國立交通大學

資訊管理研究所

碩士論文

旅遊英語會話相似主題段落發掘之研究

A Study of Discovering Similar Topic Segments in Travel English Conversation

研究生：陳彥廷

指導教授：劉敦仁 博士

中華民國九十六年七月

旅遊英語會話相似主題段落發掘之研究

A Study of Discovering Similar Topic Segments in Travel English Conversation

研 究 生： 陳 彥 廷　　Student: Yen-Ting Chen

指導教授： 劉 敦 仁　　Advisor: Dr. Duen-Ren Liu

國立交通大學

資訊管理研究所

碩士論文

A Thesis

Submitted to Institute of Information Management

College of Management

National Chiao Tung University

in Partial Fulfillment of the Requirements

for the Degree of

Master of Science in Information Management

July 2007

Hsinchu, Taiwan, the Republic of China

中華民國 九十六 年 七 月

旅遊英語會話相似主題段落發掘

研究生：陳彥廷　　指導教授：劉敦仁博士

國立交通大學資訊管理研究所

摘要

本研究希望能建立一套自動化的方法輔助使用者在英語旅遊會話上的學習。當使用者在閱讀一篇連貫的旅遊英語會話時，系統能將該篇會話切割出許多主題獨立的段落，並針對各個主題段落舉一反三地從現有的語料庫中找出相似主題的會話段落，推薦給使用者，讓使用者學習英語時得收觸類旁通之效。

本研究的研究重心在於會話段落主題相似度之比較，並提出一套以語料庫統計資訊為基礎的字根重要性權重與字根相關性權重的設定方法，以增進段落語意相似度比較之準確率。實驗結果顯示，各項權重設定方法均能有效提昇相似度比對效果，使得本研究所提出之相似段落發掘方法優於傳統以詞彙比較為基礎之相似段落發掘方法。

關鍵字:自然語言處理 Natural Language Processing、資訊擷取 Information Retrieval

A Study of Discovering Similar Topic Segments in Travel English Conversation

Student: Yen-Ting Chen **Advisor:** Dr. Duen-Ren Liu

Institute of Information Management

National Chiao Tung University

Abstract

In this study, we hope to design an automated method to help users learn travel English conversation. When a user reads a continuous travel English conversation, the system will partition it into multiple topic segments. For each segment, the system will discover similar topic segments from the corpus repository and recommend them to the user to help the user learn more about each topic segment.

The focus of the research lies in the measure of similarity between topic segments. This study proposes a set of weighting methods about the importance and correlation of word stem based on corpus statistics in order to promote the precision of the similarity measure. The experimental results show that all of the weighting methods will improve the performance of the similarity measure, making our similar segment discovery method outperforms the traditional similar segment discovery method based on lexical matching.

Keywords: Natural Language Processing, Information Retrieval

誌謝

研究是一條陌生而漫長的路，走到今天，終於算是告一段落了。這一路走來，讓我明白自己能力的侷限，也讓我體會到人情的溫暖。雙親、師長、同學、朋友、學長姊與學弟妹們給予的支持，不管是實質上或是精神上，都是我突破每一道瓶頸的關鍵。

在這裡，我感謝所有在我生活上或是學業上曾經給予過幫助與陪伴的人，你們任何一點點的關懷或恩惠，都化作了我繼續向前的動力。碩班生涯將盡於此，雖然有短短的兩年，但因為你們，這些日子將成為我最珍貴的回憶。

目錄

圖目錄

表目錄

1 緒論

1.1 研究背景與動機

在這個全球化蓬勃發展的年代，做為國際語言的英語，重要性不言可喻。由於數位學習是當今世界的趨勢，電腦輔助語言教學（Computer Aided Language Learning，簡寫為 CALL）的研究，也因而應運而生。CALL 的發展方向，不單單只是將英語教材數位化，更重要的是讓電腦以更有智慧的方式輔助使用者，增進學習的成效。

本研究所著重的範疇為「旅遊英語會話」的學習。市面上的旅遊英語會話書籍，對於某個情境，通常都只能提供一兩篇範例會話。然而，實際出國旅行時，我們所遭遇的情況卻是千變萬化。因為語言是活的，同樣的對話主題，就算是在相似的情境下也常常是豐富而多變的。

本研究希望能建立一套自動化的方法，讓使用者在閱讀一篇連貫的會話時，電腦能將該篇會話切割出許多小主題，並針對每一個主題段落舉一反三地從既有的語料庫中找出相似的會話段落推薦給使用者，讓使用者學習英語時得收觸類旁通之效。

評估文章間的語意相似度的相關研究，在資訊擷取（Information Retrieval）以及自然語言處理（Natural Language Processing）領域中已經有相當長久的研究，也產生出許多經典的相似度比對方法，如 TF-IDF 的單字權重設定、機率模型與向量模型等等。然而，學者大部分都將研究對象設定為具有相當長度的文章，所比較的文章主題上也具有一定的差異，此與本研究所面臨的問題有許多本質上的不同。

首先，本研究進行相似度比對的會話段落常常只由兩三個句子所組成，相對於一篇完整的文章，可使用來判斷語意的資訊十分稀少。另外，為了推薦相似段落給使用者，語料庫中蒐集大量「情境相同」之範例會話，使具有相同主題的會話段落一再出現於不同的範例會話之中，造成許多過去學者使用之相似度比對方法將無法適用於本研究。例如，TF-IDF 權重設定就不適合應用在同質性高的語料庫之上。

基於上述討論，本研究提出不同於過去的方法進行短文語意相似度的比較，並將其實際運用於英語語言教學之上。

1.2 研究目的

本研究的目的為設計一個相似段落發掘的方法。當使用者瀏覽一篇旅遊會話時，電腦能自動將此會話切割為許多主題獨立且完整的段落，並針對每個段落舉一反三地從現有的語料庫中找出內容不同，但「主題相似」的會話段落，推薦給使用者，讓使用者學習英語會話時能觸類旁通，達事半功倍之效。

如表 1，左手邊「原始對話」為使用者從資料庫中選取的一篇海關入境的旅遊會話，電腦自動將原始對話切割成兩個語意完整的段落，並且從資料庫裡找出右手邊與原會話段落具有相似主題的「系統推薦對話」，提供給使用者參考。

表 1 對話段落推薦範例

	原始會話	系統推薦會話（一）	系統推薦會話（二）
段落一	How long are you planning to be here? *About three weeks.*	How long do you plan to stay here? *About one week.*	Are you traveling Alone? *I am with a friend. We plan to stay here for two weeks.*
段落二	And what is your purpose for visiting Australia? *My wife and I are on vacation.*	And are you here for business or pleasure? *Just a vacation.*	Is your visit for business or pleasure? *Pleasure, sir.*

2 文獻探討

本論文的研究大致可以分為兩個部份：1） 將一篇完整的對話切割成許多主題獨立且完整的段落。2） 比對各段落間的相似度，從而找出主題相似的段落。

2.1 主題分割

關於第一部份「主題分割（Topic Segmentation）」的工作，相關研究由來已久。此研究議題的定義為「在一篇連續的文稿中，偵測出意義完整的段落，每個段落各自代表一個獨立的主題。」此領域中，大部分的研究都將背景資料設定為長達五六頁以上，未標示文章主題的文章或對話紀錄 [Hearst 1997; Galley et al. 2003; Utiyama and Isahara 2001]。

其中最著名的研究為 1997 年，學者 Marti A. Hearst 所描述的 TextTiling 方法 [Hearst 1997]。此方法假設人們在討論一個主題時，常會重複某些跟主題有密切相關的詞語，使詞語的分佈產生叢集的現象（Lexical Coheshion）。TextTiling 的方法內容為：「文章中的每一個句子的結束時，計算句子結束點前後特定長度的內文的相似度，若相似度突然在某個結束點突然大幅降低，則視此句的結束為一個段落的結束，亦即下一句為新的段落的開始。」TextTilling 中的作法是在句子（論文中所使用的為虛擬句）的結尾，往前與往後拉出一段固定長度的區間，並利用向量相似度來計算兩區間詞彙的相似度。

如圖 1，橫軸為文章中的句子，縱軸為句點前後兩區間的詞彙相似度。當詞彙相似度劇烈降低又迅速爬升時，即為主題轉換的信號(以虛線標示)。

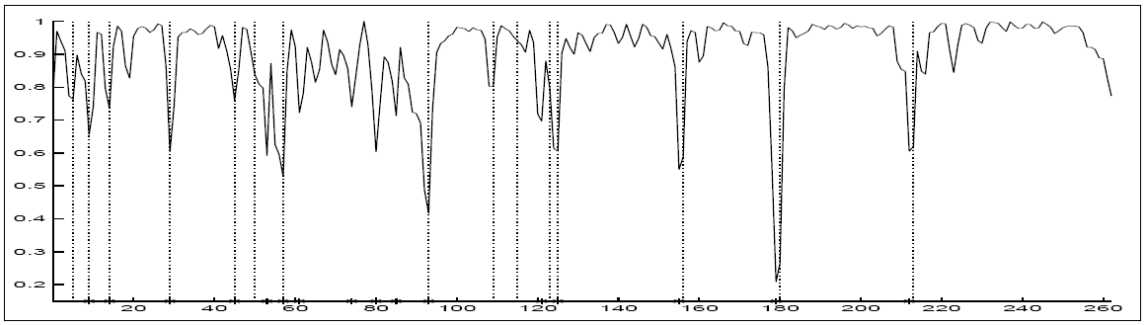

圖 1 TextTilling 段落分割方法

在[Galley et al. 2003]中，學者把研究目標放在對話紀錄的分割。作者加入了 Lexical Chain 的觀念來計算語句結束的前後區間的相似度。他的假設為：「主題轉換的發生點，通常位於強烈且明顯的詞彙重複的開始或結束。」作者以 Lexical Chain 來表示單一詞彙的分佈。給予短而密度高的 Lexical Chain 高分；反之，長而密度低的則給予低分。最後以 Lexical Chain 為特徵，比較兩兩連續區間彼此的相似度，當相似度驟降時，表示該處為段落的分割點。

以「字彙的重複」為分割段落線索的研究為數眾多，除了上述的方法以外，其他的想法還包括以語句之間的相似度矩陣作為分的割依據 [Choi 2000]或是採用機率統計模型 [Utiyama and Isahara 2001; Reynar 1999]等等。

此外，在分割對話紀錄時，也有學者將對話中一些語言上的特徵作為分割對話主題的依據。例如口語上暗示新對話主題開始的關鍵字詞、對話中的停頓或是主要發言角色的轉換等等 [Hsueh et al. 2006; Galley et al. 2003]。

2.2 短文相似度比較

文件主題相似度比對方法，在資訊擷取領域中已有長久而豐富的研究。然而由少量的句子所構成的短文由於字彙稀少，其語意主題相似度比對相較於文件將更為困難。過去學者的研究方向大致可分為下列幾項：

1) 字彙的重複

此方法假設兩篇短文共用的字彙越多，相似度越高 [Hatzivassiloglou et al. 1999]。此種方法包括了許多不同的變形，其中包括是否考慮單字的大小寫之別，以及是否將單字轉換為字根的形式後再比較短文間的字彙重複。另外也有學者在比較字彙的重複之前，先將短文中資訊意義較低的停用字（Stopwords）移除，以減少相似度比對時的雜訊。此方法最簡單的實做方式是將語句以向量模型表示，向量的一個維度代表一個字彙，利用向量內積求得短文間的相似度。然而，語言的使用千變萬化，同樣語意的短文，使用的字彙卻可能有相當大的差異。據此，短文間的語意相似度不一定與字彙重複的多寡成正比。

2) 字義相似度

由於字彙重複，非有即無，無法考量到不同字彙字義上的相似程度。有學者為了彌補此一缺憾，將專家所建立的階層式語意知識本體導入字彙相似度比較之中[[Dolan et al. 2005; Li et al 2006; Corley and Mihalcea 2005; Hatzivassiloglou et al. 1999; 鄭守益和梁婷 2005]。

此類知識本體以階層式的架構表示人類對於單字意義上的分類，位於越上層的單字，語意上的概念越抽象；反之，越下層單字的語意越為明確。中文的單字

語意知識本體如哈爾濱工業大學所發展之同義詞詞林，英文則以普林斯頓大學建立的 WordNet 最為有名。此階層式的知識架構，為判斷單字間語意相似度的重要依據。

以單字語意知識本體為基礎的字義相似度計算的方法眾多，知名的包括 Wu and Palmer （1994）、Resnik （1995）、Jiang & Conrath （1997）與 Lin （1998）[Corley and Mihalcea 2005; 顏偉和荀恩東 2004; Pedersen]。

由階層式知識本體計算得單字字義相似度之後，便可將此一資訊帶入語句相似度計算之中，如圖 2 [Li et al 2006]。其假設為，若兩短文所包含的字彙，字義上的相似度越高，則短文越相近。在[Li et al 2006; Corley and Mihalcea 2005; 鄭守益和梁婷 2005]的研究皆使用到此一方法。

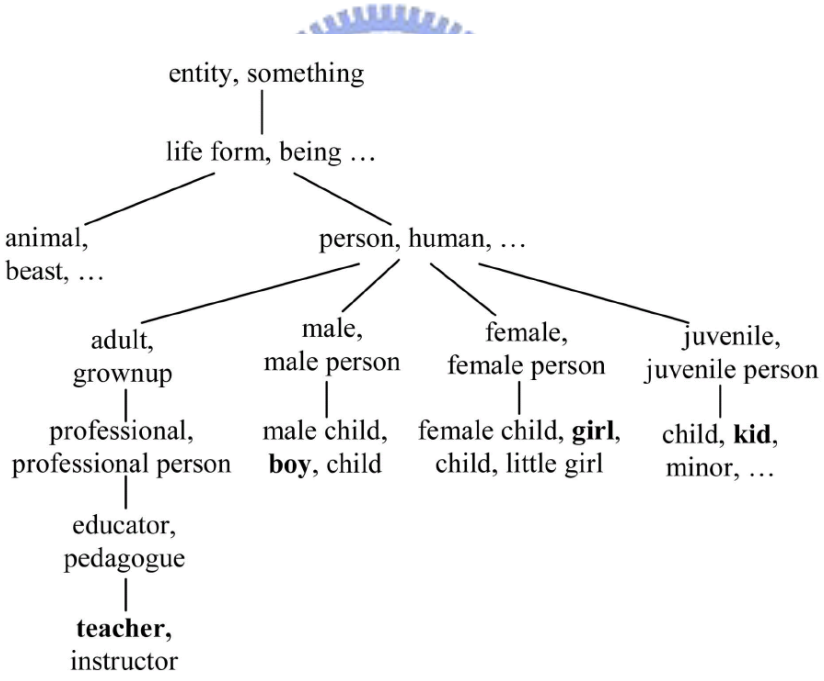

圖 2 階層式語意知識體

此方法有幾項缺點，首先，人類所使用的字彙其中的概念錯綜複雜，難以完整且清楚地將其表示為階層式的知識本體。其次，不管是 WordNet 或是同義詞詞林，這些知識本體都是由專家所訂立，因此字義的分類難免會有主觀的成份存在。

3) 單字詞性分類

在短文中，不同詞性的單字具有不同的角色與功能。[Corley and Mihalcea 2005] 只取短文中的名詞、動詞、副詞及序數，四種詞性之單字，並獨立處理不同詞性的字彙。[Hatzivassiloglou et al. 1999] 則特別針對短文中的「名詞片語」與「專有名詞」進行比對，統計兩篇短文共有的名詞片語與專有名詞的數量。

4) 單字資訊內容

短文中每個單字含有多寡不同的資訊內容（Information Content），一般而言，資訊內容越高的單字越能幫助辨識短文的語意主題。過去學者時常以平衡語料庫（balance corpus）計算單字資訊內容。所謂的平衡料庫，義為均衡地包含各種主題及各式文體的語料庫。單字在平衡語料庫中出現機率越小者，資訊內容越高；反之，出現機率越高則資訊內容約低。利如介係詞”to”和”for”的資訊內容就不如名詞”travel”與”business”來的高。在[Li et al 2006]的研究中，即依照每個單字的資訊內容設定不同的字彙權重值;短文中資訊內容越高的單字給予越高的權重。

5)　潛在語意分析（Latent Semantic Analysis ）

潛在語意分析可說是向量模型的延伸，一種以整個語料庫為依據的文件或短文相似度比較方法[Foltz et al. 1998; Landauer et al. 1997; 汪若文 2004]。進行潛在語意分析時，首先需建立一個字彙對情境的矩陣（word by context matrix）。經過 singular value decomposition （SVD）分解，會將該矩陣切為三個矩陣的乘積。將此三個矩陣做降階處理後，再重新建構起原本的字彙對情境矩陣。經過這樣的處理，LSA 可將散佈在情境之中的知識表現在重新建構的字彙對情境的矩陣當中。

然而，由於字彙對情境的維度是固定的，當情境限定為由少數幾個句子構成的短文時，矩陣可能會非常的稀疏。此外，潛在語意分析也沒有考量到字彙的排列等任何語法上的資訊 [Li et al 2006]。

3 相似主題段落發掘方法論

本章將詳述相似主題段落發掘之方法。3.1 節為本研究使用之會話語料的簡介，接著在 3.2 節大致介紹整個研究的架構及流程。3.3 節為落分割方法，而 3.4 則為本研究所提出之段落主題相似度比較的一系列方法。最後在 3.5 節將系統找出之相似段落推薦給使用者。

3.1 會話語料

本研究所使用的會話語料為許多篇內容連貫且彼此獨立的旅遊英語會話，為遊客與當地人之間的對談，內容通常包含明確的目的，並非一般性的閒聊。

蒐集語料時，特地選取四個性質相異的對話情境，分別為海關入境、提領行李、速食店點餐與餐廳會話，如表 2。本研究將會話資料庫中包含的四個情境，分別獨立做處理，不考慮跨情境尋找相似的段落。

表 2 會話語料之情境及內容

會話情境	海關入境	提領行李	速食店點餐	餐廳會話
會話內容	入境審查、海關提問和攜帶物品的申報。	提領行李與通報行李遺失。	速食店點餐與結帳。	訂位、點餐與用餐問題

3.2 流程架構

系統流程如圖 3。首先是段落分割，將連貫的會話分為小段落。接著進行文字的前置處理、字根各項權重的計算與設定（包括詞性權重、情境權重與相關度權重）。最後以段落中的字根為特徵，進行段落相似度比較，把段落之間的相似度排序結果儲存於資料庫內。

當使用者瀏覽一篇連貫的英語會話時，我們便可自動辨識出其中包含的會話的段落，並從資料庫裡選出主題最相近的段落推薦給使用者參考。

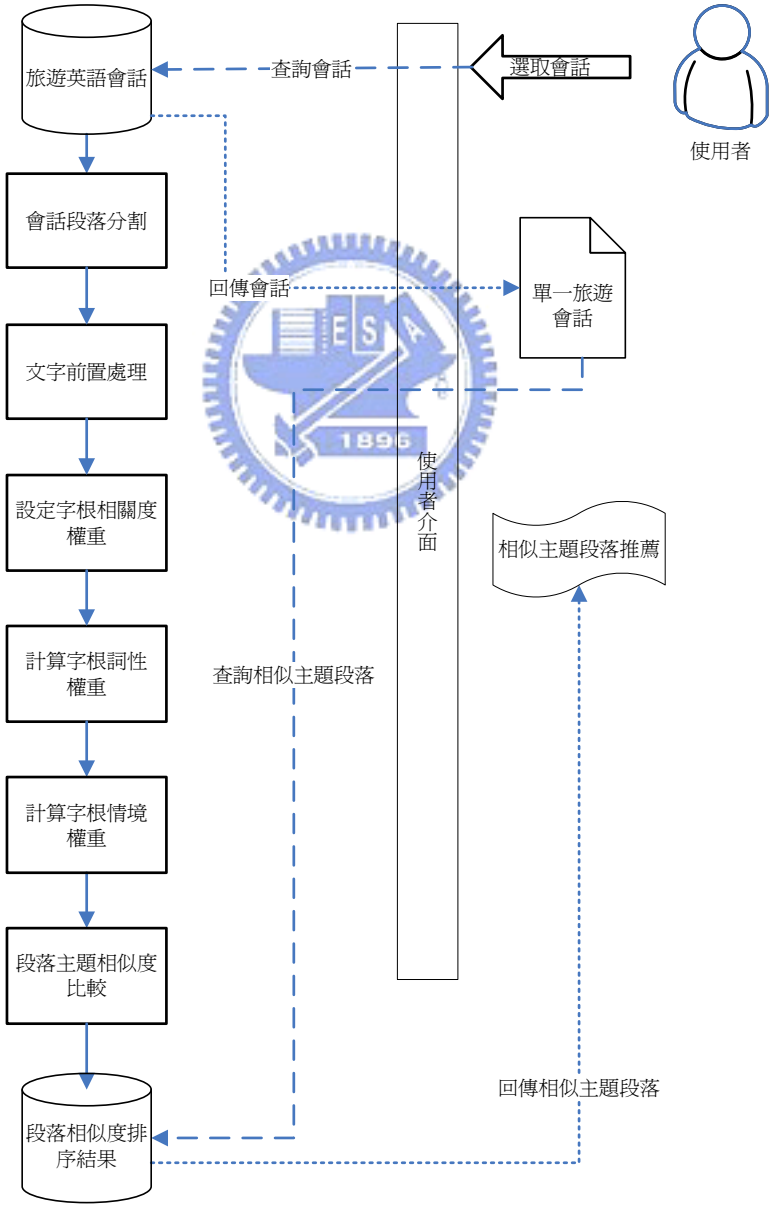

圖 3 系統架構

3.3 段落主題分割

一篇連貫的旅遊會話，其實還可以再細分出許多主題獨立的小段落。例如表 3 為一篇遊客與海關之間的對話，其中包含了三個會話主題，分別為「請求出示護照」、「詢問旅行目的」以及「詢問停留時間」，而這些各自包含不同主題的段落，就是我們系統中相似度比較的基本單位。

表 3 會話分段範例

段落主題	會話內容
請求出示護照	海關：Your passport, please. 旅客：Yes. Here you are.
詢問旅行目的	海關：What is the purpose of your trip? 旅客：I am here for sightseeing.
詢問停留時間	海關：How long are you going to stay here? 旅客：I'll be staying here for a month.

然而，利用電腦自動辨識出如此簡短的主題段落，並不是一件容易的事。目前學術界有關主題分割的研究，分割的對象限於有一定長度的文章或談話紀錄，主要是利用字彙的叢集現象或語言上的特徵來做段落分割[Hearst 1997; Galley et al. 2003; Utiyama and Isahara 2001]。而在本研究的語料中，一個主題段落可能只包含了短短的兩句話，字彙叢集的現象幾乎是看不到的。此外，旅遊英語範例會話並不包含一般談話紀錄中經常會看到的語言特徵（例如口語會話的發語詞”okay”、”alright”、”anyway”…等等），因此過去以語言特徵為依據之分割方法無法適用於此。

據我們的觀察，在旅遊會話的情境裡遊客與當地人通常是以「問句接答句（直述句）」的方式進行對話，而每一回的「問答」，也常代表一個主題完整的對話段落。例如表 3 中各個段落皆由海關與旅客間的「一問一答」所構成。

因此，本研究假設每一回的問答即代表一個主題完整且獨立的段落，並以「問答」作為段落切割的依據。除了會話開頭的段落可能為連續的直述句以外，其餘的段落皆包括「一個或多個連續的問句」與「一個或多個連續的直述句」兩部份，如表 4。

表 4 段落種類範例

會話開頭 （連續直述句）	海關：Next please. 旅客：Here's my ticket passport and the entry card. 海關：Thank you.
連續問句接 連續直述句	海關：Have you got anything to declare? Any gifts for people in this country? Have you got any spirits or tobacco? 旅客：I have bought some small gifts for friends, and here is a carton of cigarettes I bought on the plane. 海關：One carton of cigarettes is your duty-free allowance.

在會話中，以此方法分割的段落，常常只有一問一答兩個句子。此外，同主題的段落之間，使用的字彙也常會有相當大的差異。如表 5，段落 A 與段落 B 的主題皆為海關詢問遊客旅遊目的，其字彙上卻有很高的差異性，顯示段落間的「主題相似度」難以單純地從段落間的「字彙的重複」得知。因此，段落相似度比對是個十分具有挑戰性的工作，也是本研究的重心所在。我們將從下一節開始，介紹本研究的段落主題相似度比較方法。

表 5 相似主題段落範例

段落 A	段落 B
海關：And what is your purpose for visiting Australia? 遊客：Sightseeing. I am a tourist.	海關：Is your visit for business or pleasure? 遊客：Pleasure, sir.

3.4 段落主題相似度比較

本研究中，段落主題相似度主要是採用向量相似度（Cosine Similarity）方式計算，以單字字根為向量，並將本研究所提出之字根詞性權重、字根情境權重與字根相關度權重加入相似度比對中。3.4.1 為文字前置處理，3.4.2 到 3.4.4 介紹本研究各項字根權重設定方法，最後在 3.4.5 進行段落間的向量相似度比較。

3.4.1 文字前置處理

本研究的會話前置處理，包括數字一般化、單字小寫化與字根擷取三個部份。本研究並未採用資訊擷取領域時常見的停用字濾除（stop word removing），因為會話段落的長度十分簡短，若把段落中所有的停用字移除將使段落所包含的資訊量過於稀少，難以比較段落間的相似度。

1. 數字一般化

在會話中，某些對話主題常常會伴隨著數字的出現，例如餐廳訂位時會討論到訂位時間，速食店結帳時會提及餐點費用。這些數字，不論是阿拉伯數字或是英文單字，都是辨別會話主題的一項重要線索。然而，兩個討論相同主題的段落其中包含的數字，不管是時間或是金額，兩兩相同的情況並不多見。因此，我們將會話中的數字一般化；在段落相似度比對時，將所有的「數字」皆視為相同的元素，忽略其數值大小上的差異。例如某段落的主題為餐廳訂位，會話內容如下：

遊客：I would like a table for ***two*** at ***seven*** tonight.

服務生：Ok, no problem.

我們將段落中的所有數字以一個通用元素"IsANumber"替代，結果如下：

遊客：I would like a table for ***IsANumber*** at ***IsAnumber*** tonight.

服務生：Ok, no problem.

2. 單字小寫化

以英語的語言特性而言，依單字本身的特性與出現的位置，單字時常會有大小寫的區別。然而，大小寫資訊在口語對話中所包含的語意訊息十分有限，兩個相同的單字以不同的大小寫形式出現，其語意上，經常還是相當接近，甚至是完全相同的。因此，我們在相似度比對的過程中，去除大小寫資訊，將所有單字皆以小寫表示。

3. 字根擷取

在英語中，具有相同字根的字彙，通常包含著十分接近的語意。據此，我們

將單字的字根（stem）擷取出來，用字根來代表單字，以期在不使準確率大幅降低的情況下，能更有效地比對出主題相似的段落。本研究所使用的字根處理方法是 M. F. Porter 所提出的「字尾移除演算法（Suffix removal stemming algorithm）」[Porter 1980]。如表 6，上方的資料為原始的對話內容，下方則為經過小寫化及字根擷取後的會話。

表 6 字根擷取範例

A: Could you bring me a disembarkation card, please? B: Sure, I think we should still have some. B: What are the duty-free limits for Taiwan? A: A liter of spirits, 200 cigarettes and 500ml of perfume. B: Can you give me a multiple-entry visa? A: Certainly, but it will be 20 dollars more.
A: could you bring me a ***disembarkat*** card , ***pleas***? B: sure, i think we should still have some. A: what ***ar*** the ***duty-fre*** ***limit*** for taiwan? B: a liter of ***spirit***, 200 ***cigarett*** and 500ml of ***perfum***. A: can you give me a ***multiple-entri*** visa? B: ***certainli***, but it will be 20 ***dollar*** more.

3.4.2 字根詞性權重

在進行其他文字前置處理前，我們先將會話中的單字進行詞性的標記（Part of Speech），以提供後續設定詞性權重之用。本研究使用的詞性標記軟體為史丹佛大學所開發出的 Stanford Log-linear Part-Of-Speech Tagger [Stanford 2006]。此軟體依據 Penn Treebank English POS tag set [Marcus et al. 1993]的英文詞性分類標示英語單字詞性。

詞性標記的範例如表 7，每個單字的詞性皆標記於其後。例如：第一句的第二個字”is”在會話中的詞性為 VBZ(第三人稱單數動詞現在式)，第三個字”inside”則為 IN（介係詞）。

表 7 詞性標記範例

A: What /WP is /VBZ inside /IN the /DT bag /NN ? /.
B: It /PRP is /VBZ alcohol /NN . /.
A: You /PRP have /VBP to /TO pay /VB tax /NN for /IN over /IN three /CD bottles /NNS . /.
A: How /WRB much /JJ is /VBZ the /DT duty /NN ? /.
A: It /PRP is /VBZ 60 /CD U.S /NNP dollars /NNS . /.
B: Wow /UH , /, so /RB much /JJ . /.

在英語會話中，不同詞性的單字，往往具有不同的段落主題辨識力。過去資訊擷取領域的研究常單獨取出文章中的名詞進行處理 [Hatzivassiloglou et al. 1999]，因為相對於其他詞性，「名詞」含帶的資訊量通常是最高的。而「動詞」則因通用性較高，語意辨識度不如名詞，如：各式 Be 動詞和”do”、”get”或”take”等等常見的動詞。這些動詞常出現於各式各樣、不同主題的段落之中，資訊辨別度較差。專有名詞雖然屬於名詞的一類，不過其所反應的語意資訊往往過於狹隘，不能普遍地代表某一會話主題。如下面兩個對話片段：”I would like to visit *America* for a month.”與”I would like to visit *Australia* for two weeks.”兩句的主題皆為旅行時間長度，至於遊客拜訪的國家是美國或是澳洲，對於會話主題並無影響。

基於上述討論，我們為會話中的字根，依其原來所代表之單字的詞性，設定不同的權重，如表 8。

表 8 單字詞性權重表

詞性	權重
名詞（不含專有名詞）	高
動詞	中高
專有名詞	中低
其他詞性	低

3.4.3 字根情境權重

傳統資訊擷取常以 TF-IDF 做為單字權重設定的依據。然而在本研究的語料庫中，許多重要的主題關鍵字，往往散佈於多篇對話與段落之中，因此 IDF 指標在此並不適用。少了有效的 IDF 指標，TF-IDF 也將無法發揮作用。

為了辨識出對話的主題關鍵字，本研究提出以下假設：「相對於平衡語料庫，在某一會話情境中，單字出現的相對機率越高者，該字在此情境中的重要性越高；反之，相對出現機率越低者，重要性也越低。」本研究所使用的平衡語料庫為布朗語料庫[Brown Corpus]，整個語料庫約包含一百零一萬四千三百個英語字彙。本研究以下列公式計算每個單字的情境權重：

$$CP_{i,j} = \frac{\text{word i count in context j}}{\text{total word count in context j}} \quad (3\text{-}1)$$

$$BP_i = \frac{\text{word i count in balance corpus}}{\text{total word count in balance corpus}} \quad (3\text{-}2)$$

$$w_{i,j} = \log\left(\frac{CP_{i,j}}{BP_i} + 1\right) \quad (3\text{-}3)$$

$$w_{r,j} = \frac{\sum_{k=1}^{n} w_{k,j}}{n} \quad (3\text{-}4)$$

$CP_{i,j}$：單字 *i* 出現在會話情境 *j* 的機率

BP_i：單字 *i* 出現在平衡語料庫的機率

$w_{i,j}$：單字 *i* 在會話情境 *j* 的情境權重

k：字根 *r* 所代表的單字

n：字根 *r* 所代表的單字總數

$w_{r,j}$：字根 *r* 在會話情境 *j* 的情境權重

首先，計算單字 *i* 在指定情境與平衡語料中的出現機率，將機率相除後得兩機率的比值。由於不同單字的機率比值差異相當大，為縮小情境權重值的差距及避免負數權重的出現，實驗中將機率比加一取對數後得情境權重。

由於本研究以字根代表單字，實驗中我們將字根的情境權重為設定為其所代表的單字的情境權重的平均。如字根”visit”的情境權重為”visiting”與”visit”兩字的平均。

表 9 字根情境權重範例

情境	字根	情境權重($w_{i,j}$)
海關入境	purpos	1.558
海關入境	inspect	1.594
提領行李	tag	1.940
提領行李	lost	1.487
速食店點餐	dollar	2.154
速食店點餐	on	0.163
餐廳會話	drink	1.557
餐廳會話	most	0.209

如表 9，海關入境情境中，經常出現海關詢問旅行目的的會話段落，如：”What is the purpose of your visit?”或”What is the purpose of you visit here?”等等，”purpose “一詞在此會話情境中出現機率明顯高於平衡語料庫，造成了代表”purpose”的字根”porpos”獲得較高的情境權重。相反地，”most”一詞在餐廳會話的情境下出現機率反而較平衡語料庫為低，使字根”most”得到較低的情境權重。

3.4.4 字根相關度權重

會話中，相似主題的段落，使用的字彙經常有相當大的差異，若單純比較段落之間的單字組成，將難以辨識出彼此間的主題相似度。例如以下兩段範例對話：

（段落一）

海關：” Are you carrying any spirits or tobacco?”

遊客：”No.”

（段落二）

海關：“Any alcohol or cigarettes?”

遊客：”No.”

兩段皆為海關詢問旅客是否攜帶菸酒入境，然而，包含的詞彙卻大相逕庭。在前段，「酒類」海關使用的詞彙為 spirits，後段為 alcohol；前段中的「香煙」為 tobacco，後段則為 cigarettes。

上面的例子為同義字交替使用的情況，除此之外，兩兩主題相似的段落，也常常會使用「字義相似度不高」，但字彙「相關度高」的單字。

例如當海關詢問旅遊目的時，前來觀光的遊客可回答”I am here for pleasure.” 或是”I am on vacation.”。兩句話所描述的主題相似，字彙交集卻不甚明顯。其中，pleasure 與 vacation，字義相差甚大，難以找出兩單字之間任何字義上的相似之處。但是在海關入境的情境中，pleasure 與 vacation 卻時常出現在相鄰的語句裡，顯示兩單字間具有某種「相關性」。

據此，若我們能找出在某個情境下單字間的相關性，並在段落相似度比較時依此相關性擴充段落的特徵向量，將能有效提昇段落主題相似度比較的效果。

在本研究裡，我們假設「兩個相異的單字，共同出現在相同主題單元的情形越明顯，其相關度越高。」其中「主題單元」意指「只包含單一主題的會話片段」。對於英語會話，我們假設一個句子只討論一個主題，問句與緊接於其後的答句也只包含一個主題。據此，我們將「單一語句」與「相鄰的問答句」設定為主題單元，如表 8 為一篇英語旅遊會話，總共包含六個句子與四個主題單元。

表 10 主題單元範例

語句型態	內容	主題單元
（1）直述句	Please take out your customs declaration card.	單元 1
（2）問句	What's inside this luggage?	單元 2
（3）直述句	Some clothes and personal items.	單元 2
（4）問句	Do you have anything that needs to be declared?	單元 3
（5）直述句	I have nothing to declare.	單元 3
（6）直述句	You may go now.	單元 4

因本研究以字根代表單字，故我們計算的為字根之間的相關度。我們以概似比估計值檢定（Likelihood Ratio Test），判斷兩個字根之間的相關度 [Manning and Schütze 1999]。 $P(r_2 \mid r_1) = p_1$

我們提出以下兩項假設： $P(r_2 \mid \neg r_1) = p_2$

$H1: P(r_2 \mid r_1) = p = P(r_2 \mid \neg r_1)$

$H2: P(r_2 \mid r_1) = p_1 \neq p_2 = P(r_2 \mid \neg r_1)$

r_1、r_2 為兩個相異的名詞字根。在 H1 中，我們假設 r_2 出現在某一主題單元的機率與 r_1 是否出現於該主題單元無關。而 H2 則假設某一主題單元是否包含 r_1，將會影響 r_2 出現於該主題單元的機率。我們假設：當 H2 為真時，p_1 遠大於 p_2，即 r_1 出現時 r_2 經常會跟著一起出現。p_2 遠大於 p_1 的情況十分少見，因而忽略之。

我們以最大可能估計（Maximum Likelihood Estimate）求出 H1 中的 p 及 H2 中的 p_1 與 p_2，如以下公式。

$$p = \frac{c_2}{N} \quad (3\text{-}5)$$

$$p_1 = \frac{c_{12}}{c_1} \quad (3\text{-}6)$$

$$p_2 = \frac{c_2 - c_{12}}{N - c_1} \quad (3\text{-}7)$$

c_1：包含 r_1 的主題單元數

c_2：包含 r_2 的主題單元數

c_{12}：同時包含 r_1 與 r_2 的主題單元數

N：情境中主題單元總數

其中，c_i 為包含 r_i 的主題單元數，例如字根"claim"在「提領行李」的會話情境裡共出現於 15 個相異的主題單元中。我們假設字根是否出現於某主題單元為二項分佈（Binomial Distribution），如公式（3-8）。

$$b(k, n, x) = C_k^n x^k (1-x)^{(n-k)} \quad (3\text{-}8)$$

因此，代入我們實際觀察到的 c_1、c_2 與 c_{12} 值，H1 與 H2 的可能性分別為：

$$L(H1) = b(c_{12}, c_1, p)\, b(c_2 - c_{12}, N - c_1, p) \quad (3\text{-}9)$$

$$L(H2) = b(c_{12}, c_1, p_1)\, b(c_2 - c_{12}, N - c_1, p_2) \quad (3\text{-}10)$$

我們將以上兩項估計值的比值設為 λ。

$$\lambda = \frac{L(H1)}{L(H2)} \quad (3\text{-}11)$$

在[Mood et al. 1974: 440]中提到-2log λ 接近於卡方分佈，故我們將原有的比值 λ，轉換為-2log λ。

$$-2\log\lambda = -2\log\frac{L(H1)}{L(H2)} = -2\log\frac{b(c_{12}, c_1, p)\, b(c_2 - c_{12}, N - c_1, p)}{b(c_{12}, c_1, p_1)\, b(c_2 - c_{12}, N - c_1, p_2)} \quad (3\text{-}12)$$

自由度為 1 時的卡方臨界值如表 11。

表 11 自由度為 1 時的卡方臨界值

α	臨界值（critical value）
0.99	0.00016
0.95	0.0039
0.10	2.71
0.05	3.84
0.01	6.63
0.005	7.88
0.001	10.83

本研究將 95%與 99.9%的信心水準設定為門檻值，若-2log λ 大於或等於 10.83（$\alpha \leq 0.001$，即超過 99.9%的信心水準），我們將兩字根的相關度設定為 1。若-2log λ 低於 3.84（$\alpha > 0.05$，小於 95%的信心水準），我們將兩字根間的相關度設定為 0，以濾除相關性不明顯的字根對。若-2log λ 介於 3.84 與 10.83 之間，我們將兩字根的相關度設定為 $\frac{-2\log\lambda}{10.83}$，字根相關度設定整理於公式（3-13）。

$$\begin{aligned}
&\text{if } -2\log\lambda \geq 10.83 && \text{Correlation}(r_1,r_2)=1 \\
&\text{if } 3.84 < -2\log\lambda < 10.83 && \text{Correlation}(r_1,r_2)=\frac{-2\log\lambda}{10.83} \\
&\text{if } -2\log\lambda < 3.84 && \text{Correlation}(r_1,r_2)=0
\end{aligned} \qquad (3\text{-}13)$$

Corelation(r_1,r_2)：r_1,r_2 之相關度

經過多次嘗試後發現，動詞、介係詞、代名詞等詞性，因為通用性較高，常會找出一些較不具意義的關聯。據此，本研究中只計算名詞字根之間的相關度，因名詞相較於其他詞性，包含的語意通常較為明確，是辨識語意主題較佳的線索。

計算完某情境中各字根間的相關度後，即可依據字根間的相關度，在段落向量相似度比對時，擴充段落的字根向量。

假設進行相似度比較的兩段落，其所包含的字根分別為集合 S_1 與 S_2，我們需將此二集合轉為以字根為特徵的向量以比較兩段落主題上之相似度。首先，我們將 S_1 與 S_2 中所有相異字根取出，放入集合為 S 中，如下：

$S = S_1 \cup S_2 = \{r_1, r_2, \ldots, r_n\}$

接著，我們以集合 S 中的字根為向量維度，將 S_1 與 S_2 轉換為向量 $\vec{S}_1$ 與 $\vec{S}_2$。若 $\vec{S}_k$（$k = 1,2$）包含字根 r_i，則將 $\vec{S}_k$ 在 r_i 維度的分量值設為 1。若 $\vec{S}_k$ 不包含字根 r_i，則找出段落 S_k 中與 r_i 相關度最高之字根 r_j，將 $\vec{S}_k$ 在 r_i 維度的分量值設為 r_i 與 r_j 相關度權重值，如公式（3-14）。

$$\text{if } r_i \in S_k \quad \vec{S}_{k,r_i} = 1 \qquad \text{（3-14）}$$

$$\text{if } r_i \notin S_k \quad r_j = \text{maxSim}(S_k, r_i)$$

$$\vec{S}_{k,r_i} = \text{CW}(r_i, r_j)$$

S_k：段落 *k* 之字根集合

r_i ：S 中第 *i* 個字根

$\vec{S}_{k,r_i}$ ：$\vec{S}_k$ 在 r_i 維度的分量值

$\text{maxSim}(S_k, r_i)$：S_k 中與 r_i 相關度最高之字根

$\text{CW}(r_i, r_j)$：r_i 與 r_j 之相關度權重

其中 r_i 與 r_j 相關度權重為字根相關度乘上 α（α<1），如公式（3-15）。因為兩個相異的字根，縱使彼此的相關度再高，仍無法視為同一字根，故須降低其權

重。

$$CW(r_i, r_j) = \text{Correlation}(r_1, r_2) \times \alpha \qquad (3\text{-}15)$$

回到一開始所舉的例子。海關詢問旅客攜帶菸酒與否可能會有以下兩種問法：" Are you carrying any spirits or tobacco?"或" Any alcohol or cigarettes?"。依照上述的方法，在海關入境情境中，我們可以找到如表 12 的字根相關。本研究所提出之方法雖然無法找出同義字字根間的相關性(同義的單字很少會同時出現在相同的主題單元中)，但卻可以明確地找出不同問句所使用的「煙」與「酒」的單字字根的相關性。

表 12 字根相關度範例

r_1	r_2	c_1	c_2	c_{12}	-2log λ	單字相關度
cigarette	spirit	7	4	2	11.05	1.00
tobacco	alcohol	3	3	1	6.23	0.57

c_1：r_1 於情境中出現次數

c_2：r_2 於情境中出現次數

c_{12}：r_1 與 r_2 出現於同主題單元的次數

＊海關入境情境之主題單元總數（N）為 374

我們將兩段落進行字根擷取後，可將兩句話轉換為 S_1 與 S_2 兩個字根集合。

$S_1 = (\text{ar, you, carri, ani, spirit, tobacco, no})$

$S_2 = (\text{ani, alcohol, or, cigarett, no})$

兩集合所共有的字根為

$S = (\text{ar, carri, ani, spirit, tobacco, alcohol, or, cigarett, no, you})$

在表 13 中，我們將 S_1 與 S_2 兩集合轉換為向量 $\vec{S}_1$ 與 $\vec{S}_2$，並比較設定字根相關度權重與否的差別。

表 13 設定字根相關度權重範例

未使用字根相關度權重：										
維度：	ar	carri	ani	spirit	tobacco	alcohol	or	cigarette	no	you
$\vec{S}_1$：	1	1	1	1	1	0	0	0	1	1
$\vec{S}_2$：	0	0	0	0	0	1	1	1	1	1
使用字根相關度權重：										
維度：	ar	carri	ani	spirit	tobacco	alcohol	or	cigarette	no	you
$\vec{S}_1$：	1	1	1	1	1	0.57α	0	1.00α	1	1
$\vec{S}_2$：	0	0	0	1.00α	0.57α	1	1	1	1	1

由表 13 可知，透過字根相關度設定，將可擴增段落的字根特徵，增進具有相關字根的段落彼此間的向量相似度。

3.4.5 向量相似度比較

本研究中，段落主題相似度主要是採用向量相似度（cosine similarity）方式計算，以字根特徵將段落轉換為向量，並依字根的詞性權重、情境權重和相關度權重設定各分量值。

與一般向量相似度比對不同的是，本研究只考慮字根是否出現於某段落中，而不考慮其出現的次數；也就是說，相同的字根，出現在某段落中的次數並不會影響該字根所佔之權重。

原因在於本研究進行相似度比較的對象為語句數含量稀少的段落，在這樣的簡短的段落中，「主題關鍵字詞」重複出現的情形十分少見。重複出現於段落中的單字通常是一些資訊內容含量較低的代名詞、助動詞或介係詞等等。若將詞頻加入設定字根權重，將使「主題關鍵字詞」的權重相對降低，不利於主題相似度

之比對。

我們以 3.4.4 節所描述的方法，將進行比較的兩段落之字根集合定為 S_1 與 S_2，兩集合所有之相異字根集合為 S。以 S 之字根為維度將兩段落轉為向量 $\vec{S}_k$（k=1,2），並依序設定字根相關度權重、字根詞性權重與字根情境權重。詳細流程如圖 4。

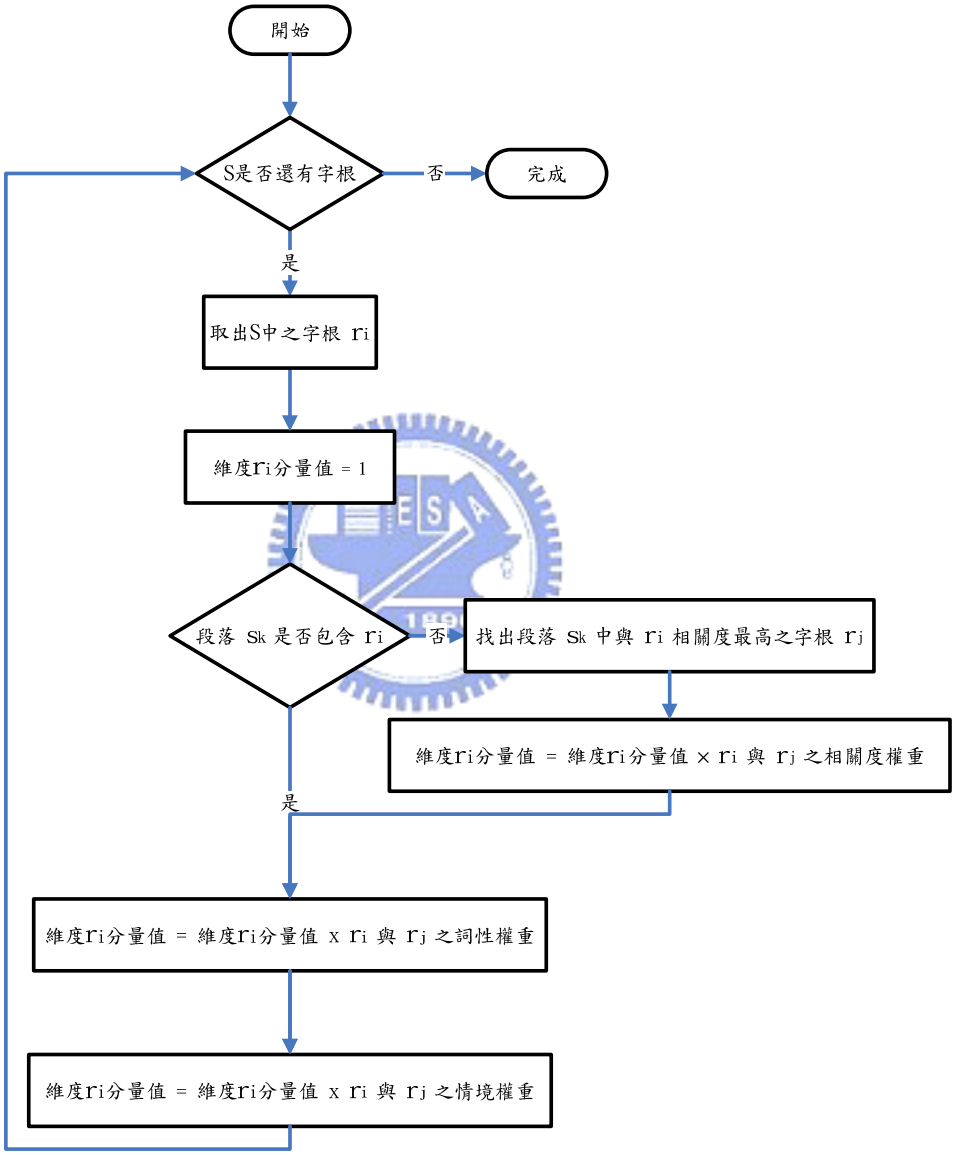

圖 4 段落向量設定流程圖

設定好 $\vec{S}_1$ 與 $\vec{S}_2$ 各分量值後，我 C 量帶入向量相似度模型中，計算出兩段落之主題相似度，如公式（3-16）。

$$\text{sim}(\text{Segment}_1,\text{Segment}_2)=\frac{\vec{S}_1\bullet\vec{S}_2}{|\vec{S}_1|\times|\vec{S}_2|}=\frac{\sum_{i=1}^{n}\vec{S}_{1,r_i}\times\vec{S}_{2,r_i}}{\sqrt{\sum_{i=1}^{n}\vec{S}_{1,r_i}\times\sum_{i=1}^{n}\vec{S}_{2,r_i}}} \quad (3\text{-}16)$$

$$0\le \text{sim}(\text{Segment}_1,\text{Segment}_2)\le 1$$

$\vec{S}_1$：段落 Segment_1 之向量

$\vec{S}_2$：段落 Segment_2 之向量

$\vec{S}_{1,r_i}$：$\vec{S}_1$ 中 r_i 維度的分量值

$\vec{S}_{2,r_i}$：$\vec{S}_2$ 中 r_i 維度的分量值

3.5 相似段落推薦

我們將段落分割結果及各段落彼此之間的相似度存於資料庫中，當使用者挑選出一篇連貫對話時，我們便可由資料庫內挑選出相似對話推薦給使用者參考，如圖 5。使用者可於畫面左方挑選會話情境並從中選出一篇範例對話。若使用者對某段落特別感興趣，可於畫面右下方輸入該段落之編號，系統將自動推薦最相似的三個段落。

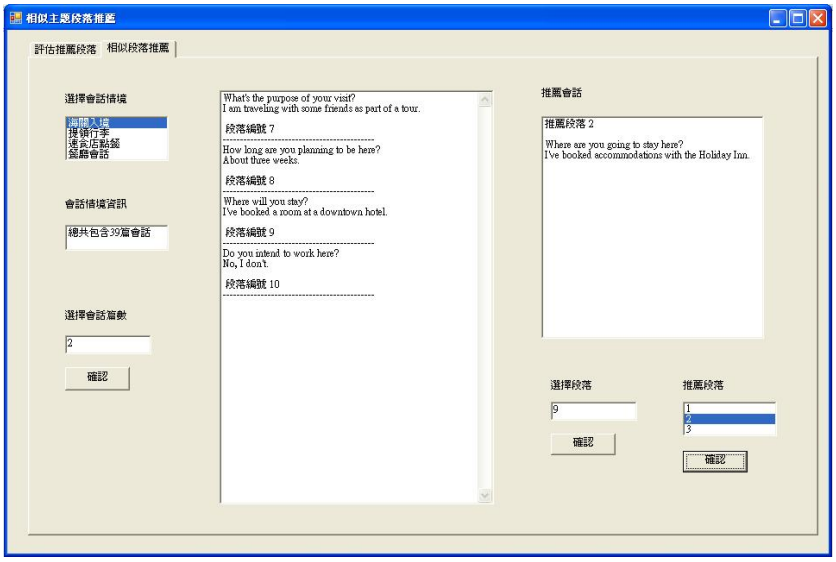

圖 5 相似主題段落推薦使用者介面

4 實驗結果與評估

本章將依序測試字根詞性權重（4.3.1）、字根情境權重（4.3.2）與字根相關度權重（4.3.3）對於相似度比對準確率之影響，並將本研究提出之方法與詞頻權重方法進行比較（4.3.4）。

4.1 實驗資料

實驗資料來源為市面上所販售之英語旅遊會話教學書籍，總共包括十六本會話書。從書中挑選出符合以下情境之範例會話後，再由人工輸入至會話語料庫。語料庫中總共包含 131 篇連貫會話，一共 1488 句。

本實驗將四個不同的會話情境視為各自獨立的資料集。選擇段落進行相似度比較時，只考慮相同情境下的會話段落，不考慮其他不同的會話情境。

表 14 實驗資料

會話情境	海關入境	提領行李	速食店點餐	餐廳會話
對話篇數	39	17	11	64
對話句數	532	191	150	615
會話內容	入境審查、海關提問和攜帶物品的申報。	提領行李與通報行李遺失。	速食店點餐與結帳。	訂位、點餐與用餐問題

4.2 實驗設計與流程

4.2.1 會話段落切割

由於是以「會話段落」做為相似度比較的基本單位，故首先以「問答」為線索，將語料庫中所有連貫的會話切割成為主題獨立的小段落。表 15 為各情境的會話篇數與包含的段落總數。

表 15 段落切割結果

會話情境	海關入境	提領行李	速食店點餐	餐廳會話
對話篇數	39	17	11	64
對話段落數	179	59	44	189

4.2.2 段落相似度比較

將會話劃分為更小的「段落」之後，即可進行段落間的相似度比對。本實驗將逐一檢驗字根詞性權重、情境權重與相關度權重設定對相似度比較的影響。此外，實驗中加入了以詞頻為權重的向量相似度比較，做為本研究之方法的比較基準。

實驗中將字根相關度權重的參數 α 設為 0.5，字根詞性權重之設定如表 16。

表 16 字根詞性權重之權重值

詞性	權重
名詞（不含專有名詞）	3
動詞	2
專有名詞	1.5
其他詞性	1

由於本實驗是以專家來評量實驗結果（4.2.3），實驗評估十分耗時耗力，故並未針對不同的參數設計額外的實驗。

4.2.3 評估準則

本研究主要採用資訊檢索領域常用的評估項目－準確率（Precision Rate）來做相關實驗的評估，準確率的意義為：「系統找到之相似段落中，與測試段落主題相似的比率。」如公式（4-1）。

$$\text{準確率 (Precision)} = \frac{|\text{系統發掘} \wedge \text{主題相似}|}{|\text{系統發掘}|} \qquad (4\text{-}1)$$

本實驗邀請三位具有大學以上英文程度的專家，評估各種不同相似度比對方法發掘之相似段落。須有超過半數以上（即兩位以上）的專家認定某方法比對出之段落與測試段落主題相似，我們才將這兩個段落視為主題相似之段落。

由於並不是每一個會話段落在語料庫中都有其他與之相似的段落，因此本實驗由人工挑選出主題較為常見的會話段落做為相似度比較的「測試段落」，各情境會話各選出二十分之一，四個情境測試段落數總共 24 個段落。

針對每一個測試段落，各方法將找出三個與之最為相似的段落（Top-3），並由專家判定主題是否與測試段落相關。比對段落相似度的系統畫面如圖 6。

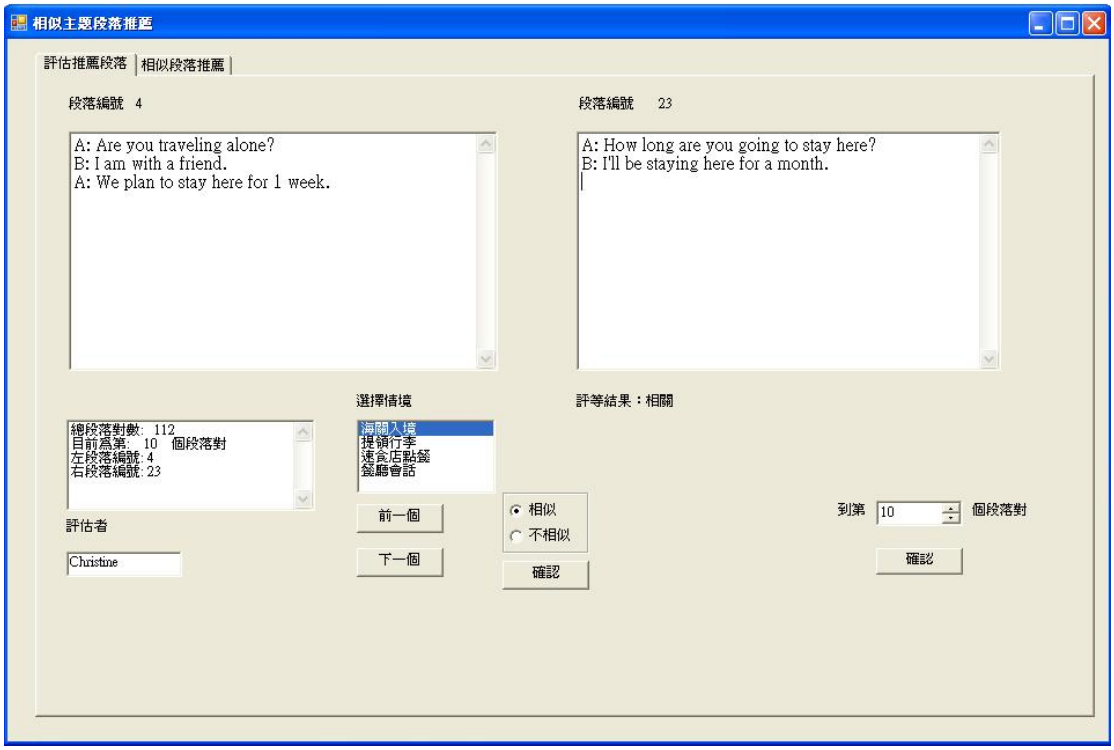

圖 6 評估系統畫面

4.3 實驗結果與分析

4.3.1、4.3.2 及 4.3.3 這三節中，我們將逐一檢視字根詞性權重、字根情境權重與字根相關度權重之設定對於語意主題相似度判別的效用。4.3.4 我們將比較本實驗所提之方法與以詞頻為權重之向量相似度比較準確率之差異。

本實驗分為兩個部份，首先我們分別獨立測試語料庫中四個會話情境下之相似度比較準確率，接著考慮四個會話情境中所有的測試段落，計算綜合四個情境的準確率。

4.3.1 字根詞性權重

首先，我們測試「字根詞性權重」的設定在不同的情境下，對於相似度比對的準確率是否有所影響。我們以 Top-3 準確率做為評估之依據，實驗結果如表 17 與圖 7。圖 7 之橫軸為各種不同的會話情境，縱軸為相似度比較準確率。

表 17 不同情境下設定字根詞性權重之 Top-3 準確率比較

段落相似度比對方法	海關入境	提領行李	速食店點餐	餐廳會話
包含字根詞性權重	0.740741	0.555556	0.666667	0.666667
不包含字根詞性權重	0.740741	0.444444	0.666667	0.633333

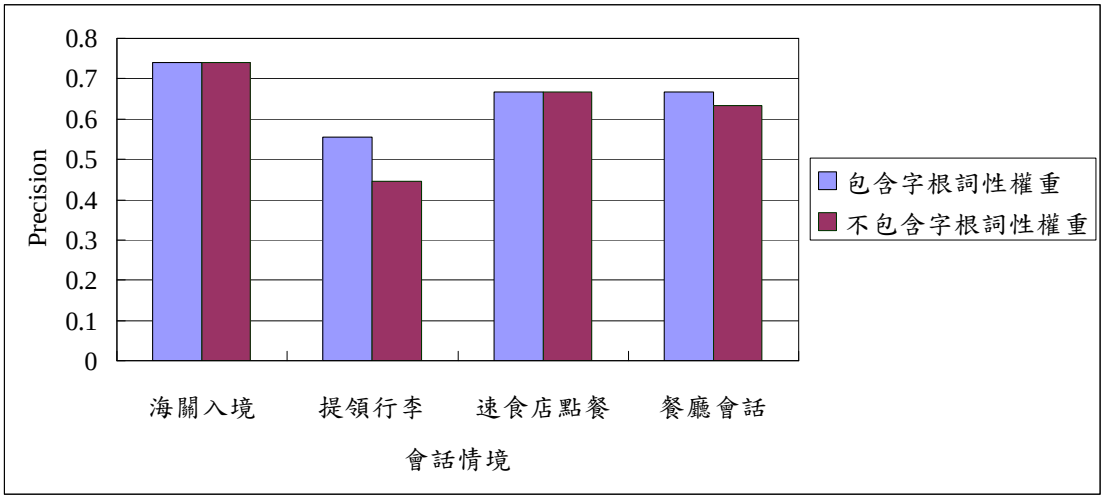

圖 7 不同情境下設定字根詞性權重之 Top-3 準確率比較

由圖 7，詞性權重在「提領行李」與「餐廳會話」中有較明顯的效果，顯示在這兩個會話情境中，名詞與動詞等詞性的主題辨識力較高。在海關入境及速食店點餐中，Top-3 準確率並無顯著差距。

綜合所有情境，兩個方法之 Top-1、Top-2 與 Top-3 準確率如表 18 與圖 8。

表 18 設定字根詞性權重之準確率比較

段落相似度比對方法	Top-1 Precision	Top-2 Precision	Top-3 Precision
包含字根詞性權重	0.7916667	0.666667	0.680556
不包含字根詞性權重	0.75	0.645833	0.652778

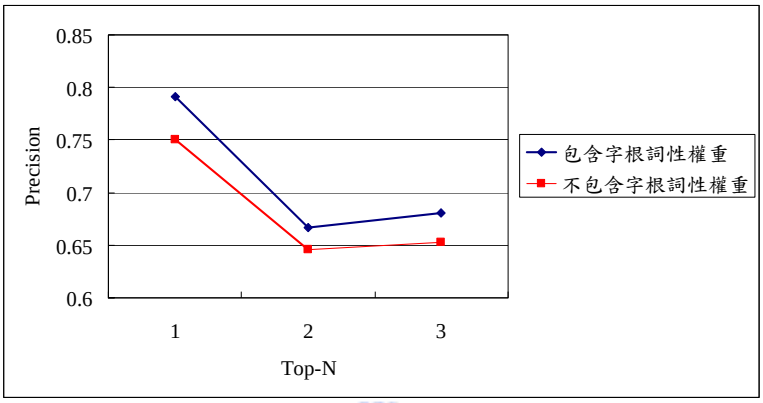

圖 8 設定字根詞性權重之準確率比較

由圖 8 之結果可發現，提高名詞、動詞與專有名詞之字根之權重並降低其他詞性字根權重，能有效地提昇段落語意主題比較地準確率。由此我們可以推知，段落語意之主題往往反應在名詞等關鍵詞性之上，其他詞性對於主題的鑑別力並不高，甚至可能是影響準確率的雜訊。

4.3.2 字根情境權重

在此我們測試「字根情境權重」之設定，在不同情境下對 Top-3 準確率之影響，結果如表 19 與圖 9。

表 19 不同情境下設定字根情境權重之 Top-3 準確率比較

段落相似度比對方法	海關入境	提領行李	速食店點餐	餐廳會話
包含字根情境權重	0.740741	0.555556	0.666667	0.666667
不包含字根情境權重	0.703704	0.444444	0.333333	0.7

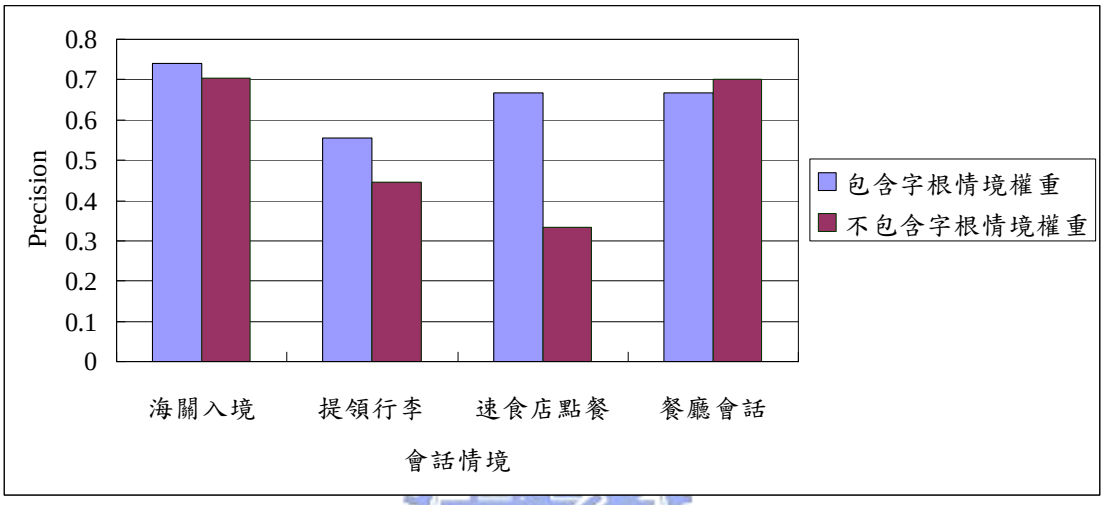

圖 9 不同情境下設定字根情境權重之 Top-3 準確率比較

在海關入境、提領行李與速食店點餐中，字根情境權重都有良好的效果，尤其是速食店點餐。我們推測其原因為速食店點餐情境所挑選出之測試段落，皆含有許多情境權重很高的單字字根，如”drink”、”dollars”與”cents”等等。這些單字的主題辨識力都相當高。例如包含”drink”的段落主題通常都是點選飲料，而包含”dollars”或”cents”的段落主題幾乎都是與結帳有關。

在餐廳會話的情境下，加入情境權重後準確率卻有小幅度的下降。我們推測原因在於餐廳會話中情境權重較高的單字主題辨識力並不佳。例如：具有高度情境權重之單字”dessert”會在服務生介紹套餐內容時出現，也會在遊客點餐時出現。綜合所有情境之準確率如表 20 與圖 10。

表 20 設定字根情境權重之準確率比較

段落相似度比對方法	Top-1 Precision	Top-2 Precision	Top-3 Precision
包含情境權重	0.7916667	0.666667	0.680556
不包含情境權重	0.7083333	0.625	0.638889

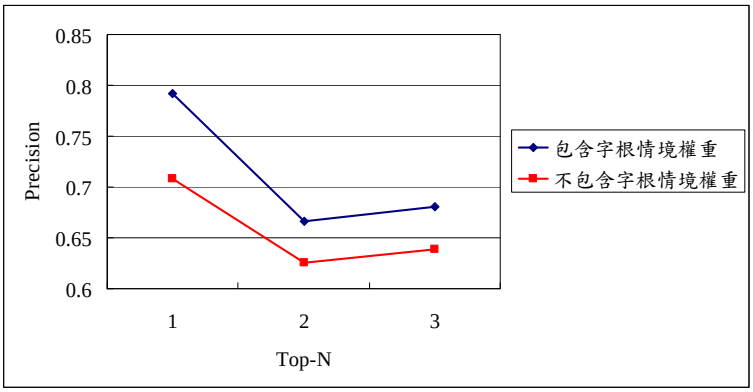

圖 10 設定字根情境權重之準確率比較

由圖 10 我們可以明顯看到加上字根情境權重之後，不論在 Top-1、Top-2 或 Top-3 時，皆可讓準確率有一定程度地提昇。

4.3.3 字根相關度權重

本節我們將測試「字根相關度權重」在段落語意主題相似度比較時，是否能增加比對的準確率，實驗結果如表 21 及圖 11 所示。

表 21 不同情境下設定字根相關度權重之 Top-3 準確率比較

段落相似度比對方法	海關入境	提領行李	速食店點餐	餐廳會話
包含字根相關度權重	0.740741	0.555556	0.666667	0.666667
不包含字根相關度權重	0.703704	0.555556	0.333333	0.733333

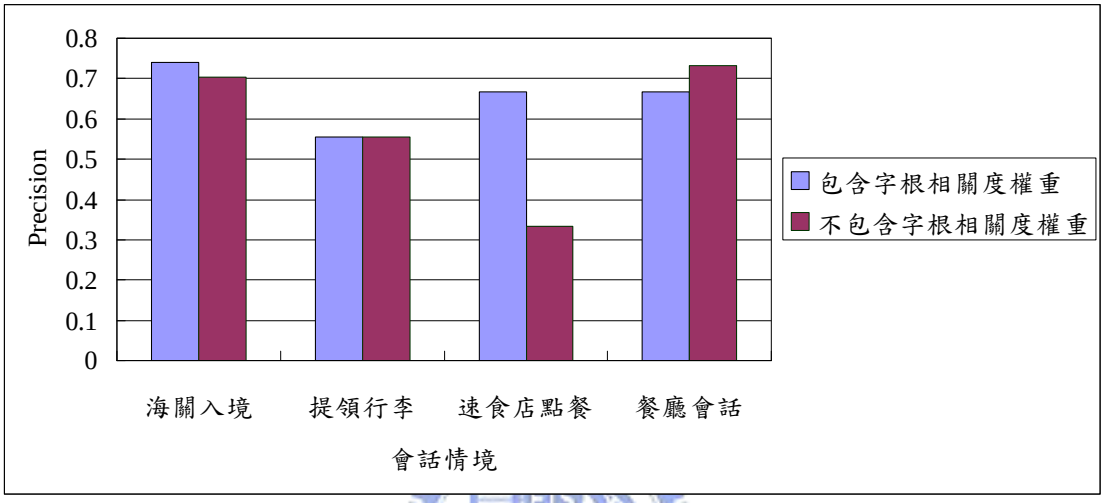

圖 11 不同情境下設定字根相關度權重之 Top-3 準確率比較

由圖 11 可發現，在速食店點餐之情境下包含字根相關度權重將使準確率大幅度上升，效果最為顯著；而餐廳會話中準確率卻有小幅下降。

其原因推測為在餐廳會話中，不同類型的餐點名稱常常出現在同一個主題單元之中。例如服務生介紹今日特餐時，就可能包括了沙拉、開胃菜、主餐、甜點與飲料等等餐點，遊客在點餐時也可能一次就點了湯、主餐和飲料等等。這種情況將造成各種不同類型的餐點字彙之字根如：”salad”、”steak”、”dessert”等，被認定為彼此相關之字根。若某段落主題為點主餐，另一段落主題為點甜點，這些相關之字根將造成相似度比對之誤差。在其他一般的會話情境下，這種情況就相對少見。

綜合所有情境之準確率如表 22 與圖 12。

表 22 設定字根相關度權重之準確率比較

段落相似度比對方法	Top-1 Precision	Top-2 Precision	Top-3 Precision
包含字根相關度權重	0.7916667	0.666667	0.680556
不包含字根相關度權重	0.6666667	0.666667	0.666667

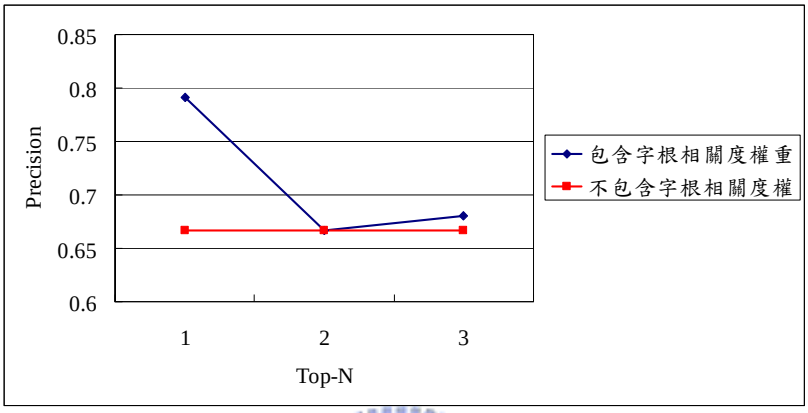

圖 12 設定字根相關度權重之準確率比較

在 Top-1 時，我們很明顯地可以看到設定字根相關度權重之效果，然而在 Top-2 及 Top-3 時，設定此權重與否無明顯差異。我們推測原因在於我們只計算「名詞」的相關度，並且只有在字根相關度（-2log λ）大於門檻值（3.84）時才給予相關度權重。因此，包含與測試段落相關字根的段落數量有限，在比較其他不包含相關字根之段落的相似度時，是否設定字根相關度權重將不會有任何影響，也就因而造成了實驗結果在 Top-2 及 Top-3 時並無明顯的差異。

4.3.4 與詞頻權重之比較

最後，我們要將本研究所提之方法與以「詞頻」為權重的向量相似度比對方法進行比較。「詞頻權重」方法首先以字根取代段落中的單字，並以字根為特徵，字根出現次數為特徵權重，套入向量模型比較不同段落間的相似度。實驗結果如及表 23 與圖 13。

表 23 不同情境下本研究之方法與詞頻權重之 Top-3 準確率比較

段落相似度比對方法	海關入境	提領行李	速食店點餐	餐廳會話
本研究之方法	0.740741	0.555556	0.666667	0.666667
詞頻權重	0.666667	0.555556	0.166667	0.8

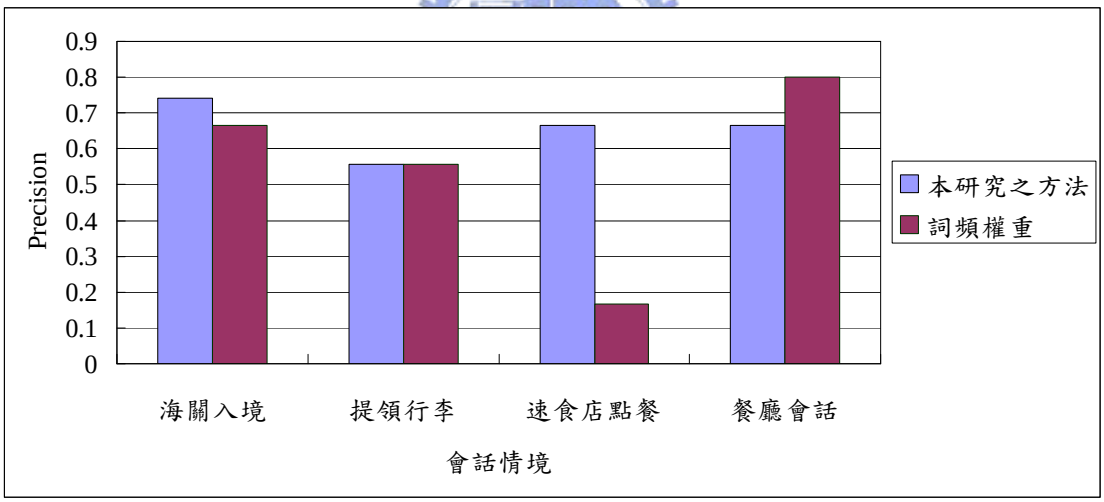

圖 13 不同情境下本研究之方法與詞頻權重之 Top-3 準確率比較

本研究之方法在海關入境及速食店點餐時明顯優於以詞頻為權重之相似度比對。餐廳會話情境下，因字根情境權重與相關度權重之設定，造成準確率下降，使本研究之方法準確率略低於詞頻權重相似度比對。

綜合所有情境之準確率如表 24 與圖 14。

表 24 本研究之方法與詞頻權重之準確率比較

段落相似度比對方法	Top-1 Precision	Top-2 Precision	Top-3 Precision
本研究之方法	0.7916667	0.666667	0.680556
詞頻權重	0.7083333	0.666667	0.666667

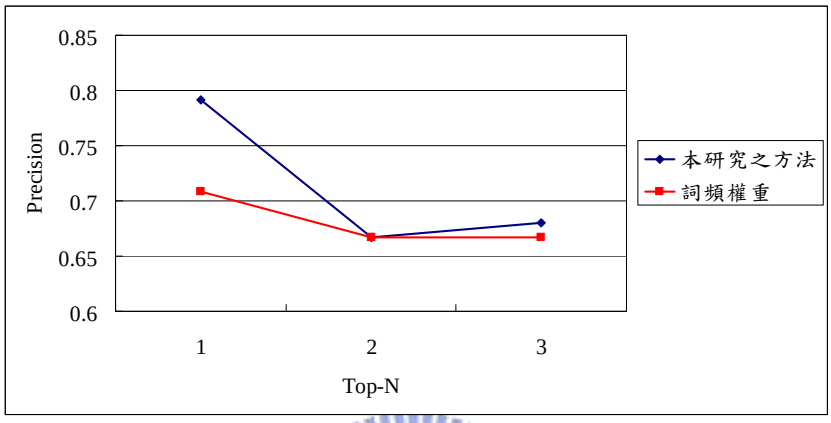

圖 14 本研究之方法與詞頻權重之準確率比較

在圖 14 可發現，本研究之方法在 Top-1 與 Top-3 時皆優於以詞頻為權重之相似度比對。

5 結論與建議

5.1 結論

本研究希望設計一套自動化的方法，輔助使用者在閱讀英語旅遊會話時，可隨時針對某會話主題，學習更多不同的對談內容。本研究的核心在於段落主題相似度的比較，提出三個權重設定方法以增進相似度比對之準確率。

由實驗得知，本研究之方法在不同情境下具有不同的效果，但綜合所有情境之 Top-3 準確率仍高達 68%，效果十分良好。

5.2 未來發展方向

本研究之會話語料受限於人工輸入，語料量有限。未來若能直接與英語會話書籍之出版公司接洽，採用該公司電子資料庫內之會話，將可大幅增加會話語料。

此外，若將來能投入更多的人力，先由專家標記各會話段落間的相關性成為訓練資料（Training Data），並將此資訊加入相似度比對之中，相信能大幅提昇相似度比對的準確率。

參考文獻

[Brown Corpus] Brown Corpus manual

http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

[Choi 2000] Choi, F., “Advances in domain independent linear text segmentation,” *Proceedings of NAACL’00, 2000.*

[Corley and Mihalcea 2005] Corley, C., and Mihalcea, R., “Measuring the Semantic Similarity of Texts,” *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, pages 1318, Ann Arbor, June 2005.*

[Dolan et al. 2005] Brockett, C., and Dolan, W. B., “Automatically Constructing a Corpus of Sentential Paraphrases,” *Proceedings of The Third International Workshop on Paraphrasing (IWP2005),Jeju, Republic of Korea, 2005*

[Dunning 1993] Dunning, T., “Accurate methods for the statistics of surprise and coincidence,” *Computational Linguistics 19:71-74*

[Foltz et al. 1998] Foltz, P.W., Kintsch, W., and Landauer, T.K., “The Measurement of Textual Coherence with Latent Semantic Analysis,” *Discourse Processes, vol. 25, nos. 2-3, pp. 285-307, 1998.*

[Galley et al. 2003] Galley, M., McKeown, K., Fosler-Lussier, E., and Jing, H., “Discourse Segmentation of Multi-Party Conversation,” *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL’03) (Sapporo, Japan, July 7-12, 2003).*

[Hatzivassiloglou et al. 1999] Hatzivassiloglou, V., Klavans, J., and Eskin, E., “Detecting text similarity over shortpassages: Exploring linguistic feature combinations via machine learning,” *Proceed-ings of the Joint SIGDAT Conference on Empirical Methods in Natural LanguageProcessing and Very Large Corpora 1999.*

[Hearst 1997] Hearst, M., “TextTiling: Segmenting text into multi-paragraph subtopic passages,” *Comput. Linguist. 23, 33-64, 1997.*

[Hsueh et al. 2006] Hsueh, P.Y., Moore, J., and Renals, S., “Automatic segmentation of multiparty dialogue,” *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, 2006.*

[Landauer et al. 1997] Landauer, T.K., Laham, D., Rehder, B., and Schreiner, M.E., “How Well Can Passage Meaning Be Derived without Using Word Order? A Comparison of Latent Semantic Analysis and Humans,” *Proc. 19th Ann. Meeting of the Cognitive Science Soc., pp. 412-417, 1997.*

[Li et al 2006] Li, Y., McLean, D., Bandar, Z.A., O'Shea, J.D., and Crockett, K., "Sentence Similarity Based on Semantic Nets and Corpus Statistics," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18,no.8, pp. 1138-1150, Aug., 2006.

[Lin and Och 2004] Lin, C., and Och, F., “Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics,” *Proceedings of ACL 2004, pp. 606-613.*

[Manning and Schütze 1999] Manning, C., and Schütze, H., “Foundations of Statistical Natural Language Processing,” *Cambridge, MA, MIT Press: 172-175, 1999.*

[Marcus et al. 1993] Marcus, M.P., and Marcinkiewicz, M.A., Santorini, B., “Building a large annotated corpus of English: the penn treebank,” *Computational Linguistics, v.19 n.2, June 1993*

[Mood et al. 1974: 440] Mood, A.M., Graybill, F.A., and Boes, D.C., “Introduction to the theory of statistics,” *New York: McGraw-Hill. 3rd edition, 1974*

[Pedersen] Pedersen, T., website http://www.d.umn.edu/~tpederse/similarity.html

[Porter 1980] Porter, M.F., “An Algorithm for Suffix Stripping,”*Program, 14, pp.*

130-137, 1980

[Reynar 1999] Reynar, J., “Statistical models for topic segmentation,” *Proceedings of the ACL, 1999.*

[Stanford 2006] Stanford Log-linear Part-Of-Speech Tagger, version 2006-05-21 http://nlp.stanford.edu/software/tagger.shtml

[Utiyama and Isahara 2001] Utiyama, M., and Isahara, H., “A statistical model for domain-independent text segmentation,” *Proc. Of the ACL, 2001.*

[汪若文 2004]汪若文，2004，運用潛在語意索引的自動化文件分類，國立交通大學，碩士論文。

[鄭守益和梁婷 2005] 鄭守益，梁婷，中文句子相似度之計算與應用，第十七屆自然語言與語音處理研討會，*Tainan, Taiwan, 2005 Proceedings of ROCLING XVII pp. 113-124.*

[顏偉和荀恩東 2004] 顏偉，荀恩東，基於 WordNet 的英語詞語相似度計算，第二屆全國學生計算語言學研討會，2004。