行政院國家科學委員會專題研究計畫　成果報告

多語言複合式文件自動摘要之研究(3/3)

計畫類別：個別型計畫
計畫編號：NSC94-2213-E-009-013-
執行期間：94 年 08 月 01 日至 95 年 07 月 31 日
執行單位：國立交通大學資訊科學學系(所)

計畫主持人：楊維邦

計畫參與人員：柯皓仁教授，葉鎮源，謝佩原，&#63866；哲瑋，鄭佳彬，&#63943；政璋，顧世彥，王瓊婉，&#63988；昕潔，張家&#63839；

報告類型：完整報告

處理方式：本計畫可公開查詢

中　華　民　國 95 年 10 月 31 日

行政院國家科學委員會補助專題研究計畫 ☑ 成 果 報 告
□期中進度報告

多語言複合式文件自動摘要之研究

計畫類別：☑ 個別型計畫　　□ 整合型計畫
計畫編號：
(1/3) NSC-92-2213-E-009-126-
(2/3) NSC-93-2213-E-009-044-
(3/3) NSC-94-2213-E-009-013-
執行期間：　92 年 08 月 01 日至　95 年 07 月 31 日

計畫主持人：楊維邦 教授
共同主持人：
計畫參與人員：　柯皓仁 教授, 葉鎮源, 謝佩原, 梁哲瑋, 鄭佳彬, 劉政璋, 顧世彥, 王瓊婉, 林昕潔, 張家寧

成果報告類型(依經費核定清單規定繳交)：□精簡報告　☑完整報告

本成果報告包括以下應繳交之附件：
□赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位：國立交通大學資訊科學學系(所)

中　華　民　國　**95** 年　**10** 月　**31** 日

中文摘要

自動化文件摘要(Automated Text Summarization)之研究，探討如何分析、統整與萃取文件中重要的資訊，並以簡明的形式呈現，可作為使用者或其他資訊系統的判斷與決策依據。過去關於文件摘要的研究，大多著眼於單文件摘要。近來，多文件摘要愈顯得重要。多文件摘要與單文件摘要最大的差異，在於過濾重複的資訊(Anti-redundancy)，同時須避免重要資訊的流失。

本計畫為三年期研究案，目的在於研究多文件自動摘要(Multidocument Summarization)、跨語言多文件摘要(Cross-Language Multidocument Summarization)與使用者問答導向摘要(Query-Focused Multidocument Summarization)技術。

- 第一年計畫: 多文件摘要技術之研究

計畫目的在於研究並發展多文件自動摘要的技術，探討如何導出文件的結構、如何組織文件的內容及如何表示所抽取出文件抽象涵義等議題。我們考量相關文獻所提多文件摘要的方法之優缺點，並以先前我們所發展之單文件摘要方法為基礎，加以改進以適用於多文件摘要。

我們利用潛在語意分析(Latent Semantic Analysis, LSA)與主題關係地圖(Text Relationship Map)作為多文件分析模型。此模型將每個語句視為圖(Graph)中的一個點，兩兩語句若相似度大於臨界值，則彼此間具有一連結，其中語句相似度的計算採用 LSA 所倒出的概念空間模型，以達到具有語意判別的相似度計算。根據主題關係地圖模型，我們提出兩種段落重要性評估模型，分別為 1) Global Bushy Path；2) Aggregate Similarity。依據上述兩種計算重要性的模型，我們修改 Maximal Marginal Relevance，提出新的摘要段落挑選方式，以達到多文件中去除重複性(Redundancy)的要求。

實驗評估採用 ROUGE 來計算專家所產出的摘要與機器所產出摘要中，平均包含有多少個相同的字。我們以 ROUGE-1 為評估標準，實驗結果顯示 Model 1 獲得平均 ROUGE-1 分數為 0.3564，Model 2 獲得平均 ROUGE-1 分數為 0.3497。比較 DUC 2003 的比賽結果，我們的方法亦有不錯的表現，約在中等以上。

- 第二年計畫: 中英文多文件摘要技術之研究

計畫目的在於發展多語多文件摘要(Multilingual Multidocument Summarization)技術。研究內容著重於中英文混合式多文件摘要，研究議題為如何跨越語言型態及結構的障礙，提出計算中英文語句相似度的方法。

我們提出一套中英概念空間擷取的方法。基於中英雙語平行語料庫(Chinese-English Parallel Corpus)，進行中英文詞分群(Word Clustering)，以建立階層式概念空間(Hierarchical Concept Space)。接著，將中英文語句對應至該階層式概念空間中，透過此對應關係，可將關鍵詞向量空間表示式(Keyword-Level Vector Representation) 轉換至概念向量空間表示式 (Concept-Level Vector Representation)。最後，於相同的概念空間中，便可計算任兩中中、中英及英英語句的概念相似度。

中英混合式詞分群可抽取出中文詞與英文詞間的關連。當分群的群數目越多時，位於詞群中的中英文詞可互為翻譯；當分群的群數目越少時，位於詞群中的中英文詞可視為相關。透過階層式概念空間相似度的組合，可得不錯的中英語句對應。實驗結果顯示，考慮 Top 10 組中英對應段落，平均 Precision 約 57%為相關。摘要內容的評估，則由專家進行評比。平均而言，資訊量涵蓋度為 7.06，可讀性為 6.04。(其中，1 代表最差，5 代表普通，10 代表最好。)

- 第三年計畫: 問答式多文件摘要技術之研究

第三年的研究內容著重於參考使用者給定的自然語言查詢問題及設定檔(Profile)[1], 以產生可以回答使用者問題的摘要。

我們提出一套計算文件集中每個語句與使用者所下的查詢間的相關度計算方式。考量傳統 Vector Space Model 與 Latent Semantic Analysis 的優缺點，我們提出以線性方式結合上述兩種模型的相似度。此外，我們亦研究過去應用於文件摘要的 shallow features 是否依然於問答摘要環境下扮演重要角色。於本計畫中，我們考慮以下五個語句特徵，分別為 1) Sentence Position; 2) Sentence TF-IDF Weight; 3) Sentence's Similarity to the Title; 4) Sentence's Similarity to the Document; 5) Sentence's Similarity to the Topic Cluster。

針對我們所提出的語句與查詢相似度計算方法，並結合五個語句特徵值，我們修改 Maximum Marginal Relevance，提出新的摘要段落挑選方式，以達到多文件中去除重複性(Redundancy)的要求。

實驗中，我們以 DUC 2005 的資料集測試，該年度評估以 ROUGE-2 及 ROUGE-SU4 為官方標準。實驗結果顯示，我們所提出的方法有不錯的表現，在最佳的情況下，獲得 ROUGE-2 的分數為 0.0757，ROUGE-SU4 的分數為 0.1299。

[1] 該設定檔定義摘要內容爲 specific 或 general。

英文摘要

Automated text summarization investigates the process of extracting the most important information from a source (or sources), and presenting a summary to the user. In the past, much related work only focuses on single-document summarization. Multi-document summarization obtains increasing attentions in recent years. The distinction between single and multi-document summarization is that the latter has to handle anti-redundancy and to keep salient information meanwhile.

This project is composed of three parts and was conducted in the past three years. The goal of this research is to investigate and develop the techniques focusing on multidocument summarization, cross-language multidocument summarization, and query-focused multidocument summarization.

- The First-Year Project: A Study on Multidocument Summarization

This principal objective of this project is to develop new approaches to address multi-document summarization. The research includes 1) Conceptual Modeling and Representation for Multiple Documents, 2) Paragraph Significances Measurement, and 3) Content Ordering for the summary.

We exploit latent semantic analysis and text relationship map to derive conceptual model (or a graph model) for multiple documents. Based on the model, we propose two approaches to measure the significance of a paragraph. They are 1) Global Bushy Path, and 2) Aggregate Similarity. Moreover, we propose a novel paragraph re-ranking approach on the basic foundation of Maximal Marginal Relevance to extract salient paragraphs.

The evaluation metric used in this project is ROUGE, which is the official evaluation tool at DUC 2004. We report ROUGE-1 as the official score for each proposed models. Model 1 obtained a value of 0.3564 while Model 2 obtained a value of 0.3497 for ROUGE-1. When compared with the official results at DUC 2003, our proposed methods are at the average position and obtained competitive results.

- The Second-Year Project: A Study on Cross-Language Multidocument Summarization

The principal objective of this project is to develop novel techniques to address multilingual, multidocument summarization. The main issue is to propose a method to compute the similarity between two short passages which are written in different languages.

With a Chinese-English parallel corpus, we propose a framework to derive a hierarchical concept space by word clustering. On the basis of the concept space, a Chinese or an English paragraph can be mapped into the space and represented as a concept-level vector representation. Since a Chinese or an English paragraph is mapped into the same concept space, it becomes easy to compute the similarity between two paragraphs. Once the similarity between a Chinese and an English passage is obtained, the summary is generated using the techniques developed in the first-year project.

The preliminary results show that the larger the number of concept clusters, words in the same concept can be regarded as corresponding translations; while the smaller the number of concept clusters, words in the same concept are loosely-related. In the set of Chinese-English paragraph pairs with Top 10 similarities, approximately 57% are judged as related by humans. Regarding the quality of the summary, in average, a score of 7.06 and 6.04 was obtained in terms of information coverage and readability respectively, in which 1 the worst, 5 good, and 10 the best.

- The Third-Year Project: A Study on Query-Focused Multidocument Summarization

The principal objective of this project is to develop a summarization method to answer user-specific questions. In this report, we try to merge techniques of multidocument summarization and question-answering to generate a brief, well-organized fluent summary to provide more relevant information for the purpose of answering real-world complicated questions. The problem is addressed as a query-biased sentence retrieval task.

We propose a hybrid relevance analysis to evaluate sentence relevance to the query. This is achieved by combining similarities computed from the vector space model and the latent semantic analysis. Surface features are also examined to know the impacts of low-level features for query-focused multidocument summarization. In other words, the summary is created by including sentences with the topmost significances which are measured in terms of sentence relevance and surface feature salience. In addition, a modified Maximal Marginal Relevance is proposed to reduce redundancy by taking into account sentence shallow feature scores.

The experimental results showed the proposed method obtained competitive results when evaluated with the DUC 2005 corpus. Regarding ROUGE-2 and ROUGE-SU4, the proposed method obtained a value of 0.0757 and 0.1299 respectively.

目錄

圖目錄

表目錄

計畫概述

本計畫主要目的在於研究多文件自動摘要(Multidocument Summarization)、跨語言多文件摘要(Cross-Language Multidocument Summarization)與使用者問答導向摘要(Query-Focused Multidocument Summarization)技術。

本計畫共分為三年期計畫，研究範疇及目的分述如下：

- 第一年：多文件摘要技術研究

  研究範疇包含：1) 文件模型建構(Modeling)及表示(Representation)的方法，主要分析文件的主題、判斷主題的重要性及排序(Ordering)等，將抽取出的主題結構化及組織化；2) 偵測多個文件主題模型間的相關性，並透過相關結構的連結，以得到完整的資訊模型；3) 探討由文件模型導出摘要的方法。

- 第二年：跨語言、多媒體摘要技術研究

  研究範疇包含：1)中英文文件內容的轉譯及相似度的評量；2)探究中英文轉譯應用在自動摘要的可能性，並基於語意的連結來評估中英文的相似度。研究內容著重於中英文混合式多文件摘要，研究議題為如何跨越語言型態及結構的障礙，提出計算中英文語句相似度的方法。

- 第三年：客製化摘要呈現技術研究

  研究範疇包含：1) 參考使用者給定的自然語言查詢問題及設定檔(Profile)[2], 以產生可以回答使用者問題的摘要；2) 偵測自然語言問題與文件集中相關語句；3) 研究過去應用於文件摘要的 shallow features 是否依然於問答摘要環境下扮演重要角色。

[2] 該設定檔定義摘要內容爲 specific 或 general。

執行成果

第一年成果

第一年的研究內容著重於如何判別文件主題及抽取具有代表性的語句作為候選摘要語句。

我們考量相關文獻所提多文件摘要的方法之優缺點，並以先前我們所發展之單文件摘要方法為基礎，加以改進以適用於多文件摘要。首先，我們利用潛在語意分析(Latent Semantic Analysis, LSA)與主題關係地圖(Text Relationship Map)作為多文件分析模型。此模型將每個語句視為圖(Graph)中的一個點，兩兩語句若相似度大於臨界值，則彼此間具有一連結，其中語句相似度的計算採用 LSA 所倒出的概念空間模型，以達到具有語意判別的相似度計算。

根據主題關係地圖模型，我們提出兩種段落重要性評估模型，分別為 1) Global Bushy Path；2) Aggregate Similarity。其中 Global Bushy Path 以圖中各個點所具有的連結數當作其重要性(Significance)；Aggregate Similarity 類似於 Global Bushy Paht，不同之處在於其計算所有連結的相似度總合，作為每個點重要性。根據上述兩種計算重要性的模型，我們修改 Maximal Marginal Relevance，提出新的摘要段落挑選方式，以達到多文件中去除重複性(Redundancy)的要求。

實驗評估採用 ROUGE 來計算專家所產出的摘要與機器所產出摘要中，平均包含有多少個相同的字。我們以 ROUGE-1 為評估標準，實驗結果顯示 Model 1 獲得平均 ROUGE-1 分數為 0.3564，Model 2 獲得平均 ROUGE-1 分數為 0.3497。比較 DUC 2003 的比賽結果，我們的方法亦有不錯的表現，約在中等以上。

第二年成果

第二年的研究內容著重於中英文混合式多文件摘要，研究議題為如何跨越語言型態及結構的障礙，提出計算中英文語句相似度的方法。

我們提出一套中英概念空間擷取的方法。基於中英雙語平行語料庫(Chinese-English Parallel Corpus)，進行中英文詞分群(Word Clustering)，以建立階層式概念空間(Hierarchical Concept Space)。接著，將中英文語句對應至該階層式概念空間中，透過此對應關係，可將關鍵詞向量空間表示式(Keyword-Level Vector Representation) 轉換至概念向量空間表示式 (Concept-Level Vector Representation)。最後，於相同的概念空間中，便可計算任兩中中、中英及英英語句的概念相似度。

中英混合式詞分群可抽取出中文詞與英文詞間的關連。當分群的群數目越多時，位於詞群中的中英文詞可互為翻譯；當分群的群數目越少時，位於詞群中的中英文詞可視為相關。透過階層式概念空間相似度的組合，可得不錯的中英語句對應。

實驗結果顯示，考慮 Top 10 組中英對應段落，平均 Precision 約 57%為相關。摘要內容的評估，則由專家進行評比。平均而言，資訊量涵蓋度為 7.06，可讀性為 6.04。(其中，1 代表最差，5 代表普通，10 代表最好。)

第三年成果

第三年的研究內容著重於參考使用者給定的自然語言查詢問題及設定檔(Profile)[3]，以產生可以回答使用者問題的摘要。

我們提出一套計算文件集中每個語句與使用者所下的查詢間的相關度計算方式。考量傳統 Vector Space Model 與 Latent Semantic Analysis 的優缺點，我們提出以線性方式結合上述兩種模型的相似度。此外，我們亦研究過去應用於文件摘要的 shallow features 是否依然於問答摘要環境下扮演重要角色。於本計畫中，我們考慮以下五個語句特徵，分別為 1) Sentence Position; 2) Sentence TF-IDF Weight; 3) Sentence's Similarity to the Title; 4) Sentence's Similarity to the Document; 5) Sentence's Similarity to the Topic Cluster。

針對我們所提出的語句與查詢相似度計算方法，並結合五個語句特徵值，我們修改 Maximum Marginal Relevance，提出新的摘要段落挑選方式，以達到多文件中去除重複性(Redundancy)的要求。

另外，摘要的產出方式必須符合使用者所給定的 Profile 限制。目前我們僅考量摘要內容的豐富性，分別為 Specific 與 General。Specific 主要著重在回答使用者的問題；General 則除了回答問題外，還會提供相關的背景資訊。為了達到提供 Specific 的要求，我們認為當使用者要求 Specific 的摘要內容時，實則要求具有較多人、地、時的資訊，因此挑選候選語句時，會依據其涵蓋的人、地、實資訊給予不同的權重。

實驗中，我們以 DUC 2005 的資料集測試，該年度評估以 ROUGE-2 及 ROUGE-SU4 為官方標準。實驗結果顯示，我們所提出的方法有不錯的表現，在最佳的情況下，獲得 ROUGE-2 的分數為 0.0757，ROUGE-SU4 的分數為 0.1299。

[3] 該設定檔定義摘要內容爲 specific 或 general。

參與人員

Table 1: 參與人員及執行工作列表

計畫項目	參與人員	服務單位	職稱	擔任工作
第一年計畫 第二年計畫 第三年計畫	楊維邦	國立交通大學 資訊科學學系 國立東華大學 資訊管理學系	教授	計畫主持人
第一年計畫 第二年計畫 第三年計畫	柯皓仁	國立交通大學 資訊管理研究所	教授	計畫參與人員
第一年計畫 第二年計畫 第三年計畫	葉鎮源	國立交通大學 資訊科學學系	博士生	文獻蒐集研讀、理論分析、演算法設計及評估方法設計
第一年計畫	謝佩原	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第一年計畫	梁哲瑋	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第一年計畫	鄭佳彬	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第二年計畫	劉政璋	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第二年計畫	顧世彥	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第二年計畫	王瓊婉	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第三年計畫	林昕潔	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作
第三年計畫	張家寧	國立交通大學 資訊科學學系	碩士生	文獻蒐集研讀、協助演算法設計與實作

多語言複合式文件自動摘要之研究 (1/3)

計畫類別: 個別型計畫

計畫編號: NSC-92-2213-E-009-126-

執行期間: 2003/08/01 – 2004/07/31

計畫主持人: 楊維邦 教授

計畫參與人員: 柯皓仁 教授, 葉鎮源, 謝佩原, 梁哲瑋, 鄭佳彬

中文摘要

自動化文件摘要(Automated Text Summarization)之研究，探討如何分析、統整與萃取文件中重要的資訊，並以簡明的形式呈現，可作為使用者或其他資訊系統的判斷與決策依據。過去關於文件摘要的研究，大多著眼於單文件摘要。近來，多文件摘要愈顯得重要。多文件摘要與單文件摘要最大的差異，在於過濾重複的資訊(Anti-redundancy)，同時須避免重要資訊的流失。另外，摘要內容排序(Content Ordering)，亦是多文件摘要必須探討的議題。

本計畫為三年期研究案『多語言複合式文件自動摘要之研究』之第一年計畫。本年度計畫之目的，在於研究並發展多文件自動摘要的技術，將探討如何導出文件的結構、如何組織文件的內容及如何表示所抽取出文件抽象涵義等相關議題。研究範疇包含：1) 文件模型的建構及表示；2) 文件主題偵測及重要性評估；3) 摘要內容組織及排序。我們利用潛在語意分析(Latent Semantic Analysis)與主題關係地圖(Text Relationship Map)作為多文件分析模型，提出兩種段落重要性評估模型，分別為 1) Global Bushy Path；2) Aggregate Similarity。根據上述模型，我們採用 Maximal Marginal Relevance，提出新的摘要段落挑選方式。

實驗評估採用 ROUGE 來計算專家所產出的摘要與機器所產出摘要中，平均包含有多少個相同的字。我們以 ROUGE-1 為評估標準，實驗結果顯示 Model 1 獲得平均 ROUGE-1 分數為 0.3564，Model 2 獲得平均 ROUGE-1 分數為 0.3497。比較 DUC 2003 的比賽結果，我們的方法亦有不錯的表現，約在中等以上。

關鍵詞：多文件自動摘要；潛在語意分析；主題關係地圖

英文摘要

Automated text summarization investigates the process of extracting the most important information from a source (or sources), and presenting a summary to the user. In the past, most related work on text summarization focuses on single-document summarization. Multi-document summarization obtains increasing attentions in recent years. The distinction between single and multi-document summarization is that the latter has to handle anti-redundancy as well as to keep salient information meanwhile. Moreover, content ordering, which is to provide a cohesive and coherent summary, is an important issue to be examined.

As the first part of the project "The Research on Cross-Language, Composite and Multi-Document Automated Text Summarization", the principal objective is to develop new approaches to address multi-document summarization. Our researches include 1) Conceptual Modeling and Representation for Multiple Documents, 2) Topic Detection and Paragraph Significances Measurement, and 3) Content Ordering for the summary.

We exploit latent semantic analysis and text relationship map to derive conceptual model for multiple documents. Based on the model, we propose two approaches to measure the significance of a paragraph. They are 1) Global Bushy Path, and 2) Aggregate Similarity. Besides, we propose a novel paragraph re-ranking approach on the basic foundation of Maximal Marginal Relevance to extract salient paragraphs.

The evaluation metric used in this project is ROUGE which is the official evaluation tool at DUC 2004. In the experiment, we report ROUGE-1 as the official score for each proposed models. Model 1 obtained a value of 0.3564 while Model 2 obtained a value of 0.3497 for ROUGE-1. When compared with the official results at DUC 2003, our proposed methods are at the average position and obtained competitive results.

Keywords: Multi-document Summarization; Latent Semantic Analysis; Text Relationship Map

1. 研究背景及目的

今日電腦與資訊技術蓬勃發展的數位時代，網際網路已成為現代生活中不可或缺的重要角色，更帶動人類文明往新的資訊紀元(Information Era)推進。拜科技之賜，各種媒體資料的數位化；透過網際網路管道，大量且豐富的數位內容(Digital Content)得以無遠弗屆地傳播。就現況而言，各式各樣的資訊於網際網路中流通，資訊的傳播不再單純藉由傳統平面媒體，人們亦漸漸習慣經由網路找尋所要的資料。資訊的蒐集變得方便，然而亦衍生相關問題，如「資訊爆炸(Information Explosion)」。

龐大的資訊量，使得搜尋及辨別有用資訊的困難度大幅提昇，如何快速且有效地獲得真正符合自身需求的資訊，亦是目前熱門的研究議題。為解決此類問題，使用者藉由輔助工具的幫助，得以快速獲知資料的意涵，期能正確地判斷是否符合自身的需求。相關的輔助工具有：1) 搜尋引擎(Search Engine)及 2) 自動摘要系統(Automated Summarization)。其中，搜尋引擎扮演『資訊過濾器(Information Filter)』的角色，其功用乃是分析檢索條件(Query)，搜尋與檢索條件相關的資料；自動摘要系統則扮演『資訊監督者(Information Spotter)』的角色，其功用在於分析、統整相關的資料，以簡明的形式呈現，以幫助使用者在最短時間得知資料內容的意義[20]。

自動摘要系統依原始資料之性質可分為：

- 文件摘要(Text Summarization) – 原始資料為純文字。
- 多媒體摘要(Multimedia Summarization) – 原始資料為影像及聲音。
- 複合性摘要(Hybrid Summarization) – 原始資料綜合純文字和多媒體。

本年度研究計畫主要著重於文件摘要技術的研發。以下深入針對何謂文件自動化摘要，技術起源與發展，與摘要類型等作一詳盡介紹。

1.1. 文件摘要介紹

根據 Mani 與 Maybury 為文件自動摘要所下的定義[34]，自動化文件摘要乃是從原始資料中精鍊出最重要資訊的過程，其結果即為該原始資料的精簡化版本，且可作為人們或其他資訊系統的判斷與決策依據。

> *Text summarization is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks).*

文件摘要研究起源於 1950 年代，所使用的技術可大致分為下列幾種：

- 1950至1960年代，研究方法著重於寫作格式(Genre)的分析。相關研究，如[16][32]。舉例來說，語句中含有提示片語(Cue Phrase)，如『In Summary』或『In Conclusion』，則該語句可視為摘要語句。此類技術之優點在於簡單容易，然而卻與文件類型相關，技術重複利用性並不高。

- 1970至1980年代初期，研究方法轉而利用人工智慧來建構知識的表示法(Knowledge Representation)，藉以達到分析文件主題及涵義的目的。相關研究，如[53][14]。此類技術以模板(Template)來辨認人物、地點及時間等事件基本要素(Entity)，並透過知識模型的推演偵測主題及產生摘要。其缺點在於模板的定義不夠詳盡，容易導致摘要意義上與原內容可能有所出入，且模板的定義必須依賴專家，耗時又耗人力。

- 1990年代開始，資訊擷取(Information Retrieval, IR)技術被廣泛應用。相關研究，如[4][8][20][24][27][45][48]。然其分析只著重於字詞層面(Word-level)，並沒有考慮同義詞(Synonymy)、一詞多義(Polysemy)及字詞依屬(Term Dependency)關係等語意層面(Semantic-level)分析。資訊擷取技術的應用，發展至今已經越來越多人朝著語意層面深入研究。

由摘要系統的輸入來看，文件摘要可分為單文件摘要(Single Document Summarization)及多文件摘要(Multdocument Summarization)。單文件摘要將文件精簡化、重點化，著重於從單一文件中刪減無用資料。相關研究，如[1][2][4][20][24][27][30][40][48][52]。多文件摘要則處理多篇探討相似主題的文件(Topically-related Documents)，著重於刪減、過濾無用且重複的資料。相關研究，如[7][19][23][31][33][37][45][46]。一般來說，理想的多文件摘要系統必須滿足以下要求[19]：

1. Clustering – 具有將相似的文件或段落群集成相關資訊的能力。
2. Coverage – 摘要應涵蓋不同文件中所有主題的能力。
3. Anti-redundancy – 刪減摘要中各段落間重複資料的能力。
4. Summary Cohesion Criteria – 將各項資料結合成一致性高且適合閱讀的摘要，包含各段落的排序。
5. Quality – 摘要結果須具有高可讀性，並具有相關性高且內容豐富的資訊。
6. Identification of Source Inconsistence – 辨認不同文件所提供不同資訊、錯誤及不一致性的能力。
7. Summary Updates – 追蹤時間性事件發展能力，以提供使用者最新的資訊。
8. Effective User Interface – 提供與使用者互動的介面，如個人化摘要及呈現。

由語言的角度來看，文件摘要亦可分為單語言文件摘要(Mono-lingual Summarization)與多語言文件摘要(Multi-lingual Summarization)。多語言文件摘要，如[55][10][57][29]。多語言摘要係指文件來源可能為不同語言或單文件中包含不同語言等，此類摘要著重於克服各語言在型態、結構及用語習慣的差異，並提供各語言間互相轉譯的能力。

1.2. 研究目的與範疇

本計畫為三年期研究案『多語言複合式文件自動摘要之研究』之第一年計畫。本年度計畫之目的，在於研究並發展多文件自動摘要的技術，將探討如何導出文件的結構、如何組織文件的內容及如何表示所抽取出文件抽象涵義等相關議題。研究範疇包含：

1. 文件模型的建構及表示
2. 文件主題偵測及重要性評估
3. 摘要內容組織及排序

針對以上所述研究項目，我們採用先前研究單文件摘要所提出的方法 – LSA-based T.R.M. Approach [52]為基礎，加以改良以適用於多文件摘要的研究，同時提出兩種評估段落(或語句)重要性的模型。同時，考慮到多文件摘要必須去除重複性(Redundancy)的議題，我們利用 Maximum Marginal Relevance (MMR) [8]來達成此目的。於第三章中，將對我們所提的摘要技術作詳盡的介紹。以下僅就上述研究項目，整理所對應的採用方法。

Table 2: 研究項目與相關技術

研究項目	相關技術
文件模型的建構及表示	1) Latent Semantic Analysis 2) Text Relationship Map
文件主題偵測及重要性評估	1) Global Bushy Path 2) Average Similarity
摘要內容組織及排序	MMR

2. 相關研究

近年來網際網路興起後，文件摘要從 80 年代的蟄伏期復甦又蓬勃發展起來。總括而言，大部分文件摘要之研究著重於單文件摘要。目前的研究涵蓋資訊擷取(Information Retrieval)與自然語言處理(Natural Language Processing)等技術，除詞性分析、詞組分析，更運用 WordNet [39]等領域知識(Domain Knowledge)

輔助，進行資訊萃取(Information Extraction)等較複雜的語意分析。

相關的研究單位很多，著名的有卡內基美濃大學(Carnegie-Mellon University)、康乃爾大學(Cornell University)、南加州大學(Southern California University)、密西根大學(Michigan University)及哥倫比亞大學(Columbia University)等；國內大學，如台灣大學、清華大學都有相關的研究。除此之外，美國國防部高等研究計畫機構(DARPA)亦舉辦大型的文件摘要比賽，如SUMMAC [50]與DUC [15]，皆有詳細的介紹文件摘要方法及評估的標準。

由此可知，文件自動摘要是具有實際運用價值與挑戰性的研究，以下我們介紹幾個著名的相關研究成果，然後針對相關的文獻作深入的討論。

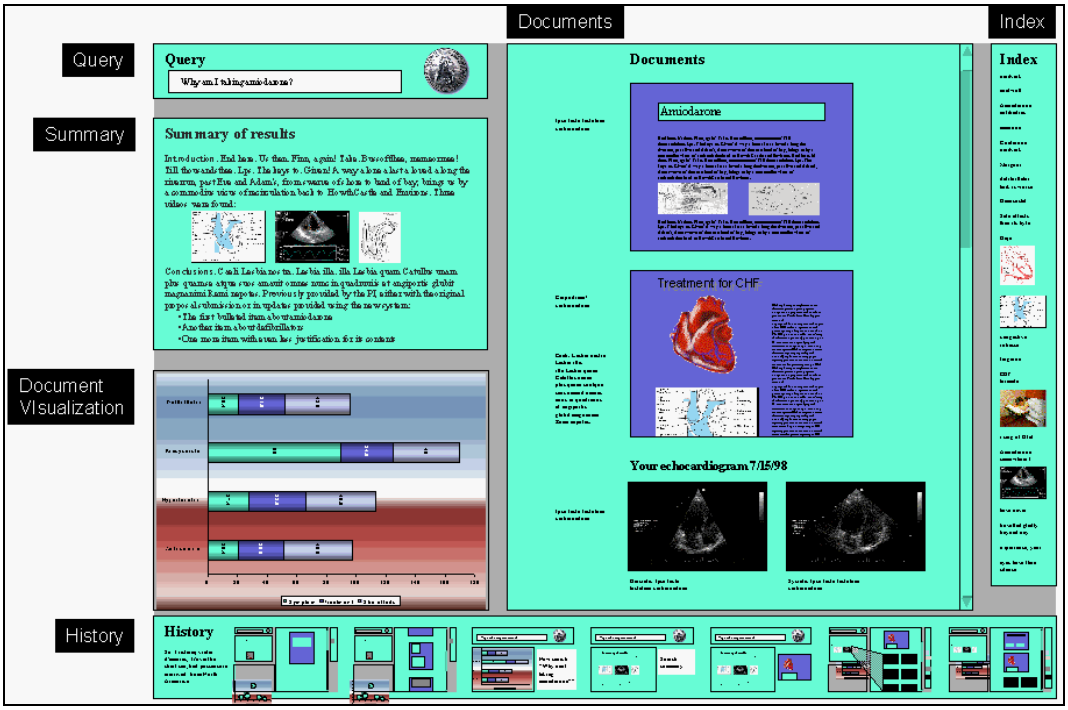

Figure 1: PERSIVAL 使用者介面 [43]

哥倫比亞大學發展PERSIVAL系統[43]，將病人就醫紀錄與資料庫中相關的醫學影像和聲音作關聯，並透過適當的版面設計將結果以摘要的形式呈現出來。如Figure 1所示，文件摘要不再只侷限於純文字資訊摘取，更包括內容上下文中影像與聲音的關聯及呈現。該系統亦建置個人化環境(Personalized Environment)，提供不同層面的摘要內容。舉例而言，病人及親屬所看到的摘要內容與醫療人員所看到的摘要內容，其深度及廣度皆有所不同。該系統更提供專業術語(Terminology)的轉換，使摘要的內容與使用詞彙，隨使用者背景知識與興趣的差異而有所不同。同時，提供文件的視覺化(Visualization)摘要，以方便使用者快速了解各文件所提及的主題與各文件的差異性。

哥倫比亞大學亦針對新聞文章發展 Columbia Summarizer [36]。該系統依據不同類型的文件整合不同的摘要技術，定義文件類型分為 1) 單一事件(Single Event)；2) 相關多事件(Multiple Related Event)；3) 傳記(Biography)；4) 相關事件的討論議題(Discussion Issue)等。Figure 2 為該系統的系統架構圖，共分為三個部分：Preprocessing、Routing 及 Summarizer Module。Preprocessing 將輸入的文件轉換成統一的 XML 格式；Router 則依據輸入文件的類型轉送給適當的摘要器；MultiGen [37]處理具有相同事件的文件集；DEMS (Dissimilarity Engine for Multi-document Summarization) [49]則依據輸入文件的特徵分析，處理多事件及傳記等類型的摘要。目前，該系統已整合於線上新聞摘要系統 NewsBlaster [11]，提供每日新聞主題偵測、追蹤及摘要服務。

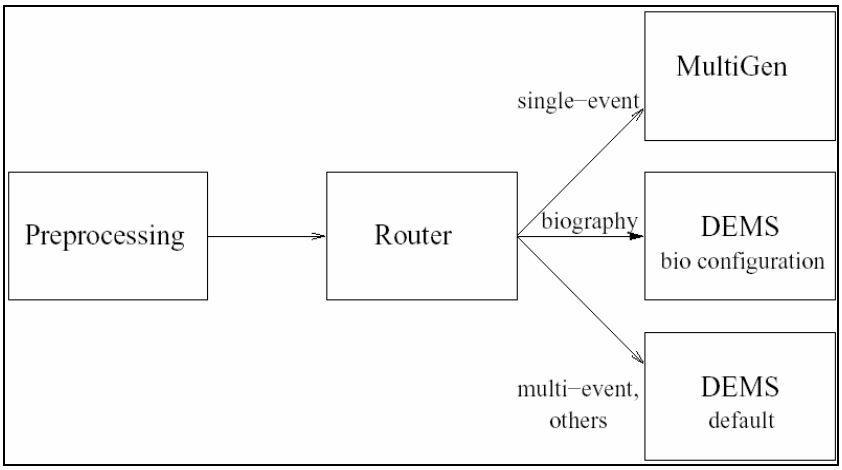

Figure 2: Columbia Summarizer 系統架構[36]

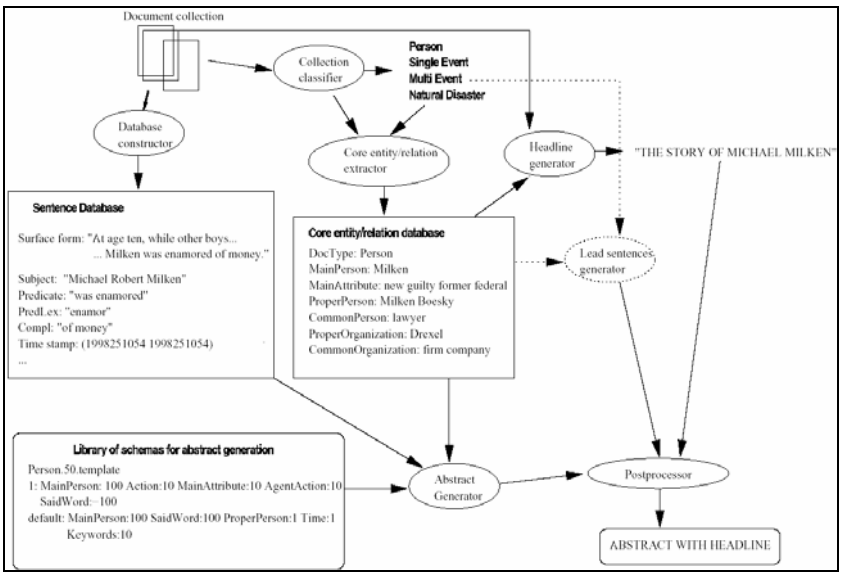

Figure 3: GLEANS 系統架構圖[13]

南加州大學發展一套摘要系統 GLEANS [13]，Figure 3 為其系統架構。該系統將文件集中所有文件所描述的物件(Entity)及物件間關聯(Relationship)抽取出來，並以資料庫表格的表示法儲存。同時，利用分類技術將文件集分為四種不同的類別，分別為 1) 單人物(Single Person)；2) 單事件(Single Event)；3) 多事件(Multiple Event)；4) 天然災害(Natural Disaster)。根據不同的類別，依據事先定義好的模板(Template)產生長度較短的內容提要。最後，依照不同的類別，考慮內容一致性(Coherence)的關係以產生最後的摘要內容。GLEANS 之特點在於利用模板產生品質較佳的摘要(Abstract)。

南加州大學同時發展 NeATS (Next Generation Automated Summarization) [31]。該系統分析字詞的重要性，包含單字詞(Unigram)、二字詞(Bigram)及三字詞(Trigram)，利用 log likelihood ratio 計算字詞與主題的相關性，自動擷取出文件集中與主題相關的部分(即 Topic Signature)。NeATS 系統可產生一般性摘要(Generic Summary)，且其結果會依照使用者喜好的主題有所調整。其運作步驟如下：1) 擷取主題特徵(Topic Signature) [24]及主題語句[24]，並依照語句的重要性排序(Ranking)；2) 利用 OPP (Optimal Position Policy) [24]將不具重要性之語句移除；3) 強化一致性(Cohesion)及連貫性(Coherent)；4) 利用 MMR [8]篩選語句以減少重複性，並保留下固定的摘要語句數目的語句；5) 強化時間順序(Chronological)的一致性；6) 將摘要結果格式化並輸出。

密西根大學對於多文件摘要提出各種不同的摘要技術，包含 Centroid-based Approach [45]、Cross-Document Structure Theory [54]、Revision-based Approach [42]及 Event-based Approach [12]。同時，亦實作三個不同類型之線上摘要系統，分別為 MEAD [38]、NewsInEssence [41]及 WebInEssence [44]。目前，MEAD 已經發展為適用於一般領域的多文件摘要模組，其研究目的為 1) 發展中英文的多文件摘要模組；2) 發展單/多文件摘要系統的評估工具；3) 實驗並評估四種不同的摘要標準，包含 Co-Selection、Content-based、Relative Utility 及 Rank Preservation。NewsInEssence 為應用於新聞領域的摘要系統，提供新聞文章的主題群集(Topic Clustering)、即時搜尋、文章摘要及使用者互動(User Interaction)等功能。WebInEssence 則整合文件摘要技術於搜尋技術中。

德州大學開發 GISTexter 摘要系統[21]，Figure 4 為 GISTexter 的系統架構圖。該系統利用資訊萃取(Information Extraction, IE)系統 – CICERO [22]與外在的知識，如 WordNet [39]，抽取文件中所提到的物件與事件，並且建構物件與事件的關聯模型。摘要的產生方式則是套用既有的模板來產生內容連貫性與一致性高的摘要。對單文件來說，主要是透過語句抽取(Sentence Extraction)的方式；對多文件來說，則是將分散於多文件中的相同主題(Shared Topics)抽取出來。

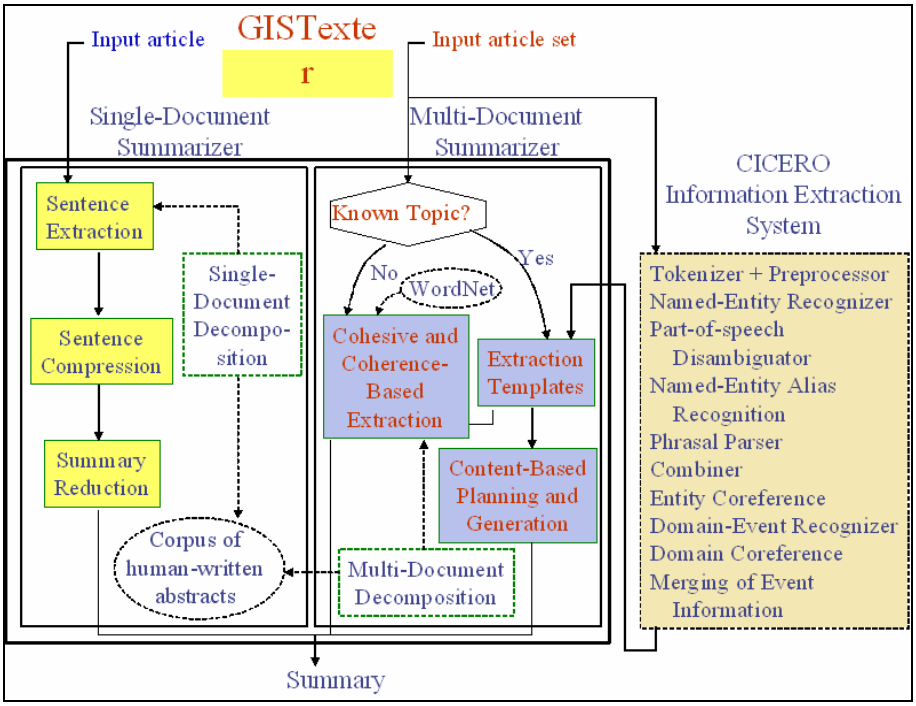

Figure 4: GISTexter 系統架構[21]

馬里蘭大學發展適用於大型文件集(Large Corpus)的摘要系統，名為 XDoX [23]，其處理的文件集大小約為 50-500 篇文章。該系統利用分群技術(Clustering)將文件集分為幾個有意義的主題，接著以段落為單位，依據段落與主題群集的相關程度作分類，最後依據不同的群集產生摘要，其流程如 Figure 5 所示。另外，XDoX 提供使用者兩種不同的摘要結果，一為詳細的摘要，提供較豐富的資訊；另一個則依據使用者的需求，如壓縮比等等，提供資訊量較少的摘要。

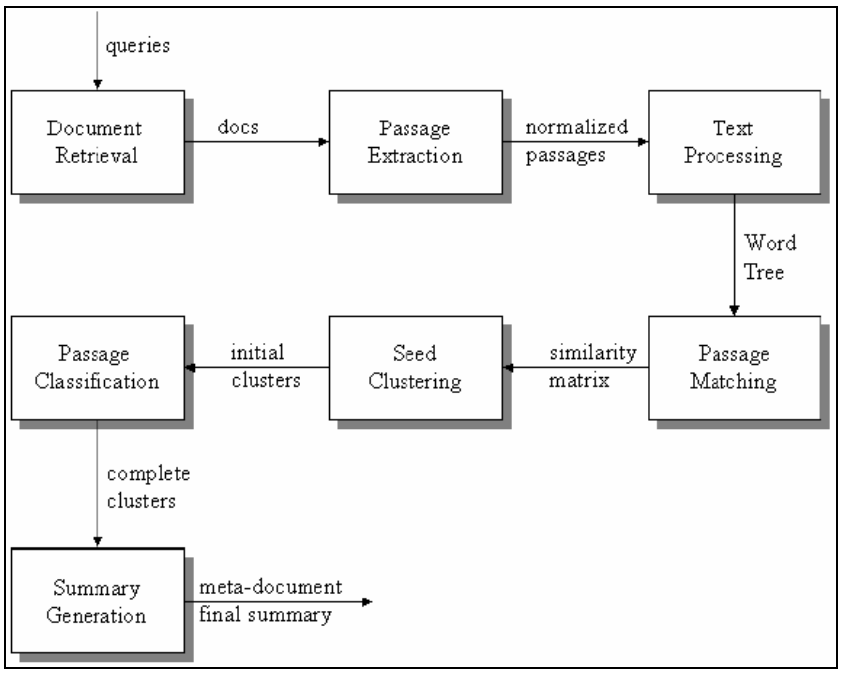

Figure 5: XDoX 系統架構[23]

微軟劍橋研究中心利用語彙鏈結(Lexical Bonds)技術產生摘要[25]。其方法共分為三個步驟，分別為 1) 分析(Analysis)；2) 轉換(Transformation)；3) 合成(Synthesis)。首先，將文件的特徵抽取出來；共考慮 12 種文件的特徵，如語句在文件中的位置等。接著，利用 SVM (Support Vector Machine)將文件中的重要語句挑選出來。最後，依照語句出現在文件中的順序排序以產生摘要。該方法的好處在於利用統計與機器學習的技術，並且透過語彙鏈結將文字的語意納入考量，可產生連貫性較佳的摘要。

其他相關研究，如[46]針對同一事件的新聞文件作摘要。他們利用專有名詞識別技術(Named Entity Identification)擷取人名、地名及組織名等資訊，並由新聞摘要的語料庫中學習摘要的產生方式；同時，利用事先定義好的摘要模板來產生摘要。McKeown et al. [37]將機器學習與統計的技術整合應用於多文件摘要的研究。他們的方法可分為三個部分：1) 主題辨識(Theme Identification) [18]: 透過分群技術將文件中的主題(Theme)抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [5]: 將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) [37]: 利用 FUF/SURGE [17][47]將所摘錄出來的重要字詞重新組合以產生流暢的摘要。

國內相關研究，如台灣大學[56]，將中文新聞文件拆解成長度較短的句子(Sub-sentence)，同時萃取名詞與動詞關鍵詞。接著，利用關鍵詞計算任兩小句間關聯強度，將關聯度大於門檻值之小句作成連結，最後評估小句連結並將重要的小句取出作為摘要。清華大學[55]，提出一個可調式中文文件摘要系統，該系統包含三個部分，分別為 1) 文件分群；2) 文件內容分析；3) 摘要呈現。其概念乃是將語句分群，以抽取出代表某事件的重要語句；接著，去除重複的資訊，同時標示重點後將新聞摘要呈現給使用者。

本計畫研究團隊亦對於單文件摘要進行相關研究，提出兩種新的文件摘要方法，以摘錄文件中重要語句，分別為 Modified Corpus-based Approach (MCBA) [52]及 LSA-based T.R.M. Approach (LSA+T.R.M.) [52]。

MCBA 基於統計模型與特徵分析，以評估語句的重要性。考慮的特徵，分別為語句位置(Position)、正面關鍵詞(Positive Keyword)、負面關鍵詞(Negative Keyword)、向心性(Centrality)及與標題相關度(Resemblance to the Title)。我們提出三個新的想法：1) 利用語句位置重要性分級提高不同語句位置的重要性；2) 利用詞彙關聯(Word Co-occurrence)分析文件中新詞，並將新詞加入關鍵詞的重要性計算；3) 利用基因演算法(Genetic Algorithm)找出適合之語句權重計算方式(Score Function)。實驗結果顯示，我們所提的方法有良善的表現。當壓縮比(Compression Ratio)為 30%時，平均而言，F-measure 為 0.5151。

LSA+T.R.M.利用潛在語意分析(Latent Semantic Analysis)技術，以擷取文件概念結構(Conceptual Structure)，即語意矩陣(Semantic Matrix)，可達到進行語意層面分析之目的。同時，利用語意矩陣導出語句表示式(Sentence Representation)，以建構主題相關地圖(Text Relationship Map)。最後，透過主題相關地圖，篩選重要的語句成為摘要。針對語意矩陣的建構，我們考量單文件層面(Single-document Level)及文件集層面(Corpus Level)，並比較兩種模式之適用性。實驗收集 100 篇關於政治類的中文文件。實驗結果顯示，我們所提的方法有良善的表現。當壓縮比(Compression Ratio)為 30%時，平均而言，F-measure 為 0.4242。

2.1. 文獻探討

2.1.1. MEAD[38]

MEAD [38]接受分群過後的文件集[4]，以語句(Sentence)為單位，針對每個文件群(Document Cluster)抽取出具有代表性的語句為摘要。方法如 Figure 6 所示。MEAD 考慮文件群中每個語句與群中心(Centroid)的相關度、語句的位置及該語句與所屬文件中首句的相似度，以評估每個語句的重要性。同時對每個文件群抽取出 n_i*r 個語句，以組成摘要；其中，n_i 代表 $Cluster_i$ 中語句的總數，r 代表壓縮比。

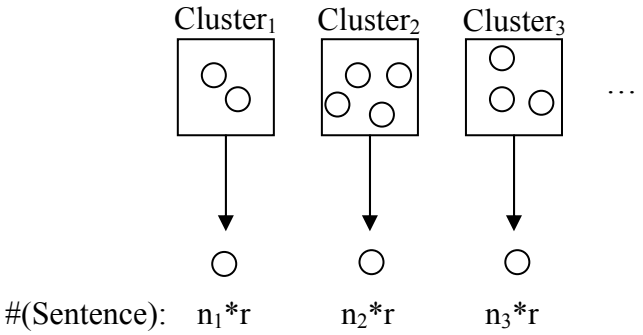

Figure 6: MEAD 運作原理示意圖

評估語句的重要性，MEAD 考慮以下三個特徵，分別為 1) 語句所在的文件群；2) 語句於文件中的位置，通常出現在文件中的首句可視為代表整篇文章，因此加重這些語句的重要性；3) 語句如果與首句有關的話，亦加重該語句的重要性。最後，MEAD 以線性組合(Linear Combination)綜合地評估語句的重要性，如 Eq. (1)：

[4] MEAD 接受相關的文件集，以產生摘要。然此處所提及之相關文件集，實為考慮 loosely-related documents。

$$S_i = w_1 \times C_i + w_2 \times F_i + w_3 \times L_i \quad \text{Eq. (1)}$$

其中，C_i 代表語句所在文件群的群中心(Centroid)的相關度，F_i 代表跟所屬文件之首句的相似度，L_i 代表該語句是否為所屬文件的首句。一般而言，MEAD 使用的首句加重計分法，比較適用於藝術類的文章或是新聞文章[5]；如果文件集是為其他領域，例如技術類的文件，則首句加重計分法要再調整才合適。

2.1.2. Theme Recognition [37]

McKeown et al. [37]認為主題相關的文件集中，存在有許多不同的主題(Theme)；依著此假設將機器學習與統計的技術整合應用於多文件摘要的研究。他們的方法，分為三個部分：1) 主題辨識(Theme Identification) [18]: 透過分群技術將文件中的主題(Theme)抽取出來，同時辨識文件間相似及差異的部分；2) 資訊融合(Information Fusion) [5]: 將討論相關主題的段落融合，並去除重複的資訊；3) 摘要生成(Text Reformulation) [37]: 利用 FUF/SURGE [17][47] 將所摘錄出來的重要字詞重新組合以產生流暢的摘要。

首先，考慮以下特徵以決定兩段落的相似度，進而利用分群法將找出主題，即相似段落的集合。

- Word co-occurrence：假如段落中有許多相似的字，則兩個段落可視為相似。
- Matching noun phrases：利用 LinkIt [51]判斷是否互相關聯的名詞片語群組。
- WordNet synonyms：使用 WordNet [39]找出同義詞組。
- Common semantic classes for verb：判斷具有同一語意的動詞詞組。

接著，利用 Information Fusion 的技術，從主題中萃取出具有代表性的詞組或片語。接著，依照出現在文章中的次序，對片語排序。最後，藉由 FUF/SURGE 自然語言產生器生成完整語句。FUF(Functional Unification Formalism)利用 SURGE 產出句法樹(Parsing Tree)，接著，藉由句法樹的轉換，以產生新的語句。

2.1.3. MMR[8]及 MMR-MD[19]

MMR (Maximal Marginal Relevance) [8]適用於單文件摘要，可用於降低摘要中具有相同涵義的語句，即減少重複性資訊。其概念乃是對所挑選出與 Query 相關的語句重新排序，以符合具有最大相關度及最大差異度的特性。排序方式如 Eq. (2)：

[5] 此類文章通常於第一段第一句說明整篇文章的重點。因此，首句之重要性必須加重考慮。

$$MMR \overset{def}{=} Arg \max_{S_i \in R \setminus S} [\lambda Sim_1(S_i, Q) - (1-\lambda) \max_{S_j \in S} Sim_2(S_i, S_j)] \quad \text{Eq. (2)}$$

其中，S 代表以挑選出的語句集合，S_i 代表某個語句，Q 代表 Query，$Sim_1(S_i,Q)$ 計算 S_i 與 Q 的相似度，$Sim_2(S_i,S_j)$計算 S_i 與 S_j 的相似度。

Table 3: MMR-MD 中 Sim_1 及 Sim_2 的計算方式[19]

$$Sim_1(P_{ij}, Q, C_{ij}, D_i, D) = w_1 * (P_{ij} \cdot Q) + w_2 * coverage(P_{ij}, C_{ij}) + w_3 * content(P_{ij}) + w_4 * time\_sequence(D_i, D)$$

$$Sim_2(P_{ij}, P_{nm}, C, S, D_i) = w_a * (P_{ij} \cdot P_{nm}) + w_b * clusters\_selected(C_{ij}, S) + w_c * documents\_selected(D_i, S)$$

$$coverage(P_{ij}, C) = \sum_{k \in C_{ij}} w_k * |k|$$

$$content(P_{ij}) = \sum_{W \in P_{ij}} w_{type}(W)$$

$$time\_sequence(D_i, D) = \frac{timestamp(D_{maxtime}) - timestamp(D_i)}{timestamp(D_{maxtime}) - timestamp(D_{mintime})}$$

$$clusters\_selected(C_{ij}, S) = |C_{ij} \cap \bigcup_{v,w:P_{vw} \in S} C_{vw}|$$

$$documents\_selected(D_i, S) = \frac{1}{|D_i|} * \sum_{w} [P_{iw} \in S]$$

where
Sim_1 is the similarity metric for relevance ranking
Sim_2 is the anti-redundancy metric
D is a document collection
P is the passages from the documents in that collection (e.g., P_{ij} is passage j from document D_i)
Q is a query or user profile
$R = IR(D, P, Q, \theta)$, i.e., the ranked list of passages from documents retrieved by an IR system, given D, P, Q and a relevance threshold θ, below which it will not retrieve passages (θ can be degree of match or number of passages)
S is the subset of passages in R already selected
$R \setminus S$ is the set difference, i.e., the set of as yet unselected passages in R
C is the set of passage clusters for the set of documents
C_{vw} is the subset of clusters of C that contains passage P_{vw}
C_v is the subset of clusters that contain passages from document D_v
$|k|$ is the number of passages in the individual cluster k
$|C_{vw} \cap C_{ij}|$ is the number of clusters in the intersection of C_{vw} and C_{ij}
$w_{\cdot}$ are weights for the terms, which can be optimized
W is a word in the passage P_{ij}
$type$ is a particular type of word, e.g., city name
$|D_i|$ is the length of document i.

MMR-MD [19]延伸 MMR 的概念，針對多文件摘要提出適合的排序方式。主要目標是使摘要語句對文件的主題和 Query 有極高的相似度，同時能夠降低摘要中具有重覆意思的段落數目。MMR-MD 同時考慮到時間順序、專有名詞、對主題的相似度以及代名詞的 Penalty。其挑選段落的方式，如 Eq. (3)：

$$MMR-MD \overset{def}{=} Arg \max_{P_{ij} \in R/S} [\lambda Sim1(P_{ij}, Q, C_{ij}) - (1-\lambda) \max_{P_{ij} \in S} Sim_2(P_{ij}, P_{nm}, C, S)] \quad \text{Eq. (3)}$$

其中，$Sim_1(P_{ij},Q,C_{ij})$計算 P_{ij} 與 Q 的相似度，同時衡量與段落所在的文件群的相關度；$Sim_2(P_{ij},P_{nm},C,S)$計算 P_{ij} 與 P_{nm} 的相似度，其中 P_{nm} 為一以挑選出之段落。上述兩相似度的計算方式，整理於 Table 3。

MMR-MD 希望能使的摘要中的段落儘可能的相似於 Query，但其所選到的段落間要儘可能的不相似。λ則是用來控制要取與 Query 相似度高的段落，但彼此之間的重複性可能也高，或是要與段落相似度稍低的段落，但彼此之間的重複性也低。適當的λ值可以找到兼具主題但又不會有過多重複性段落為摘要。

2.1.4. Graph Matching [33]

Mani et al. [33]將文件表示成圖形(Graph)，其中，每個節點代表一個關鍵詞(Term)，節點與節點間用不同的關係連接起來，包含 1) 片語關係(PHRASE)；2) 形容詞關係(ADJ)；3) 同義關係(SAME)；4) 關聯關係(COREF)。文件圖形模型，如 Figure 7 所示。

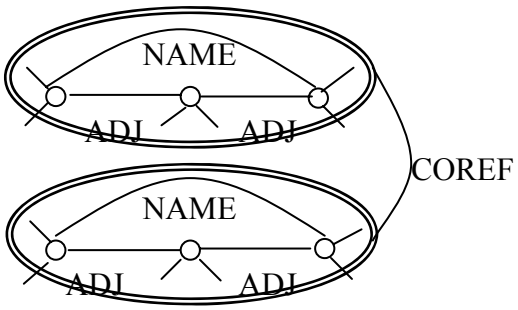

Figure 7: 文件圖形模型範例

首先，賦予每個節點一權重(Weight)，權重值初始為該關鍵詞的 TF-IDF [3]值。接著，利用 Spreading Activation [9]演算法，透過節點間相連的連結權重變更節點的權重值，以找出與 Query 相關的節點。接著，比較兩兩文件圖形模型的相似度(Commonality)及差異性(Difference)。他們提出 FSD (Find Similarities and Differences)演算法，以找出兩圖形中相似或差異的節點。透過以下演算法，將節點分成兩群，一群為相似的節點，另外一群則為具有差異的節點。

1) Common = {c | concept_match(c, G1) & concept_match(c, G2)}
2) Differences = (G1 ∪ G2) – Common
concept_match(c, G) is true iff c1 ∈ G such that c1 is a topic term or c and c1 are synonyms.

最後，透過分析 Common 及 Difference 中的關鍵詞，計算語句的重要性，並挑選出重要的語句當成摘要結果。語句重要性的計算方式，如 Eq. (4)：

$$score(s)=\frac{1}{|c(s)|}\sum_{i=1}^{|c(s)|} weight(w_i), \quad where\ c(s)=\{w \mid w \in Common \cap s\} \qquad \text{Eq. (4)}$$

3. 研究方法

我們以先前研究單文件摘要所提出的方法 – LSA-based T.R.M. Approach [52]為基礎，加以改良以適用於多文件摘要的研究，同時提出段落重要性評估的三種模型。本節中，首先介紹潛在語意分析(Latent Semantic Analysis) [28]與主題關係地圖(Text Relationship Map) [48]，最後說明我們所提出的多文件摘要技術模型 – LSA-based MD-T.R.M. Approach。

3.1. 潛在語意分析 (Latent Semantic Analysis)

潛在語意分析 (Latent Semantic Analysis) [28]為以數學統計為基礎的知識模型，其運作方式與類神經網路(Neural Net)相似。不同的是類神經網路以權重的傳遞(Propagation)與回饋(Feedback)修正本身的學習；潛在語意分析則以奇異值分解(Singular Value Decomposition, SVD)與維度約化(Dimension Reduction)為核心作為邏輯推演的方式，其原理如 Figure 8 所示。

潛在語意分析將文件或文件集表示為矩陣，透過 SVD 將文件所隱含的知識模型，抽象轉換到語意空間(Semantic Space)，再利用維度約化萃取文件知識於語意空間中重要的意涵。整個過程除可以將隱含的語意顯現出來外，更能將原本輸入的知識模型提升到較高層次的語意層面。

潛在語意分析的應用非常廣泛，包含資訊擷取、同義詞建構、字詞與文句相關性判斷標準、文件品質優劣的判別標準及文件理解與預測等各方面的研究。

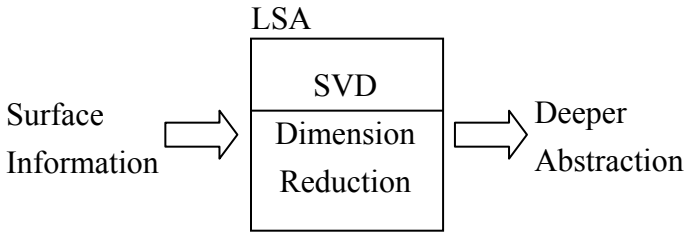

Figure 8: LSA 工作原理

做法上，首先將文件集(Corpus)中所有文件的 Context[6]建構為 Word-by-Context 矩陣(A)。矩陣中的每個元素($a_{i,j}$)，即某關鍵詞(W_i)在某 Context (C_j)中的權重或出現頻率。接著，透過奇異值分解將 A 分解轉換成三個矩陣乘積，即 $A=USV^T$。其中，S 代表語意空間(Semantic Space)，U 代表關鍵詞於此語意空間中的表示法，V^T 則代表 Context 於此語意空間中的表示法。再利用維度約化可更精確地描述語意空間的維度，並重建矩陣 $A'=U'S'V'^T$，可更進一步導出 Word-Word、Word-Context 或 Context-Context 的關聯強度。值得一提的是，潛在語意分析具有知識推演的能力；如果將原始矩陣中的任一數值改變，其結果會影響到最後重建的矩陣，且影響的範圍不單為原先經過改變的數值，更會影響到矩陣中的其他數值。

3.2. 主題關係地圖 (Text Relationship Map)

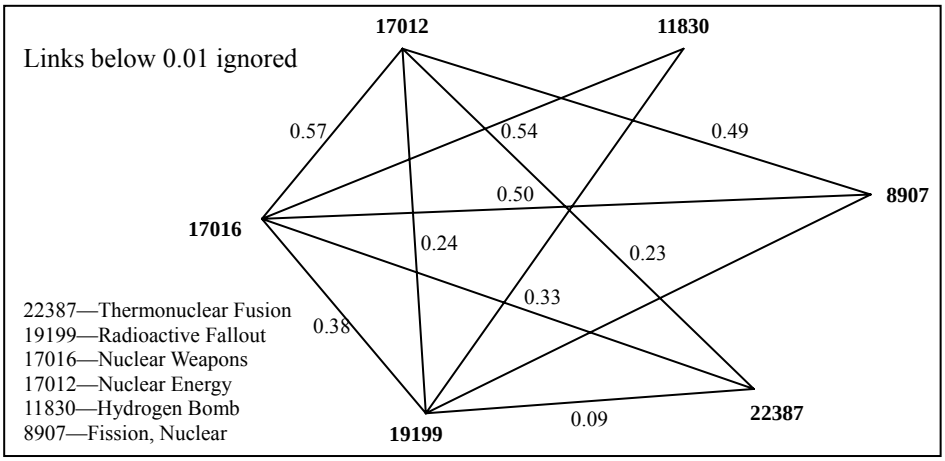

Figure 9: 主題關係地圖的範例[48]

主題關係地圖(Text Relationship Map) [48]將文件集中文件間關聯度表示成關係地圖。作法上將每篇文件以關鍵詞的向量表示法(Vector)表示，計算兩兩文件的相似度(Similarity)；當相似度大於臨界值時，表示此兩篇文件存在連結關係

[6] Context 可視需求定義為語句(Sentence)，段落(Paragraph)，或文件(Document)的層面來考量。

(Semantic Related Link)。依此原則可以建構出所有文件間的關係地圖。舉例來說，Figure 9 中編號 17012 及 17016 的文章，二者的相似程度約 0.57，大於臨界值 0.01，所以存在連結關係；而 8907 與 22387 的相似度則低於臨界值，因此於主題關係地圖中並不存在連結。一般來說，具有連結的文章，可說它們之間具有關聯性。

[48]將主題關係地圖的概念應用於單文件摘要研究。以每個段落(Paragraph)為單位計算兩兩段落的相似度，建構主題關係地圖[7]。當某個節點具有的連結數愈多，則代表該節點所對應的段落和整篇文件中主題的相關度愈高。[48]依據連結數目的多寡來決定摘錄段落順序，並提出以下三種方法以產生單文件摘要：

1. Global Bushy Path

首先定義任一節點的 *Bushiness* 為該節點與其他節點的連結數目；擁有越多關聯連結的節點，表示該節點所對應的段落與其他段落所討論的主題相似，因此，該段落可視為討論文件主題的段落。Global Bushy Path 將段落依照原本出現在文件中的順序以及其連結個數由大而小的排列。接著，挑選排名前 *K* 個段落(Top-*K*)，即為該文件的摘要。

2. Depth-first Path

Depth-first Path 選取某個節點 – 可能為第一個節點或是具有最多連結的節點，接著每次選取於原始文件中順序與該節點最接近且與該節點相似度最高的節點當作下一個節點，依此原則選取出重要而且連續的段落以形成文件摘要。

3. Segmented Bushy Path

Segmented Bushy Path 分為兩個步驟，首先分析文件結構進行文件結構切割(Text Segmentation)。接著針對每個 Segmentation 個別利用 Global Bushy Path 來選取重要的段落。為了保留所有 Segmentation 的內容，每個 Segmentation 至少要挑選出一個段落納入最後的摘要。

[7] 地圖上每個節點為文件中的某個段落；兩節點的連結，則表示兩節點的相似度大於臨界值。

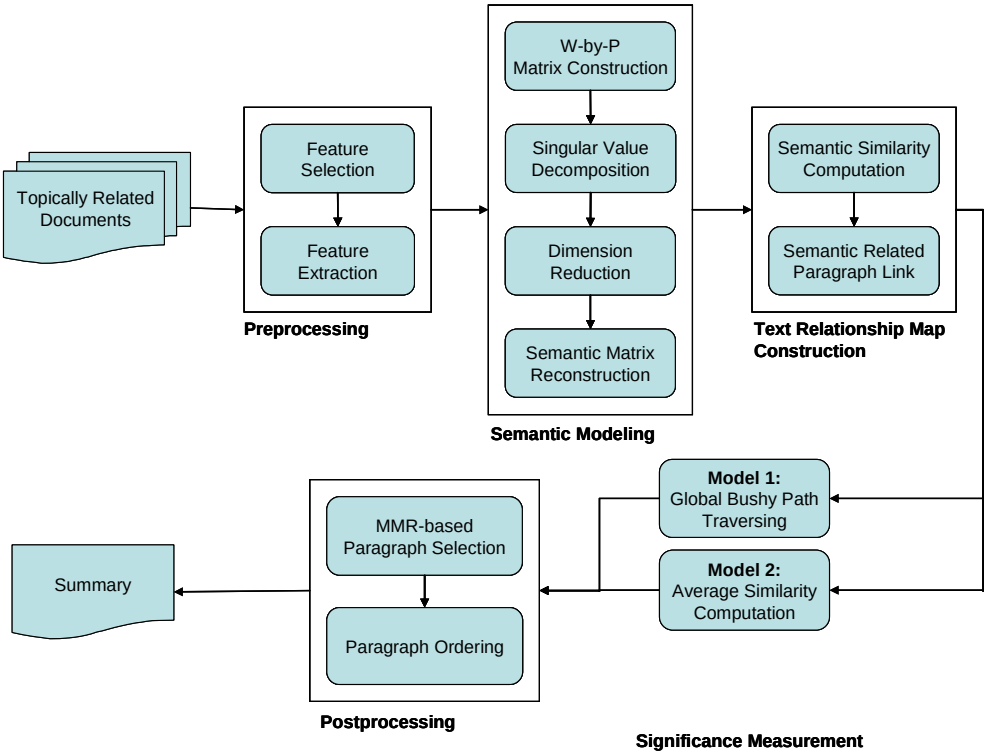

Figure 10: 本研究所提之多文件摘要架構

3.3. Proposed LSA-based MD-T.R.M. Approach

本節以我們先前對於單文件摘要所提出的方法 LSA-based T.R.M. Approach [52]為基礎，加以改進以適用於多文件摘要，並提出段落重要性評估的三種模型。系統架構如 Figure 10 所示[8]，共包含五個模組，分別為前處理(Preprocessing)、語意模型建立(Semantic Modeling)、主題關係地圖建構(Text Relationship Map Construction)、段落重要性評估(Significance Measurement)及後處理(Post-processing)。以下分別說明各個模組之功用。

3.3.1. 前處理(Preprocessing)

前處理包含兩個步驟，分別為特徵選取(Feature Selection)及特徵擷取(Feature Extraction)。

■ 特徵選取

我們以段落(Paragraph)為單位，考慮所有的單字詞(Unigram)、二字詞(Bigram)及三字詞(Trigram)。針對二字詞及三字詞，利用 Mutual Information [35]計算其代

[8] 本計畫所提之多文件摘要架構，乃延伸先前研究所提出適用於單文件摘要之 LSA-based T.R.M. Approach [52]，利用潛在語意分析(LSA) [28]與主題相關地圖(Text Relationship Map) [48]作為文件分析模型。

表性，以篩選不具代表性之特徵，計算方式如 Eq. (5)[9]：

$$MI(x,y)=\frac{P(x,y)}{P(x)p(y)} \quad \text{Eq. (5)}$$

其中，x 與 y 為相鄰之兩個單字詞[10]，$P(x)$為 x 出現於文件集的個數，$P(y)$為 y 出現於文件集的個數，$P(x, y)$則為 x 與 y 共同出現的個數。為了更進一步篩選出具有代表性的特徵，針對每個特徵計算其 IDF (Inverse Document Frequency) [3]，其計算如 Eq. (6)所示：

$$IDF(w_j)=\log\frac{N}{n} \quad \text{Eq. (6)}$$

其中，w_j 為一特徵關鍵詞，N 為文件集中段落的總數，n 為 w_j 出現的段落總數。當 IDF 值大於預設的臨界值，則表示該特徵具有代表性。

■ 特徵擷取

每個特徵的重要性，除了考慮每個特徵關鍵詞於段落出現的頻率外，亦考慮每個特徵關鍵詞於文件集中的重要程度。定義其權重為 K_{ij}，其計算如 Eq. (7)：

$$K_{ij}=G_i * L_{ij} \quad \text{Eq. (7)}$$

假設文件集中段落的集合為 $P=\{P_j \mid P_j$ 代表某一 $P_{k,l}$，即文件 D_l 中 P_k 段落$\}$，G_i 代表特徵關鍵詞 W_i 於 P 集合中的分佈權重，L_{ij} 代表 W_i 在 P_j 中的分佈權重。假設 c_{ij} 為 W_i 出現在 P_j 中的次數，t_j 為 W_i 出現在 P 集合中的次數，則 W_i 在 P_j 中的相對頻率計算方式如 Eq. (8)：

$$f_{ij}=\frac{c_{ij}}{t_i} \quad \text{Eq. (8)}$$

接著，考慮 P 集合中 W_i 的資訊分佈量(Entropy)，計算方式如 Eq. (9)：

$$E_i=-\frac{1}{\log(N)}\sum_{j=1}^{N} f_{ij} * \log(f_{ij}) \quad \text{Eq. (9)}$$

由 Eq. (9)可知當 f_{ij} 等於 1 的時候，E_i 的值為 0；當 f_{ij} 等於 $1/N$ 的時候，E_i 的值為 1。當 E_i 的值越接近於 1 的時候，表示 W_i 在 P 集合中的分佈越平均，W_i 的重要性便會降低；相反地，如果 E_i 的值越接近 0 的時候，表示 W_i 只出現在某些

[9]計算三字詞之 MI 值則為 MI(x, y, z)。
[10] 考慮二字詞或三字詞時，以段落為 window。

段落， W_i 的重要性便比平均分布在 P 集合中的特徵關鍵詞來得高。最後，定義 W_i 於 P 中的總體權重 G_i，如 Eq. (10)：

$$G_i = 1 - E_i \qquad \text{Eq. (10)}$$

另外，定義 W_i 於 P_j 中的權重 L_{ij}，如 Eq. (11)，其中 n_j 代表 P_j 中所含的特徵關鍵詞總數。

$$L_{ij} = \log_2\left(1 + \frac{c_{ij}}{n_j}\right) \qquad \text{Eq. (11)}$$

3.3.2. 語意模型建立(Semantic Modeling)

我們以建構 Word-by-Paragraph 的矩陣作為代表文件集之語意模型。假設該矩陣為 A，其中 a_{ij} 代表 W_i 於 P_j 的權重值[11]。接著，將矩陣 A 作奇異值分解(SVD)，使得 $A=USV^T$。對於 S 進行維度約化(Dimension Reduction)，同時取適當的維度後重新建構矩陣 $A'=U'S'V'^T$。此時，便得到具有語意的 Word-by-Paragraph 矩陣表示法，其中，每個列向量(Row-Vector)代表該關鍵詞在每個段落中的權重，而每個行向量(Column-Vector)代表該段落由各個關鍵字所組成的意義。

先前提及潛在語意分析(LSA) [28]能將文章中的隱性語意(Latent Semantic)表現出來。若以潛在語意分析所導出之段落表示式計算任兩段落的相似度，其結果會比單純使用關鍵字出現頻率權重的表示法來得好。基於這個想法，我們以潛在語意分析所得到的段落表示式 – 行向量(Column Vector)套用在主題相關地圖(Text Relationship Map) [48]，並衡量潛在語意分析對於摘要結果的影響。

3.3.3. 主題關係地圖建構(Text Relationship Map Construction)

以潛在語意分析重建之後得到的行向量當作段落的表示法，並計算任兩向量的 Cosine 值來衡量計算任兩段落的相似度。建構主題相關地圖時，只保留約 1.5 倍語句數目的連結；亦即，若有 n 個段落的話，那麼總共的連結數目 C(n, 2)個，而最後只保留相似度高的前 1.5*n 個連結。

3.3.4. 重要性評估(Significance Measurement)

我們提出兩種評估方式，以評估主題相關地圖上節點(即段落)的重要性。分別敘述如下：

[11] a_{ij} 的值可透過 Eq. (7)的公式計算。

- Model 1: Global Bushy Value

Global Bushy Value (GBV)[12]為主題相關地圖上任一節點與其他節點間的連結數目；定義如 Eq. (12)所示，其中，P_i 為主題地圖上一節點。由此可知，擁有越多關聯連結的節點，表示該段落與其他段落的寫作與用字方式相似，並且討論的主題也相似，因此，該段落視為討論主題的段落。

$$GBV(P_i) = \sum_{\forall P_j, P_j \text{ has a link with } P_i} 1 \qquad \text{Eq. (12)}$$

- Model 2: Average Similarity

相較於 Model 1 只考慮到主題相關地圖上每個節點的連結個數，我們參考[26]，並考慮每個連結權重的方式，以 Aggregate Similarity 計算每個節點的重要性，Aggregate Similarity 的示意圖如 Figure 11：

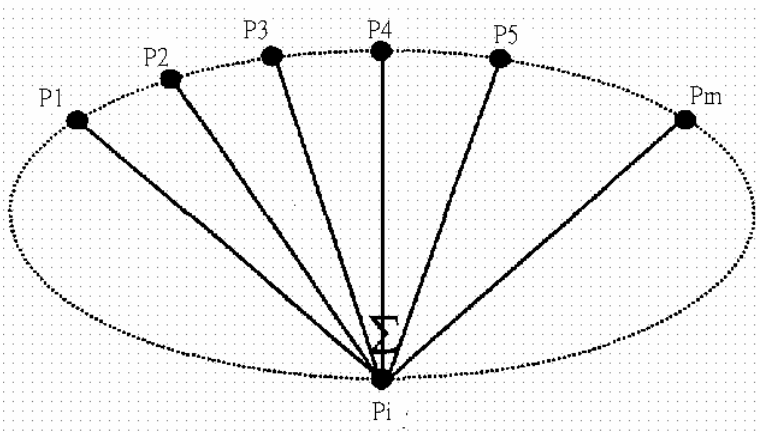

Figure 11: 計算 Aggregate Similarity 的概念圖示[26]

圖中的每個節點代表某個段落的向量表示法，每個連結代表兩個語句間的相似度，任兩個語句的相似度即是計算相對應向量間的內積值。Aggregate Similari 之計算如 Eq. (13)：

$$AvgSim(P_i) = \sum_{\substack{\forall P_j \neq P_i \\ P_j \text{ has a link with } P_i}} sim(P_i, P_j) \qquad \text{Eq. (13)}$$

其中，$sim(P_i, P_j)$為兩個節點間的相似度，即是計算相對應向量間 Cosine 值。計算每個節點的 Aggregate Similarity，其好處在於除了考慮到每個節點的連結個數，同時亦考慮到每個連結的權重值。

[12] 即[48]中定義之 Bushiness 值。

3.3.5. 後處理(Post-processing)

後處理包含兩個步驟，分別為段落選取(Paragraph Selection)及段落排序(Paragraph Ordering)。

■ 段落選取

我們參考 Maximal Marginal Relevance (MMR)[13] [8]的概念，提出段落選取的方法，如 Eq. (14)所示：

$$PS \overset{def}{=} Arg \max_{P_i \in R \setminus S} [SIG(P_i) - \lambda * \max_{P_j \in S} REL(P_i, P_j)] \quad \text{Eq. (14)}$$

其中，S 代表已被選到之段落的集合，$SIG(P_i)$ 代表 P_i 的重要性，$REL(P_i, P_j)$ 代表 P_i 與 P_n 的關聯強度。

做法上，PS 依序選取出組成摘要的段落，同時評估目前衡量的段落與先前取出的段落間相關程度；此機制乃是為了去除重複性(Anti-redundancy)，以提供使用者更多的資訊。Table 4 整理針對前一步驟所提出的兩個模型中 *SIG* 與 *REL* 的計算方式。

Table 4: *SIG* 與 *REL* 的計算方式

	$SIG(P_i)$	$REL(P_i, P_j)$
Model 1	$GBV(P_i)$	$\begin{cases} \alpha & \text{if } P_i \text{ has a link with } P_j \\ 0 & \text{otherwise} \end{cases}$
Model 2	$ASim(P_i)$	$sim(P_i, P_j)$

■ 段落排序

段落排序之目的，將挑選出來的段落依據內容一致性(Cohesion)及連貫性(Coherent)重新排序，以提供使用者適合閱讀的摘要。本研究中，我們的重點在於探討如何有效的評估段落或語句的重要性，以作為挑選摘要語句的依據。因此，本研究中有關段落排序的方法，僅將段落依其所屬文章的發佈日期(Publication Date)排序。

[13] 請參考相關研究工作中所提及之 MMR [8]及 MMR-MD [19]。

4. 結果與討論

4.1. 測試資料集

實驗中以 DUC (Document Understanding Conference)的 DUC 2003 資料為測試對象，該年度的評比共分為四個項目：1) Task 1 – Very short summaries; 2) Task 2 – Short summaries focused by events; 3) Task 3 – Short summaries focused by viewpoints; 4) Task 4 – Short summaries in response to a question。其中 Task 2 及 Task 3 為多文件摘要評比，然而因為我們所發展的摘要技術針對新聞事件多文件摘要而設計，因此，我們使用 Task 2 的資料作為評估演算法好壞的資料集。

Task 2 所提供的資料共有 30 個文件群(Document Clusters)，每個文件群中各有約 10 篇新聞文章，且這 10 篇新聞文件皆是討論相同的新聞事件發展。Task 2 要求每個參與比賽的系統針對這 30 個文件群各產生約 100 字的摘要[14]。

4.2. 評估方法

DUC 從 DUC 2004 起採用 ROUGE (Recall-Oriented Understudy for Gisting Evaluation)[58]為評估工具，因此我們亦已該工具來評估演算法的好壞。DUC 2003 的資料中，每個文件群皆由四個專家閱讀過所有文章後，針對該文件群所討論的事件主題等資訊，以人工方式產出約 100 字的短摘要。ROUGE 以專家所產出的摘要當參考答案，對照機器所產出的摘要內容，主要計算有平均有多少個 Word 被專家與機器所產生的摘要所共同包含。

ROUGE 的評估主要有 ROUGE-1, ROUGE-2, …, ROUGE-N, ROUGE-L, ROUGE-WL。ROUGE-N 以 N 個字為單位計算 Recall 值, 如 ROUGE-2 為以 Bigram 為單位時所得到的 Recall。ROUGE-L 以 common string 為單位計算 Recall，ROUGE-WL 則為加權過後的 ROUGE-L。目前已經證明 ROUGE-1 的評估分數比較接近專家所給定的評估分數，因此我們僅列出 ROUGE-1 的分數。

4.3. 實驗結果

實驗的參數設定，Model 1 的 α 值(亦即 Table 4 中的 α 值)動態設定為該 P_i 與目前所產生摘要段落集合中具有連結的 P_j 個數。Eq. (14)中的 λ 值則設定為 0 到 50，以比較各種不同 λ 值下對於摘要好壞的影響。

[14] 小於 100 字不予加分；多於 100 字則需先縮減為 100 字後再由評估工具計算分數。

Table 5 列出 Model 1 的 ROUGE-1 平均值；Table 6 則列出 Model 2 的 ROUGE-1 平均值。Figure 12 則將 Table 5 與 Table 6 整合製圖。由圖中可知，當 λ=0 時(亦即沒有考慮到 Anti-Redundancy)，此時的結果最差。另外，Model 1 不管在任何 λ 值的設定下，皆比 Model 2 有較佳的表現。這是因為 Model 2 考慮語句與其他語句的平均相似度值，導致失去了連結個數的表現。因而使得結果變得較差。

Table 5: Model 1 的實驗結果 (ROUGE-1 Average)

	λ=0	λ=10	λ=20	λ=30	λ=40	λ=50
ROUGE-1	0.3405	0.3579	0.3544	0.3521	0.3532	**0.3564**

Table 6: Model 2 的實驗結果 (ROUGE-1 Average)

	λ=0	λ=10	λ=20	λ=30	λ=40	λ=50
ROUGE-1	0.3402	0.3457	**0.3497**	0.3434	0.3429	0.3466

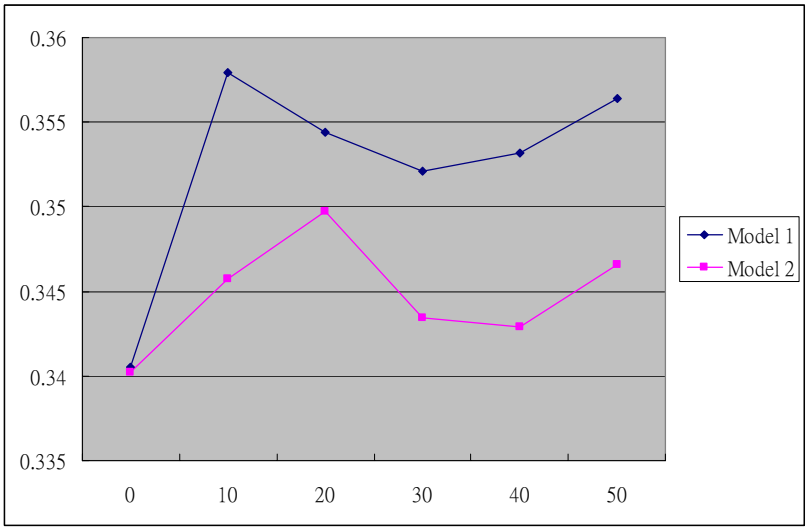

Figure 12: Model 1 與 Model 2 的 ROUGE-1 比較

Table 7 中列出 DUC 2003 年的比賽結果，其中 SYSID 代表不同的系統，且字母 A-J 為專家所作的摘要與其他專家的摘要比較結果，數字代表當年度參加比賽的系統代號。由此表並比較我們的所提出的演算法結果，可知我們的方法亦有不錯的結果，評估所獲得的分數約為中等以上。

Table 7: DUC 2003 的部分 Official Results

SYSID	ROUGE-1	SYSID	ROUGE-1
B	0.4467	17	0.3606
C	0.4451	6	0.3598
D	0.4358	20	0.3475
E	0.4201	14	0.3362
A	0.4149	23	0.3339
H	0.4109	18	0.3297
I	0.4001	21	0.3267
G	0.3956	3	0.3170
12	0.3918	2	0.3090
F	0.3918	11	0.3068
J	0.3883	19	0.3057
16	0.3747	15	0.2943
13	0.3698	10	0.2905
26	0.3627	22	0.2565

5. 結論

本計畫為三年期研究案之第一年計畫，我們考量相關文獻所提多文件摘要的方法之優缺點，並以先前發展之單文件摘要方法[52]為基礎，加以改進以適用於多文件摘要，同時對於段落重要性評估提出兩種模型。

以下整理我們所提出的摘要架構中各個模組及所採用的技術。

1. 前處理(Preprocessing)
   - Feature Selection
   - Feature Extraction
2. 語意模型建立(Semantic Modeling)
   - Latent Semantic Analysis
   - Semantic Matrix
3. 主題關係地圖建構(Text Relationship Map Construction)
   - Text Relationship Map
   - Similarity Matrix
4. 重要性評估(Significance Measurement)
   - Model 1：Global Bushy Path
   - Model 2：Average Similarity

實驗部分以 DUC [15]提供的資料進行實驗，並以過去 DUC 結果評估我們所提之多文件摘要方法的優劣。評估採用 ROUGE 來計算專家所產出的摘要與機器所產出摘要中，平均包含有多少個相同的字。我們以 ROUGE-1 為評估標準，實驗結果顯示 Model 1 獲得平均 ROUGE-1 分數為 0.3564，Model 2 獲得平均 ROUGE-1 分數為 0.3497。比較 DUC 2003 的比賽結果，我們的方法亦有不錯的表現，約在中等以上。

參考文獻

[1] Aone, C., Okurowski, M. E., Gorlinsky, J., & Larsen, B. (1999). A trainable summarizer with knowledge acquired from robust NLP techniques. Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[2] Azzam, S., Humphreys, K., & Gaizauskas, R. (1999). Using coreference chains for text summarization. In *Proceedings of the ACL'99 Workshop on Coreference and Its Application*, Baltimore.

[3] Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. Harlow, UK: Addison-Wesley Longman Co Inc.

[4] Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain (pp. 10-17).

[5] Barzilay, R., McKeown, K. R., & Elhadad, M. (1999). Information fusion in the context of multi-document summarization. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA (pp. 550-557).

[6] Bellegarda, J. R., Butzberger, J. W., & Chow, Y. L. (1996). A novel word clustering algorithm based on latent semantic analysis. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vol. 1 (pp. 172-175).

[7] Boros, E., Kantor, P. B., & Neu, D. J. (2001). A clustering based approach to create multi-document summaries. In *Proceedings of the Document Understanding Conference* (*DUC-2001*), New Orleans, LSA, USA.

[8] Carbonell, J., & Goldstein, J. (1999). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'98*), Melbourne, Australia (pp. 335-336).

[9] Chen, C. H., Basu, K., & Ng, T. (1994). An algorithmic approach to concept exploration in a large knowledge network. *Technical report*, MIS Department, University of Arizona, Tucson, AZ, USA.

[10] Chen, H. H., & Lin, C. J. (2000). A multilingual news summarizer. In *Proceedings of the 17th Conference on Computational Linguistics*, Saarbrücken, Germany (pp. 159-165).

[11] Columbia Newsblaster: summarizing all the news on the web. Available at http://www1.cs.columbia.edu/nlp/newsblaster.

[12] Daniel, N., Radev, D. R., & Allison, T. (2003). Sub-event based multi-document summarization. In *Proceedings of the HLT-NAACL-03 Text Summarization Workshop*, Edmonton, Alberta, Canada (pp. 9-16).

[13] Daumé III, H., Echihabi, A., Marcu, D., Munteanu, D. S. & Soricut, R. (2002). GLEANS : a generator of logical extracts and abstracts for nice summaries. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[14] DeJong, G. F. (1979). Skimming stories in real time: an experiment in integrated understanding. *Doctoral dissertation*, Computer Science Department, Yale University, New Haven, CT, USA.

[15] Document Understanding Conference (DUC). Available at http://duc.nist.gov.

[16] Edmundson, H. P. (1969). New methods in automatic extracting. *Journal of the ACM* (*JACM*), 16(2), 264-285.

[17] Elhadad, M. (1993). Using argumentation to control lexical choice: a functional unification implementation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.

[18] Eskin, E., Klavans, J., & Hatzivassiloglou, V. (1999). Detecting similarity by applying learning over indicators. In *Proceedings of the 37th Conference on Association for Computational Linguistics* (*ACL'99*), College Park, Maryland, MD, USA.

[19] Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 40-48).

[20] Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'01*), New Orleans, LA, USA (pp. 19-25).

[21] Harabagiu, S. M., & Lacatusu, F. (2002). Generating single and multi-document summaries with GISTEXTER. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[22] Harabagiu, S. M., & Maiorano, S. (2000). Acquisition of linguistic patterns for knowledge-based information extraction. In *Proceedings of the 2nd International Conference on Language Resources and Evaluation* (*LREC-2000*), Athens, Greece.

[23] Hardy, H., Shimizu, N., Strzaliowski, T., Ting, L., & Zhang, X. (2001). Cross-document summarization by concept classification. In *Proceedings of the Document Understanding Conference* (*DUC-2001*), New Orleans, LSA, USA.

[24] Hovy, E., & Lin, C. Y. (1999). Automated text summarization in SUMMARIST.

Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[25] Karamuftuoglu, M. (2002). An approach to summarization based on lexical bonds. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[26] Kim, J. H., Kim, J. H., & Hwang, D. (2000). Korean text summarization using an aggregate similarity. In *Proceedings of the 5th International ACM Workshop on Information Retrieval with Asian Languages,* Hong Kong, China (pp. 111-118).

[27] Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'95*), Seattle, WA, USA (pp. 68-73).

[28] Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.

[29] Lenci, A., Bartolini, R., Calzolari, N., Agua, A., Busemann, S., Cartier, E., Chevreau, K., & Coch, J. (2002). Multilingual summarization by integrating linguistic resources in the MLIS-MUSI project. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (*LREC'02*), Las Palmas, Canary Island, Spain.

[30] Lin, C. Y. (1999). Training a selection function for extraction. In *Proceedings of the 8th International Conference on Information and Knowledge Management* (*CIKM'99*), Kansas City, MO, USA (pp. 55-62).

[31] Lin, C. Y., & Hovy, E. (2002). NeATS in DUC 2002. In *Proceedings of the Document Understanding Conference* (*DUC-2002*), Philadelphia, PA, USA.

[32] Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2), 159-165.

[33] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.

[34] Mani, I., & Maybury, M. (eds.) (1999). *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[35] Maosong, S., Dayang, S., & Tsou, B. K. (1998). Chinese word segmentation without using lexicon and hardcrafted training data. In *Proceedings of the 17th International Conference on Computational Linguistics* (*COLING'98*) *and the 36th Annual Meeting of the Association for Computational Linguistics* (*ACL'98*) (*COLING-ACL'98*), Montreal, Quebec, Canada (pp. 1265-1271).

[36] McKeown, K. R., Evans, D., Nekova, A., Barzilay, R., Hatzivassiloglou, V., Schiffman, B., Blair-Goldensohn, S., Klavans, J., & Sigelman, S. (2002). The Columbia Multi-document Summarizer for DUC 2002. In *Proceedings of the*

Document Understanding Conference (*DUC-2002*), Philadelphia, PA, USA.

[37] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence* (*AAAI'99*), Orlando, FA, USA (pp. 453-460).

[38] MEAD. Available at http://tangra.si.umich.edu/clair/mead.

[39] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: a on-line lexical database. *Lexicography*, 3(4), 235-312.

[40] Myaeng, S. H., & Jang, D. (1999). Development and evaluation of a statistically based document system. Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[41] NewsInEssence: Interactive Multi-source News Summarization. Available at. http://www.newsinessence.com/nie.cgi.

[42] Otterbacher, J. C., Radev, D. R., & Luo, A. (2002). Revisions that improve cohesion in multi-document summaries: a preliminary study. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*, Philadelphia, PA, USA (pp. 27-36).

[43] PERSIVAL: Personalized Search and Summarization over Multimedia Information. Available at http://persival.cs.columbia.edu.

[44] Radev, D. R., Fan, W., & Zhang, Z. (2001). WebInEssence: a personalized web-based multi-document summarization and recommendation system. In *Proceedings of the NAACL Workshop on Automatic Summarization*, Pittsburgh, PA.

[45] Radev, D. R., Jing, H., & Budzikowska, M. (2000). Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *Proceedings of the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 21-30).

[46] Radev, D. R., & McKeown, K. R. (1998). Generating natural language summaries from multiple on-line sources. *Computational Linguistic*, 24(3), 469-500.

[47] Robin, J. (1994). Revision-based generation of natural language summaries providing historical background: corpus-based analysis, design, implementation and evaluation. *Ph.D. Thesis*, Department of Computer Science, Columbia University, New York, NY, USA.

[48] Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information Processing & Management*, 33(2), 193-207.

[49] Schiffman, B., Mani, I., & Concepcion, K. J. (2001). Producing biographical summaries: combing linguistic knowledge with corpus statistics. In *Proceedings of European Association for Computational Linguistics*.

[50] TIPSTER Text Summarization Evaluation Conference (SUMMAC). Available at http://www.itl.nist.gov/iaui/894.02/related_projects/tipster_summac.

[51] Wacholder, N. (1998). Simplex NPs clustered by head: a method for identifying significant topics in a document. In *Proceedings of Workshop on the Computational Treatment of Nominals*, *COLING-ACL*, Montreal, Canada (pp. 70-79).

[52] Yeh, J. Y., Ke, H. R., & Yang, W. P. (2004). Text summarization using a trainable summarizer and latent semantic analysis. *Information Processing & Management: Special Issue on ICADL 2002*. Accepted and to be appeared. (Recommended by ICADL 2002).

[53] Young, S. R., & Hayes, P. J. (1985). Automatic classification and summarization of banking telexes. In *Proceedings of the 2nd Conference on Artificial Intelligence Applications* (pp. 402-408).

[54] Zhang, Z., Blair-Goldensohn, S., & Radev, D. R. (2002). Towards CST-enhanced summarization. In *Proceedings of the 18th National Conference on Artificial Intelligence* (*AAAI'02*), Edmonton, Alberta, Canada (pp. 439-445).

[55] 陳鈺瑾 (2000). 可調式之中文文件自動摘要. *碩士論文*, 國立清華大學資訊工程研究所, 新竹, 台灣.

[56] 黃聖傑 (1999). 多文件自動摘要方法研究. *碩士論文*, 國立台灣大學資訊工程研究所, 台北, 台灣.

[57] 蘇哲君 (2001). 中英雙語多文件自動摘要系統研究. *碩士論文*, 國立台灣大學資訊工程研究所, 台北, 台灣.

[58] Lin, C-Y. 2004. ROUGE: a Package for Automatic Evaluation of Summaries. *Proc. of the Workshop on the Text Summarization Branches Out (WAS)*, Barcelona, Spain, 2004.

多語言複合式文件自動摘要之研究 (2/3)

計畫類別: 個別型計畫

計畫編號: NSC-93-2213-E-009-044-

執行期間: 2004/08/01 – 2005/07/31

計畫主持人: 楊維邦 教授

計畫參與人員: 柯皓仁 教授, 葉鎮源, 劉政璋, 顧世彥, 王瓊婉

中文摘要

本計畫為三年期研究案 – 多語言複合式文件自動摘要之研究之第二年計畫，其研究目的在於發展多語多文件摘要(Multilingual Multidocument Summarization)技術。研究內容著重於中英文混合式多文件摘要，研究議題為如何跨越語言型態及結構的障礙，提出計算中英文語句相似度的方法。

研究方法中，我們提出一套中英概念空間擷取的方法。基於中英雙語平行語料庫(Chinese-English Parallel Corpus)，進行中英文詞分群(Word Clustering)，以建立階層式概念空間(Hierarchical Concept Space)。接著，將中英文語句對應至該階層式概念空間中，透過此對應關係，可將關鍵詞向量空間表示式(Keyword-Level Vector Representation)轉換至概念向量空間表示式(Concept-Level Vector Representation)。最後，於相同的概念空間中，便可計算任兩中中、中英及英英語句的概念相似度。

中英混合式詞分群可抽取出中文詞與英文詞間的關連。當分群的群數目越多時，位於詞群中的中英文詞可互為翻譯；當分群的群數目越少時，位於詞群中的中英文詞可視為相關。透過階層式概念空間相似度的組合，可得不錯的中英語句對應。實驗結果顯示，考慮 Top 10 組中英對應段落，平均 Precision 約 57%為相關。摘要內容的評估，則由專家進行評比。平均而言，資訊量涵蓋度為 7.06，可讀性為 6.04。(其中，1 代表最差，5 代表普通，10 代表最好。)

關鍵詞：多語多文件摘要；中英文語句相似度；階層式概念空間；概念對應

英文摘要

As the second part of the project titled “The Research on Cross-Language, Composite and Multi-Document Automated Text Summarization,” the principal objective of this project is to develop novel techniques to address multilingual, multidocument summarization. The main issue is to propose a method to compute the similarity between two short passages which are written in different languages.

In this technical report, with a Chinese-English parallel corpus, we propose a framework to derive a hierarchical concept space by word clustering. On the basis of the concept space, a Chinese or an English paragraph can be mapped into the space and represented as a concept-level vector representation. Since a Chinese or an English paragraph is mapped into the same concept space, it becomes easy to compute the similarity between two paragraphs. Once the similarity between a Chinese and a English passage is obtained, the summary is generated using the techniques developed in the first-year project.

The preliminary results show that the larger the number of concept clusters, words in the same concept can be regarded as corresponding translations; while the smaller the number of concept clusters, words in the same concept are loosely-related. In the set of Chinese-English paragraph pairs with Top 10 similarities, approximately 57% are judged as related by humans. Regarding the quality of the summary, in average, a score of 7.06 and 6.04 was obtained in terms of information coverage and readability respectively, in which 1 the worst, 5 good, and 10 the best.

Keywords: Multilingual Multidocument Summarization; Chinese-English Sentence Similarity; Hierarchical Concept Space; Concept Mapping

1. 研究背景及目的

文件自動摘要(Automated Text Summarization)乃是從原始資料中精鍊出最重要資訊的過程，其結果即為該原始資料的精簡化版本，且可作為人們或其他資訊系統的判斷與決策依據[15]。

Text summarization is the process of distilling the most important information from a source (*or sources*) *to produce an abridged version for a particular user* (*or users*) *and task* (*or tasks*).

相關的研究起源於 1950 年代，初期以分析文件格式(Genre)為主[5], [13]，接著以模板(Template)來辨認人、事、時、地、物，以達到分析文件涵蓋主題的目的[22]。從 1990 年代開始，資訊擷取(Information Retrieval)的技術開始被應用於文件摘要上[4], [8], [9], [27]，此時期亦是單文件摘要(Single Document Summarization)研究的高峰期。單文件摘要將文件精簡化，著重於刪減無用資訊。

自從 2000 年開始，因為網路的發達，使得人們每天所接受的資訊量過於龐大，更有賴於文件自動摘要以幫助過濾資訊。此時期的研究主要偏重於多文件自動摘要(Multidocument Summarization)；多文件摘要針對多篇描述相似主題的文件(Topically-Related Documents)，研究議題為偵測多文件中所提及的相異及相同主題，並過濾出現於多文件中的重複性資訊。具代表性的研究有[2], [7], [14], [16]。2002 年開始，多語言文件摘要(Multi-lingual Document Summarization)的研究開始進行，如[3], [4], [6], [28]。多語言文件摘要係指文件來源可為不同語言所描述的文件，此類研究克服各語言型態、結構及用語習慣的差異，並提供語言間互相轉譯的能力。

本計畫為三年期研究計畫『多語言複合式文件自動摘要之研究』之第二年計畫。研究內容為探討中英文雙語混合式文件自動摘要的可行性。中文與英文之型態與結構皆有很大的差異，如何跨越語言的差異性，以分析中英文文件相似度乃是本計畫主要的研究議題。

多文件自動摘要中最重要的研究議題在於找出不同文件中相異及相同的部份，透過分群的方法，將相似度高的語句聚集在一起。同一語句群中，所涵蓋的資訊可視為相同；該語句群並可視為多文件中所提及的一個重要概念。同樣地，跨語言多文件自動摘要的研究中，重要的議題在於如何計算任兩相同或相異語言所構成語句間相似度，以便跨越語言的隔閡，達到將相似度高的語句聚集的目的。

本計畫提出利用中英平行語料庫(Chinese-English Parallel Corpus)，透過單字/詞分群(Word Clustering)以找出相關的中英詞群集。同一中英詞群集中，所包含的中文詞或英文詞即可視為相對應的翻譯。中英詞群集的集合，我們稱為該中英平行語料庫所分析出來的概念空間(Concept Space)。換句話說，透過概念群(Concept Cluster)的對應，便可將任一語句對應到概念空間上，並以不同的概念群當成向量空間的維度，進而得到該語句位於概念空間中的向量表示式(Vector Representation)，便可計算任兩相異語言所構成語句間相似度。

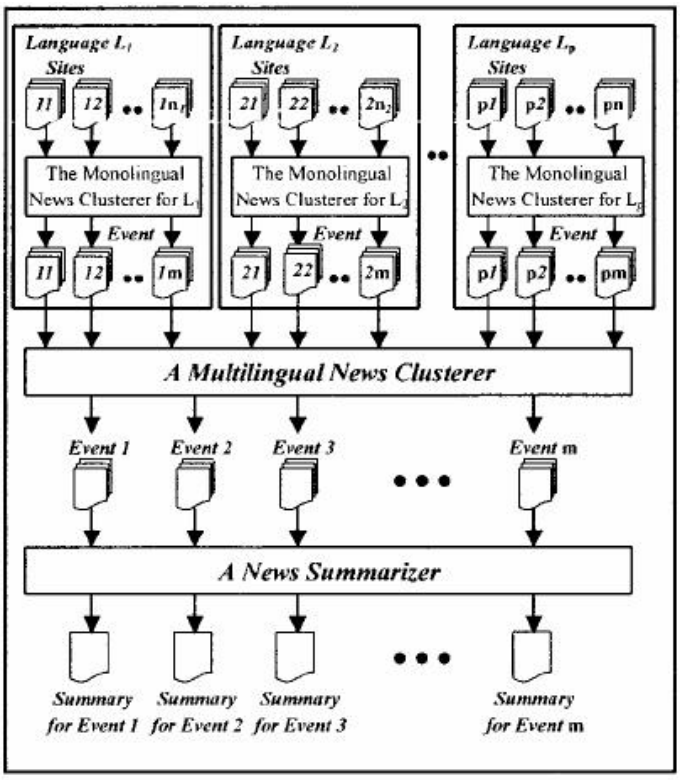

Figure 13: 多語言多文件摘要系統架構

2. 相關研究

多語言多文件摘要的研究，著重於如何解決不同語言間型態及結構的差異性，並決定文件的相似度，以偵測出不同文件中相異及相同的部份，進而根據這些資訊決定摘要內容。[3], [4], [28]嘗試由單字、語句與文件三個不同角度的分析比較，去除重複的資訊，並降低語言的障礙，而達到多語文件分析統整的目的。Figure 13 為其提出之多語多文件自動摘要系統架構。

此架構中，第一步驟乃是分別針對各個不同的語言計算語句的相似度，並利用分群的方法以達到從不同語言文件中抽取事件群(Event)的目的。一個事件群，便代表圍繞相同人、地、時、物所發生的事件。接著，多語新聞分群器(Multilingual

News Cluster)以各個語言文件所分析得到的事件群當輸入，將語意相同的事件群連結，以達到多語架構下事件群的分群。

[4], [28]中對於中英文雙語事件群的連結有深入的探討。Figure 14 為其中英文混合語句分群架構。其基本想法為中英文語句各自分群後，再利用群中中英文語句的對應，以建立群間連結關係。

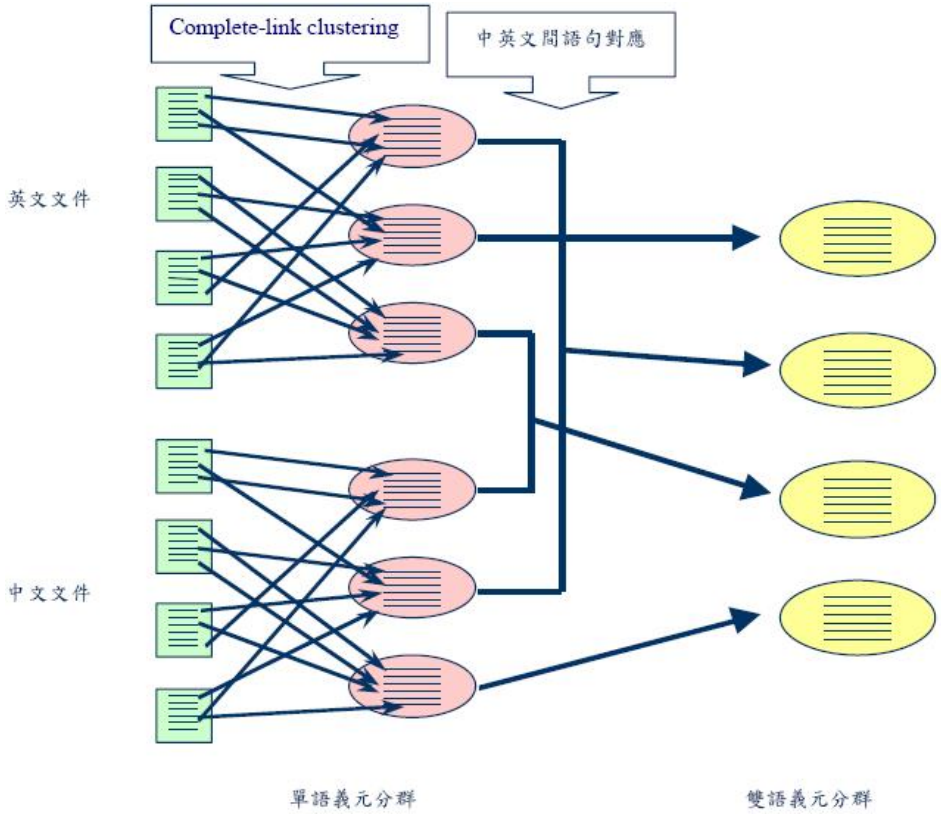

Figure 14: 中英文混合語句分群架構

他們提出五種中英文語句對應的策略：

■ 策略一：完全不考慮位置及字詞歧義性的問題

對於存在於英文語句與中文語句中所有名詞及動詞，利用中英字典翻譯，並參考字詞於同義詞詞林[27]及 WordNet[17]中的關係，將所有相似的字詞對個數加總即為相似度。

■ 策略二：採取先佔原則，即每個字詞只能產生一個相似連結

針對每個詞，考慮其對應到其他語言的詞間相似個數。如 Figure 15 所示，C_2 的比對對象為所有 E_1 至 E_n。

■ 策略三：採取先佔原則，並考慮位置關係

同策略二，然比對的範圍縮為一個設定的 Window 內。如 Figure 16 所示，C2 比對的對象為 E_1 至 E_3。

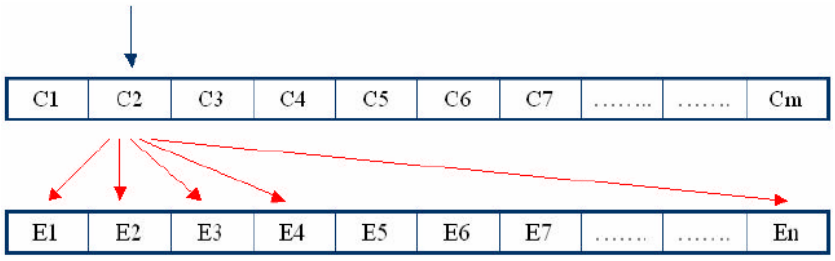

Figure 15: 策略二單一字詞相似比對示意圖

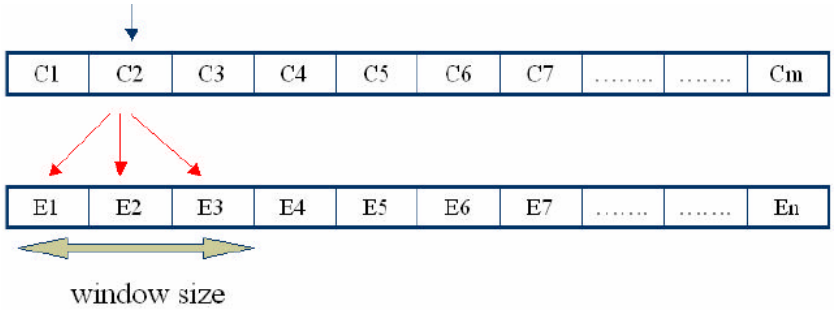

Figure 16: 策略三單一字詞相似比對示意圖

■ 策略四：以沒有歧義性的詞優先產生連結，並決定鄰近詞的位置關係

優先以翻譯沒有歧義性的詞產生連結，如 Figure 17 中，C_2 與 E_5 為沒有歧義詞。針對相鄰詞 C_3，則考慮 E_3, E_4, E_6 與 E_7。

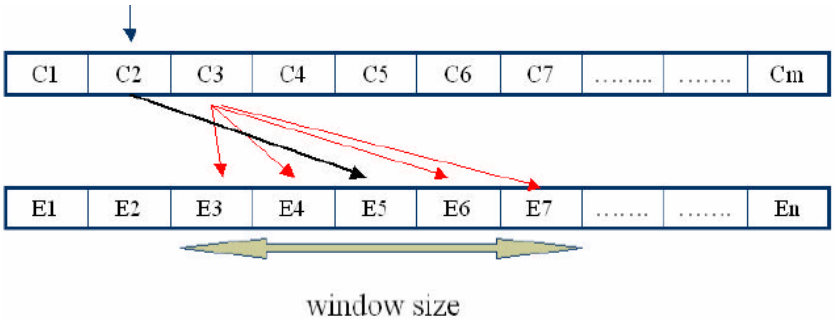

Figure 17: 策略四單一字詞相似比對示意圖

■ 策略五：沒有歧義性的詞優先產生連結，並決定鄰近詞的區間位置關係

以沒有歧義性的詞優先連結，並以兩兩無歧義詞作為比較區間。如 Figure 18 中，C_2 的比對對象為 E_2 至 E_5。

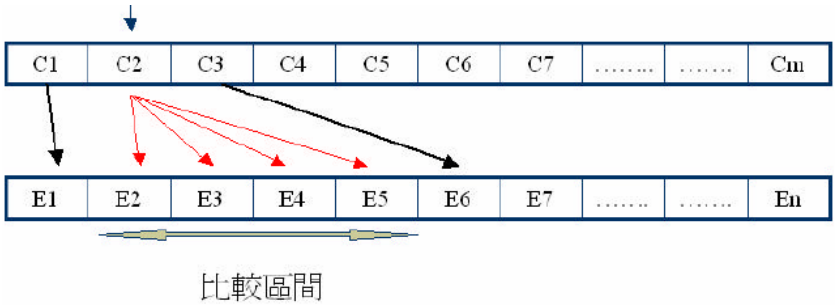

Figure 18: 策略五單一字詞相似比對示意圖

不同於[3], [4], [28]，分別針對不同語言產生事件群，再利用語句對應的方式連結不同語言的事件群；[6]直接利用翻譯軟體將不同的語言先翻譯為英文，在同一語言的架構下，進行事件群的分群。其困難在於解決翻譯時所導入的雜訊及錯誤，對分群時所造成的影響。解決的方法乃是利用 WordNet[17]，透過同義詞、上位詞及下位詞等關係計算語句的相似度。Figure 19 為哥倫比亞大學 Newsblaster 系統所提出的多語多文件摘要架構。

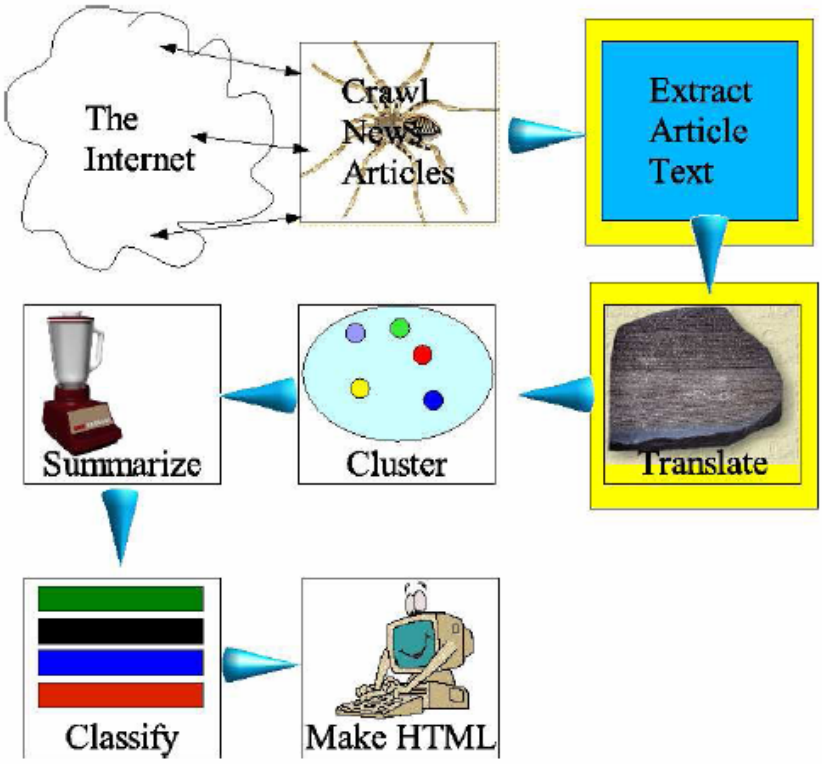

Figure 19: Columbia Newsblaster 多語言摘要系統架構

本計畫中，我們的方法類似於[4], [28]所提及的方法，即直接計算中英文語句間的相似度，並沒有事先將英文翻譯成中文，或將中文翻譯成英文。不同於[4], [28]，我們所提出的方法沒有利用查字典的方式來比對中英文詞，乃是事先利用中英雙語平行語料庫，建立中英文詞群。每個詞群可視為語料庫中所導出之概念，詞群中的中英文詞可視為相關或相對應的翻譯；所有詞群的集合，即為一雙語概念空間(Bilingual Concept Space)。對於測試文件中中英文語句的比對，我們分別將中英文語句對應至雙語概念空間中，並以詞群作為概念空間中的維度，以描述中英文語句為概念空間的向量表示式。透過這樣的轉換，便可以越過語言的隔閡，以達到計算中英語句相似度的目的。

3. 研究方法

多文件自動摘要中最重要的研究議題在於找出不同文件中相異及相同的部份，透過分群的方法，將相似度高的語句聚集在一起。同一語句群中，所涵蓋的資訊可視為相同；該語句群並可視為多文件中所提及的一個重要概念。同樣地，跨語言多文件自動摘要的研究中，重要的議題在於如何計算任兩相同或相異語言所構成語句間相似度，以便跨越語言的隔閡，達到將相似度高的語句聚集的目的。

本章提出利用中英平行語料庫[15](Chinese-English Parallel Corpus)，透過單字/詞分群以找出相關的中英詞群集。同一中英詞群集中，所包含的中文詞或英文詞即可視為相對應的翻譯。中英詞群集的集合，我們稱為該中英平行語料庫所分析出來的概念空間(Concept Space)。換句話說，透過概念群(Concept Cluster)的對應，便可將任一語句對應到概念空間上，並以不同的概念群當成向量空間的維度，進而得到該語句位於概念空間中的向量表示式(Vector Representation)。有了向量表示式，接著便可計算任兩相異語言所構成語句間相似度。

3.1 中英雙語混合式文件自動摘要架構

Figure 20:為我們所提出之中英雙語混合式文件自動摘要架構，共包含五個模組： 1) 前處理(Pre-processing)；2) 階層式概念空間對應(Hierarchical Concept Mapping)；3) 相似矩陣計算(Similarity Matrix Computation)；4) 以主題相關地圖(Text Relationship Map)為基礎之多文件摘要(T.R.M. Multi-Doc Summarization)；5) 概念空間建構(Concept Space Training)。其中，概念空間建構除包含前處理模組外，另有詞分群(Word Clustering)模組。

[15] 平行語料庫爲一中英對照之語料庫，且每個中文語句都有其相對應之英文語句，因此可由字出現於語句中的相對位置，學習該字出現在另一語言的相對應翻譯。

3.2 概念空間建構

概念空間建構利用中英雙語平行語料庫，透過單字/詞分群以找出相關的中英詞群集。同一中英詞群集中，所包含的中文詞或英文詞即可視為相對應的翻譯。

3.2.1 前處理(Pre-processing)

前處理對中英雙語平行語料庫進行斷詞切字的工作，並計算每個中英文詞於相對應的段落[16]中所出現的頻率給予其特徵值，同時建構 Word-By-Paragraph 的矩陣作為詞分群模組的輸入。此步驟中，我們分別利用中央研究院資訊科學所詞庫小組所開發的中文斷詞系統[23]及 LT POS[12]進行中文與英文斷詞及詞性標記工作。對於英文斷詞結果，同時去除停用詞(Stop-Word)並利用 Porter's Stemming[18]還原字根。

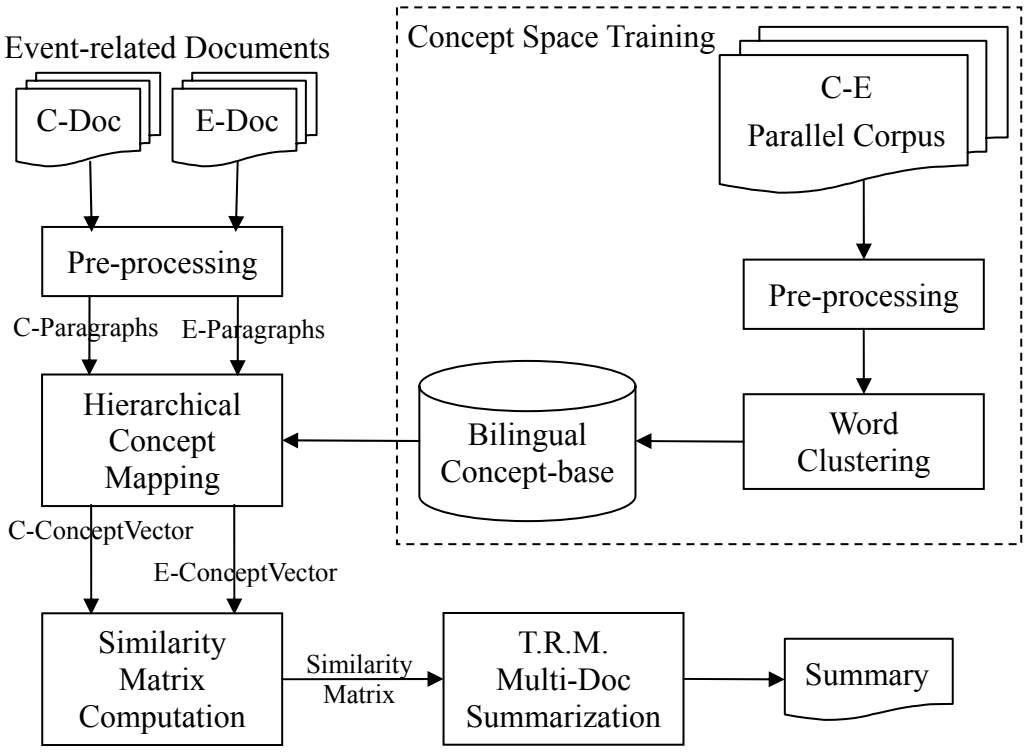

Figure 20: 中英雙語混合式文件自動摘要架構

[16] 本計畫以段落(Paragraph)爲一有意義的處理單位，亦可爲語句(Sentence)或文件(Document)。

■ 關鍵詞選取

大多數語句都是『主詞-述詞-受詞』的結構[26]。其中，主詞(Subject)與受詞(Object)為名詞；述詞則為動詞(Verb)。另外，就關鍵詞的代表性而言，名詞與動詞的重要性比冠詞、副詞或介系詞高出很多。因此，我們只保留名詞及動詞作為建構 Word-by-Text 矩陣的關鍵詞。

■ 特徵值計算方式

假設中英雙語平行語料庫中所有段落的集合為 $P=\{P_j \mid P_j$ 代表 P_C 或 P_E, 其中 P_C 及 P_E 分別為相對應的中英文段落$\}$，$W=\{w_i \mid w_i$ 可為中文詞或英文詞$\}$。則特徵值的計算方式如 Eq. (15)，其中 $w_{i,j}$ 代表 w_i 於 P_j 中的權重值，tf 為 w_i 出現於 P_j 中的頻率，maxtf 代表 P_j 中出現最多次數之 w_k 的頻率。

$$w_{i,j} = 0.5 + 0.5 \times \left(\frac{tf}{\max tf} \right) \qquad \text{Eq. (15)}$$

■ Word-by-Paragraph 矩陣

假設中英雙語平行語料庫中共有 n 個段落，l 個中文詞及 m 個英文詞，則 Word-by-Paragraph 的矩陣如 Figure 21。其中，P_j 代表相對應的 P_C 或 P_E，c_i 列代表中文詞 w_i 出現於所有 P_C 中的權重，e_i 列代表英文詞 w_i 出現於 P_E 中的權重。

	P_1	⋯	⋯	P_n
c_1	$w_{1,1}$	⋯	⋯	$w_{1,n}$
⋮	⋮	⋯	⋯	⋮
c_l	$w_{l,1}$	⋯	⋯	$w_{l,n}$
e_1	$w_{l+1,1}$	⋯	⋯	$w_{l+1,n}$
⋮	⋮	⋯	⋯	⋮
e_m	$w_{l+m,1}$	⋯	⋯	$w_{l+m,n}$

Figure 21: 中英詞混合之 Word-by-Paragraph 矩陣

3.2.2 詞分群(Word Clustering)

詞分群的基本假設為當兩個中英詞出現於同一相對應段落中的頻率越相近，則代表他們於中英雙語語料庫中所表現的意義越相像。基於此假設，概念空間建構利用中英雙語平行語料庫，透過單字/詞分群以找出相關的中英詞群集。同一中英詞群集中，所包含的中文詞或英文詞即可視為相對應的翻譯。

我們採用 CLUTO[10]分群工具，利用 bisecting *K*-means 演算法[21]進行詞分群；同時，我們將分群的結果分割成數個階層。越上層表示所選取的 *K* 值越小，代表每個群所包含的中英詞越多，其表現的概念涵蓋越廣，中英詞間的關係可視為概念層面(Concept-Level)相關；相反地，越下曾表示所選取的 *K* 值越大，代表每個群所包含的中英詞越少，同一群集中的中英詞越有可能互為翻譯。

Figure 22: (a) 說明詞分群利用中英詞出現頻率的特性，達到中英詞混合的分群結果。其中，e_1 與 c_2 可視為相關或相對應的中英詞翻譯。Figure 22: (b)說明利用詞分群所得到的概念空間，可將段落對應至以 C_1 及 C_2 兩概念所描述的向量表示式。

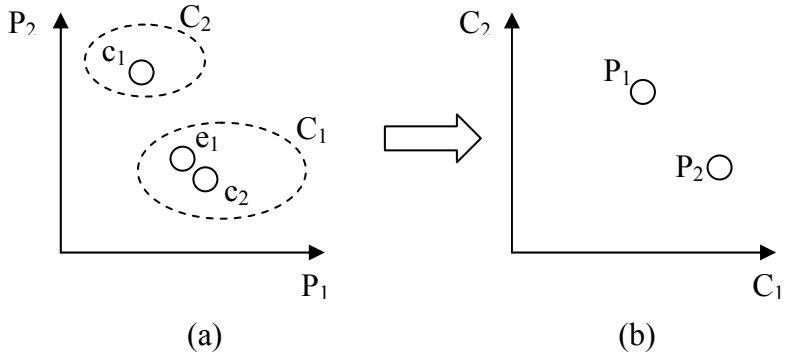

Figure 22: (a) 詞分群概念空間建構; (b) 段落對應於概念空間示意圖

此外，對於出現於某個詞群中所有中英詞 w_j，我們分別 Eq. (16)與 Eq. (17)計算其內在 z-score (Internal z-score)及外在 z-score (external z-score)。其中，s_j^I 為 w_j 與其他所有位於同一群的詞間的平均相似度，u_j^I 為所有位於同一群的詞之 s_j^I 平均值，σ_j^I 為其標準差；s_j^E 為 w_j 與其他所有位於不同群的詞間的平均相似度，u_j^E 為所有位於同一群的詞之 s_j^E 平均值，σ_j^E 為其標準差。

$$I\text{-}zscore(w_j) = (s_j^I - u_l^I)/\sigma_l^I \quad \text{Eq. (16)}$$

$$E\text{-}zscore(w_j) = (s_j^E - u_l^E)/\sigma_l^E \quad \text{Eq. (17)}$$

對 w_j 而言，當 $I\text{-}zscore(w_j)$的值越大及 $E\text{-}zscore(w_j)$越小時，則表示 w_j 越能代表該群的特性。根據此特性，我們設計權重公式 Eq. (18)，給予位於某個詞群 C 中的詞 w_i 不同的權重，以代表其位於該群的重要性。其中，min($I\text{-}zscore(w)$)為該群集中最小的 I-zscore 值，min($E\text{-}zscore(w)$)為該群集中最小的 E-zscore 值。1+max(min($I\text{-}zscore(w)$), min($E\text{-}zscore(w)$))乃是為了確保 $I\text{-}zscore(w_j)$及 $E\text{-}zscore(w_j)$的值大於 0。

$$w_{i,c} = \frac{1 + I - zscore(w_i) + \max(\min(I - zscore(w)), \min(E - zscore(w)))}{1 + E - zscore(w_i) + \max(\min(I - zscore(w)), \min(E - zscore(w)))} \quad \text{Eq. (18)}$$

有了 $w_{i,c}$，則每個群 C_j 利用 Eq. (19)表示。

$$C_j = < w_{1,c_j}, ..., w_{l,c_j}, w_{l+1,c_j}, ..., w_{l+m,c_j} > \quad \text{Eq. (19)}$$

3.3 中英雙語混合式文件摘要

利用概念空間建構模組所得到的階層式概念群(Hierarchical Concept Cluster)，可將中英文文件中任一段落或語句對應到概念空間上，並以不同概念群當成向量空間的維度，進而得到該段落位於概念空間的向量表示式。有了向量表示式，便可以跨越語言的隔閡，計算任兩相同或相異語言所構成的段落間相似度，以達到將多文件摘要中將相似度高的段落聚集的目的。

3.3.1 前處理(Pre-processing)

此步驟與概念空間建構模組中前處理步驟大同小異。乃是分別對中英文件中的段落進行斷詞切字的工作，並計算每個中英文詞於相對應的段落[17]中所出現的頻率給予其特徵值。同樣地，關鍵詞的選取僅保留名詞與動詞；特徵值的計算亦採用 Eq. (15)的計算公式。則任一中文段落 P_C 可描述為 Eq. (20)，任一英文段落 P_E 可描述為 Eq. (21)。

[17] 本計畫以段落(Paragraph)爲一有意義的處理單位，亦可爲語句(Sentence)或文件(Document)。

$$P_C =< w_{1,P_C},...,w_{l,P_C},0,...,0 >$$ Eq. (20)

$$P_E =< 0,...,0,w_{1,P_E},...,w_{m,P_E} >$$ Eq. (21)

3.3.2 階層式概念空間對應(Hierarchical Concept Mapping)

概念空間建立模組中，我們分析不同層級的分群結果，並且建立階層式詞分群。針對不同階層，可以把某個段落 P (P_C 或 P_E)對應到該階層之概念空間中。

依照 Eq. (20)或 Eq. (21)，$P =< w_{1,P},...,w_{l,P},w_{l+1,P},...,w_{l+m,P} >$。假設階層 h 之概念空間共有 k 個概念群，分別為 C_1, …, C_k，根據 Eq. (19)，C_j 可表示為 $C_j =< w_{1,c_j},...,w_{l,c_j},w_{l+1,c_j},...,w_{l+m,c_j} >$，則我們定義 P 對應至 C_j 的權重為 Eq. (22)。

$$w_{P,C_j} = \max_i (w_{i,P} \times w_{i,C_j})$$ Eq. (22)

則 P 對應至階層 h 之概念空間表示式為 Eq. (23)。

$$P =< w_{P,C_1},...,w_{P,C_k} >$$ Eq. (23)

計算不同階層的相似度考量，乃是因為當考慮越低階層時，即分群數目越多時，所包含的概念群中雖然中英文的對應卻正確，但是當段落對應到該概念空間時，因為概念空間維度被過度細分，可能導致原本有關係的兩段落，其概念相似度變得越低。相反地，當考慮越高階層時，即分群數目越少時，其概念所涵蓋的詞較多較廣，當段落對應到該概念空間時，可能導致原本沒有關係的兩段落，反而變得有關係。為了解決上述問題，我們計算不同階層的相似度，並用線性組合方式求得相似度。當階層越低時，我們給予較低的權重；相對地，階層較高時，給予較高的權重。

3.3.3 相似矩陣計算(Similarity Matrix Computation)

概念空間建立模組中，我們分析不同層級的分群結果，並且建立階層式詞分群。針對不同階層，可以把 P_i 或 P_j 對應到概念空間中，透過 Eq. (23)，可得到 P_i 或 P_j 對應於階層 h 的概念空間表示式，並計算位於不同階層中的概念空間相似度，最後利用線性組合(Linear Combination)的方式，可得到任兩段落 P_C 與 P_C

或 P_C 與 P_E 或 P_E 與 P_E 的相似度。示意圖如 Figure 23，針對兩兩段落，其相似度為 Eq. (24)。有了相似度的計算公式，便可以針對任兩段落，建立相似度矩陣，以提供多文件摘要模組使用。

3.3.4 以主題相關地圖為基礎之多文件摘要(T.R.M. Multi-Doc Summarization)

多文件摘要模組，我們採用第一年計畫中所提之以主題相關地圖為基礎之多文件摘要方法(T.R.M. Multi-Doc Summarization)；不同之處為第一年計畫中，我們使用 LSA(Latent Semantic Analysis)作為概念空間的建構模組，然而 LSA 的計算複雜度太高，對於較大的訓練文件集並不適用，因此我們改採用詞分群的方式，以導出中英雙語平行語料庫中的概念空間。

$$sim(P_i, P_j) = \sum_l w_{h_l} \cdot sim_{h_l}(P_i, P_j) \qquad \text{Eq. (24)}$$

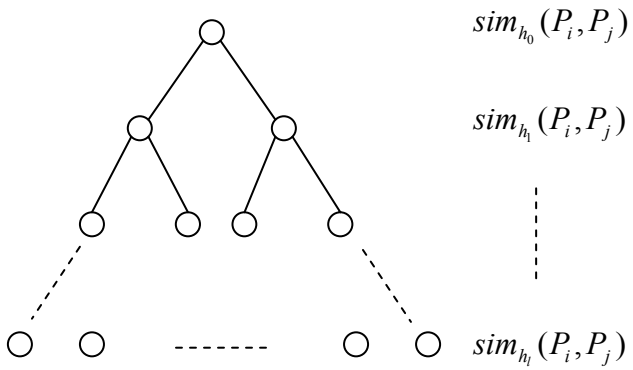

Figure 23: 計算不同階層概念空間相似度示意圖

主題相關地圖乃由[20]所提出，適用於單文件摘要。地圖上每個節點為一段落或語句，兩兩節點間的相似度如果大於某個臨界值，則存在有相連的關係。計算每個節點的連結數，當其連結數越大時，表示其節點與越多節點相關，可視為討論相關的主題，因此越重要。然而，應用於多文件摘要上，必須解決重複性(Redundancy)的問題。去除重複性的機制，我們採用第一年計畫中 Model 1，僅考慮主題相關地圖上連結數目，如 Eq. (25)。

$$PS \overset{def}{=} Arg \max_{P_i \in R \setminus S} [\lambda SIG(P_i) - (1-\lambda) \max_{P_j \in S} REL(P_i, P_j)] \qquad \text{Eq. (25)}$$

假設 S 為已被選到當摘要之段落集合，SIG(Pi)代表 Pi 的連結數，REL(Pi, Pj)為 Pi 與 Pj 關聯強度，其中 Pj 屬於 S 集合；當 Pi 與 Pj 於地圖上有連結時，便需要付出代價α。調整α，可去除不同層級的重複。

4. 結果與討論

4.1 概念空間建構

4.1.1 中英雙語平行語料庫

我們收集民視英語新聞[25]之中英文對照新聞，涵蓋範圍從 2003 年 5 月至 2005 年 3 月，共 6,821 組中英文對照的段落。斷詞切字的結果，中文部份共 16,506 個相異名詞及 10,950 個相異動詞；英文部份共 10,687 個相異名詞及 2,767 個相異動詞。

4.1.2 中英文混合詞分群結果

Table 8 舉例說明某一 General 階層的分群的結果。此表格中，中英文對應的部份乃為人工對應的結果，就原本的詞群而言，並沒有該對應關係。由此表格可知，當所選取的階層越高時，亦及所分群的群數目越少時，該詞群所涵蓋的概念越廣，中文詞與英文詞間雖然有關係，但是並不能明確地知道其相關的對應翻譯。

Table 8: General 詞群結果舉例

中文詞	英文詞
仙妮雅唐恩	shania
仙妮雅唐恩	twain
席琳狄翁	celine
席琳狄翁	dion
密西艾略特	missy
密西艾略特	elliot
傑西辛普森	–
鄉村	–
實力派	–
嘻哈教母	–
戴安基頓	dian
戴安基頓	keaton

Table 9: Specific 詞群結果舉例

詞群	中文詞	英文詞
0	克斯勒	quastler
1	腰身	curvacious
2	槍聲	bang
3	鄭明修	ming-hsiu
4	薛樂儀	le-yi
5	陳玉鳳	yu-feng

Table 9 列出某一 Specific 階層的分群中 6 個詞群結果。由此表可知，當所選取的概念階層越低時，亦及所分群的群數目越多時，該詞群所涵蓋的概念越細，中文詞與英文詞間已經可以視為相對應的翻譯。然而，詞群中亦有可能包含少數的雜訊；如詞群 5，該群共包含三個詞，分別為『陳玉鳳』、『餐飲店』及『yu-feng』。原文中，陳玉鳳為該餐飲店的老闆。

4.2 中英文段落對應分析

4.2.1 測試集介紹

我們從中央通訊社[24]收集 2005/5/8 至 2005/5/13 一周內的所有中英文新聞，先分別分對中英文文件作文件分群，在由人工將中英文文件群對應，以得到討論相同事件的中英文新聞，共得到 10 個文件群。Table 10 列出 10 個文件群所包含的事件及相關中英文文件數目。

Table 10: 測試集文件群分析

	事件	中文文件數	英文文件數
1	國代選舉民進黨獲 127 席 贊成修憲陣營大勝	6	4
2	謝揆：政府釋利多　打造良好經濟環境	6	3
3	參與 WHO 外交部：不接受矮化台灣地位安排	6	5
4	國代選舉蘇貞昌籲投民進黨完成國會席次減半	8	5
5	國代選舉 民進黨對選情審慎樂觀	8	4
6	宋楚瑜抵北京 重申兩岸兄弟一家親	8	7
7	兩岸一中 宋楚瑜：兩岸各表一中與憲法一中	7	6
8	修憲辯論 正方：關鍵性修憲 讓憲法更嚴謹	7	5
9	民進黨團提選罷法修正案　加重賄選刑責	4	5
10	學者：宋胡會共識 壓縮政府談判空間	8	7

Table 11 為測試集中相異詞數目及其於中英雙語平行語料庫中涵蓋的比例。平均而言，中文詞及英文詞分別約 74%及 86%涵蓋於平行語料庫中。

Table 11: 測試集中相異詞數目及其於平行語料庫中的涵蓋比例

語言及詞性	測試集中相異詞數目	涵蓋於平行語料庫數目	百分比
中文名詞	3,408	2,452	72.0%
中文動詞	2,870	2,281	79.5%
英文名詞	1,993	1,614	81.0%
英文動詞	893	811	91.0%

4.2.2 段落對應結果

Table 12 列出各個事件群中，考慮相似度最高的前 10 組中英對應段落，以人工的方式標記是否相關的結果。該表格顯示平均約 57%的中英對應段落為相關的對應(亦即，平均的 Precision 為 57%)[18]。同時，我們發現當事件群中中英文件的關係越相關時，如事件群 5 與 6，其對應的結果越正確；當事件群中中英文件的關係越不相關時，如事件群 3，其正確對應的數目便越低。

Table 12: Top10 相似度高之中英段落中正確對應數目

事件群	相關中英對應段落數目
1	6
2	5
3	3
4	6
5	8
6	8
7	6
8	5
9	4
10	6
平均	5.7

[18] 初始實驗結果因測試集雜亂導致結果不佳。我們縮減測試集，使得中英文文章的數量相近得到比較好的結果。

Table 13、Table 14 及 Table 15 列出三組正確的中英對應段落範例以供參考。

Table 13: 中英對照例一

中文	他強調，堅持體驗一個中國的九二共識，堅持反對台獨，是兩岸對話、協商的政治基礎，也是兩岸關係和平穩定發展的政治基礎。
英文	Hu stressed that insisting on the "one China" policy, the "1992 consensus, " and opposing Taiwan independence would be the premise on which the resumption of cross-strait dialogue and negotiations would be based on.

Table 14: 中英對照例二

中文	七個反對修憲案的政黨或聯盟為：台灣團結聯盟、親民黨、無黨團結聯盟、建國黨、新黨、王廷興等二十人聯盟、張亞中等一百五十人聯盟；反對陣營得票率百分之十六點八六，共獲五十一席。
英文	Seven other parties and groups that oppose the proposed amendments won 51 seats altogether and are not expected to be able to stop the Constitution-amending juggernaut pushed by the two major parties.

Table 15: 中英對照例三

中文	陳總統昨天表示，中國利用在野黨，介入干涉台灣五一四選舉；向美方施壓指台灣憲改是法理台獨，要在野黨、美方阻擋台灣憲改。總統力陳五一四選舉重要性 A 憲改是台灣民主深化鞏固工程。
英文	According to the president, the Chinese official has requested that the United States and Taiwan's opposition parties try to stop the constitutional amendments from being adopted since Beijing considers that the constitutional reform process is aimed at achieving Taiwan independence.

4.3 中英混合式多文件摘要結果

中英混合式多文件摘要的生成，我們提供兩種摘要表現方式。一為以中文為主並附加英文於中文摘要後；另外一種則為相反。目前中英混合式多文件摘要的評估方式，我們以人工問卷的方式，由每個測試者閱讀每個事件群及摘要內容，並評比該摘要內容的好壞。評比的維度，包含 1) 摘要內容的資訊量涵蓋程度；2) 摘要內容的可讀性。

實驗設計共有 5 位專家，針對上述維度對每個事件群所產生的摘要內容進行評比，給予不同的分數。分數的範圍為 1~10，1 代表最差，5 代表普通，10 代表最好。Table 16 為人工評估的結果。平均而言，資訊量涵蓋度為 7.06，可讀性為 6.04。

Table 16: 摘要資訊量涵蓋度及可讀性評估

事件群	資訊量涵蓋度	可讀性
1	7.2	5.5
2	6.9	6.2
3	5.5	5.4
4	6.5	5.8
5	8.2	6.8
6	7.7	7.0
7	7.3	6.5
8	7.0	5.8
9	7.5	5.6
10	6.8	5.8
平均	7.06	6.04

Table 17 僅列出事件群 6 由系統所產生的摘要結果以供參考。

Table 17: 事件群 6 之摘要內容範例

他說，所謂搭橋，是為兩岸搭起互信之橋、合作之橋與溝通之橋。宋楚瑜也不是任何人的信差。 (pno: 3; article: 6; pdate: 2005-05-25) 親民黨主席宋楚瑜今天抵達北京，他感謝中共中央總書記胡錦濤的邀請，讓親民黨打破五十多年來兩岸政治禁忌，和中國共產黨進行黨與黨對話。他說，親民黨相信 只要兩岸兄弟一家親，一定可以找到方法，解決兩岸過去的誤解。 (pno: 1; article: 78; pdate: 2005-05-25) 宋楚瑜的專機約在下午四時三十分抵達北京，中國國台辦主任陳雲林等人到場迎接。 (pno: 2; article: 78; pdate: 2005-05-25) 他說，以親民黨主席身分和親民黨和平工作團身分到北京，感謝中共黨中央與總書記胡錦濤的邀請，讓親民黨打破五十多年來兩岸政治禁忌，親民黨能和中國曳?珩 i 行黨與黨的對話。 (pno: 4; article: 78; pdate: 2005-05-25) 親民黨主席宋楚瑜今天與中國共產黨總書記胡錦濤會面時表示，親民黨三點基本立場堅定不移，堅定支持九二共識「一個中國」基本原則、從不認為台獨應是台灣選項、以及主張和平。 (pno: 1; article: 145; pdate: 2005-05-25) 胡錦濤說，國親兩黨主席連戰、宋楚瑜來訪，大陸同胞和台灣同胞都給予支持與歡迎，表明兩岸同胞認為這些做法符合他們的心願。 (pno: 9; article: 146; pdate: 2005-05-25) Soong, who arrived in Shanghai Friday on the third leg of his current nine-day visit to China, said when he meet with Chinese President Hu Jintao in Beijing May 12, he will urge China to adopt concrete measures to protect "taishang's" rights and interests. (pno: 2; article: 4; pdate: 2005-05-25) Soong flew from Hunan Province to Beijing on the fifth and most important leg of his nine-day "bridge-building" visit to China, where he will hold talks with Chinese President Hu Jintao, who serves concurrently as general secretary of the Communist Party of China. (pno: 2; article: 46; pdate: 2005-05-25)

5. 結論

本計畫為三年期研究『多語言複合式文件自動摘要之研究』之第二年計畫。透過詞分群的方式，我們將中英雙語平行語料庫中的中英文詞進行分群分析，並建構階層式概念空間。對於測試的文件集，我們以段落為單位，將每個段落由關鍵詞的表示式(Word-Level Representation)轉換成概念表示式(Concept-Level Representation)，並組合不同階層概念空間所得到的相似度，以計算任兩中中、中英及英英段落的概念相似度，最後得到多文件摘要結果。

由實驗結果中可知當所選取的階層越高時，亦及所分群的群數目越少時，詞群所涵蓋的概念越廣，中文詞與英文詞間雖然有關係，但是並不能明確地知道其相關的對應翻譯。當所選取的概念階層越低時，亦及所分群的群數目越多時，詞群所涵蓋的概念越細，中文詞與英文詞間已經可以視為相對應的翻譯。

中英文段落對應方面，目前實驗結果 Top 10 的平均 Precision 為 57%。我們亦發現如 Table 13、Table 14 及 Table 15 所示之結果確實驗證我們方法的可行性。就摘要結果的好壞評估，我們以人工問卷的方式，由每個測試者閱讀每個事件群及摘要內容，並評比該摘要內容的好壞。評比的維度，包含 1) 摘要內容的資訊量涵蓋程度；2) 摘要內容的可讀性。實驗設計共有 5 位專家，針對上述維度對每個事件群所產生的摘要內容進行評比，給予不同的分數。分數的範圍為 1~10，1 代表最差，5 代表普通，10 代表最好。平均而言，資訊量涵蓋度為 7.06，可讀性為 6.04。

參考文獻

[1] Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. In *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain (pp. 10-17).

[2] Boros, E., Kantor, P. B., & Neu, D. J. (2001). A clustering based approach to create multi-document summaries. In *Proceedings of the Document Understanding Conference* (*DUC-2001*), New Orleans, LSA, USA.

[3] H.-H. Chen, J.-J. Kuo, S.-J. Huang, C.-J. Lin and H.-C. Wung, "A Summarization System for Chinese News from Multiple Sources," *Journal of the American Society for Information Science and Technology*, 54(13), 1224-1236, 2003.

[4] H.-H. Chen and C.-J. Lin, "A Multilingual News Summarizer," *Proceedings of the 18th International Conference on Computational Linguistics*, pp. 159-165.

[5] H. P. Edmundson, "New Methods in Automatic Extracting," *Journal of ACM* (*JACM*), 16(2), 264-285, 1969.

[6] D. Evans, J. L. Klavans, K. R. McKeown, "Columbia Newsblaster: Multilingual News Summarization on the Web," *Proceedings of Human Language Technology* (*HLT*), Boston, MA, 2004.

[7] Goldstein, J., Mittal, V., Carbonell, J., & Kantrowitz, M. (2000). Multi-document summarization by sentence extraction. In *Proceedings the ANLP/NAACL Workshop on Automatic Summarization*, Seattle, WA (pp. 40-48).

[8] Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'01*), New Orleans, LA, USA (pp. 19-25).

[9] Hovy, E., & Lin, C. Y. (1999). Automated text summarization in SUMMARIST. Mani, I., & Maybury, M. (eds.), *Advances in automated text summarization*. Cambridge, Mass.: MIT Press.

[10] G. Karypis, "CLUTO: Software Package for Clustering High-Dimensional Datasets," http://www-users.cs.umn.edu/~karypis/cluto/index.html.

[11] Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. In *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (*SIGIR'95*), Seattle, WA, USA (pp. 68-73).

[12] Language Technology Group, "LT POS," http://www.ltg.ed.ac.uk/software/pos/.

[13] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, 2(2), 159-165, 1958.

[14] Mani, I., & Bloedorn, E. (1999). Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1-2), 35-67.

[15] I. Mani and M. Maybury (eds.), "Advances in automated text summarization," *MIT Press*, Cambridge, Mass, 1999.

[16] McKeown, K. R., Klavans, J. L., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence* (*AAAI'99*), Orlando, FA, USA (pp. 453-460).

[17] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: a on-line lexical database. *Lexicography*, 3(4), 235-312.

[18] M. F. Porter, "An Algorithm for Suffix Stripping," *Program*, 14(3), 130-137, 1980.

[19] G. Salton and M. J. McGill, "Introduction to Modern Information Retrieval," *McGraw-Hill*, 1983.

[20] G. Salton, A. Singhal, M. Mitra, and C. Buckley, "Automatic Text Structuring and Summarization," *Information Processing & Management*, 33(2), 193-207, 1997.

[21] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," *Proceedings of the KDD Workshop on Text Mining*, 2000.

[22] Young, S. R., & Hayes, P. J. (1985). Automatic classification and summarization of banking telexes. *Proceedings of the 2nd Conference on Artificial Intelligence Applications* (pp. 402-408).

[23] 中央研究院資訊科學所詞庫小組，"中文斷詞系統," http://ckipsvr.iis.sinica.edu.tw/.

[24] 中央通訊社, http://www.cna.com.tw/.

[25] 民視英語新聞, http://englishnews.ftv.com.tw/index.asp.

[26] 陳光華, "新資訊時代的啟發性資訊服務," 21 世紀資訊科學與技術的展望學術研討會, 桃園, 1998.

[27] 梅家駒等, "同義詞詞林," *上海辭書出版社*, 上海, 1982.

[28] 蘇哲君, "中英雙語多文件自動摘要系統研究," *碩士論文*, 國立台灣大學資訊工程研究所, 台北, 台灣, 2001.

多語言複合式文件自動摘要之研究 (3/3)

計畫類別: 個別型計畫

計畫編號: NSC-92-2213-E-009-126-

執行期間: 2005/08/01 – 2006/07/31

計畫主持人: 楊維邦 教授

計畫參與人員: 柯皓仁 教授, 葉鎮源, 林昕潔, 張家寧

ABSTRACT

Much research for question-answering aims to answer factoid, definitional and biographical questions. In most cases, the answers are given as a name, date, quantity, and so on. In this paper, we try to merge techniques of multidocument summarization and question-answering to generate a brief, well-organized fluent summary to provide more relevant information for the purpose of answering real-world complicated questions. The problem is addressed as a query-biased sentence retrieval task. Formally, query-focused multidocument summarization is to synthesize from a set of topic-related documents a brief, well-organized, fluent summary for the purpose of answering a need for information that cannot be met by just stating a name, date, quantity, etc.

In this report, we propose a hybrid relevance analysis to evaluate sentence relevance to the query. This is achieved by combining similarities computed from the vector space model and the latent semantic analysis. Surface features are also examined to know the impacts of low-level features for query-focused multidocument summarization. In other words, the summary is created by including sentences with the topmost significances which are measured in terms of sentence relevance and surface feature salience. In addition, a modified Maximal Marginal Relevance is proposed to reduce redundancy by taking into account sentence shallow feature scores. The experimental results showed the proposed method obtained competitive results when evaluated with the DUC 2005 corpus.

Key-Words: - Query-focused summarization; Hybrid relevance analysis; Sentence feature salience; Latent semantic analysis; Modified Maximal marginal relevance;

1. Introduction

Automated text summarization has been in existence since the 1950's. By definition, it is the process of distilling the most important information from a source (or sources) to produce an abridged version for a particular user (or users) and task (or tasks) [18]. The technology can potentially ease the burden of information overload because only summaries need to be read or analyzed. For example, a search engine could provide summaries (e.g., snippets of texts) to help users spot desired documents in a short time.

Over the past few years, text summarization has drawn tremendous interest from both the natural language processing and the information retrieval communities. DUC (Document Understanding Conference) [10] is one of the active forums for large-scale evaluations of summarization systems. Since 2001, DUC has held several evaluations, including *single-document summarization*, *multidocument summarization*, *cross-lingual summarization* and *query-focused summarization*. Query-focused multidocument summarization was first formally proposed and evaluated at DUC 2005. The task is, given a set of topic-related documents (i.e., a topic cluster), a query topic, and a user profile, to generate a brief, well-organized fluent summary for the purpose of answering users' information needs. The query topic, illustrated in Table 18, models the real-world complex question-answering proposed in [2]. A user profile with a value of "general" or "specific" specifies the granularity required for the output summary. In essence, query-focused multidocument summarization integrates techniques of multidocument summarization and question-answering. From the viewpoint of multidocument summarization, it needs to generate not only a theme-relevant but also a query-relevant summary; while from the perspective of question-answering, the output summary can not just state a name, date, quantity, etc., which makes it more challenging.

In this report, we propose an approach to address query-focused summarization. The method considers as the task a query-biased sentence retrieval task. In other words, only relevant sentences are included in the summary. It measures the relevance of a sentence to the query using a novel hybrid relevance analysis which linearly combines relevance measures from the vector space model and latent semantic analysis (LSA). The output summary is generated according to sentence salience which is estimated in terms of sentence relevance and low-level feature significance.

Furthermore, a modified redundancy reduction module based on Maximal Marginal Relevance (MMR) [6] is proposed for anti-redundancy.

This report is organized as follows: Section 2 introduces previous related work. In Section 3, an overview of the proposed system is presented. In Section 4, the proposed approach to the query-focused multi-document summarization task is described. Experimental results are given in Section 5. Finally, Section 6 concludes this work.

Table 18: DUC 2005 example queries

Dataset: d332h
Granularity: general
Question: What kinds of non-tax crimes have led to tax evasion prosecutions (failure to file, inaccurate filing), instead of or in addition to prosecution for the non-tax crimes themselves?

Dataset: d313e
Granularity: specific
Question: In what countries are MAGLEV rail systems being proposed? Are the proposals for short or long haul? Is government financing required for construction?

2 Related Work

Much previous research has regarded query-focused summarization as a sentence extraction task which identifies sentences that are relevant to the query topic. Approaches to determining sentence relevance vary greatly. For example, Daumé and Marcu [8] employed a Bayesian language model to estimate sentence relevance for ranking sentences. They found that the Bayesian model consistently works well even when there is significantly less information in the query. Jagadeesh *et al.* [14] combined feature scores obtained from relevance-based language modeling, latent semantic indexing, and special words to determine the relevance of a sentence to the information need. Hachey *et al.* [11] presented a system which measures relevance and redundancy using a latent semantic space constructed over a very large corpus.

Approaches based on the graphic model have also been tried. For example, Mani and Bloedorn [17] used a document graph to formulize relations between sentences inside a document. A spreading activation algorithm is applied to perform a query-biased summarization. Bosma [5] created a graphical representation for a document using the Rhetorical Structure Theory (RST) [19]. Based on the graph model, a graph search algorithm is exploited to identify relevant sentences.

Some approaches use shallow semantic analysis. D'Avanzo and Magnini [7] exploited key-phrase extraction to identify relevant terms and used machine learning to select significant key-phrases. Summaries are generated according to relevance and coverage of keyphrases of a certain topic. Li *et al.* [15] built a query-oriented multidocument summarization system under the framework of MEAD [20] by integrating entity-based, pattern-based, term-based, and semantic-based features. Hovy *et al.* [13] proposed a method on the basis of the extraction of basic elements. A basic element, defined as a head-modifier-relation representation, is regarded as a basic unit to determine the salience of a sentence.

In contrast, some approaches depend on deep discourse analysis. For instance, Schilder *et al.* [22] investigated a tree matching algorithm to obtain a tree similarity of dependency parse trees between a question and candidate answer sentences. Sentences with the highest similarities are extracted as the summary. Ye *et al.* [24] handled query-focused summarization by computing sentence semantic similarity via concept links. In their work, concept links were shown to outperform word co-occurrence since they highlight words that are semantically-related. A modified MMR for anti-redundancy was proposed by introducing semantic similarity into the original MMR [6]. Zhou *et al.* [26] combined lexical chains and document index graphics by adding verbs to chains. Sentences are scored according to the integrated chain structure.

Last but not least, Seki *et al.* [23] proposed a summarizer that focuses on subjectivity analysis. Subjectivity refers to aspects of language description that are formed to express an author's opinions, evaluations, and speculations. The summarizer generates summaries to reflect information needs based on subjectivity clues. Berger and Mittal [3] introduced a statistical model for query-relevant summarization. The parameters were learned with a collection of FAQs using maximum-likelihood estimation. Blair-Goldensohn [4] adapted a system originally designed to answer definitional and biographical questions and enhanced it with sophisticated question parsing, topic term identification, and passage retrieval. In

addition, they experimented with several schemes for including the content of nearby sentences to help determine sentence relevance.

3 System Overview

The overview of the proposed system is shown in Figure 24. The system first evaluates the relevance of each sentence to the query and its sentence significance on the basis of surface features, and then applies sentence selection to generate a summary of approximate 250 words to reflect the information need defined in the query topic and the level of granularity specified in the user profile. There are eight modules in the proposed system:

1) Document Analysis

Several document preprocessing steps are conducted in this module, including sentence boundary detection, tokenization, Part-of-Speech (POS) tagging, stopword removal, word stemming, and named entity extraction. After preprocessing, words tagged as NN (noun), VB (verb), JJ (adjective), or RB (adverb) are regarded as significant unigrams and are used to generate bigrams and trigrams. Bigrams and trigrams occurring in more than three sentences as well as all unigrams are kept to build a vector representation for each sentence using a term weighting scheme proposed by Allan *et al.* [1], as shown in Eq. (26). In Eq. (26), $tf_{t,s}$ is the frequency of a term t occurring in a sentence s, sf_t is the number of sentences in which term t appears, and N is the total number of sentences in the document collection.

$$w(t,s) = \log(tf_{t,s} + 1)\log(\frac{N+1}{0.5 + sf_t}) \quad \text{Eq. (26)}$$

2) *Text Feature Extraction* (refer to Section 4.2)

This extracts surface features which are useful for query-focused summarization. These features are further exploited to measure the significance of a sentence in the sentence scoring module.

3) *Query Analysis*

The same procedure of document analysis is applied to the given query. A query is represented as a vector, except that the term weighting scheme uses Eq. (27).

$$w(t,q) = \log(tf_{t,q} + 1) \quad \text{Eq. (27)}$$

where $tf_{t,q}$ the frequency of term t in the query q.

4) *Summary-Type Detection* (refer to Section 4.3)

This module determines whether the desired summary is specific or general according to the user profile. If a specific summary is expected, the number of named entities in a sentence is regarded as another feature and integrated into the sentence scoring function as well.

5) *Relevance Analysis* (refer to Section 4.1)

Sentences are evaluated to obtain their relevance to the query by measuring their similarity.

6) *Sentence Scoring* (refer to Section 4.3)

Sentence relevance and surface feature salience are combined to estimate the significance of a sentence. Sentences are ranked according to their scores for selection.

7) *Redundancy Reduction* (refer to Section 4.4)

A modified MMR is employed to reduce information redundancy in the summary.

8) *Summary Generation* (refer to Section 4.5)

This module selects salient sentences on the basis of sentence scores and re-orders sentences according to their original time and date of publication to form a summary.

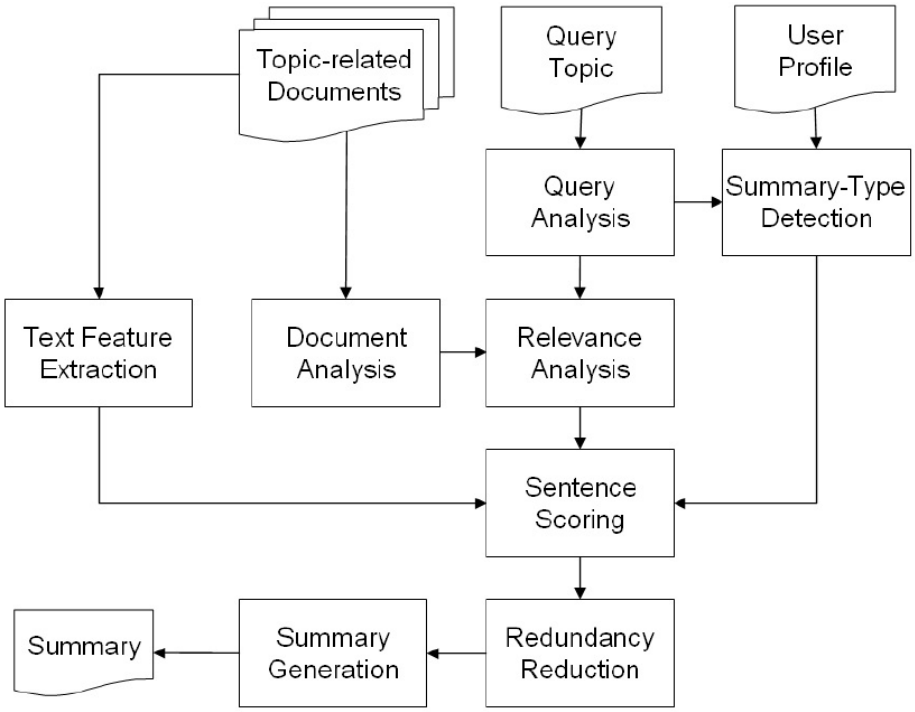

Figure 24: System overview

4 Query-Focused Summarization

This section presents details for modules introduced in the previous section.

4.1 Relevance Analysis

Given a query q, for each sentence s, the degree of relevance between s and q is evaluated based on their similarity, $sim(s, q)$. Three approaches are introduced in this section to compute $sim(s, q)$. The first is the similarity measured in the traditional vector space model (VSM); the second employs latent semantic analysis (LSA) to derive semantic-level similarity; and the third integrates similarities from VSM and LSA in a linear manner.

4.1.1 Similarity Based on VSM: $sim_1(s, q)$

In the vector space model, since s and q are both represented as weighted vectors using Eq. (26) and Eq. (27) respectively, the similarity between s and q is computed as the inner product of the two corresponding vectors. More specifically, the relevance of s given q is defined as Eq. (28). This model has been proven successful for query-biased sentence retrieval [1] and is used in this work as a competitive baseline.

$$sim_1(s,q) = \vec{s} \cdot \vec{q} \qquad \text{Eq. (28)}$$
$$= \sum_{t \in q} \log(tf_{t,q} + 1) \log(tf_{t,s} + 1) \log(\frac{N+1}{0.5 + sf_t})$$

4.1.2 Similarity Based on LSA: $sim_2(s, q)$

In recent years, LSA [9] has been profitably employed in information retrieval to derive inherent semantic structure from a corpus. We employ LSA to measure the semantic relevance of a sentence *s* to the query *q*.

First, a *word-by-sentence* matrix, *A*, is built from all sentences, as presented in Eq. (29). In this matrix, columns represent sentences and rows denote unique words found in the collection. (Note: without loss of generality, *m* is greater than or equals to *n*.) The cell of row *i* and column *j* (i.e., $a_{i,j}$), computed by Eq. (26), signifies the weight of a term t_i in a sentence s_j.

$$A = \begin{array}{c|cccc} & s_1 & s_2 & \cdots & s_n \\ \hline t_1 & a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ t_2 & a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_m & a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{array} \qquad \text{Eq. (29)}$$

Singular Value Decomposition (SVD) is then performed on *A*. The SVD projection of *A* is the product of *U*, *Z* and *V*: $A = UZV^T$ where *U* is an $m \times n$ matrix of left singular vectors, *Z* is an $n \times n$ matrix with a diagonal (σ_1, …, σ_n)[19] and zeros elsewhere, and *V* is an $n \times n$ matrix of right singular vectors. In theory, *U* and *V* are both orthogonal matrices; there exists a property that $U^TU=V^TV=I$, where *I* is the identity matrix; *Z* could be interpreted as a semantic space (or the topic structure) derived from the corpus; *U* and *V* could be viewed as semantic representations of words and sentences in *Z* respectively.

Finally, *Dimension Reduction* is applied to *Z* by keeping only *k* ($k \leq r$) singular values to obtain an approximate Z_k. *A* new matrix, $\widetilde{A}$, which denotes the semantic representation of *A* in Z_k could be obtained by folding *A* into the reduced space Z_k using Eq. (30).

[19] If rank(*A*) = *r*, *Z* satisfies $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$.

$$\widetilde{A} = A^T U_k Z_k^{-1} \quad \text{Eq. (30)}$$

Similarly, a query $q=<t_{q,1}, \ldots, t_{q,m}>$ can be mapped into the same semantic space Z_k with Eq. (31).

$$\widetilde{q} = q U_k Z_k^{-1} \quad \text{Eq. (31)}$$

Thus, the semantic similarity between s and q, i.e., $sim_2(s, q)$, can be obtained by Eq. (32).

$$\begin{aligned} \widetilde{q} \cdot \widetilde{A}^T &= (q U_k Z_k^{-1})(A^T U_k Z_k^{-1})^T \\ &= q U_k (Z_k^{-1})^2 U_k^T A \\ &= < sim_2(s_1, q), \ldots, sim_2(s_n, q) > \end{aligned} \quad \text{Eq. (32)}$$

Conceptually, LSA is the process of discovering relationships among word co-occurrences. Regarding Z_k, the SVD analysis provides information about how words and sentences are related to k latent semantics. Furthermore, in Eq. (32), q times $U_k (Z_k^{-1})^2 U_k^T$ can be viewed as query expansion, where synonymy is essentially defined by similarity of word co-occurrences derived in the semantic space.

4.1.3 Hybrid Relevance Analysis: $sim_3(s, q)$

The hybrid relevance analysis is practically proposed as a linear combination of $sim_1(s,q)$ and $sim_2(s,q)$ to take the advantage of the effectiveness of both approaches. In this combination, the noises introduced from both models could be reduced by model averaging, which is expected to obtain more robust sentence relevance. As a result, the proposed hybrid similarity metric is defined as Eq. (33).

$$sim_3(s,q) = \alpha \cdot sim_1(s,q) + (1-\alpha) \cdot sim_2(s,q) \quad \text{Eq. (33)}$$

4.2 Surface Feature Extraction

In the literature of text summarization, surface features, such as position and tf-idf weighting, have been studied and proven useful for extracting significant sentences (e.g., [16], [25]). In this work, we aim to re-examine these features in order to understand whether they could be salient evidences for sentence scoring in the query-focused summarization task. For a sentence s, we consider five surface features

to obtain its feature score. They are: 1) position; 2) average tf-idf weight of significant words; 3) similarity with title; 4) similarity with document centroid; and 5) similarity with topic centroid.

f_1 – *position*: it is believed that important sentences are usually located in particular positions in a document. Take a news article for instance, the first sentence always introduces the main topic or summarizes the whole story. The position score is defined in Eq. (34), which was proposed by Hirao, *et al.* [12]. In Eq. (34), $|D|$ is the number of words in the document D that contains s and $NC(s)$ is the number of words appearing before s in D. On the basis of this mechanism, the first sentence obtains the highest score and the last has the lowest one.

$$f_1(s) = 1 - \frac{NC(s)}{|D|} \quad \text{Eq. (34)}$$

f_2 – *avg. tf-idf weight of significant words*: generally, terms with higher term frequency and tf-idf values are more important, implying that a sentence with a higher summation of tf-idf values from its constituent words tends to be a substantial sentence. In previous studies, all words in a sentence are taken into account and the total score is averaged over the length of the sentence. In our work, in order to obtain a more precise weight for a sentence s, we only consider significant words in s. The average tf-idf score is computed as shown in Eq. (35), where $w(t, s)$ is the weight in Eq. (26).

$$f_2(s) = \underset{t \in s, t \text{ is significant}}{Average}\ w(t,s) \quad \text{Eq. (35)}$$

A significant word is defined as a keyword t which satisfies the criterion shown in Eq. (36), where u is the mean and σ is the standard deviation of all $w(t, C)$, and $w(t, C)$ is the summation of all *tf-idf* values for t from all sentences in the document collection C.

$$u + 0.5\sigma \le w(t,C) \quad \text{Eq. (36)}$$

f_3 – *similarity with title*: there is no doubt that the title always sums up the main theme of a document. In other words, the more similar a sentence is with the title, the more important it is. This similarity is measured by Eq. (37) where s_{title} is the title sentence.

$$f_3(s) = sim(s, s_{title}) = \frac{\vec{s} \cdot \vec{s}_{title}}{|\vec{s}||\vec{s}_{title}|} \quad \text{Eq. (37)}$$

f_4 – similarity with document centroid: this measures the centrality of a sentence with the whole document. The centrality means the similarity between *s* and the centroid of the document. Generally speaking, if a sentence contains more concepts identical to those of other sentences in the same document, it tends to be more significant. This feature score is obtained with Eq. (38), in which $\vec{D}_{centroid}$ is the average vector representation of all sentences in the document *D*.

$$f_4(s) = sim(s, D_{centroid}) = \frac{\vec{s} \cdot \vec{D}_{centroid}}{|\vec{s}||\vec{D}_{centroid}|} \quad \text{Eq. (38)}$$

f_5 – similarity with topic centroid: similar to f_4, this feature estimates the similarity of a sentence with the centroid of the topic cluster. The score is computed as Eq. (39), where $T_{centroid}$ is the average representation of all sentences in the document collection.

$$f_5(s) = sim(s, T_{centroid}) = \frac{\vec{s} \cdot \vec{T}_{centroid}}{|\vec{s}||\vec{T}_{centroid}|} \quad \text{Eq. (39)}$$

4.3 Sentence Scoring

The sentence score denotes the importance of a sentence and is exploited as a judgment to determine whether a sentence should be included in the output summary. We define the score of a sentence *s* by taking into account: 1) its relevance to the query *q*; and 2) its salience of low-level features. Eq. (40) delineates the scoring function.

$$score(s) = w_{sig} \cdot sig(s) + w_{sim} \cdot sim(s, q) \quad \text{Eq. (40)}$$

where $sig(s) = \sum_i w_{f_i} \cdot f_i(s)$, and $sim(s, q)$ could be any similarity metric proposed in Section 4.1. In *sig*(*s*), $f_i(s)$ is one feature score in Section 4.2, and w_{f_i} is the weight for f_i used for linear combination.

Recall that in Section 3, there is a module called summary-type detection. If the summary is specific, the number of named entities, denoted as $f_{NE}(s)$, is regarded as an additional feature considered in $sig(s)$. The idea is, when a specific summary is needed, information containing named entities, such as person, location, date/time, and so on, becomes more important. Therefore, in this case, $sig(s) = \sum_i w_{f_i} \cdot f_i(s) + w_{NE} \cdot f_{NE}(s)$.

4.4 Redundancy Reduction

Carbonell and Goldstein [6] proposed Maximal Marginal Relevance (MMR) to reduce redundant information in the summary. The main idea is, when including a sentence s in the summary, the summarizer measures MMR for s to satisfy the following criteria: 1) the maximum relevance of s to the query q and 2) the minimum similarity of s to previously selected sentences in the summary. That is, s has high marginal relevance iff it is both relevant to q and contains minimal overlap with previously selected sentences. The MMR is defined in Eq. (41), where R is the ranked list of sentences, S is the set of sentences which are already selected in the summary, SIM_1 is the similarity metric used in relevance ranking, and SIM_2 is the similarity measured by the cosine value of two sentence vectors.

Eq. (41)

$$MMR \overset{def}{\equiv} \arg\max_{s \in R-S} [\lambda \cdot SIM_1(s,q) - (1-\lambda) \cdot \max_{s_i \in S} SIM_2(s,s_i)]$$

One shortcoming of MMR is that it does not consider the sentence representative power (e.g., surface feature salience). In our work, we propose a modified MMR, presented in Eq. (42), to integrate surface feature salience into the original MMR. In Eq. (42), δ and λ are weights to control the impact of $sig(s)$, SIM_1, and SIM_2, $sig(s)$, as mentioned in Section 4.3, is the score obtained from feature salience, SIM_1 denotes the similarity metric proposed in Section 4.1, and SIM_2 simply computes the cosine similarity. In general, a sentence which has a high feature score and is highly relevant to the query but has a lower similarity to sentences in the summary will be ranked in the topmost position.

Eq. (42)

$$ModifiedMMR \overset{def}{\equiv} \arg\max_{s \in R-S} [\delta \cdot sig(s) + \lambda \cdot SIM_1(s,q) - (1-\lambda) \cdot \max_{s_i \in S} SIM_2(s,s_i)]$$

It should be noted that a similar form of the modified MMR was proposed by Ye, *et al.* [24]. However, unlike our work, they view as the sentence representative power the similarity of a sentence to other sentences in the document collection, which is computed via their proposed concept links.

4.5 Summary Generation

In this work, the output summary is generated by extracting salient sentences from documents. There are two ways to rank sentences. One is that sentences are ranked according to their scores computed by Eq. (40) (i.e., no anti-redundancy is considered). The other employed MMR, mentioned in Section 4.4, to avoid redundancy. This is achieved by re-ranking sentences with the new MMR scores.

Once sentences are sorted by their significances, k sentences with the topmost scores are included to form the summary. Those selected sentences are presented in order according to their publication date and time. Furthermore, if the length of the summary is greater than the required length (e.g., about 250 words), the summary is truncated to satisfy the constraint. Please note since we only focus on examining the effectiveness of the proposed hybrid relevance analysis and the impact of shallow feature salience on sentence scoring, sentence ordering is not an issue discussed in this work.

5 Evaluation

In this section, we present our experimental results.

5.1 The DUC 2005 Corpus

The DUC 2005 Corpus, consisting of 50 topics, was created by NIST assessors. Topics were created to explicitly reflect the specific interests of a potential user in a task context and to capture some general user/task preferences in a simple user profile. Each topic has approximate 25-50 documents. A user profile for each topic was specified by the assessors to define the desired granularity of the summary. Then, other NIST assessors were each given the user profile, the DUC topic, and the document cluster and were asked to create a summary that meets the needs expressed in the topic and user profile. These human-created summaries are used for evaluation as reference summaries.

5.2 Evaluation Metric

There were two evaluations conducted at DUC 2005. One automatically measured the consistence between reference summaries and machine-generated summaries using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [21]. The other was a manual evaluation which measured the quality of linguistic well-formedness and the responsiveness of each submitted summary using a set of quality questions. Since we did not participate in DUC 2005, we only report results evaluated by ROUGE.

ROUGE is an automatic evaluation tool for automated text summarization. It measures the number of overlapping units, such as *n-gram*, *word sequences*, and *word pairs* between computer-generated summary and ideal summaries created by humans. There are several measurements, including ROUGE-N, ROUGE-L, ROUGE-W, and ROUGE-S. DUC 2005 ran ROUGE-1.5.5 to compare human and system scores. Only the recalls of ROUGE-2 and ROUGE-SU4 were reported as the official ROUGE scores at DUC 2005.

5.3 Experimental Settings

In our experiments, we evaluated models with different settings to answer the following questions:

Q1: Can the proposed hybrid relevance analysis achieve better performance when it is employed to estimate the relevance of a sentence to the query?

Q2: Are surface features beneficial to improve the summarization performance?

Q3: Does the modified MMR outperform the original MMR?

Table 19 lists these models. In the table, B1 and B2 are baselines, which only exploit the similarity metric proposed in [1] and the original MMR [6]. M10 is the system that integrates all proposed methods in this work. Others are listed here as references to obtain more clearly the impact of different factors. As for parameters (as listed in Table 20), they were set manually in these experiments.

Table 19: Settings of different models

Settings	Relevance	Features	MMR
B1	sim_1	N	None
B2	sim_1	N	Original
M1	sim_1	Y	None
M2	sim_1	Y	Modified
M3	sim_2	N	None
M4	sim_2	N	Original
M5	sim_2	Y	None
M6	sim_2	Y	Modified
M7	sim_3	N	None
M8	sim_3	N	Original
M9	sim_3	Y	None
M10	sim_3	Y	Modified

Table 20 Parameter settings

Equation	Parameter Setting
Eq. (32)	k=10
Eq. (33)	α=0.5
Eq. (40)	w_{sig}=0.5; w_{sim}=0.7; w_{f1}=0.8; w_{f2}=0.3; w_{f3}=0.3; w_{f4}=0.5; w_{f5}=1.0; w_{NE}=0.3
Eq. (42)	δ=0.5; λ=0.7

5.4 Results

The recall results are given in Table 21, sorted by the recall values of ROUGE-2 and ROUGE-SU4 respectively. In this table, the recall values of the best systems at DUC 2005 are listed as well (see System 15 [24] and System 17 [15]).

First, when only sentence similarity with the query is considered (see B1, M3, and M7), as we expected, M7, which employs the hybrid relevance analysis, obtained the highest score. As mentioned before, this benefits from noise reduction by averaging different similarity metrics. M3 outperformed B1, showing that latent semantic analysis could derive the topic structure of the corpus by grouping words according to co-occurrences, which leads to higher recalls compared to the vector space model. Moreover, for all cases using the proposed hybrid relevance analysis, they outperformed all other models (e.g., M7>M3>B1; M8>M4>B2; M9>M5>M1;

M10>M6>M2). These results suggest that a hybrid relevance analysis which combines similarities computed from the vector space model and latent semantic analysis is a successful way to estimate a better sentence relevance to the query.

Second, for those models in which surface features are taken into account but without MMR (see M1, M5, and M9), it is obvious that a scoring mechanism enhanced with low-level text features will improve the performance (e.g., M1>B1; M5>M3; M9>M7). It is noted that according to the results of our experiments, we found that a slightly smaller w_{sig} obtains better results. This is because for query-relevant summarization, relevant sentences are much more important than those sentences with high shallow feature salience which might be interpreted as theme-relevant sentences.

Finally, considering cases when the modified MMR was applied (see M2, M6, and M10), they outperformed models which use the original MMR (see B2, M4, and M8). A sentence, if it has high feature score and is highly relevant to the query but has lower similarity with sentences in the summary, will be ranked in the topmost position. This demonstrates that the modified MMR is a suitable module for query-focused multidocument summarization.

Table 21: recalls of ROUGE-2 and ROUGE-SU4

	Models	R-2	Models	R-SU4
1	**M10**	**0.075690**	*System 15*	0.131633
2	M6	0.073880	**M10**	**0.129950**
3	M9	0.073780	*System 17*	0.129725
4	*System 15*	0.072510	M6	0.127110
5	M2	0.072280	M9	0.126870
6	*System 17*	0.071741	M5	0.124430
7	M5	0.071340	M2	0.124330
8	M8	0.070110	M8	0.124270
9	M4	0.070000	M4	0.123930
10	M1	0.069720	M7	0.121750
11	M7	0.068730	M1	0.121350
12	B2	0.067690	B2	0.120200
13	M3	0.067190	M3	0.119950
14	B1	0.064830	B1	0.117550

To sum up, we got the best results of 0.075690 and 0.129950 for ROUGE-2 and ROUGE-SU4 respectively (see M10). The results are comparable to System 15 and System 17, which had the best results at DUC 2005.

6 Conclusion

In this report, we propose a sentence retrieval approach to address query-focused multidocument summarization. The proposed method measures the relevance of a sentence to the query using a novel hybrid relevance analysis which linearly combines relevance measures from the vector space model and latent semantic analysis. The output summary is generated by including sentences with high sentence salience which is evaluated in terms of sentence relevance and low-level feature significances. In addition, a modified redundancy reduction module based on MMR is proposed for anti-redundancy by combining sentence representative power (i.e., surface feature salience) with the original MMR. The proposed method was evaluated using the DUC 2005 official corpus and found to perform well with competitive results.

The contributions of this work are three-fold. First, a hybrid relevance analysis is proposed to estimate sentence relevance to the query. Second, shallow features are employed for scoring sentence importance and are shown to be useful. Finally, a modified MMR was proposed and shown to be a suitable component for query-focused summarization when sentence representative power is considered.

In the future, we intend applying sentence compression techniques in order to include more useful information in the summary. We also plan to resolve anaphora references to obtain a summary with better readability. Sentence ordering is another issue that needs to be investigated to create a more fluent and coherent summary.

References

[1] J. Allan, C. Wade, and A. Bolivar, Retrieval and Novelty Detection at the Sentence Level, *Proc. of SIGIR'03*, Toronto, Canada, 2003.

[2] E. Amigo, J. Gonzalo, V. Peinado, A. Penas, and F. Verdejo, An Empirical Study of Information Synthesis Tasks. *Proc. of ACL 2004*, Barcelona, Spain, 2004.

[3] A. Berger, and V. O. Mittal, Query-Relevant Summarization using FAQs, *Proc. of ACL 2000*, Hong Kong, Chian, 2000.

[4] S. Blair-Goldensohn, From Definitions to Complex Topics: Columbia University at DUC 2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[5] W. Bosma, Query-Based Summarization Using Rhetorical Structure Theory, *Proc. of CLIN 2003*, Belgium, 2003.

[6] J. Carbonell and J. Goldstein, The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries, *Proc. of SIGIR'98*, Melbourne, Australia, 1998.

[7] E. D'Avanzo and B. Magnini, A Keyphrase-based Approach to Summarization: the LAKE System at DUC-2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[8] H. Daumé III and D. Marcu, Bayesian Query-Focused Summarization, *Proc. of COLING/ACL 2006*, Sydney, Australia, 2006.

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, Indexing by Latent Semantic Analysis, *Journal of the American Society for Information Science*, Vol. 41, No. 6, 1990, pp. 391-407.

[10] Document Understanding Conference. Available at http://duc.nist.gov/.

[11] B. Hachey, G. Murray and D. Reitter, Query-oriented Multi-document Summarization with a Very Large Latent Semantic Space, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[12] T. Hirao, K. Takeuchi, H. Isozaki, Y. Sasaki and E. Maeda, NTT/NAIST's Text Summarization Systems for TSC-2, *Proc. of NTCIR 2003,* Japan, 2003.

[13] E. Hovy, C.-Y. Lin, and L. Zhou, A BE-based Multidocument Summarizer with Query Interpretation, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[14] J. Jagadeesh, P. Pingali and V. Varma, A Relevance-Based Language Modeling Approach to DUC 2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[15] W. Li, W. Li, B. Li, Q. Chen and M. Wu, The Hong Kong Polytechnic University at DUC2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[16] C. Y. Lin, Training a Selection Function for Extraction, *Proc. of CIKM'99*, Kansas, MO, 1999.

[17] I. Mani, and E. Bloedorn, Summarizing Similarities and Differences among Related Documents, *Information Retrieval*, Vol. 1, 1999, pp. 35-67.

[18] I. Mani, and M. T. Maybury (Eds), *Advances in Automated Text Summarization*, Cambridge, MA: The MIT Press, 1999.

[19] W. C. Mann, and S. A. Thompson, Rhetorical Structure Theory: Toward a Function Theory of Text Organization, *Text*, Vol. 8, No. 3, 1988, pp. 243-281.

[20] MEAD, http://www.summarization.com/mead/.

[21] ROUGE, http://www.isi.edu/~cyl/ROUGE/.

[22] F. Schilder, A. McCulloh, B. T. McInnes, and A. Zhou, TLR at DUC: Tree Similarity, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[23] Y. Seki, K. Eguchi, N. Kando, and M. Aono, Multi-Document Summarization with Subjectivity Analysis at DUC 2005, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[24] S. Ye, L. Qiu, T.-S. Chua and M.-Y. Kan, NUS at DUC 2005: Understanding Document via Concept Links, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

[25] J. Y. Yeh, H. R. Ke, W. P. Yang, and I. H. Meng, Text Summarization Using a Trainable Summarizer and Latent Semantic Analysis, *Information Processing & Management*, Vol. 41, No. 1, 2005, pp. 75-95.

[26] Q. Zhou, L. Sun, and J. Y. Nie, IS_SUM: A Multi-Document Summarizer based on Document Index Graphic and Lexical Chains, *Proc. of DUC 2005*, Vancouver, Canada, 2005.

Publication List

Journal Papers

[1] Jen-Yuan Yeh, Hao-Ren Ke, and Wei-Pang Yang (2006). Summarizing Relevant Information for Question-Answering Using Hybrid Relevance Analysis and Surface Feature Salience. *WSEAS Transactions on Information Science and Applications*, 3(12), 2549-2556. (EI), **(NSC-94-2213-E-009-013)**

[2] Pei-Cheng Cheng, Jen-Yuan Yeh, Hao-Ren Ke, Been-Chian Chien, and Wei-Pang Yang (2005). Comparison and Combination of Textual and Visual Features for Interactive Cross-Language Image Retrieval. *Lecture Notes in Computer Science*, Vol. 3491, pp. 793-804. (Revised Selected Paper from CLEF 2004), (SCI, EI), **(NSC-93-2213-E-009-044)**

Conference Papers

[1] Jen-Yuan Yeh, Hao-Ren Ke, and Wei-Pang Yang (2006). Query-Focused Multidocument Summarization Based on Hybrid Relevance Analysis and Surface Feature Salience. *Proc. of the 2nd WSEAS International Symposium on Data Mining*, Lisbon, Portugal. **(NSC-94-2213-E-009-013)**

[2] 劉政璋, 葉鎮源, 柯皓仁, 楊維邦 (2005). 以概念分群為基礎之新聞事件自動摘要. *Proc. of the 17th ROCLING Conference on Computational Linguistics and Speech Processing* (*ROCLING XVII*), Tainan, Taiwan. **(NSC-92-2213-E-009-126)**

[3] Pei-Cheng Cheng, Jen-Yuan Yeh, Hao-Ren Ke, Been-Chian Chien, and Wei-Pang Yang (2004). NCTU-ISU's Evaluation for the User-Centered Search Task at ImageCLEF 2004. *Proc. of the CLEF 2004 Workshop of the Cross Language Evaluation Forum* (*CLEF 2004*), Bath, UK. **(NSC-92-2213-E-009-126)**

[4] 楊維邦, 葉鎮源 (2003). 多語言複合式文件摘要系統. Talk at *AI Forum 2003*, Kaohsiung, Taiwan. **(NSC-92-2213-E-009-126)**