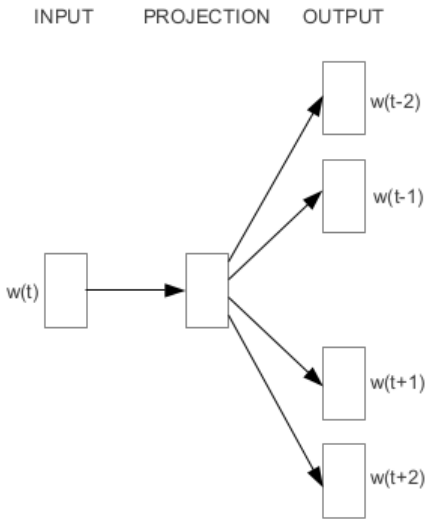

圖2：跳躍式模型

比較上述提到的兩種模型，CBOW 模型在語法分析上有較好的表現，Skip-gram 模型在語意分析上會有較好的表現，而在資料量較少的情況下，CBOW 模型就能有好的表現，但 Skip-gram 模型需要較大量的資料作訓練才能有好的表現[14]，本研究蒐集的文章數約250000篇，因此採用 Skip-gram 模型來訓練資料。

2.4 Okapi BM25

Okapi BM25演算法其中的 BM 為 Best Matching，代表最佳匹配的意思，是一種基於概率論檢索模型所提出的算法，該演算法於1970年代由英國科學家 Stephen E. Robertson[13]所提出，是一個可以計算詞彙與文章相關性的演算法，依據查詢的詞進行檢索，且可以將檢索到的結果與文件的相關程度的高低作排名匹配，是資訊檢索技術中相當經典的排序演算法，目前也是搜尋引擎常用的演算法之一，例如：Google 搜尋引擎，從搜尋網頁的結果來看，除了有支付費用的廣告網頁以外，瀏覽其他網頁可以看出來的確有進行相關程度的排序。此演算法的計算方式，主要是以詞頻的關係，也就是字詞出現的次數來分析計算並沒有將句子的語法架構考慮進去。接下來針對 Okapi BM25的公式及概念作介紹，如下面之(1)、(2)式數學公式：

$$\text{Score}(D,Q)=\sum_{i=1}^{n} IDF(q_i)\cdot\frac{f(q_i,D)\cdot(k_1+1)}{f(q_i,D)+k_1\cdot\left(1-b+b\cdot\frac{|D|}{\text{avgdl}}\right)}, \quad (1)$$

10.6924/TANET.201909.0041

$$\mathrm{IDF}(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (2)$$

D = D 這份文件
Q = Query
$f(q_i, D)$ = q_i字詞在 D 文件中出現的次數
$|D|$ = D 文件的長度
$avdl$ = 所有文件內容的平均字數長度
k_1= BM25的自由參數，用於調節飽和度變化的速率，通常預設值為2
b = BM25的自由參數，用於字串長度正規化，將文檔的長度正規化到全部文檔的平均長度上，值的範圍在0到1之間，通常預設值為0.75，而1表示全部正規化，0則是不進行正規化
N = 文件的總數量
$n(q_i)$ = 有包含q_i的文件數

3. 研究方法

3.1 系統規劃

本研究基於 python 語言，開發一個聊天機器人，利用爬蟲技術抓取 PTT 八卦板的文章，將這些資料進行自然語言處理，利用 Jieba、Word2vec 等工具，進行斷詞處理、詞向量的訓練，經過上述處理後的資料儲存至資料庫，並建構出檢索式模型(Retrieval-based model)，當使用者輸入完整語句時，系統會將該語句進行斷詞處理，並用詞向量搜尋與該句子相關的文章標題，再運用 BM25演算法找出與該句子關聯性最高的文章，將該文章底下的訊息進行分數的計算，排序出與使用者輸入之語句相關性前3高的訊息，並回覆給使用者。最後，透過 Line Developer 平台提供的 webhook 功能，將完成後的聊天機器人系統發布在 line 平台上呈現。本研究根據上述說明來規劃系統架構，如圖3所示，該架構主要由三項模組所構成，包括 Crawler、自然語言處理、Line Developer，以下將針對各模組分別詳細描述其功能：

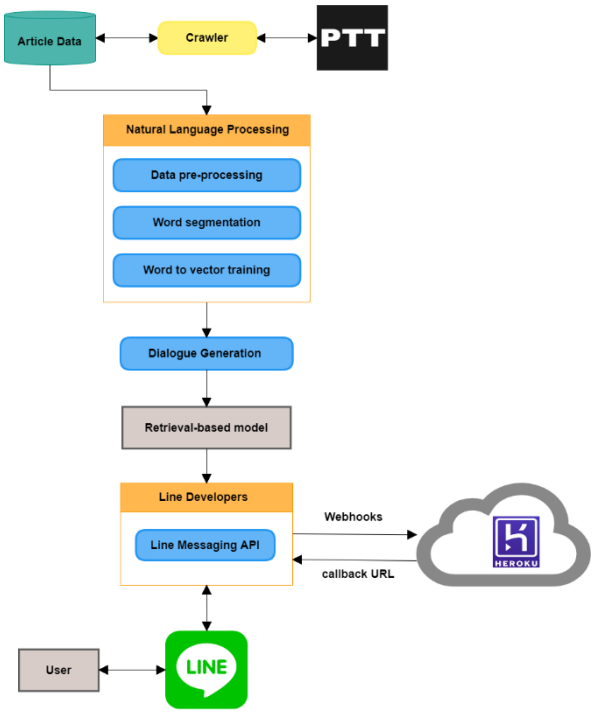

圖3：系統架構圖

3.2 Crawler

首先利用爬蟲程式抓取 PTT 八卦板的文章，並過濾文章內空白及特殊字元，抓取的內容包含：看板名稱、文章標題、文章內容、發文時間、推文(其他使用者回覆文章的訊息)，將這些資料以 JSON 格式存放在資料庫。

3.3 自然語言處理

此模組包含資料前置處理、斷詞、詞向量訓練三個元件，首先將文章過濾，篩選出需要的資料，再來使用結巴(Jieba)工具進行斷詞處理，最後使用 word2vec 工具進行詞向量訓練，以上為本研究的自然語言處理步驟。

3.4 Line Developer[11]

此模組包含了 Line Messaging API，此 API 功能是 Line 平台本身提供給開發者使用，讓開發者能將自行開發之程式與 Line 進行連動，首先，將程式佈署至 HEROKU 平台上，完成後會得到一組 URL，接著使用 Webhooks 功能即可呼叫自行撰寫的程式，並呈現在 Line 平台上。

4. 系統實作

本研究在系統整體設計方向有幾項重點：第一，首先要運用爬蟲技術抓取 PTT 八卦板的文章，並進行前置處理，過濾重複文章以及不雅字眼；第二，進行自然語言處理，包含斷詞、詞向量的訓練；第三，建構出檢索式模型，系統會分

10.6924/TANET.201909.0041

析使用者所輸入語句，將句子進行斷詞處理，並使用詞向量進行比對找出相似的文章標題，再運用 BM25演算法挑選出與使用者輸入之語句關聯性最高的文章，接著計算該文章底下的訊息，並挑選出和使用者詢問之內容相關性前3高的語句並回覆給使用者；第四，將設計好的聊天機器人系統包裝到 Line 平台上呈現，在本章節中將說明上述四個重點的實作過程，說明如下：

4.1 爬蟲程式抓取文章

如下圖4所示，為本研究使用之爬蟲程式的片段程式碼，抓取 PTT 論壇八卦板的文章，程式碼第173~187行是將抓取完的文章資料以 JSON 格式存放在資料庫裡。

```
173         # json data
174         data = {
175             'url': link,
176             'board': board,
177             'article_id': article_id,
178             'article_title': title,
179             'author': author,
180             'date': date,
181             'content': content,
182             'ip': ip,
183             'message_count': message_count,
184             'messages': messages
185         }
186         # print 'original:', d
187         return json.dumps(data, sort_keys=True, ensure_ascii=False)
```

圖4：爬蟲程式片段

4.2 自然語言處理

將文章過濾篩選出文章標題與推文回應後，接著進行斷詞處理，如圖 5 所示：輸入語句「最近有蔡英文的八卦嗎」，進行斷詞和詞性分析的測試，程式執行結果如下：ad 為副詞；v 為動詞；nr 為人名；uj 為助詞；n 為名詞；y 為語助詞。

```
#encoding=utf-8
import jieba
import jieba.posseg as pseg

jieba.set_dictionary("data/userdict.txt")

words = pseg.cut("最近有蔡英文的八卦嗎")
for word, flag in words:
    print('%s %s' % (word, flag))
```

```
最近 ad
有 v
蔡英文 nr
的 uj
八卦 n
嗎 y
```

圖5：Jieba 斷詞功能

詞向量訓練的部分，採用 Word2vec 將詞彙轉換成向量，藉此分析詞與詞之間的相似度，程式範例為輸入冰淇淋一詞，計算出與該詞彙相似度前十個字詞，如圖7所示：

```
水淇淋
詞彙相似詞前 10 排序
飲料,0.6867046356201172
洋芋片,0.6759800910949707
力醬,0.666611909866333
果汁,0.6608696579933167
荷卡,0.6607642769813538
優格,0.6598771214485168
力廣告,0.6571471095085144
力餅,0.6568143367767334
冰棒,0.6568007469177246
水果醋,0.6557888388633728
```

圖7：詞向量相似度計算

4.3 建構檢索式模型

檢索式模型的架構圖如圖7所示，本研究利用詞向量和 BM25演算法來進行檢索匹配並排序出訊息關聯性的高低，使用者可輸入完整的語句，之後系統將會分析該語句，利用詞向量找出和句子相關的文章標題，再使用 BM25演算法挑選出與該語句關聯性最高的文章標題，並將該文章底下的訊息排序出與該句子相關性最高的前3名訊息，並回覆給使用者。

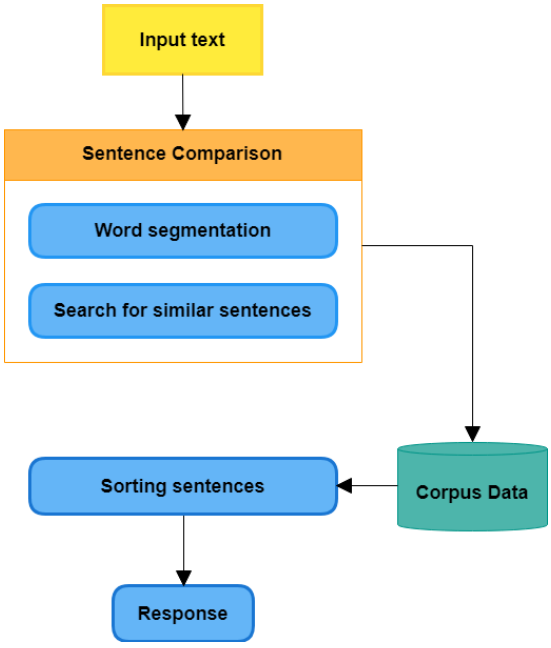

圖7：檢索式模型

4.4 將聊天機器人系統包裝到 Line 平台上呈現

完成之聊天機器人系統將佈署至 HEROKU 平台上，佈署完畢後會得到一組 URL，此 URL 是 Webhooks 功能呼叫第三方程式的關鍵，可讓開發

10.6924/TANET.201909.0041

者自行開發之程式與 line 平台進行連動，並呈現在 Line 平台上，如圖8所示：

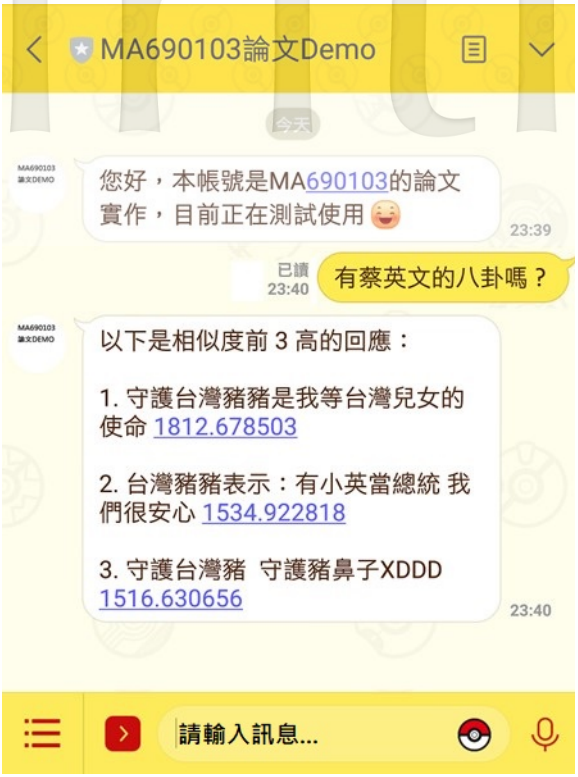

圖8：呈現畫面

5. 系統分析與比較

本系統之聊天機器人蒐集 PTT 八卦板文章作為訓練資料，利用 Jieba、Word2ve 工具進行自然語言處理與詞向量分析匹配之研究。表1為本系統與家樂福聊天機器人的功能特性分析表。

表 1　系統功能特性分析表

系統 / 特性	本系統	家樂福-福媽 [2]
使用者介面	Line 平台呈現畫面	Line 平台呈現畫面
語料來源	PTT 八卦板	地方媽媽訪談內容
完整語句查詢	支援	支援
知識庫延展性	可以蒐集新資料並訓練	不支援
訊息回覆正確率	較高，建立檢索式模型幫助分析使用者輸入之語句。	較低，輸入不同意思的語句爾偶會得到相同的回覆。

6. 結論與未來工作

本研究主要以 python 語言開發系統，運用爬蟲蒐集大量文章作為訓練資料，再搭配自然語言處理相關工具，例如：結巴(Jieba)、Word2vec，篩選文章內容並擷取推文訊息，本研究透過建構檢索式模型(Retrieval-based model)進行匹配問答，並以詞向量與 BM25演算法來達成檢索匹配的功能，有別於一般以關鍵字搜尋的系統，本系統可以讓使用者輸入完整的語句，並得到相對應的回覆。

參考文獻

[1] 石敬弘，基於類神經之關聯詞向量表示於文本分類任務之研究，國立臺灣師範大學資訊工程研究所碩士論文，2017。

[2] 家樂福-福媽，https://www.bnext.com.tw/article/50235/carrefour-line@

[3] 陳思澄，使用詞向量表示與概念資訊於中文大詞彙連續語音辨識之語言模型調適，國立臺灣師範大學資訊工程研究所碩士論文，2015。

[4] 鄭捷，自然語言處理：用人工智慧看懂中文，台灣：佳魁資訊，2018。

[5] 賴森堂、黃彥綸，LINE 通訊軟體結合 Chatbot 改善設備連線測試效率與品質，電腦稽核，38，25-36，2018。

[6] 魏彰村，運用爬蟲技術之主題導向即時通訊聊天機器人設計與實現以籃球領域諮詢結合 LINE APP 實作為例，國立中正大學通訊資訊數位學習碩士論文，2017。

[7] 蘇柏勳，對話代理人中問句補全及問答句對之自動產生，國立成功大學資訊工程學系碩士論文，2012。

[8] G.E. Hinton. Learning distributed representations of concepts. in Proceedings of the Eighth Annual Conference of the Cognitive Science Society， pages 1 – 12 ， Amherst 1986,1986. Lawrence Erlbaum,Hillsdale.

[9] L. Caviglione and W. Mazurczyk , “Understanding Information Hiding in iOS,”Computer,pp.62 – 65,2015.

[10] LINE Developers, https://developers.line.biz/en/.

[11] Mikolov, T., et al. "Distributed Representations of Words and Phrases and their Compositionality." CoRR abs/1310.4546. ,2013.

[12] Reshmi, S.Balakrishnan, Kannan. Implementation of an inquisitive chatbot for database supported knowledge bases.Sa¯dhana¯, Vol. 41, 1173–1178,2016.

[13] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu and M. Gatford,“Okapi at TREC-3”, In Proceedings of the Third Text Retrieval Conference(TREC-3), NIST, 1995.

[14] T. Mikolov， K. Chen， G. Corrado， and J. Dean. Efficient estimation of word epresentations in vector space. In International Conference on Learning Representations,2013.

[15] Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM， page36-45,1966.