# 科技部補助

# 大專學生研究計畫研究成果報告

＊ ********* ＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊ ＊

＊ 計 畫 名 稱 ： 具即時語音信息擷取功能之緊急任務救助服務匹配系統 ＊

＊ ********* ＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊ ＊


執行計畫學生： 楊凱州

學生計畫編號： MOST 105-2815-C-006-099-E

研 究 期 間 ： 105 年 07 月 01 日至 106 年 02 月 28 日止，計 8 個月

指 導 教 授 ： 盧文祥


處理方式： 本計畫可公開查詢


執 行 單 位： 國立成功大學資訊工程學系（所）


中華民國 106 年 03 月 28 日

# 摘要

隨著 Facebook 與 Line 聊天釋出了聊天機器人開發平台，短語的語意解析已成一行銷與自動化領域的重要議題。在此，本計畫以 Tomas Mikolov 提出的詞向量 [1] 為基礎，利用近似主題的詞來建立出各類主題詞向量，並將其串接為圖的結構，來進行語意解析與實體抽取。此外，我們也定義了一個多輪式的對話框架，將用戶目前對話與先前已知特徵串接起來，以行更複雜的任務處理。進行完語意解析後，本計畫建立了一個服務匹配平臺，為複雜任務提供適任服務人選。



關鍵字：詞向量、意圖抽取



# Abstract

As Facebook and Line release their chatbot develop platforms, the intent extraction about chats has became a hot topic on marketing and automation. Our research is based on the distributed word representations presented by Tomas Mikolov [1]. We pick some words which are related to the same topic to train several topic word embeddings that we use on building a semantic graph for semantic analysis and entity extraction. We define a structure for processing the memory about the former conversation. Our system is able to process a much complex task by filling the feature we extract from the former conversation into the current one. This research also build a web platform for selecting the right member to given task after the part of semantic analysis.



Keywords: word embedding, intent extraction.

# 目錄

# 一、前言

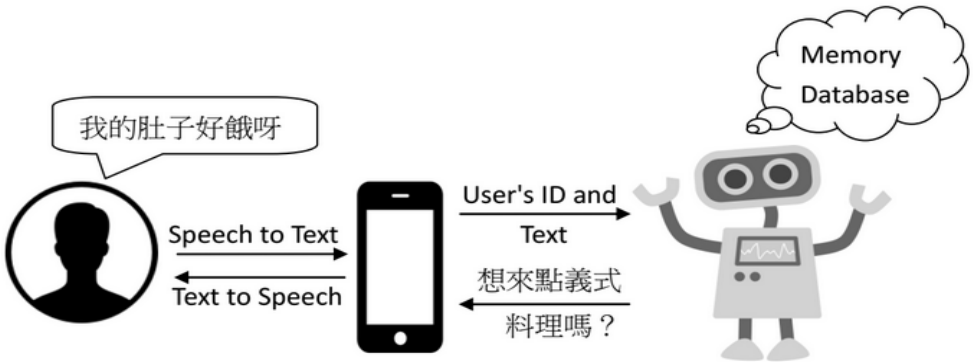
「2016 年將成為對話式商務元年。」 Uber 平台開發人 Chris Messina 的曾於 F8 大會前進行這麼說，預言今後的商務應用，將從傳統的客戶被動地查找資料，改為由聊天機器人主動並即時地為客戶提供資訊。

這個想法在 Facebook 與 Line 釋出了平台上的 Chatbot API 後更加真實，任何企業只要經過申請，便能於社交平台上創建出專屬的聊天機器人，將其作為企業與客戶間溝通的新橋樑，即時解決簡單的問答與某些個人化服務。擔當機器人核心的語意分析的服務，也成為自然語言處理中的一大熱點。本計畫提出了一個簡單且彈性的語意分析方式與多輪式對話情境，並將其連接上一個線上委託平台，提供一些如飲食、住宿推薦等生活服務。

# 二、研究動機

意圖抽取一直是自然語言處理中的重要議題，特別在目前各家聊天機器人平台釋出後，如何能有效地解析平台上使用者對話目的，並針對此提供服務、回應或客製化廣告已成企業界的關注熱點。本計畫便專注於聊天的意圖分類上，且試圖串接起前後句對話的關聯性，構成一個具備記憶性的對話語境。

本計畫結合了聊天機器人與線上服務平台，旨在將繁冗的圖型化介面簡化為一個對話框，以及能為不適應複雜操作的長者或視障者們提供服務，為此，我們提供了 Text to Speech 與 Speech to Text 功能，讓使用者可一鍵進入問答情境，並用聊天的方式，詢問一些日常的生活事物，或者將自己想要委託的複雜任務，發佈至線上服務平台。




圖一 系統操作案例

# 三、文獻探討

本系統的欲解決的核心議題鎖定在生活短語的語言處理上，包含了短語的意圖分類，以及從中抽取出情境實體。相較於文章而言，由於短語的可用特徵較少且涵蓋侷限，使得基於主題模型的分類法顯得力不從心，此外，如 LDA、SVD 等方式皆需在訓練時就決定好主題個數，對於會動態調整服務類型的聊天平臺模式，顯得彈性度過低。為因應上述兩個問題，本計畫的首要研究主體不是整個句子，而是鎖定在單詞的語意分析，以試圖最大化短語中每個特徵的可用性，於這方面的應用，我們主要採用了 Tomas Mikolov 於 [1]、[2] 中提出的詞嵌入表示法 ( Distributed Representations of Words)。

## 詞袋模型

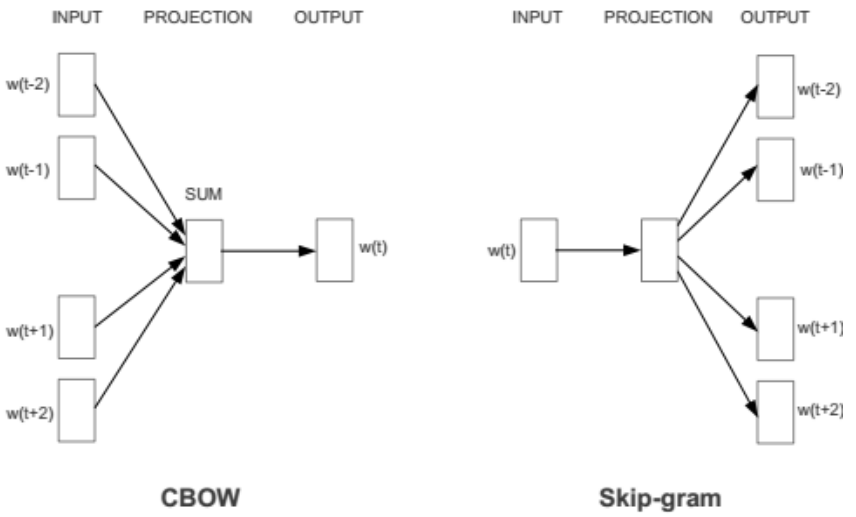
傳統上，詞的向量表示法為 One-Hot Encoding ，其將每個單詞用一個維度表示，比如資料集中有 [ eat , ate , eaten] 三者，我們可以創建一個三維的向量，第一個維度代表 eat，第二個維度代表 ate，第三個維度代表 eaten，搭配上布林值便可表示這三個單詞：

$$eat : [1,0,0] \quad ate : [0,1,0] \quad eaten : [0,0,1]$$

此方法雖然簡明易懂，卻因讓每個詞獨佔單一維度，造成每個詞之間的詞義獨立，以上例而言，eat 與 ate 的 Cosine 相似度為 0，然而 eat 跟 ate 的差異僅只時態變化，詞義上都代表吃這個行為，兩者間的相似度顯然不該為零，此外，One-Hot Encoding 的維度等同詞典中的詞總數，當詞典過大時亦容易產生維度災難(Curse of dimensionality)。

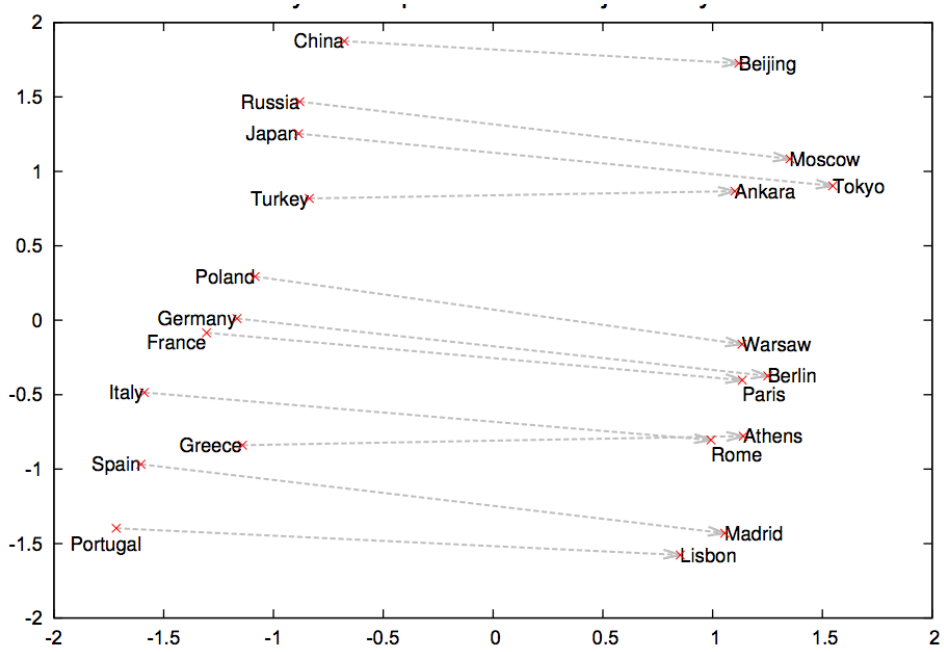
## 詞向量

為了解決 One-Hot Encoding 所造成的詞彙鴻溝，人們便開始思考該如何將稀疏的詞袋稠密化，且同時保存了詞與詞之間的相似度。此概念的實作最先起於 Yoshua Bengio 提出的 Neural Network Language Model (NNLM) [3]，其使用淺層的神經網路，輸入為若干個單詞，預測目標是接下來的單詞為何，但此時的詞向量來自隱藏層的權重，僅為訓練的副產物。目前為人所之的詞向量則是來自 [1] ， 其為 NNLM 的改良，並進一步分為 CBOW 與 Skip-gram 兩種模型：




圖二 CBOW 與 Skip-gram 模型

誠如語言學家 John Rupert Firth 所言：" You shall know a word by the company it keeps." ，一個字的詞意，會與周邊詞的詞意相關， Tomas Mikolov 的詞向量便是以此為出發點，CBOW 考量的是以周邊詞 (Wt-1, Wt-2, Wt+1, Wt+2) 的詞向量，來計算中心詞 Wt 的詞向量，Skip-gram 模型則相反，其透過給定中間詞的詞向量，去計算出周邊詞的詞向量。經此神經網路進行向量化後，詞義相近的詞會在向量空間中距離較近，並具備一些邏輯上的線性關係，如 vector(Bejing) - vector(China) + vector(Russia) ~= vector(Moscow)。

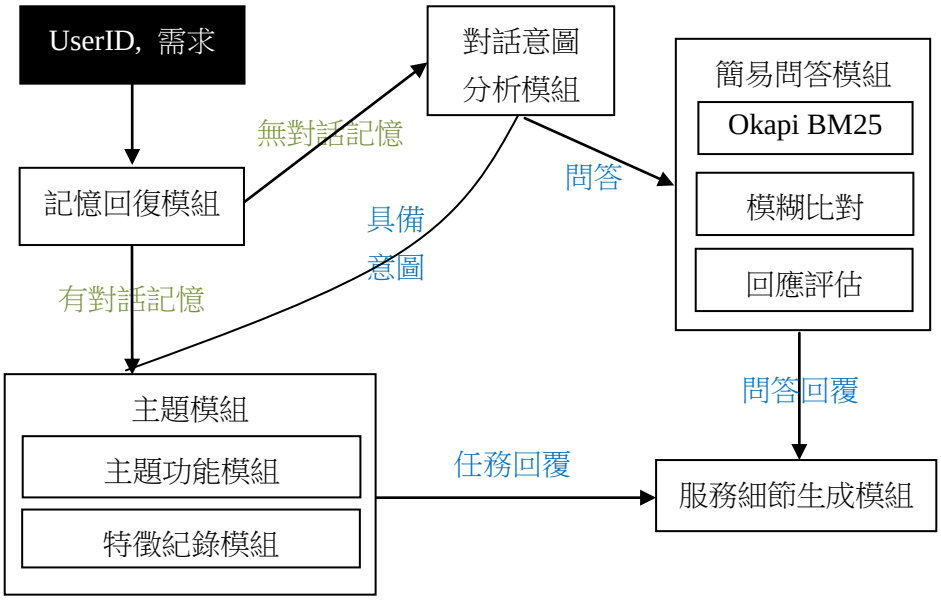



圖三 詞向量間的線性關係 - 國家與首都的對應

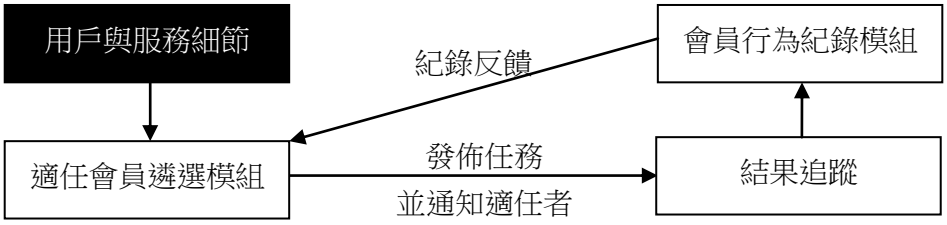
# 四、研究方法

## 專案概觀

本計畫的系統設計可分為語義解析端與服務匹配端，我們針對兩者分別定義了架構與執行流程：




圖四 語義解析端架構




圖五 服務匹配端架構

語義解析端目的是透過多輪式對話了解使用者需求，並從中對話抽取中該需求所需要的特徵，比如使用者說：「我覺得口有點渴了」，語義解析端會紀錄使用者的需求為飲食類，並反饋：「想要喝些什麼嗎？」來填補飲食類需求中缺少的「種類」特徵。

而服務匹配端則是在認知使用者的完整需求後，至本計畫的服務平台尋找適任者，以上個例子來說，語義解析端會告知服務匹配端某名用戶希望要「協助取得某個『種類』的『飲料』」（服務細節），再依照該名用戶的資訊

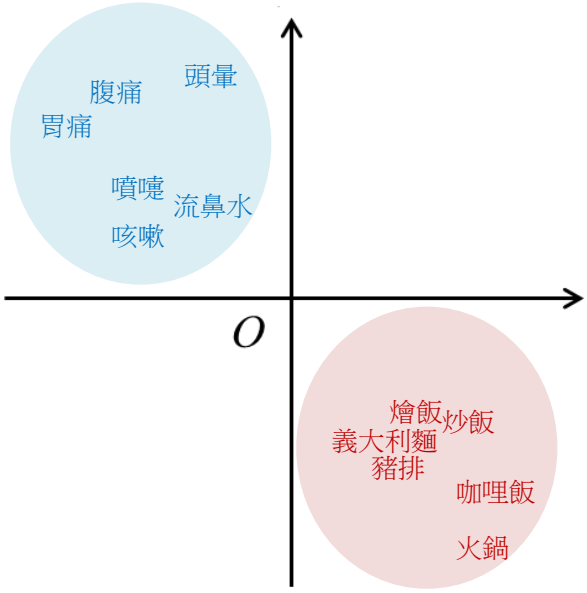
與任務細節去從平台中挑選服務者(以 E-mail 通知)。服務匹配端也會追蹤每位會員在平台上的行為記錄(如訪問頁面、承接任務類型、用戶反饋等等),以此調整遴選結果。

語義解析端的設計有兩個研究議題。其一為服務分類,要將使用者的輸入分類至購物、醫療、娛樂、飲食、生活、住宿、其他(所有類別的信心分都過低者為之) 這七個主要服務。其二為特徵捉取,要從使用者的回覆裡抽取信息,以填補服務情境中的缺少屬性(如前例中詢問飲料的「種類」)。於本計畫中,我們決定以 Word2Vec 的技術解決此二議題。首先,我們先描述模型的訓練方式與細節。

## Word2Vec 訓練與模型細節

為了能銜接上後續服務匹配模組的開發,本計畫所採用的 Word2Vec 是開源函式庫 gensim [1]提供的版本,其為 Google 先前釋出的版本提供了 python 接口,並包覆了詞跟詞之間的 Cosine 相似度比較、Top K 相似詞排序以及基於線性關係的詞意推理等常用功能。

我們將模型的訓練語料分為了一般性語料與特殊語料。一般性語料採用的是繁體中文維基百科於 2016 / 08 / 20 的文章備份,共有 2,822,639 篇文章,經 jieba [2] 斷詞後共約十萬個不重複的繁體中文單詞,該部份語料用做一般性的詞關聯度訓練。特殊語料則是偏向特定主題的文章,是為了強化詞在特定主題上的群聚性,以期後訓練能夠生成出一個明確代表該主題的詞向量。




[1] https://radimrehurek.com/gensim/models/word2vec.html
[2] https://github.com/fxsjy/jieba

圖六　詞的聚類關係

此外，將論壇與回覆文章列為語料，另一方面也是考量到維基百科於章法上過於嚴謹，以至與一般聊天情境有所區隔。為能將本系統切實地應用至生活中，一般的日常會話語料是必不可少。

將上述語料配合客製化詞典進行斷詞，並去除停用詞後，我們訓練出了一個 300 維並進行過負採樣的 skip-gram 模型，簡易評估如下：

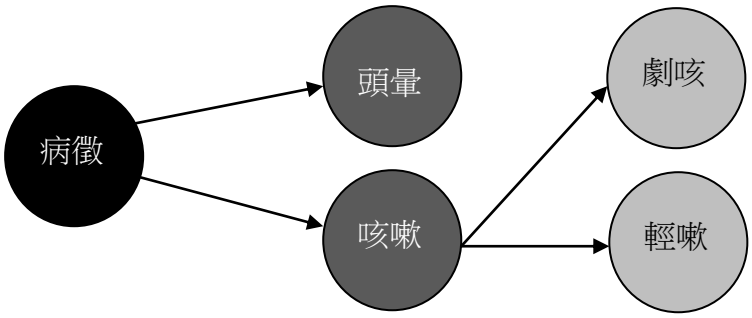

圖七、八、九 Word2Vec 模型簡評

圖七與圖八顯示概念相似的詞，有呈現出高關聯性，而圖九是從正例與反例來觀察，語意上較相似者如「感冒」、「生病」具有高相似度，而語意上無關者如「飲料」、「鬧鐘」也確實具低相似度。

## 構建語意解析圖與語意節點

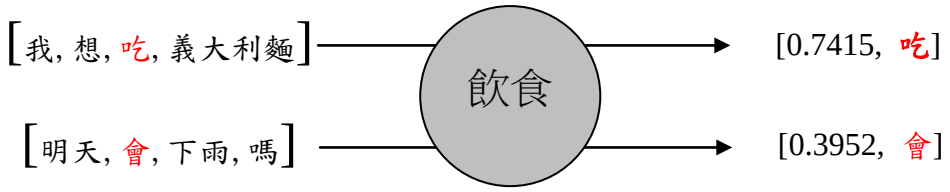
傳統上，模板式的聊天機器人往往會有出關鍵字搶佔的問題，比如定義句子中出現了關鍵字 A 就回覆 R1，出現關鍵字 B 就回覆 R2，然而在句子中同時出現了 A、B 時，當回覆 R1 或回覆 R2 即會成為問題。而為了避免該情形，我們將關鍵詞視為抽象概念，並依據每個概念的抽象程度來定義優先權，以描述疾病為例，有三種抽象概念表示為「病徵」、「咳嗽」與「頭暈」，則在語意解析時其順序應當為，「病徵」先其後再處理「咳嗽」與「頭暈」(此兩者等重)，可以將此流程考慮為系統先自使用者的對話中感知，發

現該句話與「病症」的概念有關，接下來再去觀察是跟哪種病症(是咳嗽或是頭暈)有關，是為分而治之的過程。




圖十 語意解析的過程

如圖十所示，我們能夠採用圖(Graph)的結構，來表示上述解析過程，本計畫中將該結構定義為語意圖，圖中節點稱為語意節點，每個語意節點都是一種概念的分類器，輸入一個詞袋，輸出一個信心分數(0至1) 與匹配元，表示該概念的相關性與匹配實體：




圖十一 輸入詞袋至語意節點，輸出信心分與匹配實體

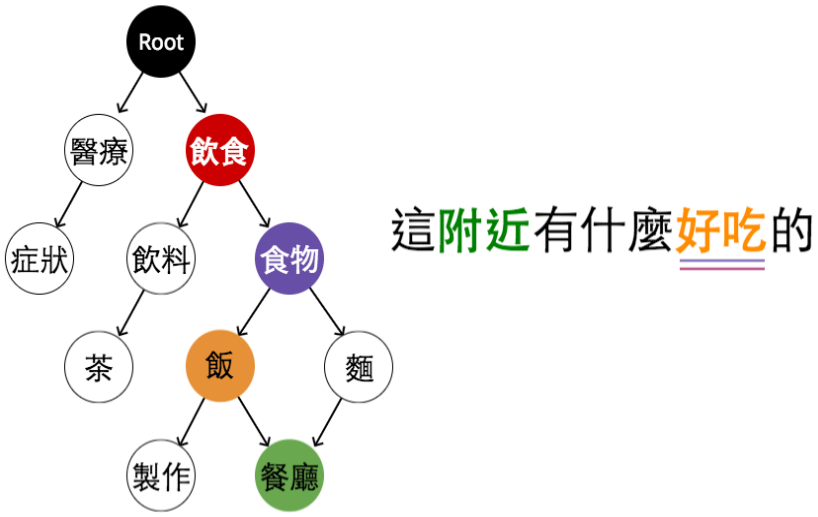
語意節點的打分是基於詞向量的 Cosine 相似度，我們先為每個概念人工定義了若干個能闡述這個概念的概念詞，比如「飲食」這個概念可包含「吃喝」、「用餐」等等。在完成了概念詞的定義後，我們透過這些概念詞的共同特徵，來生成出一個能代表目標概念的主題詞向量。有了主題詞向量，我們便可透過計算其與詞袋中每個詞的 Cosine 相似度，取其中最大值得出信心分數，而詞袋中獲得最高信心分的詞為該概念的匹配實體。

此外，我們對圖的走訪定義了兩種解析類型：

1. 語意解析：如先前「病徵 → 咳嗽 → 輕咳」的案例，此運算的目的是在將抽象概念具體化，另一個例子是「飲食 → 飲料 → 紅茶」。
2. 結構解析：目的是區隔意圖，由若干次的語意解析構成。

以「這/附近/有/什麼/好吃/的/？」為例，以定義好的語意圖(圖)來走訪，

經過第一次的語意解析「飲食→ 食物 → 飯」走至飯，移除已判定的匹配實體「好吃」，得出「這/附近/有/什麼/的/?」並以此判斷與飲食的哪方面有關， 該句經過走訪後，最高信心分 0.4318 來自［附近, 餐廳］，因語意節點餐廳沒有子節點，當次走訪結束。




圖十二 語意節點

| 第一層解析 (Root -> 飲食) | | | | | |
|---|---|---|---|---|---|
| 概念 | 匹配實體 | 最高信心 | 概念 | 匹配實體 | 最高信心 |
| 飲食 | 好吃 | 0.5147 | 醫療 | 附近 | 0.2226 |
| 第二層解析 (飲食 -> 食物) | | | | | |
| 概念 | 匹配實體 | 最高信心 | 概念 | 匹配實體 | 最高信心 |
| 食物 | 好吃 | 0.5854 | 飲料 | 好吃 | 0.5349 |
| 第三層解析 (食物 -> 飯) | | | | | |
| 概念 | 匹配實體 | 最高信心 | 概念 | 匹配實體 | 最高信心 |
| 飯 | 好吃 | 0.6109 | 麵 | 好吃 | 0.5942 |
| 第四層解析 (飯 -> 餐廳) | | | | | |
| 概念 | 匹配實體 | 最高信心 | 概念 | 匹配實體 | 最高信心 |
| 製作 | 什麼 | 0.3157 | 餐廳 | 附近 | 0.4318 |

表一 圖十二語意圖之走訪細節

經過該次走訪，我們可以判定該句話的主意圖為飲食相關且細節為尋找附近的餐廳，可以此生成一服務訊息並自動發佈至服務平台：

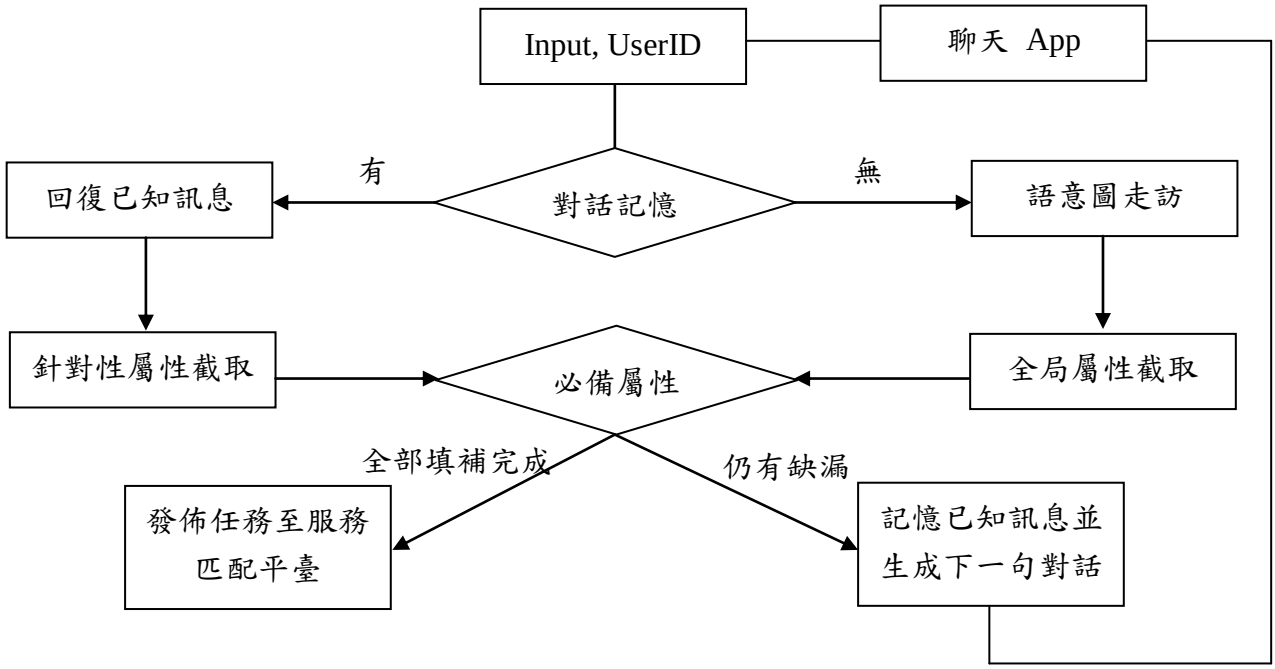
| 用戶名 | 服務類型 | 服務項目 | 屬性 |
|---|---|---|---|
| XXX | 飲食 | 詢問餐廳 | #飯 |

表二 服務訊息示例

如此一來，我們便完成了一類高彈性的語意解析方式，可依服務對象與服務提供者的不同，簡易且迅速地調整走訪方式。

## 多輪式對話設計

於表二中的屬性是由我們對該服務情境手動定義，分為了必備屬性與可選屬性，若是在匹配中無法抓取到全部的必備屬性，則會進入多輪式對話，試圖填補所有屬性。我們可以下圖呈現整個多輪式對話的運作流程：




圖十三 多輪式對話運作流程

以住宿服務為例，我們已定義的情境模板如下表：

| 服務類型 | 服務項目 | 屬性 | |
|---|---|---|---|
| 住宿 | 訂飯店 | 必備屬性 | 可選屬性 |
| | | #時間,#地點 | #特殊需求 |

表三 住宿情境模板

若是使用者說：「這附近有什麼旅館呢？」，系統透過語意走訪可以得知：

| 用戶名 | 服務類型 | 服務項目 | 屬性 |
|---|---|---|---|
| XXX | 住宿 | 訂飯店 | #地點=附近 |

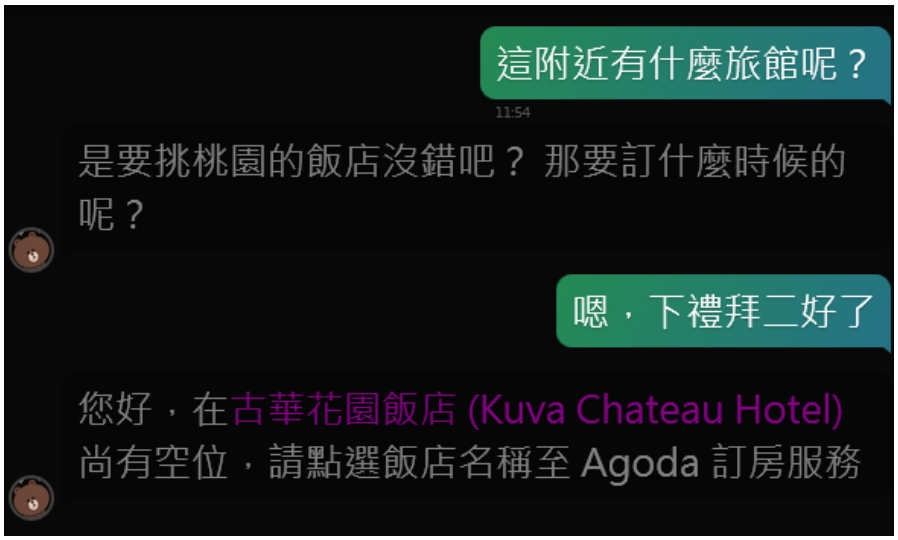
表四 第一次解析結果

由於此對話中不存在任何關於時間的陳述，故語意圖裡代表時間的語意節點無法截取到「#時間」這個必備屬性，因此這個服務並不算完成，將會被暫存至資料庫中，並記憶使用者前一句對話對話的意圖(**訂飯店**)與目前已知屬性(詢問**附近**區域)。而為了填補 # 時間屬性，系統會提示性地回覆該使用者：「想要訂什麼時候的飯店呢？」，並針對當次使用者的回覆「下禮拜二好了」，以代表時間的語意節點進行屬性截取：

| 用戶名 | 服務類型 | 服務項目 | 屬性 |
|---|---|---|---|
| XXX | 住宿 | 訂飯店 | #地點=附近、#時間=下禮拜二 |

表四 第二次解析結果

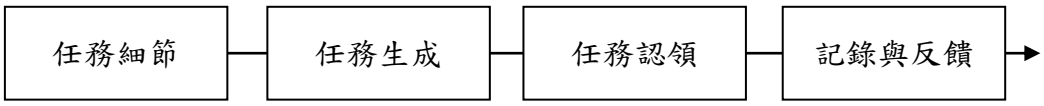
值得一提的是，在已知使用者的意圖(即查找資料庫時發現該使用者尚有未完成的服務問答時)，我們並不會去重複遍歷整張語意圖，而是只針對缺少的屬性，透過該屬性的節點去進行資訊截取，如上例我們只用了代表時間的節點去抽取出下禮拜二，這麼做的理由是，我們確定當句對話與先前對話有相依性，使得目前抽取出的對話意圖，並不一定為使用者真實的意圖(如「下禮拜三好了」會被分為「其他」而非預訂的「住宿」)，而僅是對針對我們提出的詢問所作的回答，所以不當武斷地覆蓋先前的對話意圖。此案例的整體多輪式對話呈現如下：




圖十四 多輪式對話呈現

服務認領平臺設計

在完成服務的內容解析後，會將服務細節送交至服務認領平臺並自動發佈：




圖十五 服務生命週期

匹配平台的主要設計是服務的適任會員推薦，我們以兩個基準來衡量，其一為該名會員與該服務的向性，本平台在申辦會員時會要求會員填寫偏好服務領域，此外，我們也會透過在服務平台上的連結附加參數，以 GET Method 將會員點擊的每個連結送往資料庫記錄，以常拜訪的主題頁面及常關注的任務，來推估該名會員傾向接取什麼類行的任務。第二項影響匹配的因素是時空上的向性，考量到並非所有會員都會常駐於平臺，或是特定服務需親自至當地進行(於截取任務細節區塊時完成確認)等情形，在進行向性匹配前，會先審視該名會員的活動紀錄(地區資訊以 App 端的 GPS 提供)，確認為可能人選後，再行進一步的選拔。

# 五、結果討論

在本研究中，我們利用詞向量來設計各種類型的語意解析節點，並將節點串接成圖的結構，提出了一種彈性的語意解析方式，並為其完成了與服務平台的串連。然而，目前語意節點與語意圖都來自於人工定義，於額外擴充圖時，可能因主觀因素致使訓練特徵選擇偏頗，我們建議節點的訓練特徵當採用來自同類文本之共現詞並依 TF/IDF 進行排序，如此可得較佳成效。

# 六、參考文獻

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. Advances in Neural Information Processing Systems 26, (NIPS 2013).

[2] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean. 2013. Efficient

Estimation of Word Representations in Vector Space. ICLR Workshop , 2013.

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

[4] David M. Blei. 2011. Introduction to Probabilistic Topic Models. Communications of the ACM Volume 55 Issue 4 (2012). Pages 77-84.

[5] Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M. Blei.Reading. 2009. Tea Leaves: How Humans Interpret Topic Models. Advances in Neural Information Processing Systems 22 (NIPS 2009).

[6] Arvind Neelakantan and Michael Collins. 2014. Learning Dictionaries for Named Entity Recognition using Minimal Supervision. Conference of the European Chapter of the Association for Computational Linguistic, 2014.

# 科技部補助專題研究計畫成果自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現（簡要敘述成果是否具有政策應用參考價值及具影響公共利益之重大發現）或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■ 達成目標

□ 未達成目標（請說明，以 100 字為限）

□ 實驗失敗

□ 因故實驗中斷

□ 其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形(請於其他欄註明專利及技轉之證號、合約、申請及洽談等詳細資訊)

論文：□已發表□未發表之文稿 □撰寫中 ■無

專利：□已獲得□申請中 ■無

技轉：□已技轉□洽談中 ■無

其他：(以 200 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性，以 500 字為限）。

本計畫了提出一種彈性的語意解析手段，可將其銜接上 Facebook、Line 等社群平台，提供即時的用戶意圖解析服務，並根據用戶需求的不同，從中進行對應的特徵抽取，以完成複雜的多輪式對話處理。

4. 主要發現

本研究具有政策應用參考價值： ■否 □是，建議提供機關_____
(勾選「是」者，請列舉建議可提供施政參考之業務主管機關)

本研究具影響公共利益之重大發現：■否 □是
說明：(以 150 字為限)