國立台北大學資訊工程學系專題報告

影音問答服務系統

專題組員:胡瓊文、羅雅臻、莫鎧禎、黃紫嫻

專題編號: PRJ-NTPUCSIE-107-001
執行期間:107 年 09 月 至 108 年 05 月

一、 摘要

當新生或民眾進入校園時，可能對環境不熟悉，會對校園內系館或商店位置感到疑惑。因此我們希望開發出利用影像控制、語言辨識等技術的問答服務系統,幫助學生解決問題，並藉由結合影像、語音和文字，讓問題的回覆能夠變得更淺顯易懂，方便使用者更清楚的理解問題答案。本專題計畫”影音問答服務系統”，追求開發出可以與使用者互動並回答使用者疑問的軟體，針對校園作為使用場景的影音問答系統可作為一個校園特色吸引更多人來使用，也可以減輕學校職員回答問題的負擔。

二、 簡介

(一) 研製背景

目前市面上的問答系統，有純語音的問答服務，如:小米的小愛同學、Google 的 Google Home，但是當環境吵雜無法聽清楚答案時，純語音的問答服務無法發揮功用，會造成使用時間的浪費和使用者的不方便；Messenger 的 Chatbot 則是純文字的問答服務，但是使用者使用時一定要直視螢幕並打字,無法同時做其它事，因此我們想打造一套問答系統來解決以上市面語音系統會有的問題。

(二) 研究目標

我們希望以市面上有的問答系統作為基礎，藉由結合影像、文字、語音，讓使用者可以簡單快速的獲得答案，同時也可以減少一些人力資源成本。

(1)影像:

影像可以使用真人影像或是卡通版的趣味影像，真人影像讓使用者有專業的問答體驗，並可以節省人力資源；卡通版的則可以針對年齡較小的使用者，增加趣味性的問題和角色，引發他們對產品的好奇而使用。

(2)文字:

增加字幕讓使用者在吵雜的環境下使用問答服務系統，也可以得到問題的答案，藉由字幕達到輔助的效果，讓使用者能清楚理解容易混淆的字句。

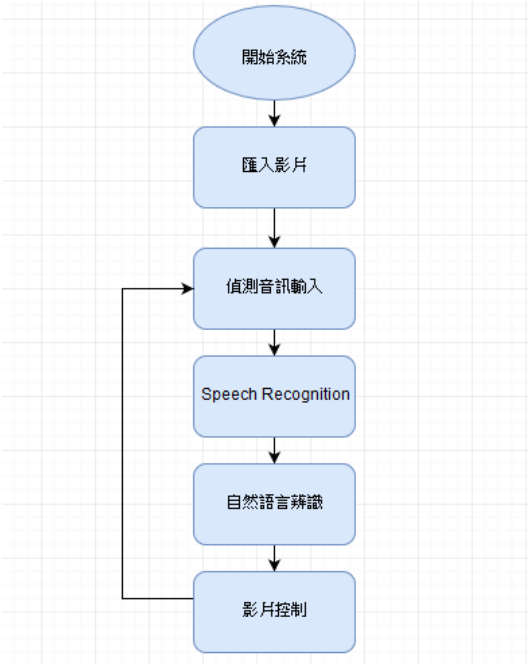
(三) 主要預期效益

預期放在校園裡面幫助大家查詢校園地理位置，能夠回答各系院以及課堂上課的位置，同時會顯示顯示到達目的地的最快速的路徑。此外臺北大學有許多三峽當地居民喜歡帶小孩、寵物到校園散步，或者是一些校園活

動會有來北大校園參觀的民眾，這些人可能不熟知臺北大學的校園環境，此時可以透過影音問答系統得知臺北大學校園內有哪些進駐商店以及商店營業時間，例如:需要買水可以詢會問哪裡有便利商店、想吃甜點可以詢問阿默蛋糕、糖匠或小木屋，我們的專題作品可以快速的解決這些校園相關的問題，讓使用者減少自己使用手機查詢學校網頁的步驟，快速且精準地回答使用者疑問。

三、 專題進行方式

本系統之架構如下：



圖一、系統架構圖

第一步將系統啟動，第二步偵測使用者音訊，輸入到系統中做辨識處理。第三步將偵測到的音訊進行特徵擷取，透過聲學模型和語言模型解碼，將音訊轉換為文字。第四步利用自然語言辨識抓取關鍵字，第五步，搜尋資料庫讀取關鍵字對應的影片時間，並調整影片位置完成輸出。系統架構圖如圖所示，詳細研究方法及步驟說明如後。

3.1 音訊的輸入

我們使用 python 調用 pyaudio 使用麥克風錄製 wav 聲音文件，數據本身的格式為 PCM 或壓縮型，屬於無損音樂格式的一種。

3.2 Speech Recognition

Speech Recognition 主要分為提取特徵、聲學模型、語言模型和訓練四部分。

- 提取特徵:

語音辨識系統使用特徵提取模組處理輸入訊號，將訊號處理盡量降低環境雜訊等因素對特徵造成的影響。聲學特徵提取將資訊大振幅壓縮，並將訊號解卷使圖形劃分器能更好劃分。

(1)線性預測分析：依據人的發聲原理，對聲道模型研究得到聲學特徵。當實際語音的採樣值和線性預測採樣值之間達到均方差最小 LMS，即可得到線性預測係數 LPC。對 LPC 的計算方法有自相關法、協方差法、格型法等。

(2)倒譜係數：利用同態處理方法，對語音訊號求離散傅立葉變換 DFT 後取對數，再求反變換 iDFT 就可得到倒譜係數。使用倒譜可以提高特徵參數的穩定性。

(3)Mel-Frequency Cepstral Coefficients：臨界頻寬指的是兩個頻率相近的音調同時發出時，人只能聽到一個音調，Mel 刻度是對兩個音調的頻率差小於臨界頻寬的度量方法。

- 聲學模型：

典型系統多採用隱馬爾科夫模型(HMM)進行建模，一個聲學模型包含數個狀態。我們使用音節或是音素作為一個聲學模型。

(1)音節：以中文來說，完整發音的單位，一個字元對應一個音節；以英文來說，一個詞彙可以對應到數個音節，例如 tomorrow 有三個音節。

(2)音素：中文「大」的發音可以拆解成ㄉ和ㄚ兩個音素，音素可分為三種方式:

I. Monophone：以單一音素作為一個聲學模型，例如ㄇ。

II. Biphone：以連續兩個音素作為聲學模型，例如:ㄇ出現於ㄇ-ㄚ和ㄇ-ㄋ將視為兩個不同的聲學模型。

III. Triphone：以連續三個音素作為聲學模型，例如:ㄕ+ㄨㄋ及ㄔ+ㄨㄤ視為兩個不同的聲學模型。

- 語言模型

語言模型對系統語言進行建模，統計語言模型使用機率統計的方法來計算語言單位的規律，n 元語法假設第 n 個詞的出現只與前面 n-1 個詞相關，而與其它任何詞都不相關，整句的機率就是各個詞出現機率的乘積，可以透過從語言中統計 n 個詞同時出現的次數得到，如果一個詞的出現依賴於它前面出現的一個詞，稱為 bigram(二元語法)。

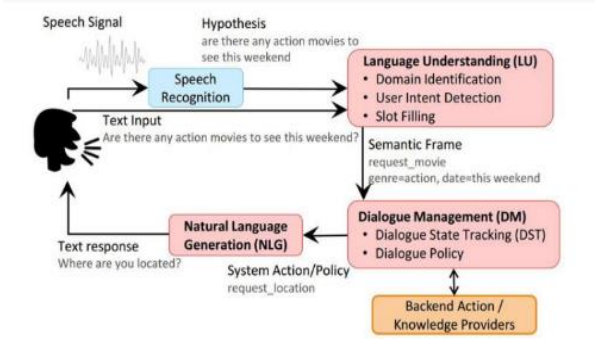
$$P(S) = P(w1, w2, w3, .., wn)$$
$$= P(W1)P(W2|W1)P(W3|W1, W2)..P(Wn|W1, W2, .., Wn-1)$$
$$\approx P(W1)P(W2|W1)P(W3|W2)..P(Wn|Wn-1)$$

若一個詞的出現依賴於它前面出現的兩個詞，稱為 trigram(三元語法)。

$$P(S) = P(w1, w2, w3, .., wn)$$
$$= P(W1)P(W2|W1)P(W3|W1, W2)..P(Wn|W1, W2, .., Wn-1)$$
$$\approx P(W1)P(W2|W1)P(W3|W2, W1)..P(Wn|Wn-1, Wn-2)$$

3.3 自然語言辨識

Chatbot 有 3 種關鍵的技術，自然語言理解、對話管理系統以及自然語言生成系統。



圖二、自然語言生成系統流程

(1)自然語言理解:

需要使用機器學習的技術來訓練出一套可以讓機器人理解人類語句背後意圖的功能。一般會利用語句中的「意圖」(Intent)和「實體概念」(Entity)，因此設計者需要提供大量擁有同一個意圖的句子，例如「籃球課在哪裡上課?」這個問句的意圖就是要前往上課地點。之後再拆解出句子中的不同條件，例如籃球課、地點，這些條件就是實體概念，Chatbot

的程式就可以根據意圖搭配所掌握的實體概念，來決定要採取的行動，之後就會進入對話管理的階段。

(2)對話管理系統:

主要有 3 種行為，要求更多資訊、確認資訊和回報資訊給使用者，如果從自然語言理解中得到的條件不完整，就需要讓 Chatbot 向使用者請求更多資訊，若得到的條件過於模糊就進行確認，如果一切無誤就執行命令並回報資訊給使用者，例如上課的地點。

(3)自然語言生成系統:

執行完自然語言理解和對話管理後，就要根據最後要回報的資料生成出符合自然語言的句子，目前要合成出符合的句子則要透過設計者設定規則。

3.4 影片控制

- Omxplayer:

Omxplayer 是針對樹莓派播放器，支持硬件解碼，屬於沒有介面的全螢幕播放器。影片須先建立資料庫儲存影片的開始時間和持續播放時間，透過自然語言辨識擷取的關鍵詞搜尋資料庫中的對應播放資料，找到起始播放時間,將影片時間進行調整播放，持續時間結束會回原始的待機畫面，繼續偵測音訊等待下一個音訊輸入。

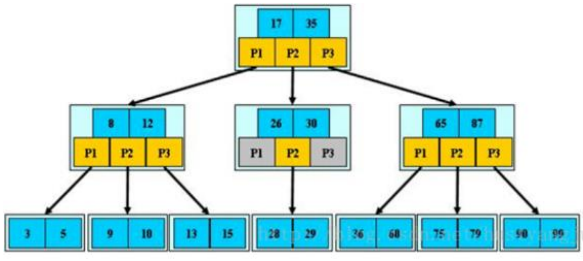
我們採用一鏡到底不需要剪輯影片的方式控制影片播放，如果要增加或修改影片不需要重新拍攝，只要加入影片再修改資料庫的數據就可以達成問答系統問題的更新。

- SQLite

資料庫使用 SQLite，它不是一個用戶端/伺服器結構的資料庫引擎，而是被整合在用戶程式中，所以應用程式經由程式語言內的 API 呼叫來使用 SQLite 的功能，這在減少資料庫存取延遲上有積極作用。本專題較適合這種形式的資料庫，因為我們只需要快速搜尋資料，而相較於跨行程通訊，SQLite 這種單一行程中函式呼叫的方式較符合我們的需求，所使用的演算法是 B-tree 和 B+tree。

(1)B-tree:

在 SQLite 中每一個表都用一個唯一的 B-tree 儲存，數據庫中有多少個表就有多少個 B-tree。與一般的 AVL Tree 或 Red-Black Tree 有一個明顯的區別那就是一個內節點可以有很多個 children。



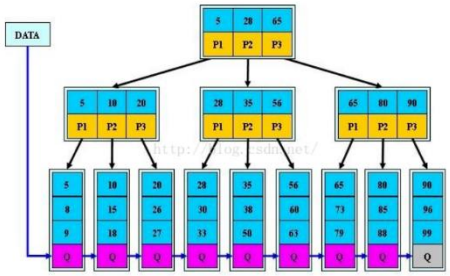
(2)B+tree:

SQLite 是根據關鍵值建立索引表的，在創建时會按照一定順序排序。B+tree 的根節點、中間節點和葉子節點也和 B-tree 一樣，對應一個 paper，只不過根節點和

中間節點的鍵值域已經排序，而且數值域不再儲存數據，而是儲存指向下一層 paper 的指標。

i. 所有的葉子節點中包含了全部關鍵字的資訊，及指向含有這些關鍵字記錄的指標，且葉子節點本身依關鍵字的大小自小而大的順序連結，而 B 樹的葉子節點並沒有包括全部需要查詢的資訊。

ii. 所有的非終端節點可以看成是索引部分，節點中僅含有其子樹根節點中最大（或最小）關鍵字，而 B 樹的非終節點也包含需要查詢的有效資訊。

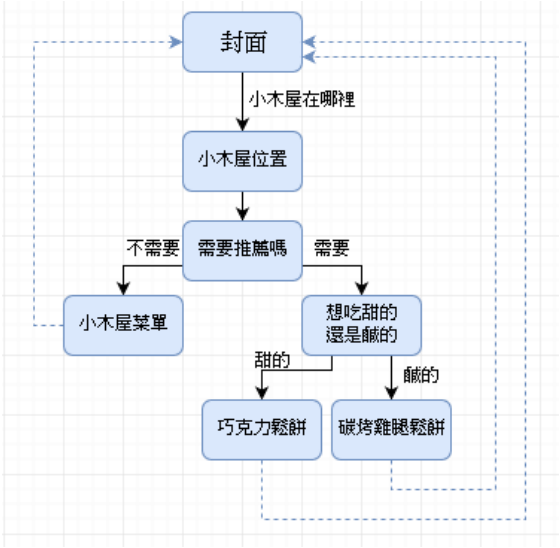


(3)使用 B+樹的原因:

B+樹查詢效率更加穩定，由於非終節點並不是最終指向檔案內容的結點，而只是葉子節點中關鍵字的索引。所以任何關鍵字的查詢必須走一條從根節點到葉子節點的路。所有關鍵字查詢的路徑長度相同，導致每一個數據的查詢效率相當。

四、 主要成果與評估

本系統之執行流程範例如下:



圖三、執行範例流程圖

在 Linux 的環境下使用 Python、Google Recognition 進行開發。

首先將影片進行匯入偵測是否有音訊輸入，音訊經由語言辨識轉換為文字，再進行自然語言辨識得到該句話的關鍵詞，搜尋資料庫擷取關鍵詞的影片時間進行影片控制和轉換。

專題實例:

使用者詢問小木屋鬆餅的位置在哪裡，系統回答位置後詢問是否需要推薦小木屋鬆餅的人氣產品(圖四)。

(圖四)

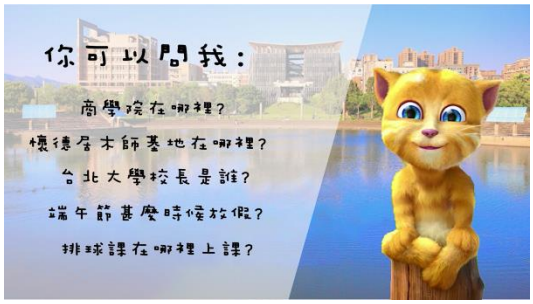
當回答為不需要時，系統會提供

小木屋鬆餅的菜單讓使用者自行參閱,當使用者回答需要推薦時，系統會再詢問需要推薦甜的商品還是鹹的商品(圖五)，並依照使用者的回覆給出相對應的答案(圖六)。

(圖五)

(圖六)

問答結束後系統皆會回到封面再重新偵測音訊，讓所有使用者能夠有一樣的問答體驗。(圖七)



(圖七)

五、 結語與展望

本次的專題研究中，成功實作出結合影像語音和文字的問答服務系統，能夠在短時間內回覆答案，減少上網查詢的繁瑣過程和時間耗費，並可以針對不同場景做問題設計和回答擴充影音問答服務系統，例如:在百貨公司裝設可以為消費者做地點導覽或優惠簡介。

未來期望可以做出針對個人化的問答服務，例如:使用學生證或教師證核對卡片、裝設攝影機進行人臉識別等方式確認個人身分，可針對個人給更合適的回答。

六、 銘謝

感謝指導教授對我們一年來的指導從一開始的專題主題討論、軟硬體架構全程參與與協助，在我們毫無頭緒時，指引我們方向，讓我們對於整個專題的執行流程有更深刻的理解。

七、 參考文獻

- *清華大學出版社《資料結構(C 語言版)》(2007 年版)，編著者嚴蔚敏，吳偉民*
- *Audio Signal Processing and Recognition (音訊處理與辨識)，張智星*
- *混合語言之語音的語言辨認 Language Identification on Code-Switching Speech，呂宜玲*
- *自然語言處理(NLP)斷開中文的鎖鍊 自然語言處理，馬偉雲*
- *〈計算語言學——人工智慧‧語言學‧認知科學的結合〉，黃居仁*