International Journal of

Computational Linguistics & Chinese Language Processing

中文計算語言學期刊

A Publication of the Association for Computational Linguistics and Chinese Language Processing

This journal is included in THCI Core, Linguistics Abstracts, and ACL Anthology.

易繫辭曰上古結繩而治後世聖人易之以書契百官以治萬民以察說文敘曰蓋文字者經藝之本宣教明化之始前人所以垂後後人所以識古故曰本立而道生知天下之至賾而不可亂也教化既萌文心雕龍則謂人之立言因字而生句積句而成章積章而成篇篇之彪炳

Vol.15 No.3-4 September/December 2010 ISSN: 1027-376X

International Journal of Computational Linguistics & Chinese Language Processing

International Journal of

Computational Linguistics & Chinese Language Processing

The editing of this journal is subsidized by Center for Humanities Research National Science Council in 2010.

Aims and Scope

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) is an international journal published by the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). This journal was founded in August 1996 and is published four issues per year since 2005. This journal covers all aspects related to computational linguistics and speech/text processing of all natural languages. Possible topics for manuscript submitted to the journal include, but are not limited to:

- Computational Linguistics
- Natural Language Processing
- Machine Translation
- Language Generation
- Language Learning
- Speech Analysis/Synthesis
- Speech Recognition/Understanding
- Spoken Dialog Systems
- Information Retrieval and Extraction
- Web Information Extraction/Mining
- Corpus Linguistics
- Multilingual/Cross-lingual Language Processing

Membership & Subscriptions

If you are interested in joining ACLCLP, please see appendix for further information.

Copyright

Cover

Calligraphy by Professor Ching-Chun Hsieh, founding president of ACLCLP
Text excerpted and compiled from ancient Chinese classics, dating back to 700 B.C.
This calligraphy honors the interaction and influence between text and language

International Journal of
Computational Linguistics and Chinese Language Processing
vol. 15, no. 3-4, September/December 2010

Contents

Papers

Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 161-180

結合長詞優先與序列標記之中文斷詞研究

A Simple and Effective Closed Test for Chinese Word Segmentation Based on Sequence Labeling

林千翔[*]、張嘉惠[*]、陳貞伶[*]

Qian-Xiang Lin[*], Chia-Hui Chang[*], and Chen-Ling Chen[*]

摘要

中文斷詞在中文的自然語言處理上，是個相當基礎且非常重要的工作。近年來的斷詞系統較傾向於機器學習式演算法來解決中文斷詞的問題。但使用傳統的作法，如隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），所以許多研究都是使用外部資源或是結合其他的機器學習演算法來幫助斷詞。然而當外部資源不易取得時，如何以簡易的方式達到準確的斷詞，則是本研究的目標。在本篇論文中我們以訓練資料所提供的詞彙建構一個辭典，並以長詞優先比對（Maximum Matching）提供正向及反向的斷詞結果做爲應用序列標記之機器學習特徵函數，用以提升隱藏式馬可夫模型（HMM）及條件隨機域（CRF）序列標記的準確率。我們發現，藉由長詞優先比對，得以在完全不修改模型之訓練及測試過程的前提下，透過辭典的遮罩（Mask）及特製化（Specialized）方式，改善斷詞的效能。實驗結果顯示，長詞優先可大幅改善馬可夫模型的斷詞效能（F-measure: 0.812→0.948）；而利用 Mask 方式則可將斷詞效能提升至 0.953;另挑選高錯誤率的字元做爲特製詞，則可再次提升斷詞效能至 0.963。若採用條件隨機域做爲序列標記模型，則僅需透過辭典遮罩，即可將系統斷詞效能提升至 0.963。

關鍵字: 自然語言處理，隱藏式馬可夫模型，中文斷詞，條件隨機域

[*]國立中央大學資訊工程學系
E-mail: chia@csie.ncu.edu.tw
The author for correspondence is Chia-Hui Chang.

Abstract

In many Chinese text processing tasks, Chinese word segmentation is a vital and required step. Various methods have been proposed to address this problem using machine learning algorithm in previous studies. In order to achieve high performance, many studies used external resources and combined with various machine learning algorithms to help segmentation. The goal of this paper is to construct a simple and effective Chinese word segmentation tool without external resources, that is, a closed test for Chinese word segmentation. We use training data to construct a vocabulary to combine maximum matching word segmentation results with sequence labeling methods including hidden Markov model (HMM) and conditional random fields (CRF). The major idea is to provide machine learning algorithm with ambiguity information via forward and backward maximum matching as well as unknown word information via vocabulary masking. The experimental results show that maximum matching and vocabulary masking can significantly improve the performance of HMM segmentation (F-measure: 0.812 → 0.948 → 0.953). Meanwhile, combining maximum matching with CRF achieves a performance with 0.953 and is improved to 0.963 via vocabulary masking.

Keywords: Chinese Word Segmentation, Maximal Matching, Hidden Markov Model, Conditional Random Field, Vocabulary Masking

1. 序論

中文斷詞在中文的自然語言處理上，是非常重要的前置處理工作。許多中文的自然語言相關的領域，例如：問答系統、自動摘要、文件檢索、機器翻譯、語音辨識…等，都需要先處理中文斷詞，可見中文斷詞是個相當基礎且非常重要的工作。

所謂的「中文斷詞」就是將一連串的中文「字串」轉換成「詞串」的組合。例如：「我昨天去台北」這個中文句子，透過中文斷詞的處理後變成「我／昨天／去／台北」，也就是將｛我、昨、天、去、台、北｝字串轉成｛我、昨天、去、台北｝的詞串組合。傳統上，處理中文斷詞會遇到的問題，大致可歸納為兩點，一是「歧義性」（ambiguity）問題，二是「未知詞」（unknown word）問題。歧義性問題即是同一個中文字串，於不同的文章當中，存在不同的斷詞結果，因此容易造成斷詞上的錯誤。歧義型態大致上可以分為兩類：

- ■ 交集型歧義（overlapping ambiguity）

  令x, y, z 代表中文字元所組成的字串，若x、z、xy 與yz 皆為辭典中的詞，則xyz的組合，於不同的文章中，可能會被斷詞成xy/z 或x/yz 等兩種不同的結果，則xyz稱為「交集型歧義字串」。例如：「不可以」三個中文字元所組成的字串，辭典

中的詞含有「不、不可、可以」，「不可以」所組成的字串，在下列句子中，因其上下文的不同而產生不同的斷詞結果：「不／可以／忘記」、「不可／以／營利／爲／目的」。

- 組合型歧義（covering ambiguity）

  令x, y 代表中文字元所組成的字串，若x、y、xy 都是辭典中的詞，xy 的組合中，可在不同的文章中，分別被斷詞成xy 或x/y，因爲詞xy 是由x 與y等兩個不同的詞所組成，因此xy 稱爲「組合型歧義字串」。例如：「才能」二個字所組成的字串，辭典中的詞有「才、能、才能」，在下列句子中「才能」組成的字串，將產生不同的斷詞結果：「他／才能／非凡」、「只有／他／才／能／勝任」。

另外，「未知詞」則指辭典中未收錄的詞，包含了人名、地名、組織名、人名地名組織名之縮寫、衍生詞、複合詞、數字型態等，由於人類所使用的語言會隨著社會不斷改變，而持續地創造出新的用語，並且詞的衍生現象也非常地普遍，因此新詞會不斷的出現，辭典永遠無法因應新詞產生的速度，所以會出現未知詞問題，斷詞系統必須能夠處理未知詞，才可提高斷詞的正確性。

近年來的斷詞系統傾向於機器學習式（machine learning-based）演算法來解決中文斷詞的問題，例如應用最大熵分類 Maximum Entropy (MaxEnt) (Xue, 2003)、向量支持機 Support Vector Machine (SVM) (Asahara, *et al*., 2003; Goh, *et al*., 2005)、Transformation-Based Learning Algorithm （TBL） (Lu, 2005)等分類演算法，另外以隱藏式馬可夫模型 Hidden Markov Model （HMM） (Asahara, *et al*., 2005; Lu, 2005; Xue & Shen, 2003; Zhang, *et al*., 2003)的序列標記演算法等等，並且顯示了使用機器學習式演算法做中文斷詞，確實可以達到很高的斷詞準確率。

本研究使用隱藏式馬可夫模型來解決中文斷詞的問題。雖然已有數篇研究同樣使用隱藏式馬可夫模型來處理斷詞問題 (Asahara, *et al*.,2003; Lu, 2005; Xue & Shen, 2003; Zhang, *et al*., 2003)，但使用傳統的作法，隱藏式馬可夫模型在解決中文斷詞的問題上，無法達到較好的斷詞效能（F-measure 約 80%），因此這些研究便結合了其他機器學習演算法，以增加斷詞的效能。

我們的研究目的是希望只使用隱藏式馬可夫模型當成主要的演算法，並且應用「特製化」（Specialization）的概念來提升隱藏式馬可夫模型的準確率。我們的作法是給予隱藏式馬可夫模型更多的資訊，在完全不修改模型之訓練及測試過程的前提下，透過兩階段特製化的方式，分別爲擴充「觀測符號」，以及擴充「狀態符號」的方式，大大地改善了隱藏式馬可夫模型的斷詞準確性。

於第一階段中，爲了擴充觀測符號，我們使用最簡單也最常被使用的辭典比對式斷詞演算法－「長詞優先法」（Maximum Matching Algorithm），來增加額外的資訊於隱藏式馬可夫模型中，使得模型擁有更多的斷詞資訊做學習。第二階段擴充狀態符號的方式，我們則使用詞彙式隱藏式馬可夫模型（Lexicalized HMM）的概念，也就是只根據某些特製詞來做特製化，將狀態做延伸，來提升系統斷詞的效能。

2. 相關研究

中文斷詞的研究已有相當歷史，但在近幾年仍陸續新的方法提出，底下我們分別就解決歧義性及未知詞兩個問題分別做文獻回顧。

首先就斷詞歧義性問題，M.Li 等人 (Li, *et al.*, 2003) 於 2003 年的研究中，提出一種非監督式（unsupervised）訓練的方法，藉由訓練 Naïve Bayes 分類器，來解決中文斷詞的交集型歧義問題，實驗結果可達到 94.13% 的準確率。另一方面，解決組合型歧義比解決交集型歧義更加困難，主要的原因是，要解決組合型歧義則需要依賴更多的內文資訊，如句法分析（syntactic）、語意分析（semantic） 以及前因後果的資訊（pragmatic information）等，才能正確的解決這類的歧義問題。1999 年 J. H. Zheng 等人 (Zheng & Wu, 1999) 使用規則式（rule-based method）的作法來處理組合型歧義，並達到 85 % 的準確率。而 2002 年 X. Luo 等人 (Luo, *et al.*, 2002) 的研究，則是使用類似於自然語言處理領域中解決「詞義消歧」（word sense disambiguation）的問題，來解決組合型歧義問題，該篇研究使用 TF.IDF 權重計算的公式，重新定義新的 TF 與 IDF 的公式，以此方式來解決組合型歧義問題，達到 96.58 % 的準確率。

解決未知詞問題是做中文斷詞的另一個重要步驟。中研院陳克健博士等人於 1997 年開始，提出了三篇關於解決未知詞問題的研究 (Chen & Bai, 1997; Chen & Ma, 2002; Ma & Chen, 2003) ，最早於 1997 的研究 (Chen & Bai, 1997)，透過統計斷詞語料庫，產生所有單一字元之已知詞的偵測規則。此階段的研究只能偵測出所有的單一字元的結果，並未真正將未知詞擷取出來。2002 年的研究 (Chen & Ma, 2002) ，則是使用人工加上一些統計的方法來建立擷取規則，將所有被偵測出屬於未知詞部分的單一字詞，透過擷取規則以合併這些單一字詞而成為未知詞。實驗中測試 1,160 個未知詞，結果達到 89 % 的擷取準確率。另外於 2003 年的研究 (Ma & Chen, 2003) 中，該研究將所有種類的未知詞的構詞方式以 context free grammar 表示出來，並搭配 bottom-up merging algorithm 來解決大部分統計特性低的未知詞擷取問題。實驗效能達到 75 % 的擷取準確率。其他解決未知詞問題的研究，如 Zhang 等人 (Zhang, *et al.*, 2002) 於 2002 年的研究，則使用類似詞性標示（part-of-speech tagging）的作法，稱為「角色標示」（roles tagging），角色指的是在未知詞的組成成分、上下文以及句子中的其他部分，並且依據句子的角色序列來辨識出未知詞。實驗部分針對中國人名以及外國翻譯名等未知詞做測試，並且達到不錯的準確率以及召回率。

近年來的研究主要趨向於機器學習式的方法來處理中文斷詞，例如最大熵分類法 Maximum Entropy（ME）(Xue, 2003) 是將斷詞轉成字元分類問題（character classification），並且使用了數種類似的特徵，如目前字元、加上前後各一字元、加上前後各兩字元等，來當作模型的屬性。而 C. L. Goh 等人則使用支持向量機 Support Vector Machine（SVM）(Goh, *et al.*, 2005) 來解決中文斷詞的問題，該篇研究結合辭典比對式方法－長詞優先法，利用長詞優先法的歧義性以及未知詞的資訊，來加強 SVM的特徵屬性以改善斷詞效能。另外也有使用感知機（Perceptron）(Li, *et al.*, 2005) 的方法做斷詞，該篇研究認為

Perceptron 方法雖然與 SVM 類似，不過效能卻較 SVM 差一些，但由於其訓練的速度非常快，因此他們系統提出的主要貢獻就是一個速度快且效能不至於差太多的斷詞方法。

另外一類則是以序列標記（sequence labeling）問題來處理中文斷詞，尤其以隱藏式馬可夫模型爲主。不過單獨使用 HMM 本身的效能並不高（約 81%）(Asahara, *et al*., 2003; Xue & Shen, 2003)，因此這兩篇研究將隱藏式馬可夫模型的斷詞結果當成是一個屬性，並分別使用 SVM (Asahara, *et al*., 2003)以及 TBL (Lu, 2005)來當成主要的演算法做斷詞，以達到較佳的斷詞結果。另外，條件隨機域 Conditional Random Fields（CRF）則是 2003 年後廣爲使用的資訊擷取（information extraction）及斷詞（segmentation）方法，如 Massechusset Amherst 大學 A. McCallum, F. Peng, F. Fang 等人在 Rocling 2004 的論文，運用了 24 個中文詞素如姓氏、國名、職稱字首、職稱字尾、地名、日期、單位、動詞、名詞、形容詞等，以及大量的辭典（Vocubulary），得以將中研院平衡語料庫(AS)的 closed test 達到 0.956；而 H. H. Tseng 等人在 Sighan Backoff 2005 的論文，則藉由罕見字的字首及字尾表所建構的特色，解決未知詞的問題，在 2,558,840 的特徵函數下，對中研院平衡語料庫的 closed set 可達到 0.97。

3. 系統架構

我們提出的系統架構如圖 1 所示，主要想法是利用訓練資料中已斷詞的文件（Segmented Texts），建立一個辭典（Vocabulary Construction），再利用長詞優先比對(Maximum Matching)提供正向及反向標記資訊，讓學習模組（Learning Module）得以學習最佳參數；實際斷詞時，即將未斷詞之文章（Unsegmented Texts），同樣利用長詞優先比對，產生與訓練資料相同的測試資料，藉由以訓練好的模型（Model），標記文件並得到斷詞結果（Segmented Texts）。我們使用兩種學習模組，一者爲隱藏馬可夫模型，一者爲條件機率域來解決中文斷詞的問題。

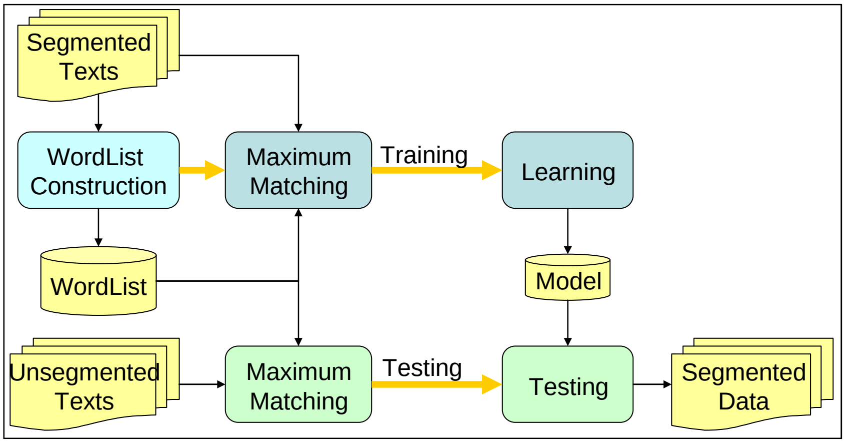


圖 1. 系統學習架構

3.1 BIES分類與序列標記問題

利用機器學習式演算法來解中文斷詞的問題時，一般的作法是將中文斷詞問題轉換成分類的問題，而最常被使用的方法就是轉換成字元分類問題（character classification problem），將每個字元都給予其對應的類別，透過字元類別來做分類，這些字元的類別由出現在中文詞當中的特定位置來決定，一個字元的位置可以分爲位於詞的開始（beginning）、位於詞的中間 （intermediate）、位於詞的結尾（end）以及由單一字元組成的詞（single-character）等四種類別，因此也稱爲「BIES 分類問題」。

理論上中文字元可以存在於中文詞的任何位置上，例如表 1 的例子，字元「中」可以存在於詞的開始（B）、詞的中間（I）、詞的結尾（E）、以及單一字元的詞（S）。所以BIES 分類所要解決的問題也就是決定每個字元的正確類別。在中文斷詞的問題上，一旦將欲斷詞字串中的所有字元都已分類完成，則也表示已經斷詞完成，例如：「今天是重要的日子」這個中文字串，利用分類問題將找出每個字元所對應的 BIES 標籤，在此例子中，也就是「BESBESBE」，則相當於是已經斷詞出｛今天、是、重要、的、日子｝等詞出來了，因此原來的中文字串便可以轉換成「今天／是／重要／的／日子」的斷詞結果。

表 1. 字元「中」可出現在詞的任何位置

B	中醫
I	國民中學
E	集中
S	在 資料庫 中

在 BIES 分類問題中，由於一個字元可出現在詞的不同位置，而導至所對應的 BIES 標籤不只一個，一旦類別標示錯誤，連帶會使得斷詞結果錯誤。但此種斷詞歧義性在 HMM 模式下，並無特殊處理方式。由於正向長詞優先與反向長詞優先在做斷詞時，遇到歧義性的句子會產生不同的斷詞結果，因此如能將正向長詞優先與反向長詞優先的資訊同時加入 HMM 模型中，相當於提供歧義性的資訊，並且長詞優先法屬於辭典比對式斷詞法，雖無法直接提供未知詞的資訊，但可間接的調整辭典大小來反應未知詞多寡。這也是我們之所以採用長詞優先比對提供 BIES 分類額外資訊的原因。

另外，BIES 字元分類雖然可以藉由前後幾個字的資訊，來完成斷詞，然而決定每個字元是類別是獨立的，因此並無法兼顧前後字元的標記，例如一個字元若被標記爲 B，則其後字元理應被標記爲 I 或 E，而不是 S 和 B，這也是序列標記問題希望能解決的問題。本篇論文採用分別採用隱藏馬可夫模型及線性條件隨機域模型做爲序列標記的演算法。

3.2 長詞優先法

長詞優先法（Maximum Matching Algorithm, MM）是最簡單也最為廣泛使用的辭典比對式的斷詞方法，其斷詞的策略是由句子的一端開始，試著比對出在辭典中最長的詞，當作斷詞結果，接著去除此詞後，剩下的部分繼續做長詞優先法斷詞，直到句子的另一端結束為止。一般來說，如果所使用的辭典夠大，長詞優先法斷詞可達到超過 90%以上的斷詞準確率。

長詞優先法依照比對方向的不同又可分為兩種不同的變形，第一種是「正向長詞優先法」（Forward Maximum Matching, FMM），即由句子開頭的第一個字元開始，由左而右逐一掃瞄，比對出在辭典中最長的詞，以當作斷詞的結果，並直到句子的結尾而結束。相反地，另一種長詞優先法的變形則是「反向長詞優先法」（Backward Maximum Matching, BMM），由句子的最後一個字元開始掃瞄，從右至左依序比對辭典中的詞，比對到最長的詞當成反向長詞優先法的斷詞結果，並直到句子的開頭而結束。

此兩種不同的長詞優先斷詞法，當斷詞的結果不同時，則表示發生交集型歧義，如表 2 中的第二個例子：「即將來臨時」字串，因為「將」可與「即」和「來」結合成｛即將、將來｝等不同的詞，因此屬於交集型歧義字串，正向長詞優先法會斷詞成「即將／來臨／時」，而反向長詞優先則斷詞成「即／將來／臨時」。

表 2. 長詞優先法的不同變形

例句	正向長詞優先	反向長詞優先
即將畢業	即將／畢業	即將／畢業
即將來臨時	即將／來臨／時	即／將來／臨時

另外，由於長詞優先法屬於辭典比對式斷詞方法，只有在辭典中的詞才有可能正確斷出，所以無法解決未知詞問題。當遇到未知詞時，正向長詞優先與反向長詞優先都將斷詞成單一中文字元。例如：「鴻海董事長郭台銘」字串，由於辭典中未收錄｛鴻海、郭台銘｝等詞，因此正向長詞優先法與反向長詞優先法都同樣會斷詞成「鴻／海／董事長／郭／台／銘」。

3.3 隱藏式馬可夫模型

隱藏式馬可夫模型可以視為一個雙層的隨機序列，包含了隱藏層的狀態序列（state sequence）和可觀察層的觀測序列（observation sequence）。隱藏層是無法直接觀察得到的，但可以從另一個可觀察的觀測序列之隨機過程的集合觀察得出。因此，隱藏式馬可夫模型是一個馬可夫鏈的機率函數，無法直接觀察的隱藏層就是一個有限狀態的馬可夫鏈，其初始的狀態機率分佈以及狀態之間的轉移機率由狀態初始機率向量Π和狀態轉移機率矩陣 A 來決定，另外還需定義觀測符號機率矩陣 B ，儲存各個觀測符號在不同的狀態下的機率值。令 S 表示所有狀態的集合，$S=\{s_1,s_2,\ldots,s_N\}$，N 表示模型中所有狀態的個數，K 表示所有觀測符號的集合，$K=\{k_1,k_2,\ldots, k_M\}$，M 表示模型中所有觀測符號的數

目，則隱藏式馬可夫模型可由三個機率分佈Π, A, B 來描述：

- Π=(π_i)代表狀態初始的機率向量，π_i=P(q_1 = s_i)，1≤i≤N，表示在 t=1 時，狀態爲 s_i 的機率，且需滿足Σπ_i=1 的條件。
- A=[a_{ij}] 代表狀態轉移機率矩陣，a_{ij}=P(q_{t+1} = s_j | q_t= s_i)，1≤i,j≤N，表示從狀態 s_i 到狀態 s_j 的機率，且滿足 a_{ij}≥0 和Σ$_j a_{ij}$=1。
- B=[$b_i(k)$] 代表觀測符號矩陣，$b_i(k)$=P(o_t = v_k | q_t= s_i)，1≤i≤N 和 1≤k≤M，表示在狀態爲 s_i 時，觀測符號爲 v_k 的機率，且滿足Σ$_k b_i(k)$=1。

給定輸入之觀察序列 $O = o_1 o_2 \cdots o_n =$ （o_t 表示在時間 t 所對應的觀測符號，且滿足 $o_t \in K$）。隱藏式馬可夫模型的目的就是要選出一個對應於觀測序列之最佳的狀態序列= $Q = q_1 q_2 \cdots q_n$ （q_t 表示在時間 t 所對應的狀態，且滿足 $q_t \in S$ ），也就是找出 $P(Q_1^n \mid O_1^n)$ 爲最大機率值時的狀態序列。

由於在馬可夫基本假設下，第 t +1 的時間狀態只和第 t 的時間狀態有關，與其他任何以前的時間狀態無關，即 $P\{q_{t+1} = s_k \mid q_1, q_2, ..., q_t\} = P\{q_{t+1} = s_k \mid q_t\}$ ，且隨機過程中的機率轉移不隨時間改變，因此 $P(Q_1^n \mid O_1^n)$ 的計算可簡化成：

$$P(Q_1^n \mid O_1^n) = \prod_{t=1}^{n} P(q_t \mid q_{t-1}) P(o_t \mid q_t) = \pi_{q1} \prod_{t=1}^{n-1} A_{q_t, q_{t+1}} \prod_{t=1}^{n} B_{q_t}(o_t) \qquad (1)$$

而取得此最大值的狀態序列 Q_1^n ，則是使用維特比（Viterbi）演算法計算得到。

隱藏式馬可夫模型當初所提出來的方法 (Rabiner, 1989) 是使用非監督式的學習方法（unsupervised approach）做訓練，也就是從未標示狀態的文件中做訓練（因而稱之爲「隱藏式」），訓練的方法則是使用 Baum-Welch 演算法做參數的更新。而近年來許多領域都已發展出大量已標示的語料庫可供訓練，隱藏式馬可夫模型同樣可以在已標示狀態的文件中來做監督式（supervised approach）訓練 (Manning & Schutze, 1999) ，訓練過程則直接利用最大概似估計法（maximum likelihood estimation）計算出模型參數則此模型，又可稱爲「可見式馬可夫模型」（Visible Markov Model, VMM）或「語言模型」（Language Model）等，但絕大部分的研究仍然稱「隱藏式」馬可夫模型。於我們的系統中，我們使用監督式的方法來訓練模型，在本論文中也直接以「隱藏式馬可夫模型」稱之。

3.3.1 觀測符號的擴充（FB+HMM）

隱藏式馬可夫模型原本的設計是只有單一個觀測符號，將長詞優先比對資訊加入隱藏式馬可夫模型，直接面臨的問題是如何計算多個觀測符號的機率。方法一是分別計算各個符號出現的機率再以觀測符號彼此獨立的假設來計算多個觀測符號的聯合機率；方法二則是直接記錄多個觀測符號的聯合機率；前者節省空間，後者機率估計較準。因此我們採用第二種方式，將正向長詞優先（FMM）與反向長詞優先（BMM）之斷詞結果(即所得的 BIES 標籤)，與原來的「字元」組成的新的觀測符號，延伸爲「字元-FMM-BMM」

等三個資訊結合而成的觀測序列。表 3 中以一個例子來針對 FB+HMM 訓練以及測試過程做個說明，在訓練階段中，原始的觀測符號序列為「研、究、生、命、起、源」，加入了長詞優先法的資訊後，新的觀測符號序列便被轉換成「研-B-B、究-I-E、生-E-B、命-S-E、起-B-B、源-E-E」。這些中文字元旁的 B、I、E、S 標籤即是由正向長詞優先與反向長詞優先法所標示的，因此新的觀測符號種類相當於增加了 16 倍，在此狀態種類並未做改變。

表 3. FB+HMM 的例子

	訓練過程		測試過程	
原始句子	研究／生命／起源		結合成分子	
HMM 訓練測試資料	觀測序列	狀態	觀測序列	狀態
	研-B-B 究-I-E 生-E-B 命-S-E 起-B-B 源-E-E	B E B E B E	結-B-S 合-I-B 成-E-E 分-B-B 子-E-E	? ? ? ? ?

3.3.2 特製隱藏式馬可夫模型

隱藏式馬可夫模型的特製化（specialization）概念，最早是由 J. D. Kim 等人於 1999 年與 2000 年等兩篇研究 (Kim, *et al*., 1999; Lee, *et al*., 2000) 所提出來的，之後於 2001 年到 2004 年間，A. Molina 及 F. Pla 等兩位學者，更是將此概念成功的應用到許多不同的領域上，如詞性標示（part-of-speech tagging）(Pla & Molina, 2001; 2004)、淺層分析（shallow parsing）(Molina & Pla, 2002)、詞義消歧（word sense disambiguation）(Molina, *et al*., 2002) 等問題上。

特製化的過程是指在不修改隱藏式馬可夫模型的訓練以及測試過程的前提下，透過狀態的延伸使得模型增加更多資訊，以提升模型準確率。其主要的作法就是給予一個特製化函式（specialization function），將原來的狀態符號產生出新的狀態符號，特製化的過程以底下式子來說明：

$$f(\langle o_i, q_i \rangle) = \langle o_i, q_i.o_i \rangle \tag{2}$$

$<o_i, q_i>$代表觀測序列中的某個觀測符號以及其對應的狀態，新的狀態符號經過特製化的過程中，由此觀測符號加上原來狀態來產生新的狀態，經過特製化過程的隱藏式馬可夫模型又稱為「特製隱藏式馬可夫模型」（Specialized HMM）。而如果不將所有的觀測符號所對應的狀態都做進行特製化，而是只針對某些較容易分類錯誤的觀測符號才做特製化，此過程則稱之為「詞彙式的隱藏式馬可夫模型」（Lexicalized HMM），此過程是屬

於特製化過程的一種特例，也被稱爲詞彙化（lexicalization），正式說來：

$$f(\langle o_i, q_i \rangle) = \begin{cases} \langle o_i, q_i.o_i \rangle & \text{if } o_i \in W \\ \langle o_i, q_i \rangle & \text{if } o_i \notin W \end{cases} \tag{3}$$

其中W 爲特製詞（specialized words），只有屬於特製詞的觀測符號才會做特製化處理，而特製詞的選擇又有許多不同的準則來選取。

表 4. 特製詞集合｛生-E-B, 起-B-B｝做詞彙化產生新的狀態

觀測符號	原來的狀態	新的狀態
研-B-B	B	B
究-I-E	E	E
生-E-B	B	B-生-E-B
命-S-E	E	E
起-B-B	B	B-起-B-B
源-E-E	E	E

在本篇論文中，透過第一階段將所有的觀測符號做延伸之後，我們進一步的以這些新的觀測符號來做詞彙化，也就是取某些特定的觀測符號來當成特製詞，將其對應的狀態做延伸的過程。舉例來說，如表 4 所示，假設觀測符號「生-E-B」、「起-B-B」屬於特製詞，則經過詞彙化的過程之後，觀測符號「生-E-B」以及「起-B-B」所對應的狀態就被轉換成「B-生-E-B」及「B-起-B-B」。也是多了兩個新的狀態：一個是由觀測符號「生-E-B」所屬的新狀態「B-生-E-B」，以及由觀測符號「起-B-B」所屬的新狀態「B-起-B-B」。因此在新的訓練資料中，狀態符號被延伸了。

此特製化過程也將牽扯到一個問題：由於隱藏式馬可夫模型的三個主要參數都與「狀態符號」有關，因此這階段的特製化過程，將增加隱藏式馬可夫模型的參數大小，因此計算量也就會跟著增加，而且過多的特製詞不見得能一直提升準確率。所以我們必須根據訓練資料來決定特製詞的大小。特製詞的選擇方式，我們是使用兩種不同的準則（criteria）來選取，此兩種不同的準則分別說明如下：

- **SWF:** (the Words with High Frequency)
  取在訓練資料中屬於最高頻率的觀測符號，當成特製詞。
- **SEF:** (the Words with Tagging Error Frequency)
  取具有高測試錯誤率 (或稱標示錯誤率) 的詞，當成特製詞。

不論是使用 SWF 或是 SEF 準則來選取特製詞，都需要決定一個門檻值（threshold），此門檻值是決定最合適的特製詞數量，我們會於實驗四中找出最佳斷詞效能的門檻值。

3.4 條件式隨機域

條件隨機域為一種無向圖(undirected graphical)模型，可被用來估算給予一觀測序列,得到相對應的狀態序列的條件機率分佈。相對於 HMM 以生成模型（generative model）描述觀察序列如何經由狀態轉移及符號產生的過程，CRF 專注於狀態序列在給定觀測序列下的條件機率分佈，屬於一種鑑別式機率模型（discriminative model）。原則上，條件隨機場的圖模型佈局是可以任意給定的，一般常用的佈局是鏈結式的架構，鏈結式架構不論在訓練、推論、或是解碼上，都存在有效率的演算法可供演算。

鏈結式條件隨機域模型（如圖 2 所示）可以定義為如下的問題：給定一組訓練資料樣本{X_1, X_2, …}，以及其相對的序列標資料{Y_1, Y_2, …}，監督式學習的目標是找到最佳的潛藏函數，使得最大似然度估計（likelihood）Π_i $P(Y_i|X_i)$最大化。

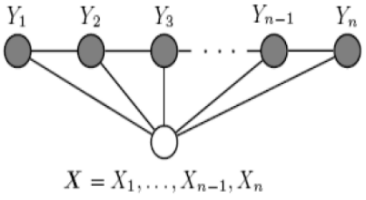


圖 2. 鏈結式條件隨機域模型

令 X^n 為一長度為 n 的觀測序列，則狀態序列 $Y=y_1, y_2, \cdots, y_n$ 的條件機率以隨機域表示如下：

$$P(y_1,\ldots,y_n \mid X) = \frac{1}{Z_X}\prod_{i=1}^{n}\psi_i^{v}(y_i, X)\psi_i^{e}(y_i, y_{i-1}, X) \tag{4}$$

其中 $\psi_i^{y}(y_i,X)$ 和 $\psi_i^{e}(y_i,y_{i-1},X)$分別為觀測序列 X 在第 i 個位置節點及邊的潛藏函數值，Z_x 則是觀測序列 X 的正規化常數，其目的在使$\Sigma_Y P(Y|X)=1$。大部份序列標記應用採用對數-線性（log-linear）的潛藏函數。

$$\begin{aligned}\psi_i^{v}(y_i, X) &= \exp\left(\sum_j \mu_j g_j(y_i, X, i)\right)\\ \psi_i^{e}(y_i, y_{i-1}, X) &= \exp\left(\sum_k \lambda_k f_k(y_i, y_{i-1}, X, i)\right)\end{aligned} \tag{5}$$

其中 $g_j(y_i,X,i)$及$f_k(y_i,y_{i-1},X,i)$分別為節點及邊的觀測屬性與標記 y 結合所得的特徵函數（feature function），而μ_j及λ_k分別為各函數的權重。雖然只是用單一指數模型來描述在給予觀察序列的條件下整個狀態序列的聯合機率，因此在不同的狀態中，不同的特徵函數所賦予的權重可以考慮到狀態彼此的情形。

在本篇論文中，我們採用 CRF++工具來做我們序列標記的工作（表 5 為 CRF++的輸入範例），並利用其樣版來產生 g_j 及f_k 等特徵函數。CRF++樣板的類型有 Unigram 與 Bigram 兩種，決定是否僅由單一狀態符號或是兩個狀態符號，與觀測符號結合來產生特徵函數。我們主要採用 Unigram 的狀態符號樣版，假設觀測符號共 6030 個字，正向與反

向長詞優先比對各可得 BIES 四種標記，若三種觀測符號均取前後各一位及當前的觀測符號共三個 unigram，兩種長詞優先比對則以目前及前一位組成 bigram，則總共可產生{(6030+4+4)*3+(8+8)*1}*4=72,584 特徵函數，其中第一項爲三種觀測符號的 unigram，第二項爲兩種長詞優先比對的 bigram，最外層的 4 則代表與狀態符號的組合。另外 Bigram 的狀態符號樣版與兩種長詞優先比對觀測符號組成(4+4)*16=128 特徵函數，故一共產生 72,716 個特徵函數。不過扣除一些不可能的 bigram 組成，最後得到 72,280 個特徵。

表 5. CRF++的輸入範例

	訓練過程		測試過程	
原始句子	研究／生命／起源		結合成分子	
CRF 訓練測試資料	觀測序列	狀態	觀測序列	狀態
	研 B B	B	結 B S	?
	究 I E	E	合 I B	?
	生 E B	B	成 E E	?
	命 S E	E	分 B B	?
	起 B B	B	子 E E	?
	源 E E	E		

4. 實驗

我們使用中央研究院平衡語料庫第 3.1 版，當成我們實驗的資料。此語料庫大小共 575 萬詞，不重覆的詞共 145,608 個，是第一個已斷好的詞並帶有詞類標記的現代漢語語料庫。我們取其中已斷詞的文章來當成我們的實驗對象，並且用隨機的方式依四比一的比例分割成兩個部分，取其中的 80% 當作訓練語料，用來訓練隱藏式馬可夫模型。而剩下的 20% 則當成我們系統的測試語料。斷詞的評估方式則是使用準確率（Precision）、召回率（Recall）以及 F-measure 等三個評估方式來驗證斷詞的效能，分別定義如下：

■ $$\text{Precision} = \frac{\text{系統正確斷出的詞數}}{\text{系統斷詞的總詞數}}$$

■ $$\text{Recall} = \frac{\text{系統正確斷出的詞數}}{\text{真正的詞數}}$$

■ $$\text{F measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

我們的實驗主要可分三階段來說明：第一階段結合長詞優先之斷詞主要分析辭典大小及未知詞對斷詞效能的影響；而第二階段則針對詞彙式隱式馬可夫模型，調整各策略使用的特製詞大小；最後階段則整體斷詞效能的比較。

4.1 結合長詞優先之斷詞效能

此部份的實驗，主要驗證隱藏式馬可夫模型結合長詞優先法之後，在觀測符號中加入更多資訊之前與加入之後的斷詞效能的比較。由於長詞優先法可以提供斷詞歧義性的資訊，同時辭典大小也可控制未知詞的多寡，因此這部分的實驗，我們先實驗辭典大小對斷詞效能的影響，接著實驗沒有未知詞影響的情形下，序列標記演算法能解決多少斷詞歧義性的問題。

由於辭典是由訓練資料產生，因此實驗時我們將訓練資料隨機分割成兩部分：訓練集合 1（set 1）以及訓練集合 2（set 2），辭典只由訓練集合 1 來產生，藉由調整兩個集合不同的分割比例，以產生出不同的數量的辭典，並對同一份測試資料下，驗證各自的斷詞效能。實驗結果如表 6 所示。

表 6. 辭典大小對斷詞效能的影響

	無未知詞	不同辭典大小				HMM
訓練資料比例(Set1/Set2)	100/0	80/0	60/20	40/40	20/60	0/80
辭典(Set 1)中的詞數	145,608	132,273	116,428	96,780	69,446	0
Set2 中的未知詞數	0	0	17,418	45,212	103,990	All
測試資料中的未知詞數	0	14,415	17,323	22,524	34,573	All
FB+HMM Recall	0.957	0.946	0.946	0.944	0.941	0.812
FB+HMM Precision	0.976	0.951	0.949	0.945	0.934	0.811
FB+HMM F-measure	**0.967**	**0.948**	**0.948**	**0.945**	**0.937**	**0.812**
BMM F-measure	**0.949**	**0.929**	**0.926**	**0.921**	**0.912**	**0.427**

表 6 的第二欄分割比例爲 100/0，相當於沒有未知詞情況下的效能；而第三～六欄使用不同比例的辭典大小，包括 80/0、60/20、40/40、20/60 等分割比例的結果，反應未知詞所佔的比例之不同時斷詞效能的表現；而最後一欄分割比例爲 0/80，代表完全不從訓練資料中建立辭典，也就是測試資料中所有的詞都屬未知詞，並且在訓練的過程中完全沒有從正向長詞優先法或反向長詞優先法中得到任何資訊，只依賴字元的資訊做斷詞。最後一列則是反向長詞優先的斷詞效能。

實驗結果顯示，在減少辭典的詞數的情況下，FB+HMM 的斷詞效能跟著減低，但是降低的幅度並不大，顯見只要有基本詞彙，即可提升 HMM 斷詞效能，但對於未知詞問題，並不能有所做爲，因此我們設計 Mask 的實驗來解決此一問題。

4.1.1 Mask模擬未知詞實驗

上述實驗在減少長詞優先法所需之辭典詞數的做法，某種程度提供了訓練資料中可能會預見未知詞的情形，但是不免犧牲了長詞優先法的正確性。因此我們引用 Mask 的作法

(Wu, Chang, & Lee, 2006)，在不犧牲訓練資料的詞的前提下，產生具有未知詞資訊的訓練資料。辭典 Mask 的概念是讓訓練過程中也有機會碰到未知詞，也就是仿造測試時真正的情形，其作法是將訓練資料分割成 K 個部分 $S_1, S_2, \cdots, S_K$，，並且每個部分都建立各自的辭典 $D_1, D_2, \cdots, D_K$，令 D 爲各辭典的聯集（$D=\cup_i D_i$），則共可產生 K+1 個辭典。接著對每一部份的訓練資料 S_i 以總辭典 D 減去 D_i，做爲長詞優先比對的辭典，用來標示 FB+HMM 所需要的觀測符號。在這過程中，訓練資料 S_i 中有些字詞會因爲未知詞的關係，會被錯標成單一字詞 S，但其狀態符號，可以讓 HMM 知道正確的標籤；如果標示結果與原來相同時，則可直接省略，以避免在一個狀態所見到的觀測符號機率不公平的增加，如此重複 K 次，加上原先以總辭典 D 對所有訓練資料所做的標記，將此 K+1 個資料形成整個 Mask 的訓練資料。圖 3 爲 K＝3 的示意圖。

我們取 K=2 至 K=10 來檢試 Mask 的效能，實驗結果如圖 4 所示，其中 K=1 表示不做分割，也就是沒有使用 Mask 的結果。實驗結果顯示，使用 Mask 的方法可提供隱藏式馬可夫模型更多未知詞資訊，使得斷詞效能有所提升，並且在 K=2 時，達到最佳的斷詞效能（F-measure =95.25%）。

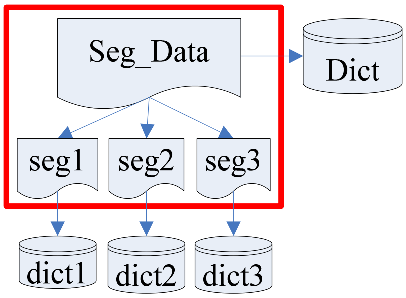


圖 3. Mask(K=3) 資料分割與建立辭典

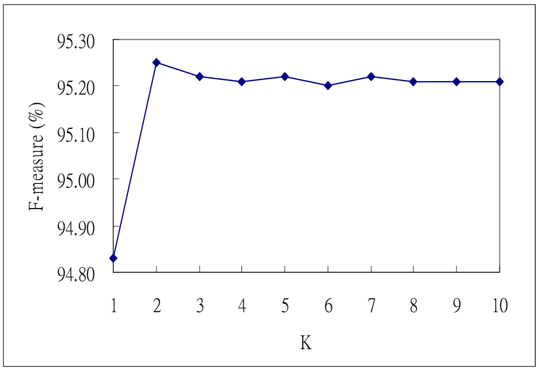


圖 4. Mask K=1 至 K=10 的實驗結果

4.1.2 斷詞歧義性的影響

接著我們針對沒有未知詞的情況下，檢視序列標記模型能解決多少歧義性問題。我們的方法是取所有訓練資料與測試資料中的所有詞，來當成長詞優先法所使用的辭典的詞（共有 145,608 個詞），使得在測試過程中不會出現未知詞，由於實際的情況下一定會遇到未知詞問題，因此這部分的實驗屬於「完美情況」的實驗。其中實驗的基準（baseline）爲正向長詞優先法（FMM）、反向長詞優先法（BMM），以及只使用字元資訊當成觀測符號的隱藏式馬可夫模型（HMM）。除了我們的系統 FB+HMM，即同時結合正向長詞優先法與反向長詞優先法資訊的實驗結果之外，我們也比較只結合正向長詞優先法資訊（F+HMM）以及只結合反向長詞優先法資訊（B+HMM）的隱藏式馬可夫模型之斷詞效能。表 7 爲實驗在無未知詞狀態下的斷詞效能。

如表 7 所示，隱藏式馬可夫模型若只有使用字元的資訊時，其結果不論是召回率、準確率或 F-measure 都只有 0.81 左右，而加入正向長詞優先法與反向長詞優先法的資訊之後，系統的斷詞效能 F-measure 便由 0.812 大幅地提升到 0.967，並且斷詞結果也勝過正向長詞優先法與反向長詞優先法等兩種基準作法。而 CRF 在使用正反向長詞優先比對所提供的歧義性資訊，得到最高的斷詞效能 0.977，顯示鑑別式機率模型較生成式機率模型有更好的預測表現 。

表 7. 無未知詞狀態下的斷詞效能

	FMM	BMM	HMM	F+HMM	B+HMM	FB+HMM	FB+CRF
Recall	0.936	0.939	0.812	0.944	0.947	0.957	0.981
Precision	0.956	0.959	0.811	0.962	0.965	0.976	0.974
F-measure	0.946	0.949	0.812	0.953	0.956	0.967	0.977

4.2 詞彙化隱藏馬可夫模型

第二階段則是 Lexicalized HMM，係根據 SWF 與 SEF 兩種不同的詞彙化策略，來調整各策略使用的特製詞大小，目的是去找出使得模型能有最佳斷詞效能的特製詞數量的門檻值（threshold）。由於這個實驗是用來調整系統用到的特製詞，而不是做斷詞效能的實驗，因此這個實驗我們只取「訓練資料」來做實驗。爲了驗證這個部分的效能，我們將全部訓練資料（佔全部資料 80%）分割成兩部分，依 7 比 1 的比例來分割（分別佔全部資料的 70% 與 10%），其中 70% 的資料（轉換成具有長詞優先法資訊的資料）用來訓練 FB＋HMM 模型，而剩下的 10% 則當成驗證效能的調整資料（validation set）。

由於 SWF 爲取訓練資料中出現頻率最高的詞當成特製詞，因此我們統計 70% 的訓練資料，取出高頻率的詞做特製詞。而 SEF 爲取高測試錯誤率的詞當成特製詞，因此我們先從 70% 的資料建立 FB＋HMM 模型，並且於調整資料中做測試，根據調整資料中高測試錯誤率的詞做特製詞。取得 SWF 與 SEF 之特製詞後，接著驗證在不同的門檻值下，調整資料的斷詞效能。實驗數據如圖 5 所示。

實驗結果顯示，我們使用 SWF 與 SEF 兩種不同的詞彙化策略，在剛開始取較少的詞當成特製詞時，兩者在調整資料下的斷詞效能都有顯著的上升，而當 SWF 準則取 292 個詞（出現頻率大於 4800 次）時，SEF 準則取 173 個詞（出現頻率大於 25 次）時，做詞彙化的斷詞效能達到最佳結果，並且再繼續隨著特製詞數的增加，斷詞結果便開始往下降，這是因為狀態增加，使得模型參數增加而導致準確率下降的緣故。

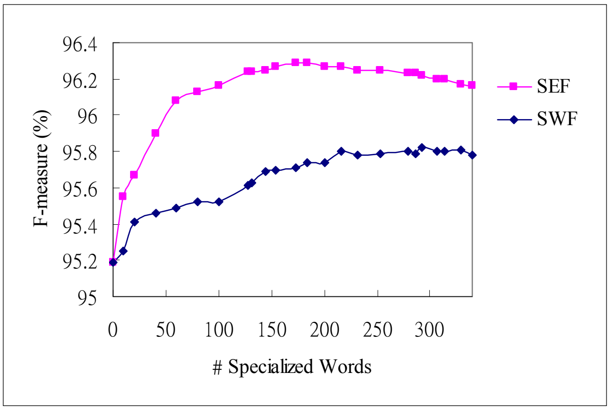


圖 5. 10% 驗證資料中 SEF 與 SWF 在不同特製詞大小下的斷詞效能

4.3 整體效能比較

最後我們比較不同學習模型在一般有未知詞情況下的斷詞效能，我們以正向長詞優先法（FMM）、反向長詞優先法（BMM）、以及只使用字元資訊之隱藏式馬可夫模型（HMM）等斷詞作法作基準，先與 FB+HMM 以及 FB+CRF 的結果做比較（如表 7 所示）。再取 Mask K=2 之最佳設定，產生訓練檔供 FB+HMM 及 FB+CRF 建立模型，同時針對特製最佳 SWF 與 SEF 特製詞（SWF 為取 292 個詞作為特製詞，而 SEF 則取 173 詞作為特製詞）來實驗，以驗證 FB+HMM 在狀態延伸前與延伸後的斷詞效能作比較。實驗結果如表 8 所示。

表 8. 有未知詞情況下 Mask 與 Specialized 效能比較

	Mask+FB+HMM	Mask +FB+CRF	SWF	SEF
Recall	0.947	0.966	0.958	0.963
Precision	0.958	0.961	0.962	0.964
F-measure	0.953	0.963	0.960	0.963

實驗結果顯示，應用長詞優先法比對可以有效的改進序列標記的效能。而在馬可夫 HMM 模型下而採用 Mask 方式產生訓練資料，可將 F-measure 由 0.948 提升至 0.953；若是 採用條件隨機域 CRF 模型，則更可將 F-measure 由 0.959 提升至 0.963。另外，不論使用 SWF 或 SEF 準則，詞彙化後的斷詞結果也可有效地將 F-measure 由 0.953 提升到 0.960 與 0.963 的結果，而且此兩種不同準則在相較之下，SEF 不但使用的特製詞數較少

且斷詞效能也較好。

表 9. 各斷詞模型的效率及使用成本分析

	HMM	FB+HMM	SEF=143	FB+CRF
Training Time	52 秒	1 分 9 秒	1 分 22 秒	37 分
Testing Time	8 秒	9 秒	58 分 6 秒	20 秒
Space	21MB	28MB	303MB	582KB

表 9 則是各種模型訓練及測試所需的時間、空間以及模型的大小。就訓練時間來說，我們可以看到 HMM 各種衍伸模型的訓練時間相較 CRF 都是相當快（65 倍）。其中 Specialized HMM 由於模型大小及測試時間遠大過其他模型，在資源有限的環境下並不實用；而 CRF 雖然訓練及測試的時間較長，但離線訓練所花的時間通常不是主要缺點，而測試時間也在合理範圍內（2 倍），在效能的考量下是實用上較好的選擇。

5. 結論

在本篇論文中，我們考慮僅從訓練語料中，如何建立一個不藉由外部資源的中文斷詞模型。首先我們結合了長詞優先法的資訊，使得觀測符號增加更多的資訊，於實驗結果顯示，結合長詞優先法可以大幅地提升馬可夫模型的斷詞效能（F-measure: 0.812→0.948）；而利用 Mask 方式也可進一步改善斷詞效能（F-measure: 0.948→0.953）；另使用詞彙式的特製化方式，挑選高錯誤的字元使得狀態增加，實驗也證明能再次提升斷詞效能（F-measure: 0.953→0.963）。若採用條件隨機率學習模組，在同樣使用正反向長詞優先比對所提供的歧義性資訊情況下，CRF 則提供比 HMM 更好的斷詞效能，F-measure 可由 0.948 提升至 0.959；若是藉由 Mask 的模擬測試，則可再將斷詞效能提升至 0.963，顯示鑑別式機率模型較生成式機率模型有更好的預測表現。因此雖然 CRF 訓練時間較久，在模型大小差異不大、但效能較佳的情形下，CRF 反而是實用上會是更好的選擇。本篇論文的效能雖未超越 H.H. Tseng 等人所達到之 0.97，但使用的特徵值相較的減少很多（1/100），以此爲基礎，加入其他特色將有機會推進中文斷詞的效能。

參考文獻

Asahara, M., Fukuoka, K., Azuma, A., Goh, C. L., Watanabe, Y., Matsumoto, Y., & Tsuzuki, T. (2005). Combination of Machine Learning Methods for Optimum Chinese Word Segmentation. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 134-137.

Asahara, M., Goh, C. L., Wang, X., & Matsumoto, Y. (2003). Combining Segmenter and Chunker for Chinese Word Segmentation. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 144-147.

Chen, K. J., & Bai, M. H. (1997). Unknown Word Detection for Chinese By a Corpus-based Learning Method. *In Proceedings of ROCLING X*, 159-174.

Chen, K. J., & Liu, S. H. (1992). Word Identification for Mandarin Chinese Sentences. *Proceedings COLING '92*, 101-105.

Chen, K. J., & Ma, W. Y. (2002). Unknown Word Extraction for Chinese Documents. *In Proceedings of COLING 2002*, 169-175.

Goh, C. L., Asahara, M., & Matsumoto, Y. (2005). Chinese Word Segmentation by Classification of Characters. *International Journal of Computational Linguistics and Chinese Language Processing*, 10(3), 381-396.

Kim, J. D., Lee, S. Z., & Rim, H. C. (1999). HMM Specialization with Selective Lexicalization. *In Proceedings of the joinSIGDAT Conference on Empirical Methods in Natural Lan-guage Processing and Very Large Corpora(EMNLP-VLC-99)*, 121-127.

Kudo, T. CRF++ 0.57: Yet Another CRF toolkit. Available from http://crfpp.googlecode.com/svn/trunk/doc/index.html

Lee, S. Z., Tsujii, J. I., & Rim, H. C. (2000). Lexicalized Hidden Markov Models for Part-of-Speech Tagging. *In Proceedings of 18th International Conference on Computa-tional Linguistics, Saarbrucken, Germany*, 481-787.

Li, M., Gao, J. F., Huang, C. N., & Li, J. F. (2003). Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 1-7.

Li, Y. Y., Miao, C. J., Bontcheva, K., & Cunningham, H. (2005). Perceptron Learning for Chinese Word Segmentation. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 154-157.

Lu, X. (2005). Towards a Hybrid Model for Chinese Word Segmentation. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 189-192.

Luo, X., Sun, M., & Tsou, B. K. (2002). Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. *In Proceedings of COLING 2002*, 598-604.

Ma, W. Y., & Chen, K. J. (2003). A Bottom-up Merging Algorithm for Chinese Unknown Word Extraction. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 31-38.

Manning, C. D., & Schutze, H. (1999). Foundation of Statistical Natural Language Processing. Chapter 9-10. 317-380.

Peng, F., Feng, F., & McCallum, A. (2004). Chinese Segmentation and New Word Detection using Conditional Random Fields. *In Proceedings of International Conference on Computational Linguistic* (COLING), 562- 568.

Molina, A., & Pla, F. (2002). Shallow Parsing using Specialized HMMs. *Journal of Machine Learning Research 2*, 595-613.

Molina, A., Pla, F., & Segarra, E. (2002). A Hidden Markov Model Approach to Word Sense Disambiguation. *In Proceedings of the VIII Conferencia Iberoamericana de Inteligencia Artificial (IBERAMIA)*, 1-9.

Pla, F., & Molina, A. (2004). Improving Part-of-Speech Tagging using Lexicalized HMMs. *Natural Language Engineering*, 167-189.

Pla, F., & Molina, A. (2001). Part-of-Speech Tagging with Lexicalized HMM. *In proceedings of International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Rabiner, L. R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE,* 77(22), 257-286.

Tseng, H. H., Chang, P. H., Andrew, G., Jurafsky, D., & Manning, C. (2005). A Conditional Random Field Word Segmenter for Sighan Bakeoff 2005. *In Proceedings of Fourth SIGHAN Workshop on Chinese Language Processing*, 168-171.

Wu, Y. C., Chang, C. H., & Lee, Y.S. (2006). A General and Multi-lingual Phrase Chunking Model Based on Masking Method. *Lecture Notes in Computer Science (LNCS): Computational Linguistics and Intelligent Text Processing*, Vol. 3878, 144-155.

Xue, N. (2003). Chinese Word Segmentation as Character Tagging. *International Journal of Computational Linguistics and Chinese*, 29-48.

Xue, N., & Shen, L. (2003). Chinese Word Segmentation as LMR Tagging. *In Proceedings of Second SIGHAN Workshop on Chinese Language Processing*, 176-179.

Zhang, H. P., Liu, Q., Zhang, H., & Cheng, X. Q. (2002). Automatic Recognition of Chinese Unknown Words Based on Roles Tagging. *In Proceedings of First SIGHAN Workshop on Chinese Language Processing*, 71-77.

Zheng, J. H., & Wu, F. F. (1999). *Study on segmentation of ambiguous phrases with the combinatorial type*. Collections of Papers on Computational Linguistics. Tsinghua University Press, Beijing, 129-134.

Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 181-192

Word Sense Disambiguation Using Multiple Contextual Features

Liang-Chih Yu*, Chung-Hsien Wu+, and Jui-Feng Yeh#

Abstract

Word sense disambiguation (WSD) is a technique used to identify the correct sense of polysemous words, and it is useful for many applications, such as machine translation (MT), lexical substitution, information retrieval (IR), and biomedical applications. In this paper, we propose the use of multiple contextual features, including the predicate-argument structure and named entities, to train two commonly used classifiers, Naïve Bayes (NB) and Maximum Entropy (ME), for word sense disambiguation. Experiments are conducted to evaluate the classifiers' performance on the OntoNotes corpus and are compared with classifiers trained using a set of baseline features, such as the bag-of-words, n-grams, and part-of-speech (POS) tags. Experimental results show that incorporating both predicate-argument structure and named entities yields higher classification accuracy for both classifiers than does the use of the baseline features, resulting in accuracy as high as 81.6% and 87.4%, respectively, for NB and ME.

Keywords: Word Sense Disambiguation, Predicate-Argument Structure, Named Entity, Natural Language Processing.

1. Introduction

A given word may have multiple meanings, and incorrect word sense recognition may reduce system effectiveness in semantic-oriented applications. Word sense disambiguation (WSD) identifies the correct sense of polysemous words, and it has emerged as a useful technique for

* Department of Information Management, Yuan-Ze University, Chung-Li, Taiwan, R.O.C.
E-mail: lcyu@saturn.yzu.edu.tw

+ Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan, R.O.C.
E-mail: chwu@csie.ncku.edu.tw

Department of Computer Science and Information Engineering, National Chiayi University, Chiayi, Taiwan, R.O.C.
E-mail: ralph@mail.ncyu.edu.tw

many applications, such as machine translation (MT) (Carpuat & Wu, 2007; Chan *et al*., 2007), lexical substitution (McCarthy, 2002; Dagan *et al*., 2006), information retrieval (IR) (Agirre *et al*., 2010), and biomedical applications (Schuemie *et al*., 2005; Stevenson *et al*., 2012). For example, in machine translation, WSD can be used to determine the correct translation for an ambiguous word. In lexical substitution, it is used to determine whether or not a target word can be replaced by another word (*e.g*., a near synonym) by determining whether both words share a common sense. Currently, WSD has been a critical component in the SemEval workshop[1] series (Kilgarriff & Palmer, 2000; Edmonds & Kilgarriff, 2002; Agirre *et al*., 2009).

Navigli (2009) provides an extensive survey of WSD approaches, investigating various features and machine learning algorithms to address specific tasks. For example, bag-of-words, n-grams, part-of-speech (POS) tags, and syntactic and semantic information have been used to build WSD systems with machine learning algorithms (Lee & Ng, 2002; Ando, 2006; Tratz *et al*., 2007; Cai *et al*., 2007; Agirre & Lopez de Lacalle, 2007; Specia *et al*., 2007). Word sense annotated corpora, such as SemCor (Miller *et al*., 1993), LDC-DSO (Ng & Lee, 1996), Hinoki (Kasahara *et al*., 2004), and sense annotated corpora constructed with the help of Web users (Chklovski & Mihalcea, 2002) are also useful resources for building WSD systems. This paper proposes the use of multiple contextual features, including the predicate-argument structure and named entities, to train two commonly used classifiers: Naïve Bayes (NB) and Maximum Entropy (ME) from the OntoNotes corpus, a multilingual corpus of large-scale semantic annotations, including word senses, predicate-argument structure, ontology linking, and coreference (Hovy *et al*., 2006; Pradhan *et al*., 2007a). We then examine whether the two proposed features can improve WSD performance.

The rest of this work is organized as follows. Section 2 gives a brief description for the OntoNotes Corpus. Section 3 presents the features used to train classifiers for WSD. Section 4 summarizes the experimental results. Conclusions are drawn in Section 5.

2. Word Sense Annotation in OntoNotes Corpus

The OntoNotes corpus contains a set of sentences with word senses annotated. In the word sense inventory, the sense definitions of words are created by manually grouping fine-grained sense distinctions obtained from WordNet (Fellbaum, 1998) and dictionaries into more coarse-grained senses. There are two reasons for this grouping instead of using WordNet senses directly. First, people have trouble distinguishing many of the WordNet-level distinctions in real text and make inconsistent choices; thus, the use of coarse-grained senses can improve inter-annotator agreement (ITA) (Palmer *et al*., 2004; 2006). Second, improved

[1] http://www.senseval.org

ITA enables machines to more accurately learn how to perform sense tagging automatically. Sense grouping in OntoNotes has been calibrated to ensure that ITA averages at least 90%. Table 1 shows the OntoNotes sense tags and definitions for the word *arm* (noun sense). Once the sense definitions are created, the sense of words in the sentences can be annotated. To accomplish this goal, the sentences containing the words in the inventory are retrieved first. For each target word (*i.e.*, a word in the inventory) in the sentences, its sense is annotated by two annotators, according to its sense definitions in the inventory. If the two annotators agree on the same sense, then their selection is stored in the corpus. Otherwise, the sense annotation is double-checked by an adjudicator for final decision. Recently, the OntoNotes corpus has been used for many applications, including the SemEval-2007 evaluation (Pradhan *et al.*, 2007b), sense merging (Snow *et al.*, 2007), class imbalance problems (Zhu & Hovy, 2007), sense pool verification (Yu *et al.*, 2007; 2010), parsing and named entity recognition (Finkel & Manning, 2009), semantic role labeling (Che *et al.*, 2010), and coreference resolution (Pradhan *et al.*, 2011).

Table 1. OntoNotes sense tags and definitions. The WordNet version is 2.1.

Sense Tag	Sense Definition	WordNet sense
arm.01	The forelimb of an animal	WN.1
arm.02	A weapon	WN.2
arm.03	A subdivision or branch of an organization	WN.3
arm.04	A projection, a narrow extension of a structure	WN.4 WN.5

3. The WSD System

The features used to build the WSD system include POS tags, local collocations, bag-of-words, named entities, and predicate-argument structure. These features are extracted from the OntoNotes corpus as follows.

Part-of-Speech (POS) tags: This feature includes the POS tags in the positions (P_{-3}, P_{-2}, P_{-1}, P_0, P_1, P_2, P_3), relative to the POS tag of the target word. For instance, the POS sequence of the constituent "…mediator in an attempt to break the…" is "NN NN IN DT TO VB DT".

Local Collocations: This feature includes single words and multi-word n-grams. The single words include (W_{-3}, W_{-2}, W_{-1}, W_0, W_1, W_2, W_3), relative to the target word W_0. Similarly, the multi-word n-grams include ($W_{-2,-1}$, $W_{-1,1}$, $W_{1,2}$, $W_{-3,-2,-1}$, $W_{-2,-1,1}$, $W_{-1,1,2}$, $W_{1,2,3}$). For instance, the multi-word n-grams of the above example constituent include {in_an, an_to, to_break, mediator_in_an, in_an_to, an_to_break, to_break_the}.

Bag-of-Words: This feature can be considered a global feature, consisting of 5 words prior to and after the target word, without regard to position.

Named Entity: OntoNotes Release 1.0[2] provides 18 types of named entities, such as PERSON, ORGANIZATION, GPE, LOCATION, and PRODUCT.

Predicate-Argument Structure: The predicate-argument structure captures the semantic relations between the predicates and their arguments within a sentence. Consider the following example sentence.

> [$_{\textbf{Arg0}}$ The New York arm of the London-based firm] auctioned off [$_{\textbf{Arg1}}$ the **estate** of John T. Dorrance Jr., the Campbell's Soup Co. heir,] [$_{\textbf{ArgM-TMP}}$ last week].

The argument label Arg0 is usually assigned to the *agent*, *causer*, and *experiencer*, while Arg1 is usually assigned to the *patient*. The ArgM-TMP represents a temporal modifier (Babko-Malaya, 2006; Palmer *et al*., 2005). The predicate-argument structure of the above sentence is illustrated in Figure 1. The semantic relations can be either *direct* or *indirect*. A direct relation is used to model a verb-noun (VN), whereas an indirect relation is used to model a noun-noun (NN) relation. Additionally, an NN-relation can be built from the combination of two VN-relations with the same predicate. Table 2 presents some examples. For instance, NN1 can be built by combining VN1 and VN2. Therefore, the two features, VN1 and NN3, can be used to disambiguate the noun *arm* [3].

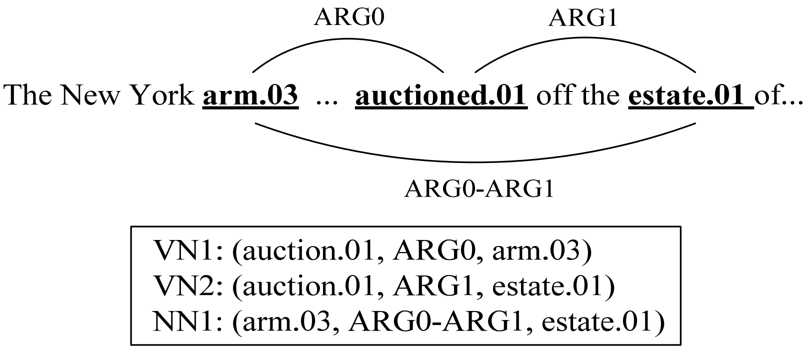

ARG0
ARG1
The New York arm.03 ... auctioned.01 off the estate.01 of...
ARG0-ARG1
VN1: (auction.01, ARG0, arm.03)
VN2: (auction.01, ARG1, estate.01)
NN1: (arm.03, ARG0-ARG1, estate.01)

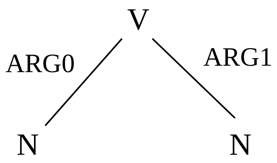
Figure 1. Example of predicate-argument structure.

Table 2. Examples of VN and NN-relations.

Relation Type	Example
VN relation V $\overset{\text{ARG1}}{\text{————}}$ N	VN1: (auction.01, Arg0, arm.03) VN2: (auction.01, Arg1, **estate.01**) VN3: (auction.01, ArgM-TMP, <DATE>)

[2] http://www.ldc.upenn.edu/Catalog/docs/LDC2007T21/ontonotes-1.0-documentation.pdf.

[3] Our WSD system does not include the sense identifier (except for the target word) for word-level training and testing.

NN relation: 	NN1: (arm.03, Arg0-Arg1, **estate.01**) NN2: (**estate.01**, Arg1-ArgM-TMP, <DATE>) NN3: (arm.03, Arg0-ArgM-TMP, <DATE>)

4. Experimental Results

4.1 Experimental Setup

OntoNotes Release 1.0 was used as the experimental corpus, with a total of 992 words in the sense inventory. Not all words, however, were polysemous, and some had a small number of sense annotated sentences. Therefore, we selected 477 polysemous words (247 nouns and 230 verbs) with at least 30 annotated sentences as the test data for the WSD task (see Table 3). The annotated sentences then were used to train two classifiers, Naïve Bayes (NB) and Maximum Entropy (ME), using the features presented in the previous section. We first trained the two classifiers using the baseline features, including the POS tag, local collocations, and bag-of-words. The named entities and predicate-argument structure then were added into both classifiers to determine whether these two features could improve WSD performance. The baseline classifier used for comparison was implemented using the principle of *most frequent sense* (MFS), with each word sense distribution retrieved from the OntoNotes corpus. The evaluation metric was accuracy, defined as the number of correctly identified senses (sentences) divided by the total number of test sentences.

Table 3. Statistics of the experimental data

		Nouns	Verbs
Num. of words (> 30 sentences)		247	230
Senses per word	Min.	2	2
	Avg.	3.26	3.58
	Max.	10	20
Sentences per sense	Min.	30	30
	Avg.	206	197
	Max.	3,053 (share.02)	2,551 (have.03)

4.2 Comparative Results

Table 4 shows the experimental results with 10-fold cross validation. The symbols B, PA, and NE in Table 4 represent the baseline features, predicate-argument structure, and named entities, respectively. For comparison of the classifiers, ME outperformed NB for all feature sets. For comparison of the feature sets, both B+PA and B+PA+NE outperformed B for both NB and ME, indicating that using both predicate-argument structure and named entities can improve performance over using the baseline features alone. Another observation is that the predicate-argument structure was more sensitive to ME than to NB because the improvement of B+PA over B in ME was greater than that in NB. Conversely, the named entity was more sensitive to NB.

Table 4. Comparative results of WSD accuracy for different features and classifiers.

Feature Types	Nouns			Verbs			ALL		
	MFS	NB	ME	MFS	NB	ME	MFS	NB	ME
B	0.820	0.810	0.865	0.764	0.797	0.856	0.793	0.805	0.862
B + PA		0.814	0.873		0.807	0.869		0.811	0.872
B + PA + NE		0.819	0.875		0.812	0.871		0.816	0.874

For more detailed analysis, Tables 5 and 6 list the WSD accuracy for parts of the nouns and verbs in the OntoNotes inventory. These words were also included in the SemEval-2007 English Lexical Sample Task (Pradhan *et al.*, 2007b).

Table 5. WSD accuracy for parts of the nouns.

Noun	# sense	MFS	NB			ME		
			B	B+PA	B+PA+NE	B	B+PA	B+PA+NE
authority	5	0.474	0.904	0.935	0.926	0.904	0.939	0.917
base	6	0.353	0.696	0.758	0.725	0.717	0.754	0.758
bill	4	0.668	0.872	0.881	0.887	0.895	0.901	0.916
carrier	8	0.765	0.704	0.808	0.815	0.758	0.792	0.819
chance	4	0.486	0.750	0.773	0.809	0.714	0.736	0.759
condition	3	0.713	0.800	0.823	0.839	0.806	0.803	0.842
defense	7	0.282	0.493	0.603	0.597	0.533	0.537	0.543
development	3	0.760	0.877	0.895	0.881	0.886	0.926	0.898
drug	2	0.684	0.783	0.811	0.845	0.791	0.789	0.800
effect	4	0.719	0.823	0.850	0.858	0.866	0.850	0.896
exchange	5	0.731	0.887	0.921	0.920	0.914	0.921	0.934

future	3	0.797	0.965	0.965	0.952	0.969	0.970	0.962
hour	4	0.854	0.847	0.880	0.882	0.863	0.888	0.873
job	3	0.780	0.738	0.757	0.768	0.809	0.849	0.845
management	2	0.618	0.837	0.853	0.866	0.821	0.840	0.806
network	3	0.605	0.750	0.788	0.824	0.705	0.750	0.736
order	8	0.722	0.871	0.877	0.892	0.876	0.869	0.883
part	5	0.702	0.915	0.900	0.907	0.944	0.931	0.940
people	4	0.912	0.917	0.915	0.918	0.925	0.933	0.937
point	9	0.737	0.853	0.851	0.875	0.885	0.877	0.870
policy	2	0.806	0.829	0.841	0.837	0.858	0.876	0.851
position	7	0.304	0.639	0.656	0.670	0.645	0.659	0.656
power	3	0.508	0.774	0.790	0.828	0.782	0.777	0.769
president	3	0.843	0.945	0.955	0.959	0.942	0.953	0.954
rate	2	0.924	0.944	0.933	0.940	0.946	0.943	0.955
source	5	0.368	0.803	0.841	0.833	0.797	0.844	0.830
space	5	0.565	0.741	0.782	0.794	0.829	0.788	0.806
state	4	0.830	0.840	0.848	0.855	0.858	0.858	0.857
system	6	0.544	0.728	0.749	0.751	0.722	0.717	0.705
Average	**4.45**	**0.657**	**0.811**	**0.836**	**0.843**	**0.826**	**0.837**	**0.839**

Table 6. WSD accuracy for parts of the verbs.

Verb	# sense	MSF	NB			ME		
			B	B+PA	B+PA+NE	B	B+PA	B+PA+NE
build	4	0.805	0.830	0.827	0.825	0.837	0.821	0.809
call	11	0.661	0.736	0.775	0.756	0.784	0.793	0.792
close	7	0.743	0.919	0.934	0.930	0.898	0.936	0.920
come	20	0.580	0.657	0.701	0.728	0.732	0.753	0.767
consider	2	0.788	0.840	0.836	0.852	0.875	0.891	0.905
cut	9	0.680	0.750	0.798	0.780	0.792	0.795	0.778
end	3	0.839	0.795	0.790	0.778	0.899	0.902	0.890
follow	7	0.666	0.766	0.795	0.804	0.756	0.825	0.825
get	14	0.447	0.656	0.682	0.676	0.721	0.748	0.748
go	18	0.275	0.545	0.599	0.585	0.617	0.672	0.664

grow	4	0.836	0.857	0.864	0.869	0.864	0.879	0.880
hold	11	0.667	0.737	0.756	0.759	0.754	0.764	0.749
keep	6	0.477	0.575	0.574	0.599	0.612	0.625	0.634
lead	3	0.417	0.859	0.882	0.886	0.870	0.895	0.891
leave	3	0.602	0.704	0.710	0.745	0.739	0.723	0.783
look	5	0.667	0.871	0.894	0.880	0.891	0.923	0.915
lose	6	0.709	0.806	0.829	0.867	0.835	0.835	0.861
make	13	0.336	0.557	0.616	0.613	0.604	0.664	0.670
put	12	0.677	0.757	0.756	0.755	0.789	0.808	0.806
raise	2	0.784	0.728	0.736	0.721	0.747	0.752	0.747
set	8	0.382	0.591	0.619	0.610	0.628	0.639	0.641
spend	2	0.700	0.878	0.972	0.969	0.885	0.987	0.991
take	20	0.663	0.611	0.619	0.616	0.670	0.684	0.683
tell	2	0.960	0.974	0.970	0.978	0.973	0.974	0.982
turn	16	0.285	0.591	0.635	0.623	0.705	0.726	0.723
Average	**8.32**	**0.626**	**0.744**	**0.767**	**0.768**	**0.779**	**0.801**	**0.802**

The “# sense” column lists the number of sense distinctions of a word, and the column “MFS” presents the sense distribution among all senses of the word. Both the number of sense distinctions and the sense distribution of words may affect WSD performance. Generally, a large number of sense distinctions with an even distribution may lead to confusion among the classifiers, hence, lower performance. For example, the noun *defense* in Table 5 has seven senses, and the proportion of the major sense is 0.282, indicating an even distribution (the distribution of the 7 senses is {.14, .18, .19, .08, .04, .28, .09} in the OntoNotes corpus), thus yielding low accuracy. The verbs *go* and *make* in Table 6 also have similar results. Conversely, a small number of sense distinctions with a skewed distribution may have better performance. For example, in Table 5, the noun *rate* with a dominant sense of 0.924 yielded high accuracy, as did the verb *tell* in Table 6.

To further analyze the effect of the sense distribution of words in the whole corpus, we ranked the 247 nouns and 230 verbs in OntoNotes in descending order based on the proportion of their major senses. Nouns and verbs with major sense proportions within a given range then were grouped together (*e.g.*, >=0.95, 0.90~0.95, 0.85~0.90, …, 0.35~0.4, and <0.35), and their average accuracy was calculated for comparison. Figures 2~4 present the results of nouns, verbs, and all words, respectively, with accuracy gradually decreasing as the sense becomes more evenly distributed. Another interesting observation is that, although ME outperformed NB, ME and NB achieved similar performance when the sense distribution became more evenly distributed (the proportion of the major sense <0.65).

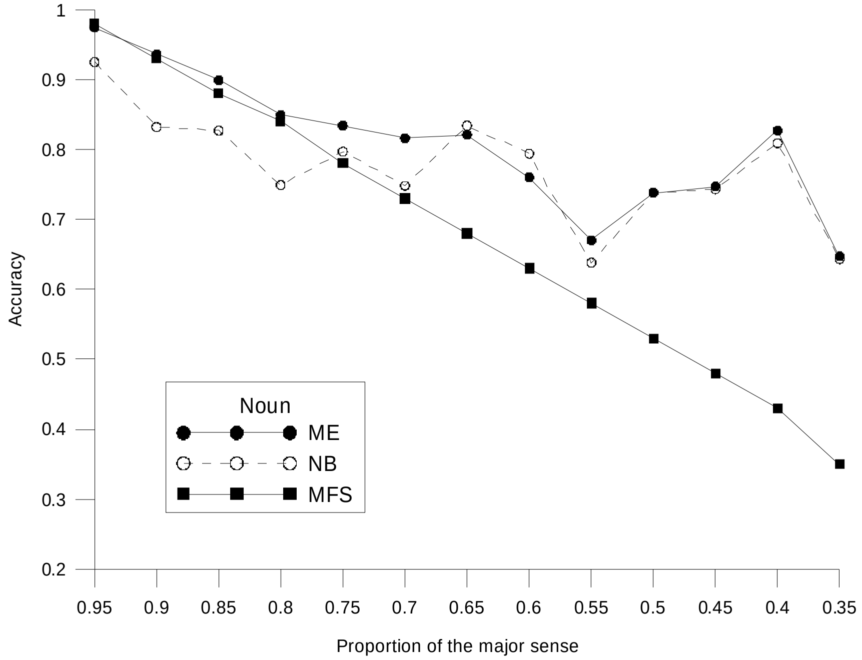


Figure 2. WSD performance against sense distribution for nouns.

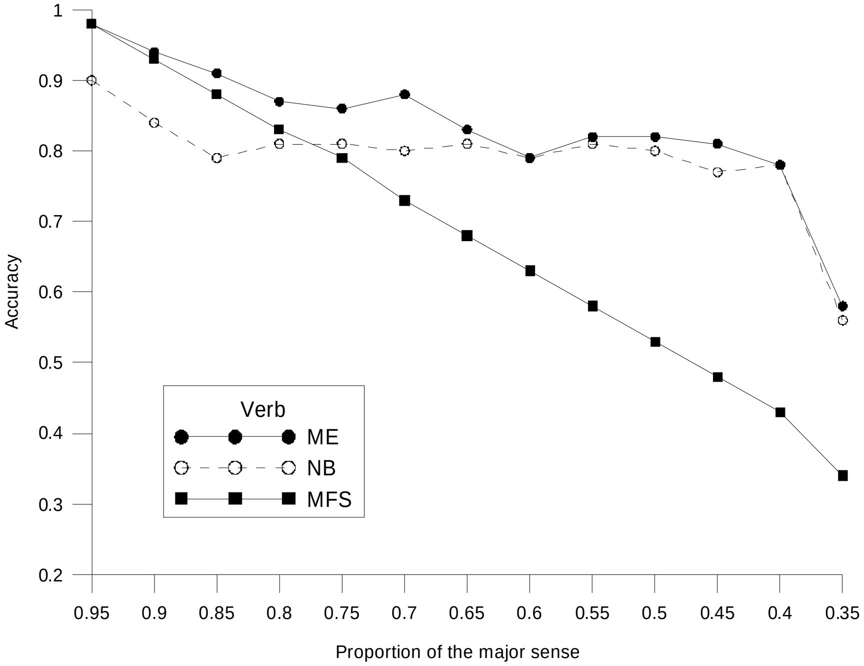


Figure 3. WSD performance against sense distribution for verbs.

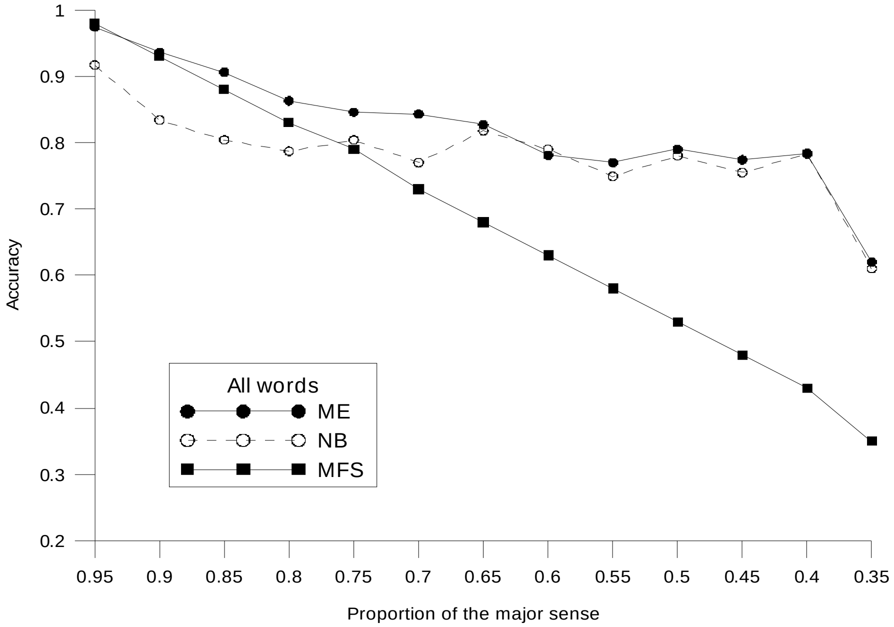


Figure 4. WSD performance against sense distribution for all words.

5. Conclusion

A WSD system was built from the OntoNotes corpus using multiple contextual features to analyze the effect of sense distribution on WSD performance. Experimental results show that both the predicate-argument structure and named entities improved WSD performance. In addition, there was a tendency for a skewed sense distribution to yield higher performance than evenly distributed word senses. Future work will focus on improving WSD performance by investigating more significant features and more effective machine learning algorithms.

Acknowledgments

This work was partially supported by the National Science Council, Taiwan, R.O.C., under Grant No. NSC 99-2221-E-155-036-MY3. The authors would also like to thank Dr. Eduard Hovy for his support.

References

Agirre, E., & Lopez de Lacalle, O. (2007). UBC-ALM: Combining k-NN with SVD for WSD. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 342-345.

Agirre, E., Ma`rquez, L., & Wicentowski, R. (2009). Computational Semantic Analysis of Language: SemEval-2007 and Beyond. *Lang Resources and Evaluation*, 43(2), 97-104.

Agirre, E., Otegi, A., & Zaragoza, H. (2010). Using Semantic Relatedness and Word Sense Disambiguation for (CL)IR. *Lecture Notes in Computer Science*, 6241, 166-173.

Ando, R.K. (2006). Applying Alternating Structure Optimization to Word Sense Disambiguation. In *Proc. of CoNLL-06*, 77-84.

Babko-Malaya, O. (2006). PropBank Annotation Guidelines.

Cai, J.F., Lee, W.S., & The, Y.W. (2007). Improving Word Sense Disambiguation Using Topic Features. In *Proc. of EMNLP/CoNLL-07*, 1015-1023.

Carpuat, M., & Wu, D. (2007). Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proc. of EMNLP/CoNLL-07*, 61-72.

Chan, Y.S., Ng, H.T., & Chiang, D. (2008). Word Sense Disambiguation Improves Statistical Machine Translation. In *Proc. of ACL-07*, 33-40.

Che, W., Liu, T., & Li, Y. (2010). Improving Semantic Role Labeling with Word Sense. In *Proc. of HLT/NAACL-10*, 246-249.

Chklovski, T., & Mihalcea, R. (2002). Building a Sense Tagged Corpus with Open Mind Word Expert. In *Proc. of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions at ACL-02*, 116-122.

Dagan, I., Glickman, O., Gliozzo, A., Marmorshtein, E., & Strapparava, C. (2006). Direct Word Sense Matching for Lexical Substitution. In *Proc. of COLING/ACL-06*, 449-456.

Edmonds, P, & Kilgarriff, A. (2002). Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems. *Natural Language Engineering*, 8(4), 279-291.

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Finkel, J. R., & Manning, C. D. (2009). Joint Parsing and Named Entity Recognition. In *Proc. of HLT/NAACL-09*, 326-334.

Hovy, E.H., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proc. of HLT/NAACL-06*, 57-60.

Kasahara, K., Sato, H., Bond, F., Tanaka, T., Fujita, S., Kanasugi, T., & Amano, S. (2004). Construction of a apanese Semantic Lexicon: Lexeed. In *IPSG SIG: 2004-NLC-159*, Tokyo, 75-82.

Kilgarriff, A., & Palmer, M. (2000). Introduction, Special Issue on SENSEVAL: Evaluating Word Sense Disambiguation Programs. *Computer and the Humanities*, 34(1-2), 1-13.

Lee, Y.K., & Ng, H.T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. In *Proc. of EMNLP*, 41-48.

McCarthy, D. (2002). Lexical Substitution as a Task for WSD Evaluation. In *Proc. of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation at ACL-02*, 109-115.

Miller, G., Leacock, C., Tengi, R., & Bunker, R. (1993). A Semantic Concordance. In *Proc. of the 3rd DARPA Workshop on Human Language Technology*, 303-308.

Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41(2), Article 10.

Ng, H.T., & Lee, H.B. (1996). Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Exemplar-based Approach. In *Proc. of ACL-96*, 40-47.

Palmer, M., Babko-Malaya, O., & Dang, H.T. (2004). Different Sense Granularities for Different Applications. In *Proc. of the 2nd International Workshop on Scalable Natural Language Understanding at HLT/NAACL-04*.

Palmer, M., Dang, H.T., & Fellbaum, C. (2006). Making Fine-grained and Coarse-grained Sense Distinctions, Both Manually and Automatically. *Journal of Natural Language Engineering*, 13, 137-163.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), 71-106.

Pradhan, S., Hovy, E.H., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2007a). OntoNotes: A Unified Relational Semantic Representation. In *Proc. of the First IEEE International Conference on Semantic Computing (ICSC-07)*, 517-524.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007b). SemEval-2007 Task 17: English Lexical Sample, SRL and All Words. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 87-92.

Pradhan, S., Ramshaw, L., Marcus, M., Palmer, M., Weischedel, R., & Xue, N. (2011). CoNLL-2011 Shared Task: Modeling Unrestricted Coreference in OntoNotes. In *Proc. of CoNLL-11*, 1-27.

Schuemie, M.J., Kors, J.A., & Mons, B. (2005). Word Sense Disambiguation in the Biomedical Domain: An Overview. *Journal of Computational Biology*, 12(5), 554-565.

Snow, R., Prakash, S., Jurafsky, D., & Ng, A.Y.(2007). Learning to Merge Word Senses. In *Proc. of EMNLP/CoNLL-07*, 1005-1014.

Specia, L., Stevenson, M., & das Gracas V. Nunes, M. (2007). Learning Expressive Models for Word Sense Disambiguation. In *Proc. of ACL-07*, 41-48.

Stevenson, M., Agirre, E., & Soroa, A. (2012). Exploiting Domain Information for Word Sense Disambiguation of Medical Documents. *Journal of the American Medical Informatics Association*, 19(2), 235-240.

Tratz, S., Sanfilippo, A., Gregory, M., Chappell, A., Posse, C., & Whitney, P. (2007). PNNL: A Supervised Maximum Entropy Approach to Word Sense Disambiguation. In *Proc. of the 4th International Workshop on Semantic Evaluations (SemEval-2007) at ACL-07*, 264-267.

Yu, L.C., Wu, C.H., Philpot, A., & Hovy, E.H. (2007). OntoNotes: Sense Pool Verification Using Google N-gram and Statistical Tests. In *Proc. of the OntoLex Workshop at the 6th International Semantic Web Conference (ISWC 2007).*

Yu, L.C., Wu, C.H., Chang, R.Y., Liu, C.H., & Hovy, E.H. (2010). Annotation and Verification of Sense Pools in OntoNotes. *Information Processing and Management*, 46(4), 436-447.

Zhu, J., & Hovy, E.H. (2007). Active Learning for Word Sense Disambiguation with Methods for Addressing the Class Imbalance Problem. In *Proc. of EMNLP/CoNLL-07*, 783-790.

Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 193-218

Using Linguistic Features to Predict Readability of Short Essays for Senior High School Students in Taiwan[1]

Wei-Ti Kuo[*], Chao-Shainn Huang[*], and Chao-Lin Liu[*]

Abstract

We investigated the problem of classifying short essays used in comprehension tests for senior high school students in Taiwan. The tests were for first and second year students, so the answers included only four categories, each for one semester of the first two years. A random-guess approach would achieve only 25% in accuracy for our problem. We analyzed three publicly available scores for readability, but did not find them directly applicable. By considering a wide array of features at the levels of word, sentence, and essay, we gradually improved the F measure achieved by our classifiers from 0.381 to 0.536.

Keywords: Computer-assisted Language Learning, Readability Analysis, Document Classification, Short Essays for Reading Comprehension.

1. Introduction

Reading is a key competence for language learners. For learners of English as a Second Language (ESL), reading provides a crucial channel for learners to integrate and exercise the knowledge of previously learned vocabulary and grammar. If we could provide appropriate material to ESL learners, they would receive individualized stimulus, maintain the motivation to learn, and benefit more from reading activities. Hence, researchers have been investigating the readability of articles and books for a long time (Flesch, 1948).

In recent decades, research about readability has not been confined to just classifying the readability of articles. In large-scale language tests that include a writing assessment, grading the writing of a large number of test takers is very time consuming. Moreover, maintaining a consistent grading standard over the group of graders is also a challenge. Hence, techniques

[1] This paper was converted from the technical report for an NSC-funded undergraduate practice project, NSC-98-2815-C-004-003-E, for Wei-Ti Kuo. This paper was revised from its Chinese version that was previously published in the ROCLING XXII conference (Huang *et al*., 2010).

[*] Department of Computer Science, National Chengchi University, Taipei, Taiwan
E-mail: {94703041, 97753007, chaolin}@nccu.edu.tw
The author for correspondence is Chao-Lin Liu.

for automated grading were studied and introduced in the Scholastic Aptitude Test (**SAT**[2]) in the USA (Burstein *et al.*, 2003; Attali & Burstein, 2006; Chang *et al.*, 2006).

In a broader sense, the problems of determining the readability of articles and judging the scores of essays are specialized instances of text classification. They are similar in that text materials are categorized based on some selected metrics, and they differ in the implications of the classification results.

Early work in readability analysis considered the frequency of words, number of sentences, and length of sentences (Flesch, 1948; Kincaid *et al.*, 1975; Chall & Dale, 1995). These methods may seem deficient nowadays, but it was not easy to consider all conceivable factors when the training corpora and the computing power were not sufficient. Other factors clearly are relevant to readability (Bailin & Grafstein, 2001), and one may consider more lexical level information, such as the hypernyms and hyponyms of words in an article, to determine the readability (Lin *et al.*, 2009). Higher levels of information, such as the structure of the articles, semantic information, and cognition-related connotation, may also be included in readability analysis (Crossley *et al.*, 2008).

Depending on the purpose of classifying the textual material, a classifier should consider factors of various aspects. Linguistic features are obvious candidates, but psycholinguistic, educational, and cultural factors are important as well. Moreover, characteristics of the readers and writers of the essays should also be considered. Classifications of articles written by native speakers and non-native speakers might be quite different. Good reading materials for second graders of native and non-native speakers would vary in terms of their vocabulary and content.

In this study, we examine short essays that were designed for reading comprehension tests at the high school level in Taiwan. Essays were classified based on a comprehensive list of lexical and syntactic features that were extracted from the words, sentences, and paragraphs in a given essay. The essays used in the experiments were realistic; therefore, they were limited in regards to the available amount. We focused on 845 tests for the first four semesters in high school, so essays were classified into four categories that corresponded to the semester of the examinee. We explored the applications of several machine learning models for the classification task, and the best F_1 measure (Witten & Frank, 2005) that we achieved was only 0.536.

We understand that there is room to improve our work, in terms of both the scale of experiments and the achieved results of accuracy. The current experience, however, supports a popular viewpoint that lexical and syntactic information about the short essays are

[2] We highlighted acronyms of phrases and special terms with boldface and blue text to help readers find their meanings.

instrumental but are not sufficient for predicting readability (Bailin & Grafstein, 2001). Some deep analysis is required to achieve better results. For instance, the set of a reading comprehension test consists of a short essay and questions for the students to answer. The set of a reading comprehension test may be considered more difficult because of its questions, not just because of its essay. Analyzing the questions is a major step for us to complete the current study.

We introduce the data source and their preprocessing in Section 2, deal with the extracted lexical features in Section 3, discuss the syntactic features in Section 4, present and compare the effects of using different combinations of the features to predict the readability in Section 5, and make some concluding remarks in Section 6.

2. Background

To make the results of this study close to reality, we obtained essays for comprehension tests for students at senior high schools in Taiwan. The essays were retrieved from the item pool that was designed for the San-Min version (三民版) of English courses, and the item pool was published in the 96th school year. The 96th school year spanned August 2007 to July 2008.

The item pool was designed for preparing competence examinations that are similar to the SAT in the USA. Students apply for college during the fifth semester in high school in Taiwan. Hence, the contents of the item pool covered only English for the first two years in senior high school and we treated a semester as a level in our experiments.

The goal of our work was to determine the level of the short essay of a given comprehension test. Namely, we classified an essay into one of four possible levels.

Table 1 shows the number of essays that we gathered from the item pool. The original essays were classified according to their levels and “tracks”. The test items were designed for three tracks of English courses. The first track was designed by Ling-Hsia Chen (陳凌霞) of National Taiwan University, and we denote this track as NTUC in Table 1. The other two tracks were designed by Kwock-Ping John Tse (謝國平) of Providence University. One of these two tracks was more recent than the other. We denote the relatively more recent one as PUTN and the older as PUTO.

The words used in the comprehension tests were chosen based on the expected competence of the students. In Taiwan, the Ministry of Education (MOE) has issued a ruling about what words middle school graduates are expected to be acquainted with (MOE, 2008). Partially because of this constraint, essays for the comprehension tests contained Chinese translations for selected words. The numbers of the essays that did not contain Chinese translations were counted, and the totals are placed under the column “No Hints”. The total number of Chinese words that appeared in the essays was placed under “Chinese Hints” in

Table 1. Chinese translations were provided in the essays for special nouns, such as names, places, and medical terms, in order to avoid the disturbance of these challenging words against comprehension.

Table 1. Data source

	NTUC	PUTN	PUTO	Row Total	No Hints	Chinese Hints (words)
Level 1	47	117	36	200	124	142
Level 2	64	127	36	227	199	45
Level 3	48	127	36	211	148	151
Level 4	45	126	36	207	198	14
Total	204	497	144	845	669	352

The appearance of Chinese translations could be considered as a noise in the original data, but it could also be considered as a feature. We took the latter position in some of our experiments and ignored the Chinese translations in some experiments. The statistics in Table 1 suggested that the appearance of Chinese translations was related to the levels. On average, there were fewer Chinese translations for the second semester of each school year.

Figure 1 shows the major steps we used to convert an essay into a feature vector. We first removed and recorded the Chinese translations from the original essay, as we discussed in the previous paragraphs. The remaining English texts were then processed by the Stanford Part-of-Speech (POS) tagger[3] and the Stanford parser[4] to extract the lexical and syntactic features. Except for the Stanford NLP tools, we relied on word lists that were selected by experts (*cf.* Section 3.1), the CMU Pronouncing dictionary[5] (*cf.* Section 3.2), and Dr.eye[6] dictionary (*cf.* Section 3.3) to broaden the types of lexical level information that we could extract.

Some linguistic features intuitively are related to the difficulty of essays, *e.g.*, the number of sentences, the number of words, the popularity (frequency) of words, the number of senses a word can carry, and the number of complex sentences. We applied tools and dictionaries for analyzing the linguistic features to create feature vectors (*cf.* Section 4).

Some basic features could be extracted easily. We calculated the number of sentences (N) in an essay and collected the following features: the number of tokens (f_1), the number of punctuations (f_2), the number of tokens and punctuations (f_3=f_1+f_2), the average number of

[3] http://nlp.stanford.edu/software/tagger.shtml

[4] http://nlp.stanford.edu/software/lex-parser.shtml

[5] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[6] http://www.dreye.com/

tokens per sentence (f_4=f_1/N), the average number of punctuations per sentence (f_5= f_2/N), and the average number of tokens and punctuations per sentence (f_6=f_3/N).

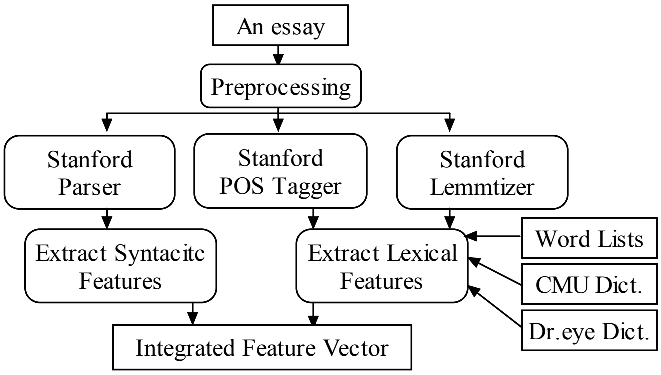


Figure 1. Converting an essay into an instance

3. Lexical Level Features

Words are the basic building blocks of essays. For ESL learners, learning basic vocabulary is an important first step into the world of English. According to the MOE's standards of course design for elementary education (MOE, 2008), graduates from middle schools should have learned and should be able to apply 1200 basic English words in daily conversations. In this section, we explain various types of lexical level features that we extracted from words in an essay.

3.1 Word Lists

Due to the crucial role of individual words in learning English, experts compiled different word lists for different purposes. We employed three lists in our work. Table 2 shows the detailed statistics of the NTNU, GETP, and CEEC word lists.

Professors at National Taiwan Normal University compiled a list of words for a competition related to English words, and we refer to this list as the **NTNU** list[7]. The NTNU list classifies words into three major groups – elementary, middle, and senior high schools – that are further divided for the targeted grades. For instance, "E34" is for the third and the fourth grades in elementary schools; and M3 is for the third year in middle school.

The General English Proficiency Test[8] (**GEPT**) is a standardized test accepted by domestic and some international institutions. To provide references for test takers, the GEPT offers word lists for different levels of test takers. Three of the lists were relevant to our work:

[7] http://vq.ie.ntnu.edu.tw/

[8] https://www.gept.org.tw/

Elementary, Intermediate, and High-Intermediate. These three lists include words that people who have graduated from middle schools, high schools, and colleges (non-English majors), respectively, should have learned.

The College Entrance Examination Center[9] (**CEEC**) is an institution for managing the college entrance examinations in Taiwan. The word list is designed for graduates of high schools and includes nearly 9000 words. This CEEC list contains 6 grades.

Table 2. Statistics about word lists

Word Lists	Level	# of Words	Total # of Words
NTNU	E34	498	6041
	E5	250	
	E6	250	
	M1	350	
	M2	350	
	M3	407	
	S1	936	
	S2	1500	
	S3	1500	
GEPT	Elementary	2184	7853
	Intermediate	2560	
	High-Intermediate	3109	
CEEC	G1	1775	8976
	G2	1490	
	G3	1472	
	G4	1350	
	G5	1543	
	G6	1346	

We employed the Stanford NLP tools to tokenize the strings in an essay, as we illustrated in Figure 1. We lemmatized the tokens and identified their POS tags. After this step, we looked up the word lists to see which level the tokens belonged to and updated the frequencies of the levels. Similar to how Dale-Chall dealt with their word list (Dale & Chall, 1995), a word not belonging to any level was considered to belong to the “difficult” level, which is an additional level not listed in Table 2.

[9] http://www.ceec.edu.tw/

We created feature vectors based on the NTNU, GEPT, and CEEC lists separately. With the above procedure, we created 10 features for an essay when we considered the NTNU list – 9 levels in Table 2 and one "difficult" level. Analogously, we had 4 and 7 features for the GEPT and CEEC lists, respectively.

We expected these features to be useful for the essay classification under the premise that, if an essay contains more words in the more advanced levels, the essay should be more difficult.

3.2 Pronunciation

For ESL learners in Taiwan, an English word with relatively more syllables is generally more difficult to remember and pronounce. This is partially due to the fact that Chinese is a tonal language and students may not be used to words with several syllables yet.

Based on this observation, we thought it might be worthwhile to explore the influence of the number of syllables on the readability of essays. Although not all long words are difficult and not all short words are easy, it was interesting to explore the intuitive impression.

After obtaining the lemmatized tokens in an essay, we looked at the CMU Pronouncing dictionary (**CMUPD**) to find the number of syllables in the tokens. The CMUPD contains more than 125000 words. The pronunciation of an English word is represented with English letters and numbers. The pronunciation of "university" is shown below.

Y UW2 N AH0 V ER1 S AH0 T IY0

Vowels and consonants are separated in CMUPD, and only vowels are followed by digits. The digits indicate stresses: 0 for no stress, 1 for primary stress, and 2 for secondary stress.

Given the CMUPD phoneme notation, we could compute the number of syllables in an English word and the total number of vowels and consonants in a word. Take "university" as an example. This token has 5 syllables and 10 vowels and consonants. In our corpus, a token may have at most 7 syllables and at most 16 vowels and consonants. If a token was not covered by CMUPD, we would record that this token had no syllables, no vowels, and no consonants.[10]

[10] We employed distributions of some random variables as features in this paper, and we generally used larger numbers to denote relatively more difficult cases. For instance, when creating features for word lists, larger indices indicated higher grade and more challenging words. Here, we converted the number of syllables of a word into a sequence of features. The first feature denoted the number of words with one syllable, the second feature denoted the number of words with two syllables, *etc*. We used the zero-th feature to denote the number of words not covered by CMUPD. This would not confuse the classifiers that we tried in Section 5 because the semantics of the order of the features was not explicit to the classifiers.

For a given essay, we would record the frequencies of tokens that have *i* syllables and *j* vowels and consonants, where *i* is in the range [0, 7] and *j* is in the range [0, 16]. Therefore, we had 25 features related to pronunciation of tokens in an essay.

3.3 Lexical Ambiguity

Ambiguity may not be just a problem for natural language processing of computers; it could be a problem for ESL learners as well. Many English words carry multiple possible meanings. If an essay contains many words with multiple possible meanings, its contents may become relatively difficult to understand. Based on this intuition, we considered the distribution of the numbers of translated senses of words in an essay as features.

Finding the number of translated senses of an English word took a little bit of work. Using the Stanford POS tagger, we could find the POS of a token. The POS tag followed the Penn TreeBank convention[11]. Also, we used Dr.eye to find the Chinese translations of English words. Dr.eye only has a very rough POS system: noun, transitive verb, intransitive verb, adjective, adverb, preposition, pronoun, conjunction, and determiner. Therefore, we had to convert a POS tag in the Penn TreeBank system into a category in Dr.eye. We employed the classification in a CEEC publication[12], and considered only 8 different POS tags. The conversion of POS tags was conducted based on the mapping listed in Table 3.

Table 3. Converting a POS in Penn TreeBank system to Dr.eye's category

POS tags	Stanford POS Tagger	Dr.eye
Noun	NN, NNS, NNP, NNPS	n.
Verb	MD, VB, VBD, VBG, VBN, VBP, VBZ	vt., vi.
Adjective	CD, JJ, JJR, JJRS	a.
Adverb	EX, RB, RBR, RBS, RP, WRB	ad.
Preposition	IN, TO	prep.
Pronoun	DT, PRP, PRP$, WDT, WP, WP$, WRB	pron.
Conjunction	CC, IN	conj.
Determiner	DT	art.

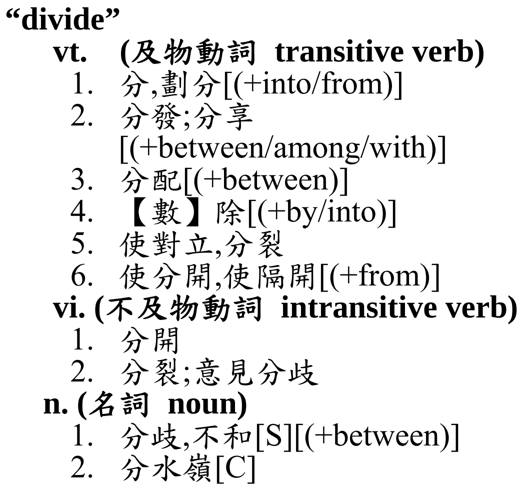
Note that the conversion was imperfect. The POS "IN" could be mapped to conjunction and preposition. When we encountered a token with "IN," we checked Dr.eye to see if the token could be used as a conjunction. If yes, that token was considered a conjunction. Otherwise, the token was considered a preposition.

[11] ftp://ftp.cis.upenn.edu/pub/treebank/doc/manual/root.ps.gz

[12] http://www.ceec.edu.tw/Research/paper_doc/ce37/6.pdf

In Dr.eye, an English word can have at most 43 translated senses. We considered the number of translated senses as a feature. A token that could not be found in Dr.eye would be considered to have no translated senses. Hence, the distribution of the number of translated senses of tokens in an essay consisted of 44 numbers.

Figure 2 shows the entry for "divide" in Dr.eye. Assuming that we have a "divide/VBD" in an essay; we would know that this "divide" was a verb and would consider that this word had 8 possible translated senses.



"divide"

vt. (及物動詞 transitive verb)

1. 分,劃分[(+into/from)]
2. 分發;分享 [(+between/among/with)]
3. 分配[(+between)]
4. 【數】除[(+by/into)]
5. 使對立,分裂
6. 使分開,使隔開[(+from)]

vi. (不及物動詞 intransitive verb)

1. 分開
2. 分裂;意見分歧

n. (名詞 noun)

1. 分歧,不和[S][(+between)]
2. 分水嶺[C]

Figure 2. The entry for "divide" in Dr.eye

4. Syntactic Level Features

We collected information not just about the words in an essay, but we also attempted to find useful syntactic information as features for the classification task. This is necessary because simple words in complex sentences may not be easy to understand.

A sentence may be complex for different reasons. We considered the **depth**s of parse trees as an indication. Figure 3 shows a parse tree for the sentence, "I liked playing basketball when I was young." Let the root, *i.e.*, ROOT, of the tree be Level 0, and its child node, *i.e.*, S, be Level 1. The deepest node in this tree is Level 9. We refer to the level of the deepest node in a tree as its depth.

We parsed sentences in our corpus with the Stanford parser (using the PCFG grammar file EnglishPCFG.ser.gz) and asked for only the parse trees with the highest score. In our corpus, the depth of the deepest tree was 31. We used features to represent the distribution of the depths of parse trees in an essay: (d_0, d_1, d_2, …, d_8). We increased d_k by 1 if the depth of a parse tree was k when k<8; and increased d_8 by 1 if the depth of a parse tree was larger than

8[13].

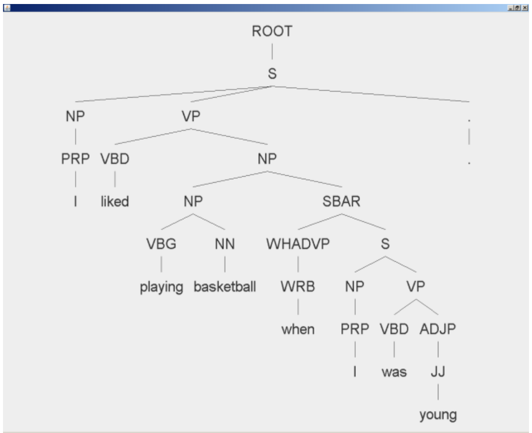


Figure 3. A sample parse tree

Given an essay, we could analyze every sentence to obtain its depth, and we recorded the average depth and the distribution of the depths of sentences in this essay.

Other than its depth, a sentence may be complex because it employs some rarely used grammatical relationships. The parse tree in Figure 3 includes several grammatical relationships: "S → NP VP .," "VP →VBD NP," "SBAR → WHADVP S," VP → VBD ADJP," *etc*. If one or more of these relationships are rare, the sentence may be difficult to read, rendering the essay not easy to understand.

We employed a corpus-based approach to determine whether or not a grammatical relationship was rare. We collected more than 7000 sentences from web sites that provide educational resources. They included "Shi Yuan You Grammar"[14], "1200 Fundamental English sentences"[15], "Learning Resources for Middle Schoolers"[16], and "I-Lan County Language Resources for Middle Schoolers"[17]. We parsed the collected sentences and recorded the frequencies of the grammatical relationships in these sentences.

We observed 985 grammatical relationships in these 7000+ sentences. Only 8 relationships occurred more than 1000 times, and 62 relationships took place more than 100 times.

As the span of the frequencies was wide and the distribution of the frequencies was

[13] Although one might expect that d_0 and d_1 should not appear in regular essays, we left these possibilities to avoid weird strings that might appear in our corpus.

[14] http://tw.myblog.yahoo.com/jw!GFGhGimWHxN4wRWXG1UDIL_XSA--/

[15] http://hk.geocities.com/cnlyhhp/eng.htm

[16] http://siro.moe.edu.tw/fip/index.php

[17] http://140.111.66.37/english/ (last visited 2010/8/14, but not functioning at the time of writing)

irregular, we quantized the ranges of the frequencies into 6 segments by the frequency binning method (Witten and Frank, 2005). The 985 relationships that we observed appeared at 127 different frequencies. We ordered them from frequent to infrequent ones and treated relationships that appeared the same number of times as the same relationship. Each segment contained 21 different frequencies (except the last segment, which covered 22 frequencies). We could consider these 6 segments of rules as "very frequent," "frequent," "slightly frequent," "slightly infrequent," "infrequent," and "very infrequent". (The choice of 6 was arbitrary. We did not try other selections.)

Given the above procedure, we could generate a vector of 7 components that considered the "rareness" of grammatical relationships in a sentence: {"very frequent," "frequent," "slightly frequent," "slightly infrequent," "infrequent," "very infrequent," "unseen"}. In a sentence containing 8 grammatical relationships, 2 very frequent, 4 frequent, 1 infrequent, 1 very infrequent, and 1 unseen, in our training corpus, we would convert it to {2, 4, 0, 0, 1, 1, 1}.

For an essay with many sentences, we could generate a 7-item vector for each sentence, and we took the average of every item to create the 7-item vector for the essay. An essay that includes a relatively larger number of rare grammatical relationships may be more difficult to read.

5. Experimental Evaluation

We classified the short essays reported in Section 2 with Weka (Witten and Frank, 2005), using different combinations of features reported in Sections 3 and 4.

Before we evaluated our methods, we acquired the SMOG scores of our essays via a Web-based service[18]. Equation (1) shows the score function for SMOG, where m represents the number of polysyllables and n is the number of sentences. A word is considered a polysyllable if it contains three or more syllables. Essays with higher SMOG scores are relatively harder to read.

$$1.043 \times \sqrt{m \times \frac{30}{n}} + 3.1291 \tag{1}$$

Table 4 shows basic statistics about the SMOG scores of our essays. The smallest, largest, and average SMOG scores increased with the levels of the essays quite impressively. This is probably a good reason for the popularity of this simple formula.

[18] Simple Measure of Goobledygook (SMOG). http://www.harrymclaughlin.com/SMOG.htm

Table 4. Basic statistics of SMOG scores

	Level 1	Level 2	Level 3	Level 4
Smallest SMOG score	6.59	7.22	6.75	7.3
Largest SMOG score	17.11	19.88	22.75	22.09
Average SMOG score	10.889	11.822	12.554	12.757

Nevertheless, if we looked into the details of the scores for individual essays, we would realize that assessing the readability of an individual essay is not easy. Figure 4 shows distributions of the SMOG scores of our essays of different levels. We quantized the SMOG with 0.5 as an interval, and accumulated the essays within an interval to draw the chart. The vertical axis shows the proportion of essays of a level for a given SMOG score interval (on the horizontal axis). Although the chart is quite complex to read, the curves clearly show that essays of easier levels may have higher SMOG scores than essays of harder levels.

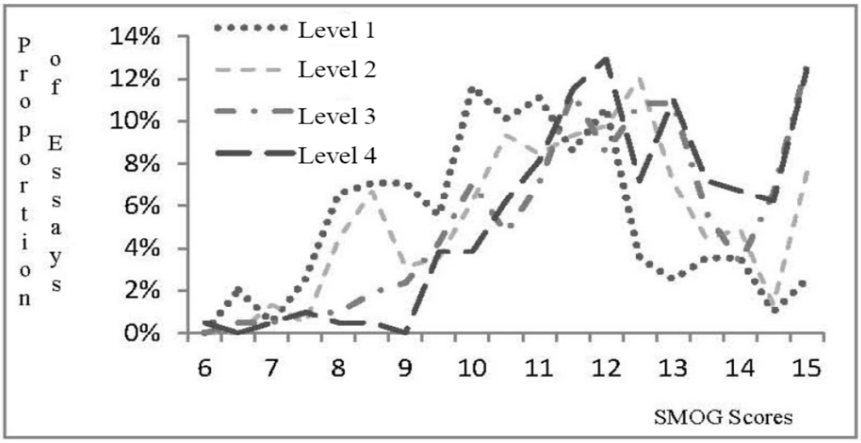


Figure 4. Distributions of SMOG scores for different levels of essays

5.1 Basic Features and Measures of the Prediction Quality

Since we had several different types of features, we grouped them to streamline our experiments. **Group A** consisted of features discussed in Sections 3.2 and 3.3: 8 features of the distribution of the number of syllables, 17 features of the distribution of the number of vowels and consonants, and 44 features of the distribution of lexical ambiguities.

Group B consisted of f_4, f_5, and f_6 in Section 2; the average depth; and the distribution of the depths of parse trees in an essay in Section 4. In total, we have 36 (=3+1+32) features in this group.

Group C consisted of the word lists in Section 3.1. We use **Ca**, **Cb**, and **Cc** to represent features generated based on the NTNU, GEPT, and CEEC word lists, respectively.

Whenever necessary, we normalized the statistics with the number of words and the number of sentences in a given essay. This is an important step to reduce the impact of different lengths of essays. In Group A, features about pronunciation were normalized by the number of words; the feature about the distribution of the number of lexical ambiguities would be normalized by the number of words. In Group B, the distribution of the depths of parse trees would have been normalized by the number of sentences in the essay. In Group C, the word counts of different levels would be normalized by the total number of words in the essay.

We ran 10-fold cross-validation with features in Group A, B, Ca, Cb, and Cc separately, using the J48 decision tree model, LMT decision tree model, Artificial Neural Networks (ANNs), and Ridor rules leaner. We did not do a random restart when we ran ANNs, and we set the number of epochs to 500 and learning rate to 0.3.

We measured the classification quality with the F_1 measure. F_1 measure is the harmonic average of recall rate and precision rate for a classification task (*cf.* Witten & Frank, 2005), and it is usually referred as the **F measure**. The **recall rate** achieved in a classification task is the proportion of instances that belong to the targeted classes captured by the classifier. The **precision rate** achieved in a classification task is the proportion of correct decisions of the classifier when it classifies instances as the targeted class.

5.2 Performance Achieved by the Basic Features

Table 5 shows the F measures achieved by individual groups of features. The best F measure was achieved when we used Cb with LMT, and the worst F measure occurred when we used Cc with J48. Table 5 also shows the column and row averages. The column averages indicate the effectiveness of a feature group, and the row averages show the effectiveness of a classifier.

Table 5. F measures achieved by individual groups of features

	A	B	Ca	Cb	Cc	Average
J48	0.297	0.270	0.297	0.335	**0.248**	0.289
LMT	0.334	0.318	0.300	**0.353**	0.264	0.314
ANN	0.278	0.291	0.340	0.323	0.268	0.300
Ridor	0.293	0.291	0.307	0.304	0.261	0.291
Average	0.301	0.293	0.311	0.329	0.260	0.299

When the feature groups were applied separately, Cc might offer inferior effects because it contained words specifically for senior-high school levels and could not provide sufficient information about relatively easier words. The column averages indicate that using Ca or Cb word lists achieved better classification quality than not using word lists, *i.e.*, A and B.

Comparing the averages, we found that Cb and LMT are, respectively, the best individual feature group and the best classifier in Table 5.

Recall that we classified essays into one of four possible levels. Hence, a purely random guess is expected to achieve only 25% in accuracy. Although there are no good ways to compare F measures and accuracy directly, the F measures listed in Table 5 were not very encouraging.

Table 6 shows the F measures that were achieved when we combined the basic features to predict the readability. Again, the results were achieved in 10-fold cross-validations. The best F measure was 0.381 when we combined Groups B, Ca, and Cb in the predication task. The worst F measure was 0.261 when we combined Groups A, B, and Ca in the task. Again, the row averages indicate that the best classifier is LMT.

Table 6. F measures achieved by combining basic features

	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc	Average
J48	0.281	0.266	0.275	0.293	0.293	0.299	0.274	0.3	0.306	0.312	0.290
LMT	0.335	0.345	0.337	0.346	0.344	0.341	0.3	0.348	0.338	0.364	0.340
ANN	0.283	0.348	0.33	0.318	0.315	0.319	0.303	0.346	0.347	0.324	0.323
Ridor	0.288	0.291	0.323	0.312	0.322	0.356	0.253	0.319	0.341	0.346	0.315
Average	0.297	0.313	0.316	0.317	0.319	0.329	0.283	0.328	0.333	0.337	**0.317**

	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc	Average
J48	**0.261**	0.303	0.301	0.291	0.307	0.325	0.303	0.295	0.304	0.286	0.298
LMT	0.331	0.327	0.35	0.341	0.321	0.35	**0.381**	0.349	0.366	0.358	0.347
ANN	0.359	0.309	0.328	0.313	0.33	0.323	0.319	0.352	0.314	0.357	0.330
Ridor	0.321	0.329	0.305	0.33	0.31	0.323	0.307	0.299	0.302	0.326	0.315
Average	0.318	0.317	0.321	0.319	0.317	0.330	0.328	0.324	0.322	0.332	**0.323**

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc	Average
J48	0.305	0.305	0.313	0.302	0.317	0.33	0.312
LMT	0.329	0.353	0.369	0.341	0.373	0.358	0.354
ANN	0.335	0.335	0.362	0.338	0.333	0.324	0.338
Ridor	0.349	0.314	0.329	0.334	0.354	0.362	0.340
Average	0.330	0.327	0.343	0.329	0.344	0.344	**0.336**

Using more features allowed us to achieve better results. The best possible F measure increased from 0.353 in Table 5 to 0.381. The overall average of Table 5 is 0.299, indicating the average performance of our classifiers when we used only one feature group. The overall average of the first (upper) part of Table 6 is 0.317, the overall average of the second part is

0.323, and the overall average of the third part is 0.336. These averages show the average performance when using two groups, three groups, and more than three groups of features. Hence, we observed that using more features groups led to steady improvement in the average prediction quality.

Figure 5 shows the trends of improving performance for our classifiers when we employed more feature groups. The legends show the number of feature groups used with the classifiers, where ">3" indicates four or five groups.

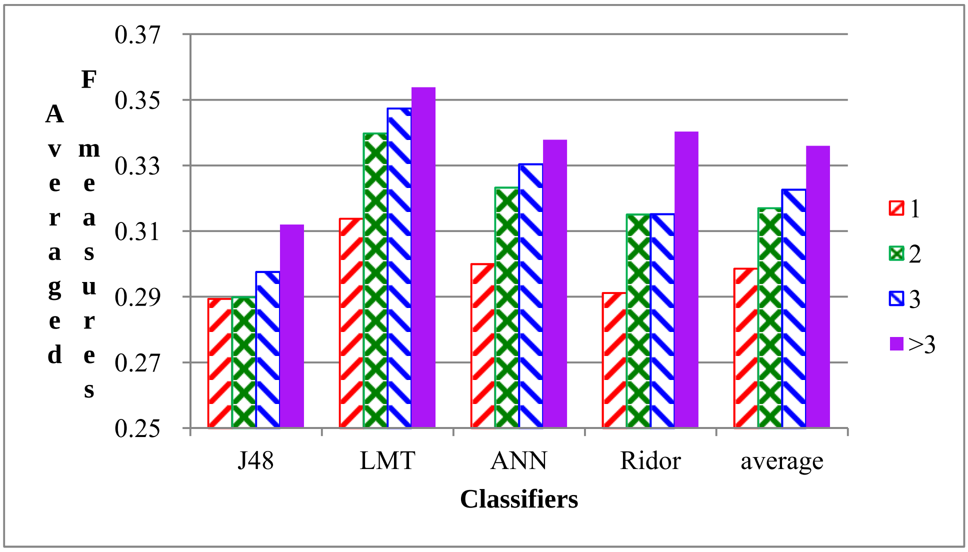


Figure 5. Using more feature groups improve the prediction quality on average

5.3 Frequencies of Grammatical Relationships

We refer to the frequency distribution of the grammatical relationships (Section 4) as **Group D**. Assume that there are two sentences in an essay, and the frequency distributions of their grammatical relationships are {0,0,2,0,0,1,3} and {1,1,0,5,5,0,4}. There are 22 grammatical relationships in this example. We add these distributions and divide each item by 22 to acquire a normalized distribution {0.045, 0.045, 0.091, 0.227, 0.227, 0.045, 0.318}.

We repeated the six experiments in Table 6. We considered the three experiments that had the best F measures (B+Ca+Cb, B+Ca+Cb+Cc, and Cb+Cc) and three word lists adding Groups A and B. The upper part of Table 7 is copied from the data in Table 6, and the lower part of Table 7 shows the F measures of the new experiments.

Table 7. Effects of including Group D

Before	A+B+Ca	A+B+Cb	A+B+Cc	B+Ca+Cb	B+Ca+Cb+Cc	Cb+Cc	Average
J48	**0.261**	0.303	0.301	0.303	0.317	0.312	0.300
LMT	0.331	0.327	0.350	**0.381**	0.373	0.364	0.354
ANN	0.359	0.309	0.328	0.319	0.333	0.324	0.329
Ridor	0.321	0.329	0.305	0.307	0.354	0.346	0.327
Average	0.318	0.317	0.321	0.328	0.344	0.337	0.327
After	A+B+Ca	A+B+Cb	A+B+Cc	B+Ca+Cb	B+Ca+Cb+Cc	Cb+Cc	Average
J48	**0.251**	0.294	0.294	0.309	0.318	0.309	0.296
LMT	0.346	0.342	0.325	0.343	**0.357**	0.345	0.343
ANN	0.327	0.339	0.306	0.308	0.351	0.327	0.326
Ridor	0.320	0.302	0.302	0.346	0.346	0.326	0.324
Average	0.311	0.319	0.307	0.327	0.343	0.327	0.322

Evidently, adding Group D in these experiments did not change the F measures significantly. Possible reasons for the observed irrelevancy include the fact that we determined the distributions based on another corpus of ours (Section 4), whose contents were designed for middle school students. The distribution of grammatical relationships in a corpus for middle schools may not be closely relevant to the readability of essays for senior high schools. Another possible reason is that Group D is in fact not relevant to readability.

5.4 Essay-Level Features

Although we normalized many features by the total number of sentences and the total number of words in an essay, we wondered about the potential contributions of the essay-level features. They include the total number of sentences, the total depth of parse trees, the number of tokens (f_1, Section 2), the number of punctuations (f_2, Section 2), the number of tokens and punctuations (f_3, Section 2), and the number of Chinese hints (Table 1 in Section 2); we refer to them as **Group E**.

We repeated the same set of experiments that we conducted for Table 7. This time, both Group D and Group E were used. Table 8 shows the F measures that we observed. The statistics suggest that using Group D and Group E helped us improve the prediction quality. As we have discussed in Section 2 about Table 1, the appearance of Chinese hints is noticeably related to the levels of the short essays. Hence, the improvement introduced by Group E was not very surprising.

Table 8. Effects of including Groups D and E

	A+B+Ca	A+B+Cb	A+B+Cc	B+Ca+Cb	B+Ca+Cb+Cc	Cb+Cc	Average
J48	0.338	0.314	0.349	0.333	0.331	0.347	0.335
LMT	0.412	0.405	0.374	0.423	**0.425**	0.412	0.409
ANN	0.370	0.353	0.345	0.352	0.402	0.363	0.364
Ridor	0.337	0.36	**0.312**	0.353	0.377	0.341	0.347
Average	0.364	0.358	0.345	0.365	0.384	0.366	0.364

5.5 Distribution of Parts of Speech

It was suggested that we explore the influence of the distribution of the POS tags of the words in an essay. We considered the eight categories of POS tags in Section 3.3 to create features. We added these new features and repeated the experiment B+Ca+Cb+Cc+D+E in Table 8. Table 9 shows a comparison of the achieved F measures before and after adding the distribution.

Table 9. Influences of distribution of POSes

	B+Ca+Cb+Cc+D+E	After adding dist. of POSes
J48	0.331	0.349
LMT	0.425	0.425
ANN	0.402	0.346
Ridor	0.377	0.343

With this limited scale of experiment, we could not reach a decisive conclusion about the effectiveness of the distribution of POS tags. The observed insignificance may result from the distribution of POS tags possibly remaining steady if we study the distribution in a large corpus (Shih, 2000) or might result from the distribution not being relevant to readability.

5.6 Articles with Chinese Hints

In Section 5.4, we investigated the contribution of using the number of Chinese hints as a feature for the classification task. Now, we explore the implications of whether an essay had Chinese hints or not on the predictability of its readability. We separated the essays into two sub-groups: those having Chinese hints and those having no Chinese hints; we then repeated the experiments for Table 5 and Table 6. Note that we removed the Chinese hints when we classified the essays that originally contained Chinese hints.

Table 10 and Table 11 show the F measures observed when we repeated the experiments with essays that originally contained Chinese hints. It was quite surprising to find that all F measures in Table 10 and Table 11 are better than their counterparts in Table 5 and Table 6, without any exceptions. The best F measure is now 0.536.

Table 10. Using individual groups for essays with Chinese hints

	A	B	Ca	Cb	Cc
J48	0.423	0.364	0.472	0.428	0.382
LMT	0.435	0.404	0.494	0.466	0.363
ANN	0.429	0.396	0.467	**0.52**	0.396
Ridor	**0.353**	0.365	0.364	0.424	0.385

Table 11. Using mixed groups for essays with Chinese hints

	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc
J48	0.342	0.335	0.406	0.364	0.427	0.349	0.343	0.437	0.443	0.406
LMT	0.4	0.479	0.432	0.458	0.487	0.507	0.402	0.49	0.493	0.493
ANN	0.404	0.406	0.424	0.471	0.457	0.389	0.422	0.475	0.406	0.439
Ridor	0.395	0.375	0.413	0.364	0.381	0.424	0.366	0.462	0.401	0.456
	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc
J48	0.345	0.342	0.342	0.412	0.348	0.394	0.353	0.441	0.391	0.429
LMT	0.46	0.477	0.391	0.449	0.489	0.457	0.485	0.438	0.507	**0.536**
ANN	0.42	0.4	0.364	0.444	0.444	0.44	0.421	0.457	0.436	0.402
Ridor	0.397	0.435	0.355	0.374	0.377	0.45	0.431	0.438	0.369	0.457

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc
J48	**0.33**	0.369	0.353	0.419	0.348	0.371
LMT	0.471	0.47	0.49	0.46	0.465	0.458
ANN	0.448	0.422	0.442	0.48	0.382	0.453
Ridor	0.412	0.387	0.473	0.35	0.424	0.416

Table 12 and Table 13 show the F measures observed when we repeated the experiments with essays that did not contain Chinese hints originally. Most of the F measures in Table 12 and Table 13 are better than their counterparts in Table 5 and Table 6, but some of them became worse. The best F measure in Table 13 is better than the best one in Table 6, but it is just 0.414.

Table 12. Using individual groups for essays without Chinese hints

	A	B	Ca	Cb	Cc
J48	0.297	**0.254**	0.344	0.315	0.297
LMT	0.356	0.295	0.378	0.349	0.308
ANN	0.358	0.286	**0.417**	0.372	0.28
Ridor	0.324	0.3	0.351	0.378	0.276

Table 13. Using mixed groups for essays without Chinese hints

	A+B	A+Ca	A+Cb	A+Cc	B+Ca	B+Cb	B+Cc	Ca+Cb	Ca+Cc	Cb+Cc
J48	0.265	0.320	0.320	0.280	0.350	0.341	**0.256**	0.386	0.345	0.382
LMT	0.345	0.381	0.401	0.375	0.387	0.366	0.324	0.378	0.400	0.392
ANN	0.327	0.413	0.368	0.339	0.323	0.337	0.267	0.403	0.383	0.366
Ridor	0.334	0.374	0.353	0.323	0.356	0.341	0.317	0.384	0.377	0.381

	A+B+Ca	A+B+Cb	A+B+Cc	A+Ca+Cb	A+Ca+Cc	A+Cb+Cc	B+Ca+Cb	B+Ca+Cc	B+Cb+Cc	Ca+Cb+Cc
J48	0.322	0.327	0.307	0.330	0.316	0.372	0.334	0.356	0.347	0.359
LMT	0.384	0.356	0.335	0.391	0.393	0.393	0.399	0.382	**0.414**	0.385
ANN	0.365	0.352	0.362	0.393	0.382	0.343	0.347	0.349	0.35	0.404
Ridor	0.349	0.36	0.382	0.340	0.335	0.368	0.391	0.333	0.33	0.367

	A+B+Ca+Cb	A+B+Ca+Cc	A+B+Cb+Cc	A+Ca+Cb+Cc	B+Ca+Cb+Cc	A+B+Ca+Cb+Cc
J48	0.352	0.294	0.348	0.380	0.310	0.345
LMT	0.386	0.388	0.369	0.381	0.396	0.400
ANN	0.380	0.379	0.349	0.390	0.353	0.389
Ridor	0.352	0.346	0.367	0.344	0.375	0.334

The F measures reported in Tables 5, 6, 10, 11, 12, and 13 suggested that the natures of essays with and without Chinese hints are different. The chart in Figure 6 shows the average performance of our classifiers when we used 1, 2, 3, and more than 3 feature groups to classify the essays that originally contained Chinese hints. The chart in Figure 7 shows the trends for predicting the levels of the essays that did not contain Chinese hints originally. The charts in Figures 5, 6, and 7 indicate that we achieved the worst performance when we mixed the essays in the corpus. If we separated those essays with and without Chinese hints, we achieved better results for both sub-groups on average. This is quite an interesting discovery, but we do not have a good explanation for this phenomenon.

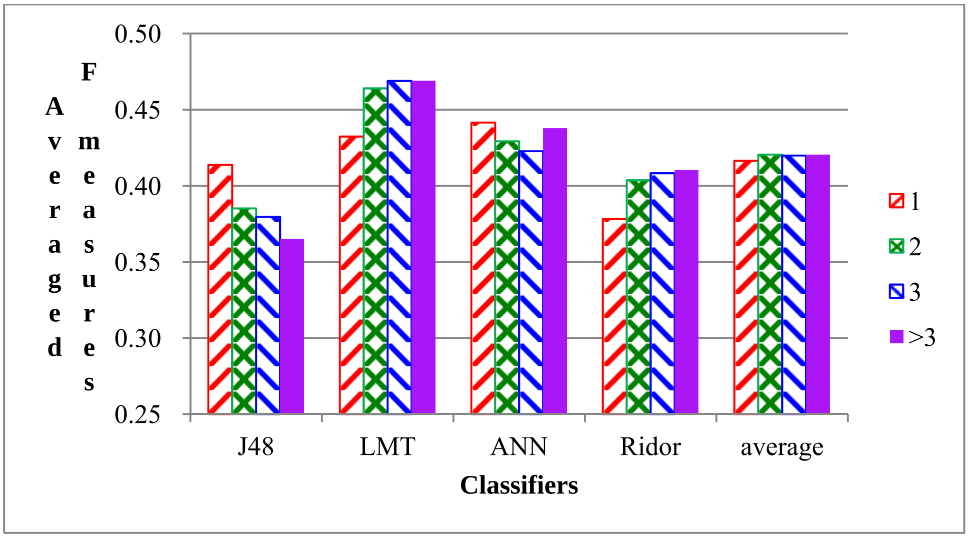


Figure 6. Predicting readability of essays with Chinese hints was easier

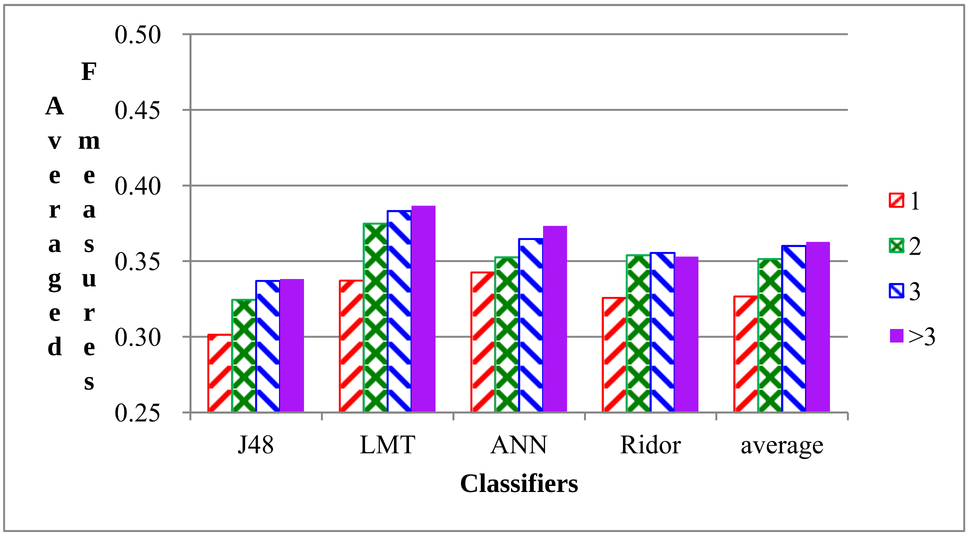


Figure 7. Predicting readability of essays without Chinese hints was harder

5.7 More Experiments with Syntactic Features

Finally, we explored some conjectural features at the syntax level, and we referred to them as **Group F**. We parsed our corpus with the Stanford parser to collect some statistics: (1) VBN appeared at most 4 times; (2) VP appeared at most 6 times; (3) MD appeared at most 3 times. Hence, we could use 16 features to describe the distributions of VBN, VP, and MD in an essay. In addition, we could use binary features to encode whether an essay contained ADJP, ADVP, and CONJP. This gave us 3 features. Adding the features for distribution of depth (9 features) and grammatical relationships (7 features), we had a total 35 features in Group F.

We repeated the experiments for Tables 11 and 13, after adding Group F to the combinations of features. Results reported in Table 14 are for essays that originally contained Chinese hints, and results reported in Table 15 are for essays that did not contain Chinese hints.

Table 14. Results of adding syntactic features for essays with Chinese hints

	F	A+F	B+F	Ca+F	Cb+F	Cc+F	A+B+F	A+Ca+F	A+Cb+F	A+Cc+F
J48	0.340	0.374	0.388	0.383	0.417	0.33	0.354	0.393	0.399	0.386
LMT	0.438	0.478	0.389	0.472	0.476	0.447	0.426	0.501	0.484	0.467
ANN	0.352	0.422	0.375	0.363	0.368	0.412	0.410	0.440	0.454	0.414
Ridor	0.390	0.426	0.373	0.328	0.385	0.298	0.385	0.403	0.386	0.377

	B+Ca+F	B+Cb+F	B+Cc+F	Ca+Cb+F	Ca+Cc+F	Cb+Cc+F	A+B+Ca+F
J48	0.321	0.348	0.345	0.386	0.390	0.428	0.321
LMT	0.450	0.529	0.385	0.461	0.457	0.484	0.459
ANN	0.376	0.400	0.363	0.380	0.418	0.435	0.408
Ridor	0.416	0.381	0.341	0.418	0.361	0.395	0.368

	A+B+Cb+F	A+B+Cc+F	A+Ca+Cb+F	A+Ca+Cc+F	A+Cb+Cc+F	B+Ca+Cb+F
J48	0.360	0.341	0.488	0.408	0.412	0.378
LMT	0.482	0.461	0.482	**0.530**	0.494	0.465
ANN	0.448	0.378	0.417	0.430	0.430	0.380
Ridor	0.415	0.373	0.423	0.384	0.393	0.384

	B+Ca+Cc+F	B+Cb+Cc+F	Ca+Cb+Cc+F	A+B+Ca+Cb+F	A+B+Ca+Cc+F
J48	0.346	0.383	0.388	0.385	**0.306**
LMT	0.452	0.522	0.496	0.443	0.482
ANN	0.417	0.427	0.417	0.427	0.415
Ridor	0.345	0.400	0.423	0.385	0.404

	A+B+Cb+Cc+F	A+Ca+Cb+Cc+F	B+Ca+Cb+Cc+F	A+B+Ca+Cb+Cc+F
J48	0.381	0.447	0.403	0.379
LMT	0.474	0.480	0.502	0.472
ANN	0.408	0.460	0.392	0.407
Ridor	0.435	0.404	0.375	0.427

Table 15. Results of adding syntactic features for essays without Chinese hints

	F	A+F	B+F	Ca+F	Cb+F	Cc+F	A+B+F	A+Ca+F	A+Cb+F	A+Cc+F
J48	0.279	0.330	0.262	0.314	0.323	0.289	0.298	0.314	0.344	0.286
LMT	0.277	0.348	0.308	0.374	0.361	0.307	0.341	0.343	0.384	0.362
ANN	0.265	0.371	0.279	0.336	0.315	0.301	0.339	0.370	0.388	0.365
Ridor	0.27	0.325	0.289	0.345	0.351	0.295	0.299	0.326	0.360	0.360

	B+Ca+F	B+Cb+F	B+Cc+F	Ca+Cb+F	Ca+Cc+F	Cb+Cc+F	A+B+Ca+F
J48	0.307	0.328	0.290	0.356	0.326	0.329	0.262
LMT	0.366	0.371	0.326	0.395	0.378	0.362	0.353
ANN	0.348	0.327	0.299	0.346	0.348	0.343	0.375
Ridor	0.342	0.320	**0.254**	0.376	0.344	0.366	0.315

	A+B+Cb+F	A+B+Cc+F	A+Ca+Cb+F	A+Ca+Cc+F	A+Cb+Cc+F	B+Ca+Cb+F
J48	0.343	0.302	0.337	0.325	0.343	0.288
LMT	0.388	0.340	0.386	0.340	0.374	0.396
ANN	0.370	0.365	0.378	**0.402**	0.384	0.350
Ridor	0.35	0.317	0.387	0.326	0.345	0.353

	B+Ca+Cc+F	B+Cb+Cc+F	Ca+Cb+Cc+F	A+B+Ca+Cb+F	A+B+Ca+Cc+F
J48	0.275	0.329	0.350	0.341	0.292
LMT	0.348	0.370	0.377	0.378	0.350
ANN	0.374	0.307	0.383	0.374	0.386
Ridor	0.338	0.314	0.362	0.372	0.321

	A+B+Cb+Cc+F	A+Ca+Cb+Cc+F	B+Ca+Cb+Cc+F	A+B+Ca+Cb+Cc+F
J48	0.351	0.346	0.326	0.344
LMT	0.380	0.370	0.389	0.393
ANN	0.376	0.406	0.338	0.378
Ridor	0.349	0.368	0.395	0.358

Although we wished to observe improved results when used these more complex features, the outcome was not encouraging. In general, the F measures in Table 14 were lower than their counterparts in Table 11. For instance, using B+Ca achieved 0.427 in Table 11, but using F+B+Ca achieved only 0.321 in Table 14. The same problem can be verified for corresponding numbers in Table 13 and Table 15. In fact, the drops from the numbers in Table 13 to the corresponding numbers in Table 15 were more severe.

Intuitively, considering syntactic features should have improved our results. Nevertheless, we probably did not choose the right features. Another possibility would be that the challenging levels of the short essays used in the comprehension tests in Taiwan simply did not relate to syntactic factors.

6. Concluding Remarks

A random classification of an essay into four categories would have achieved only 25% in accuracy on average. We considered features at the word, sentence, and essay levels in this classification task, and we found that it was possible to improve the F measure from 0.381 (Table 6) to 0.536 (Table 11). The best F measures were observed in 10-fold cross-validation tests for LMT in Weka. Not all classifiers achieved the same quality of classification. Among the four types of classifiers we used in this study, LMT performed the best on average.

The identified improvement was not small, but it was not significant enough either. The problem of determining levels of readability may not be as easy as the public scores suggested. We analyzed our corpus with the SMOG scores in Section 5.1, and found that the essays of supposedly more challenging levels may not have higher SMOG scores than the scores of the supposedly easer essays.

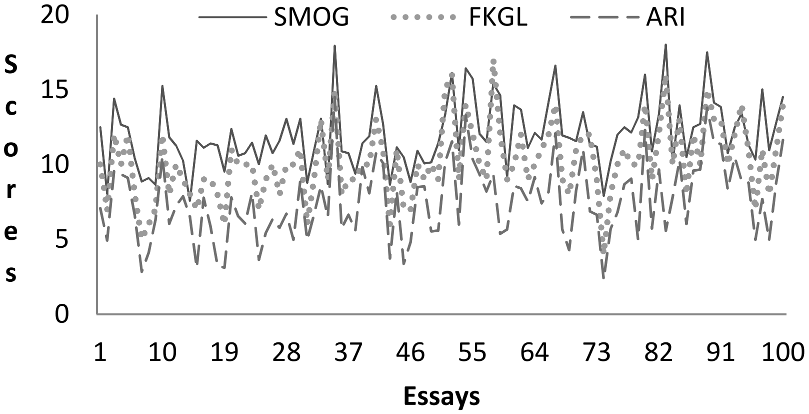


Figure 8. Readability scores of more popular formulae

We explored two additional scores for readability. In Figure 8, we show the SMOG, FKGL[19], and ARI[20] scores for 100 arbitrarily chosen essays from our corpus. The curves show rather strong similarity, which is not very surprising to us. These score functions rely mainly

[19] Flesch-Kincaid Grade Level. http://en.wikipedia.org/wiki/Flesch-Kincaid_Readability_Test

[20] Automated readability index. http://en.wikipedia.org/wiki/Automated_Readability_Index

on the word counts of different levels of words and the number of sentences in an essay. Hence, if using SMOG would not achieve good results for the classification task in our study (*cf.* Figure 4), then using the other two alternatives would not achieve much better results either.

One challenge to our work is whether we should consider only the short essays and classify the levels of the comprehension tests. A comprehension test contains the essay part and the question part. Obviously, we should take the questions into consideration in the classification task, which we have not begun yet. In addition, due to the "examination-centered" style of education in Taiwan, the same short essay may be reused in tests of students of higher classes. Such a reuse of short essays made our classification more difficult, because that made the "correct class" of an essay rather ambiguous.

Whether linguistic features were sufficient for the determination of readability of essays is also an issue. Understanding an essay may require domain-dependent knowledge that we have not attempted to encode with our features (Carrell, 1983). Culture-dependent issues may also play a role (Carrell, 1981). Hence, more features are needed to accomplish more improvement on the predication of readability, *e.g.* (Crossley *et al.*, 2008; Zhang, 2008).

A review comment suggested that there might not be sufficient differences in the short essays used in the first and the second semesters of a school year, so trying to classify the short essays into three levels (each for a school year) may be more practical. Although we did not move our work in this direction, we think the suggestion is interesting.

A reviewer noticed an interesting crossing point in Figure 4. The SMOG score at 11.5 seems to be a major point for the curves in Figure 4 to intersect. A similar phenomenon appeared in Figure 8, where approximately half of the scores of the 100 essays were above 11.5. Whether 11.5 is the watershed of the easy and difficult essays is an interesting hypothesis to verify with a larger amount of essays.

Acknowledgments

The work was supported in part by the funding from the National Science Council in Taiwan under the contracts NSC-97-2221-004-007, NSC-98-2815-C-004-003-E, and NSC-99-2221-004-007. The authors would like to thank Miss Min-Hua Lai for her technical support in this study and Professor Zhao-Ming Gao for his comments on an earlier report (Kuo *et al.*, 2009) of this paper.

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater V.2, *Journal of Technology, Learning, and Assessment*, 4(3), 3-30.

Bailin, A. & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique, *Language and Communication*, 21(2), 285-301, 2001.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays, *IEEE Intelligent Systems*, 18(1), 32-39.

Carrell, P. L. (1981). Culture-specific schemata in L2 comprehension, *Selected Papers from the Ninth Illinois TESOL/BE Annual Convention, the First Midwest TESOL Conference*, 123-132.

Carrell, P. L. (1983). Some issues in studying the role of schemata or background knowledge in second language comprehension, *Reading in a Foreign Language*, 1(1), 81-92.

Chall, J. & Dale, E. (1995). *Readability Revisited: The new Dale-Chall Readability Formula*. Brookline Books.

Chang, T.-H., Lee, C.-H., & Chang, Y.-M. (2006). Enhancing automatic Chinese essay scoring system from figures-of-speech, *Proceedings of the Twentieth Pacific Asia Conference on Language, Information and Computation*, 28-34.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices, *TESOL Quarterly*, 42(3), 475-493.

Flesch, R. (1948). A New Readability Yardstick, *Journal of Applied Psychology*, 32(3), 221-233.

Huang, C.-S., Kuo, W.-T., Lee, C.-L., Tsai, C.-C., & Liu, C.-L. (2010). Using linguistic features to classify texts for reading comprehension tests at the high school levels, *Proceedings of the Twenty Second Conference on Computational Linguistics and Speech Processing* (ROCLING XXIII), 98-112. (in Chinese)

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel, *Technical Report Research Branch Report*, 8-75.

Kuo, W.-T., Huang, C.-S., Lai, M.-H., Liu, C.-L., & Gao, Z.-M. (2009). 適用於中學英文閱讀測驗短文分類的特徵比較, *Proceedings of the Fourteenth Conference on Artificial Intelligence and Applications*. (in Chinese)

Lin, S.-Y., Su, C.-C., Lai, Y.-D., Yang, L.-C., & Hsieh, S.-K. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet, *International Journal of Computational Linguistics and Chinese Language Processing*, 14(1), 45-84.

MOE. (2008). http://www.edu.tw/eje/content.aspx?site_content_sn=15326

Shih, R. H., Chiang, J. Y., & Tien, F. (2000). Part-of-speech sequences and distribution in a learner corpus of English, *Proceedings of Research on Computational Linguistics Conference* XIII (ROCLING XIII), 171-177.

Witten, I. H. & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann.

Zhang, X. (2008). The effects of formal schema on reading comprehension – An experiment with Chinese EFL readers, *International Journal of Computational Linguistics and Chinese Language Processing*, 13(2), 197-214.

Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 219-236

Discovering Correction Rules for Auto Editing

An-Ta Huang*, Tsung-Ting Kuo∗, Ying-Chun Lai+, and Shou-De Lin∗

Abstract

This paper describes a framework that extracts effective correction rules from a sentence-aligned corpus and shows a practical application: auto-editing using the discovered rules. The framework exploits the methodology of finding the Levenshtein distance between sentences to identify the key parts of the rules and uses the editing corpus to filter, condense, and refine the rules. We have produced the rule candidates of such form, A → B, where A stands for the erroneous pattern and B for the correct pattern.

The developed framework is language independent; therefore, it can be applied to other languages. The evaluation of the discovered rules reveals that 67.2% of the top 1500 ranked rules are annotated as correct or mostly correct by experts. Based on the rules, we have developed an online auto-editing system for demonstration at http://ppt.cc/02yY.

Keywords: Edit Distance, Erroneous Pattern, Correction Rrules, Auto Editing

1. Introduction

Nowadays, people write blogs, diaries, and reports not only in their native language but sometimes in a language they are not that familiar with. During the process of writing, second/foreign language learners might make some errors, such as in spelling, grammar, and lexical usage. Therefore, how to provide editorial assistance automatically and effectively has become an important and practical research issue for NLP (Natural Language Processing) researchers. For second/foreign language learners, providing instant responses to their writing, indicating which part might be incorrect, and offering auto-editing suggestions for them to choose from would be beneficial for the improvement of their writing and other aspects of language development.

Editing plays an important part in language learning. It can be classified into human

* Department of Computer Science and Information Engineering, National Taiwan University
E-mail: r97922137@ntu.edu.tw; d97944007@csie.ntu.edu.tw; sdlin@csie.ntu.edu.tw
+ School of Applied Foreign Languages, Chung-Shan Medical University
E-mail: yingchun@csmu.edu.tw

editing and machine editing. Human editing has some limitations. Human editing is inefficient when the size of the edited articles becomes large, and it is inconvenient sometimes for people who need this service for their daily documents, like diaries, letters, and emails. Besides, human editing involves subjective opinions, which are different from the machine editing strategy that relies mostly on the objective empirical outcomes.

Despite the growing demand of editorial assistance tools, the existing ones still have considerable room for improvement. For example, the grammar checker provided by Microsoft Word has known deficiencies of being language dependent and covering only a small portion of errors without explicitly revealing the correction mechanism.

Given the importance of the need to develop editing tools, a new editing system is proposed. The current research demonstrates an auto-editing system based on the correction rules mined from online editing websites. In this paper, we focus on two research goals. First, we aim to design a strategy that identifies effective rules automatically and efficiently from editing databases. Second, we aim to design an auto-editing system based on the discovered rules.

Our method is language independent; therefore, it can be applied easily to other languages. Our evaluation reveals that, among the top 1500 rules the system found, 67.2% of them are regarded as correct or mostly correct.

The remainder of the paper is organized as follows. Section 2 describes the related work on detecting erroneous patterns. Section 3 lays out our methodology. Section 4 describes the experiment and our demo system. Section 5 concludes our study.

2. Related Works

Previous approaches can be classified into two categories. The first category detects erroneous patterns based on rules, and the second category makes use of statistical techniques for such a purpose.

2.1 Knowledge-Based Method

Some methods detecting erroneous patterns based on the manually created rules are proven to be effective in detecting grammar errors (Heidorn, 2000). Michaud, McCoy, & Pennington (2000) developed a system, including an error identification model and response generation model, using knowledge bases that cover general information about analyzing grammar structure and specific information of a user’s learning history. Also, Dan, Flickinger, Oepen, Walsh, & Baldwin (2004) presented a tutorial system based on computational grammar augmented with mal-rules for analysis, error diagnosis, and semantics-centered generation of correct forms. Nevertheless, the manually designed rules generally consume labor and time, along with requiring language experts, which limit the generalization capability of such methods. Furthermore, manually designed rules can hardly be applied to different languages.

2.2 Statistical Techniques

As discussed in Section 2.1, rule-based methods have some apparent shortcomings. Rather than asking experts to annotate a corpus, some researchers have proposed statistical models to identify erroneous patterns. An unsupervised method to detect grammatical errors by inferring negative evidence reached 80% precision and 20% recall (Chodorow & Leacock, 2000). It is reported that this system is only effective in recognizing certain grammatical errors and detects only about one-fifth as many errors as a human judge does. Some other papers focus on detecting particular errors, such as preposition errors (Hermet & Desilets, 2009), disagreement on the quantifier and misuse of the noun (Brocket, Dolan, & Gamon, 2006). Sun G. *et al.* (2007) treat the detection of erroneous sentences as a binary classification problem and propose a new feature called "Labeled Sequential Patterns" (LSP) for this purpose. This feature is compared to the other four features, including two scores produced by a toolkit, lexical collocation (Yajuan & Ming, 2004), and function word density. The results show that the average accuracy of LSP (79.63%) outperforms the other four features. Furthermore, the existence of the time words and function words in a sentence is proven to be important. In this way, one can only know whether a sentence is correct or not and would not have a clue about how to correct errors. Finally, some researchers have modeled detection of erroneous patterns as a statistical machine translation problem treating the erroneous sentences and the correct sentences as two different languages. Nevertheless, error correction could be intrinsically different from translation and there is no apparent evidence whether the existing machine translation techniques are suitable for such purpose (Guihua, Gao, Xiaohua, Chin-Yew, & Ming, 2007; Shi & Zhou, 2005).

Our work is different from the previous ones in two major respects. First, we treat error detection as a pattern mining problem to extract effective rules from an editing corpus. Second, we focus on designing a language-independent system that avoids using some language-specific features, such as not using any contextual, syntactic, or grammatical information, in this paper.

3. Methodology

3.1 Overview

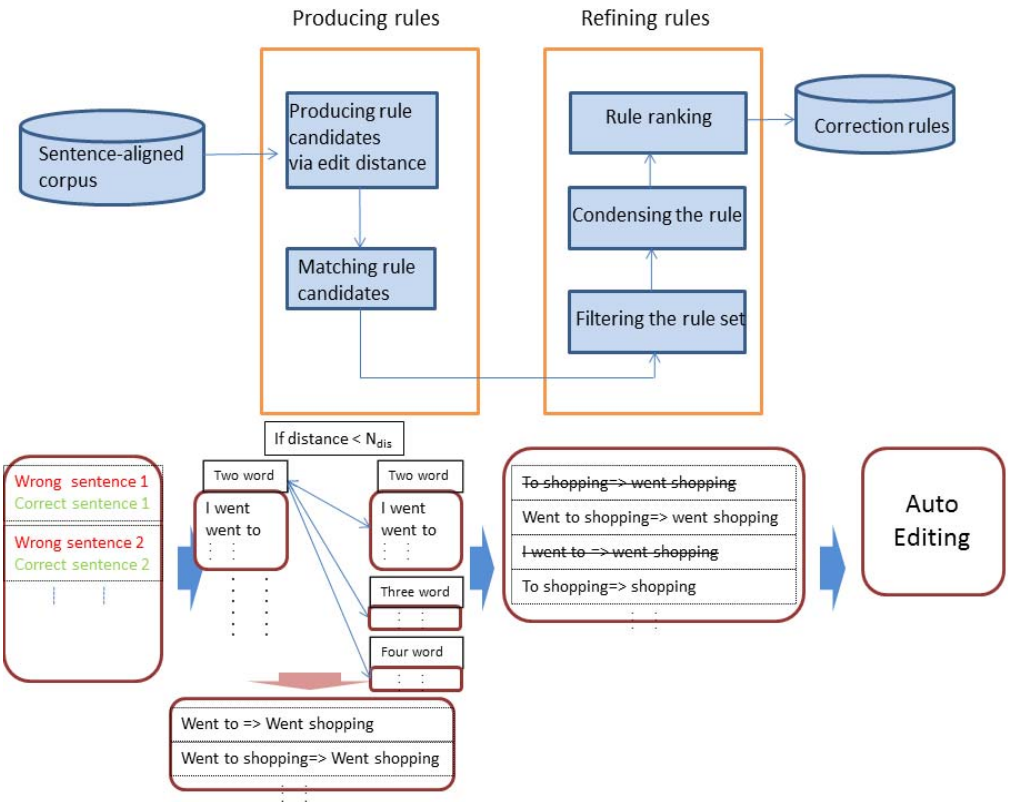


Figure 1. System Overview

Figure 1 shows that our framework consists of two parts. It produces some raw rules in the first stage and tries to refine them in the next stage.

3.2 Corpus Description

We retrieved 310967 parallel pairs of sentences (*i.e.* each pair consists of one erroneous sentence and one correct sentence) from an online-editing website Lang-8 (http://lang-8.com/). The website allows people to write diaries in their second/foreign language and the diaries (which usually contain some mistakes) would be edited by some volunteer members who are native speakers of the corresponding language. The edited part in an article is restricted to a single sentence (not cross-sentential). Consequently, we could retrieve the sentence-aligned data through crawling the website.

In the following sections, we use "W_i" to represent the erroneous sentence of the i-th pair of sentence in the corpus and "C_i" to represent the corresponding correct sentence. S_+ is defined as a collection of all correct sentences in the corpus, while S- is defined as a collection of all erroneous sentences.

3.3 Producing Rules

The following are some definitions of erroneous and correct patterns, rules, applying rules, and frequency of patterns:

Definition: (erroneous and correct) patterns: A pattern is a series of consecutive words (or characters) that belong to a subsequence of a sentence. An erroneous pattern represents such a sequence that is believed to be wrong, and a correct pattern is one that is believed to be correct.

Definition: a rule: A rule K can be written as K_L => K_R. The left-hand side of the arrow, K_L, is an erroneous pattern and the right-hand side of the arrow, K_R, is the correct pattern which K_L should be transformed to.

Definition: applying a rule to a sentence: Given a rule K : K_L => K_R, and a sentence T, if K_L exist in the T, we replaced every possible place of K_L in T to K_R. Such a process is considered as "applying rule K to a sentence T."

Definition: $fre_{S+}(K_L)$: the occurrence frequency of a pattern K_L in corpus S_+

To discover a rule A→ B from the editing corpus, we first had to identify the plausible left and right hand side of the rule. This is by no means a trivial task, and the fact that there could be various choices of such a rule made the task even more difficult. One intuitive method was to compare the word set existing in W_i and C_i and create the patterns using the difference among them. Nevertheless, such an intuitive method suffers certain deficiencies, such as the ones that appear in the following example.

Erroneous: "I with him had dinner."

Correct: "I had dinner with him."

The difference set is an empty set since the order is not considered. It is not clear how this difference set can lead to both erroneous and correct patterns. The approach we proposed was to exploit the procedure of calculating the word-level Levenshtein distance, which is often called editing distance (Levenshtein, 1966). The Levenshtein distance is defined as the minimum number of edits needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character (Levenshtein, n.d.). Similarly, the edit distance between two sentences can be defined as the minimum number of allowable operations required to transform from one of them into the other, given each unit of transformation being based on *words* rather than *characters*.

The *insert* operation inserts a word X into the erroneous sentence, which implies there is a word X that has the potential to be involved in the correct pattern K_R for a rule $K_L \rightarrow K_R$. Similarly, the *delete* operation removes one word Y from the erroneous sentence to become the correct one, and this word Y is likely to be involved in the erroneous pattern K_L. Finally, when a substitute operation is performed, the word to be replaced should appear in K_L while the replacing word shall be involved in K_R. Here, we argue that the words run through the editing-distance process from an erroneous to a correct sentence have a higher chance to be involved in the patterns of rules. For example, if we apply an editing distance approach to the following sentence pairs, multiple outputs can be acquired, such as the ones shown in Table 1 and Table 2. Levenshtein distance could calculate the difference between sentences, and we believe that rules are based on the differences.

Erroneous: "I still don't know where is it in the movie."

Correct: "I still don't understand where it is in the movie."

Based on the two editing-distance results shown in Table 1 and 2, it is possible to obtain that the four words {it, is, know, understand} are plausible words to appear in the rule $K_L \rightarrow K_R$.

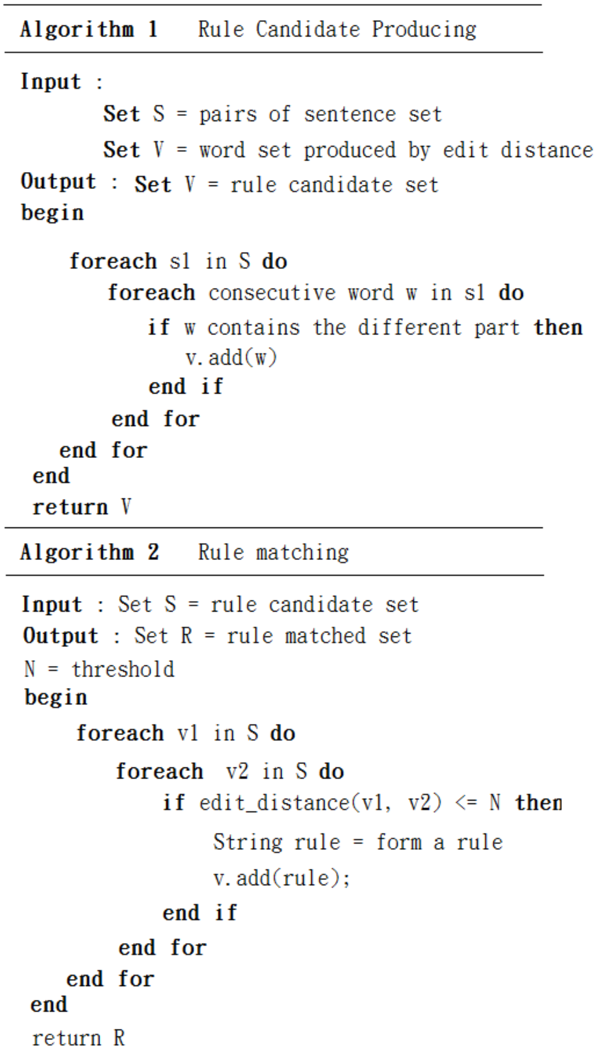
Table 1. One of the editing results for edit distance

Operation	Position	Involved word
Insert	6	It
Delete	8	It
Substitute	4	know→understand

Table 2. Another editing result for edit distance

Operation	Position	Involved word
Insert	8	Is
Delete	6	Is
Substitute	4	know→understand

For each pair of W_i and C_i, we can collect all of the involved words after producing the Levenshtein distance. Figure 2 shows the pseudo code. We exploited a dynamic programming approach to improve its efficiency.



```
Algorithm 1    Rule Candidate Producing
Input :
      Set S = pairs of sentence set
      Set V = word set produced by edit distance
Output : Set V = rule candidate set
begin
    foreach s1 in S do
       foreach consecutive word w in s1 do
          if w contains the different part then
             v.add(w)
          end if
       end for
   end for
 end
 return V
```

```
Algorithm 2    Rule matching
Input : Set S = rule candidate set
Output : Set R = rule matched set
N = threshold
begin
    foreach v1 in S do
       foreach  v2 in S do
          if edit_distance(v1, v2) <= N then
              String rule = form a rule
              v.add(rule);
          end if
       end for
    end for
 end
 return R
```

Figure 2. Pseudo code of producing rules

After applying the modified Levenshtein distance algorithm, it is possible to obtain a set of involving words R_i, as shown below.

R_i = {is , it ,understand , know}

To form a reasonable pattern, however, the words in set R_i are not sufficient. They should be combined with other terms. Ideally, K_L and K_R must consist of some words from R_i and some from the rest of the sentence. Therefore, for each pair of W_i and C_i in the corpus, we retrieved consecutive word patterns in which at least one word was from R_i. Based on R_i, the following examples are rule candidates.

Table 3. Pattern candidates for forming a rule

Candidates for K_L (Word length ≤ 4)	Candidates for K_R (Word length ≤ 4)
don't know	*don't understand*
know where	*understand where*
where is	*where it*
is it	*it is*
it in	*is in*
still don't know	*still don't understand*
don't know where	*don't understand where*
know where is	*understand where it*
where is it	*where it is*
is it in	*it is in*
it in the	*is in the*
I still don't know	*I still don't understand*
still don't know where	*still don't understand where*
don't know where is	*don't understand where it*
know where is it	*understand where it is*
where is it in	*where it is in*
is it in the	*it is in the*
it in the movie	*is in the movie*

Next, we matched each plausible candidate for K_L to each candidate for K_R to form a plausible rule(Table 3). For each plausible rule, we then checked its feasibility by applying it to W_i to see if the correct sentence C_i could be produced. The infeasible rules would be ignored.

> *Definition of feasible rule: Given a rule K : K_L => K_R . In a corpus, if at least one erroneous sentence in the corpus can be corrected using K, then K is considered a feasible rule.*

3.4 Refining Rules

So far, we have generated several rules, some of which make sense and some of which might not. In this section, we describe how to assess the quality of the rules and how to refine them.

Table 4. Observation on the frequency

	Pattern	fre_{S+}
Erroneous	Went to shopping	10
Correct	went shopping	205

	Pattern	fre_{S+}
Erroneous	am so exciting	0
Correct	am so excited	71

We believe the erroneous patterns K_L should not occur in the correct sentences too frequently (otherwise it would have been replaced by the correct one K_R); therefore, we considered fre_{S+} as a suitable metric to evaluate the quality of a rule. According to the real experiment shown in Table 4, the frequency of the erroneous patterns seems to be lower in the correct corpus, fre_{S+}, compared to the correct ones.

Next, we condensed the rules according to their fre_{S+}. The condensed rule is shorter than the original one and is supposed to be more general (*i.e.* can cover more sentences). For example, in the following sentences, the condensed rule is more general and reasonable since the subject ‘I’ has nothing to do with the erroneous pattern.

Erroneous: “I went to shopping and had dinner with my friend yesterday.”

Correct: “I went shopping and had dinner with my friend yesterday.”

Rule: “I went to shopping.” => “I went shopping.”

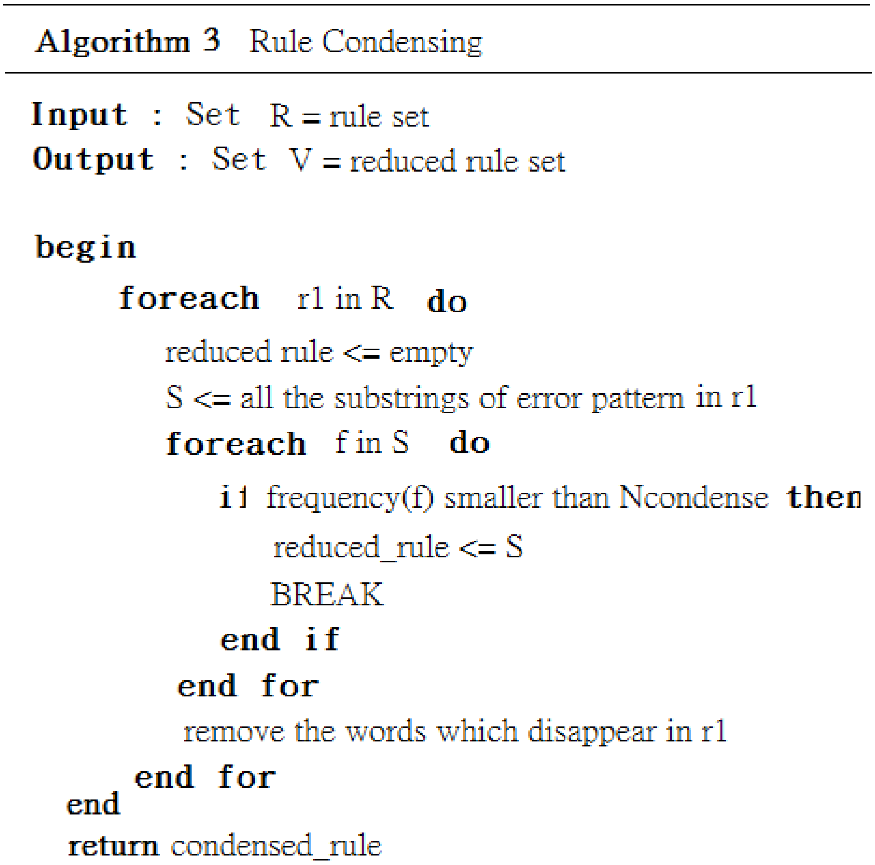
Condensed Rule: “went to shopping” => “went shopping”

To obtain the shortest possible rules for auto-editing, we proposed a simple idea to check if the left hand side K_L could be condensed to a shorter one, without boosting its fre_{S+} significantly. If yes, then it implied we had found a shorter erroneous pattern that also occurred rarely in the correct corpus. For example, for the erroneous pattern “I am surprised at.” Table 5 shows the frequency of each possible subsequence in the correct corpus. Apparently “am surprised at” is the most condensed rule that does not occur more than ten times in the correct corpus.

Table 5. An example for condensing a rule

Sentence segment	surprised	surprised at	am surprised	am surprised at	I am surprised at
Frequency	985	702	213	10	10

What follows here is the algorithm for rule condensing. If the frequency of the condensed erroneous rule is smaller than an empirically-defined threshold frequency $N_{condense}$, we will accept it as a condensed erroneous pattern. Then, we remove the same words from the K_R to produce the corresponding correct pattern. The condensing process repeats until any of the words to be removed in K_L do not occur in the K_R. The pseudo code of condensing rules is shown in Figure 3.



```
Algorithm 3   Rule Condensing

Input  :  Set  R = rule set
Output  :  Set  V = reduced rule set

begin
    foreach  r1 in R  do
        reduced rule <= empty
        S <= all the substrings of error pattern in r1
        foreach  f in S  do
            if frequency(f) smaller than Ncondense then
                reduced_rule <= S
                BREAK
            end if
         end for
         remove the words which disappear in r1
      end for
  end
  return condensed_rule
```

Figure 3. Pseudo code of condensing rules

The final step of the refinement is to rank the rules based on their qualities. We proposed two plausible strategies to rank the rules. First, it is possible to rank the rules according to $fre_{S+}(K_L)$ from low to high. In other words, a rule is less likely to incorrectly modify something right into something wrong if its fre_{S+} is low. Second, it is possible to rank the rules according to the number of sentences in the corpus that can be applied using it. The first strategy is similar to the definition of *precision* while the second is closer to the meaning of *recall*.

4. Experiments

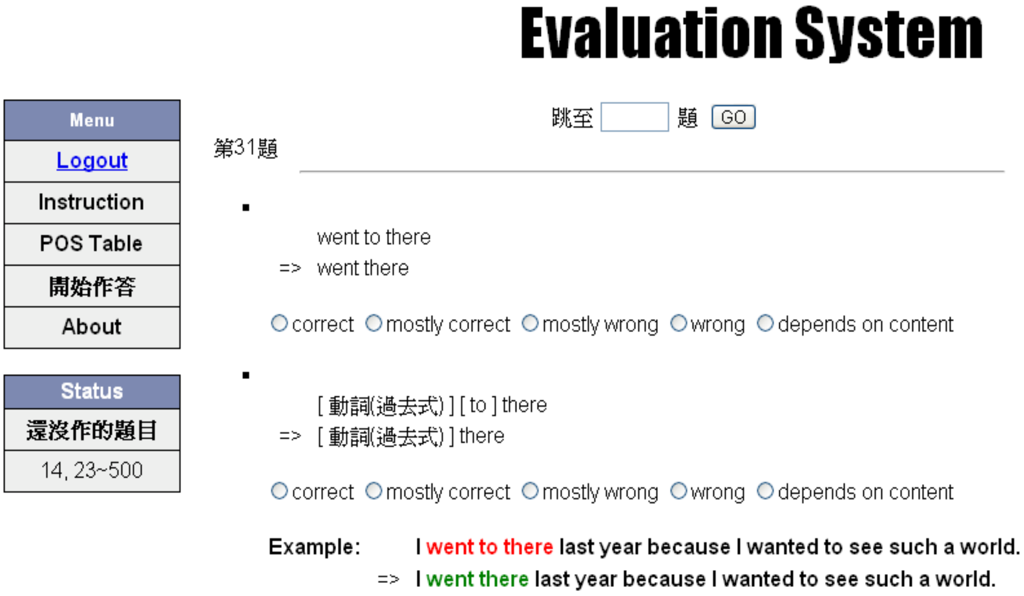
We set $N_{condense}$ as 10 and retrieved 310967 pairs of English sentences from the "Lang-8" as our parallel corpus, and the system finally generated 110567 rules. To evaluate the framework, four experts were invited to annotate the rules. Then, we demonstrated an auto-editing system to show how such rules can be applied.

4.1 Evaluation

We ranked all of the rules according to their fre$_{S+,}$ and four English majors were invited to annotate the top 1500 ranked rules. Each rule was annotated by two persons. The labels for annotations were “correct,” “mostly correct,” “mostly wrong,” “wrong,” and “depends on context”. Table 6 presents the experimental results and Figure 4 presents the evaluation system screenshot. A fair agreement was found between the two annotations, as the kappa value equals 0.49835.

Table 6. The Distribution of annotated results of the top 1500 rules

	Correct	Mostly correct	Mostly wrong	Wrong	Depends on context
R1~R1500	53.96%	12.96%	0.92%	4.5%	27.66%



第31題

- went to there
  => went there

○ correct ○ mostly correct ○ mostly wrong ○ wrong ○ depends on content

- [動詞(過去式)] [to] there
  => [動詞(過去式)] there

○ correct ○ mostly correct ○ mostly wrong ○ wrong ○ depends on content

Example: I went to there last year because I wanted to see such a world.
=> I went there last year because I wanted to see such a world.

Figure 4. Screenshot of Evaluation System

We also compared our system (using all rules or highly ranked rules), with the other two available auto-editing systems, ESL Assistant and Microsoft Word Grammar Checker. The highly ranked rules were those with fre$_{S+}$(K$_L$) smaller than 10. We retrieved 30 articles randomly from lang-8 that did not appear in our training corpus and examined their correction on the website as the gold standard. Table 7 shows the sentence-based recall and precision values.

Table 7. Evaluation results with 95% confidence

System	Recall	Precision
Our Auto-Editing System(All Rules)	20.28%±1.07%	40.16%±0.6%
Our Auto-Editing System (Highly Ranked Rules)	14.28%±0.74%	77.32%±0.55%
ESL Assistant (Claudia, Michael, & Chris, 2009)	18.4%±1.07%	42.36%±0.29%
Microsoft Word Grammar Checker	14.28%±0.72%	27.77%±1.03%

4.2 Discussion

Manual analysis of the rules was performed as well. As seen in Table 8, the results show that most of the corrections (67% of rules) are about spelling errors, collocation and phrase, and agreement of subject and verb. It is also noted that most of the incorrect rules would lead to false suggestions and 83% of the rules belonging to "*depend on context*" category are about chunks and phrases.

Table 8. Manual analysis of rules

I. Correct & Mostly Correct (67% of Rules)	**%**
1. Spelling	60%
2. Collocation and phrase (sequence of words which co-occur more often than would be expected by chance	15%
3. Agreement of subject and verb	7%
4. Choice of verb tense	5%
5. Gerund forms and infinitives	2%
6. Choice of the proper article	1%
7. Pluralization (irregular noun)	1%
8. Capitalization (use of capital letter)	1%
9. Other (use of preposition, word choice, cohesive devices, elliptical forms, punctuation, parts of speech, count and noncount nouns…etc.)	8%
II. Wrong & Mostly Wrong (0.9% of Rules)	**%**
1. Suggestions of wrong corrections	97%
2. Errors not to be spotted and corrected	3%
III. Depends on Context and/or Writers' Intention (32.1% of Rules)	**%**
1. Correctness of the chunks/phrases	83%
2. Verbal and verb tense	5 %
3. Spelling (more than one possibility)	3%
4. Word choice	2%
5. Others (use of preposition, conjunction, cohesive devices, parts of speech…etc.)	7%

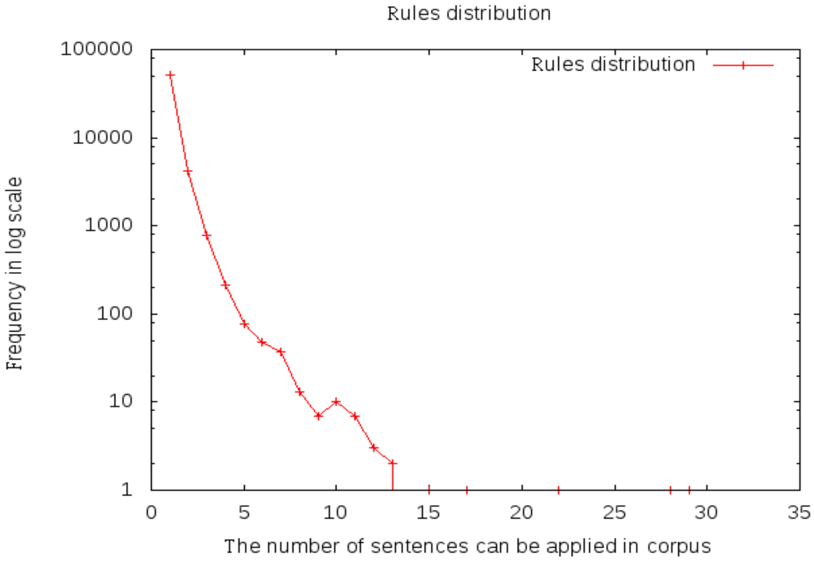


Figure 5. Rule distribution

Figure 5 shows the rule distribution. Table 9 lists some example rules discovered by our system that can hardly be detected and corrected by Microsoft Word 2007 grammar checker.

Table 9. Example rules discovered by the proposed system

Example rules
am worry about => am worried about
help me to study => help me study
I will appreciate it => I would appreciate it
went to shopping => went shopping
am so exciting => am so excited
waked => woke
look forward to read => look forward to reading
for read my => for reading my
The street name => The street's name
to playing with => to play with
He promised to me => He promised me
asked repeat => repeatedly asked
Have you listen to => Have you listened to
It's rains => It's raining
I ate a milk => I had milk
for the long time => for a long time
don't cooking => don't cook
will success => will succeed
don't know what happen => don't know what happened

4.3 Auto-editing System

We constructed an online, real-time auto-editing system and demonstrated the usefulness of our rules, which aimed to provide editorial assistance. We first tried to test whether a part of the real-time typing sentence could match the erroneous patterns. If there was a match, the chunk would be marked in red, and we applied the correction rule to suggest replacing it with the correct pattern. The user(s) was able to click the correct part (marked in green) to tell the system the given correction was accepted, and the system automatically made the change. The link to our system is: http://mslab.csie.ntu.edu.tw/~kw/new_demo.html.

4.3.1 Auto Editing

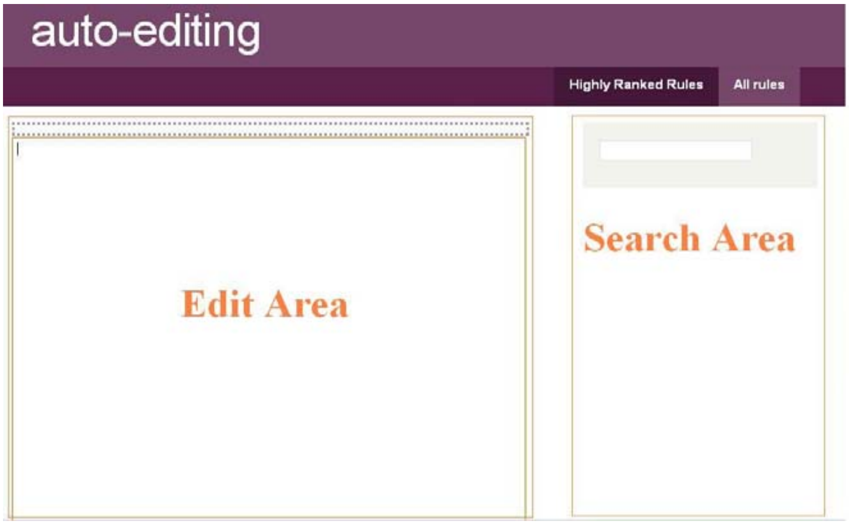
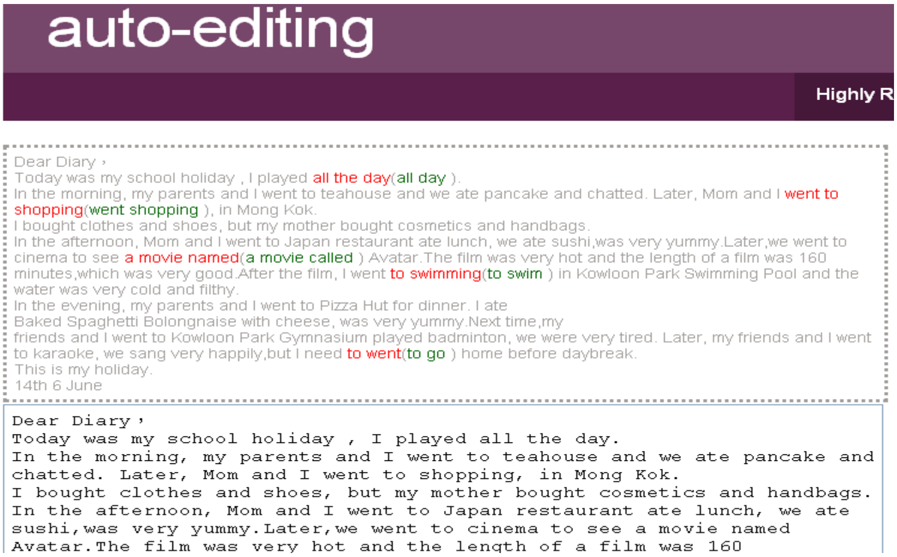


Figure 6. Screenshot of demo system

Figure 6 is the entire system view. Two kinds of rule sets can be exploited: (1) "Highly-ranked Rules" exploits only higher ranked rules and ignores lower-ranked ones; (2) "All Rules" utilizes every rule but suffers the risk of utilizing incorrect ones.



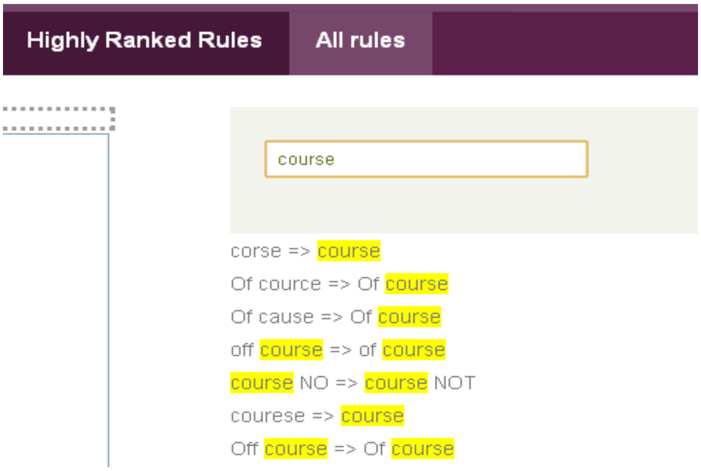
Dear Diary ,
Today was my school holiday , I played all the day(all day).
In the morning, my parents and I went to teahouse and we ate pancake and chatted. Later, Mom and I went to shopping(went shopping), in Mong Kok.
I bought clothes and shoes, but my mother bought cosmetics and handbags.
In the afternoon, Mom and I went to Japan restaurant ate lunch, we ate sushi,was very yummy.Later,we went to cinema to see a movie named(a movie called) Avatar.The film was very hot and the length of a film was 160 minutes,which was very good.After the film, I went to swimming(to swim) in Kowloon Park Swimming Pool and the water was very cold and filthy.
In the evening, my parents and I went to Pizza Hut for dinner. I ate
Baked Spaghetti Bolongnaise with cheese, was very yummy.Next time,my
friends and I went to Kowloon Park Gymnasium played badminton, we were very tired. Later, my friends and I went to karaoke, we sang very happily,but I need to went(to go) home before daybreak.
This is my holiday.
14th 6 June

```
Dear Diary ,
Today was my school holiday , I played all the day.
In the morning, my parents and I went to teahouse and we ate pancake and
chatted. Later, Mom and I went to shopping, in Mong Kok.
I bought clothes and shoes, but my mother bought cosmetics and handbags.
In the afternoon, Mom and I went to Japan restaurant ate lunch, we ate
sushi,was very yummy.Later,we went to cinema to see a movie named
Avatar.The film was very hot and the length of a film was 160
```

Figure 7. Screenshot of auto-editing

In Figure 7 shows one can type sentences in English in edit area. If any of the rules is matched, the suggested correction will appear on the above area in green. If the users agree with the corrections, they can click on the green word and the sentence will be edited accordingly.

4.3.2 Rules Keyword Search



course

corse => course

Of cource => Of course

Of cause => Of course

off course => of course

course NO => course NOT

courese => course

Off course => Of course

Figure 8. Screenshot of keywords search in rule database

On the right hand side of the page (Figure 8), the user can type a keyword to search for the related rules. Then, the system would demonstrate all of the discovered rules relevant to this keyword. The above screenshot shows the rules relevant to the keyword "course".

4.3.3 User Correction Feedback

When a user chooses a correction option from editing results, we could assume the rule receives one additional endorsement. Such information can be exploited to refine the rules. Therefore, we maintain the user feedback and use such feedback to adjust the rank of the rules. Highly endorsed rules will be promoted gradually in the ranking.

5. Conclusion

In this research, we propose a language-universal framework that is capable of producing effective editing rules. The quality of rules can be assessed using the proposed ranking strategies. Moreover, we have demonstrated the practical usage of the rules by constructing an auto-editing system to provide editorial assistance for language learners. In this paper, we demonstrated how we produced correction rules without considering syntactic structure and POS (Part-of-Speech). In the future, we would like to make use of both of the features to improve the performance of our system.

References

Heidorn, E. (2000). Intelligent Writing Assistance. in Robert, D., Hermann, M., & Harold, S.(eds.), *Handbook of Natural Language Processing*. New York: Marcel Dekker.

Michaud, L., McCoy, K., & Pennington, C. (2000). An Intelligent Tutoring System for Deaf Learners of Written English. *Proceeding of Fourth International ACM Conference on Assistive Technologies,* 92-100.

Dan, E., Flickinger, D., Oepen, S., Walsh, A., & Baldwin, T. (2004). Arboretum: Using a precision grammar for grammar checking in call. *In Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems.*

Chodorow, M., & Leacock, C. (2000). An Unsupervised Method for Detecting Grammatical Errors. *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference,* 140-147.

Hermet, M., & Desilets, A. (2009). Using First and Second Language Models to Correct Preposition Errors in Second Language Authoring. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications,* 64-72.

Brocket, C., Dolan, W., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics,* 249-256.

Sun,G., Liu, X., Cong, G., Zhou, M., Xiong, Z., Lin, C. Y., & Lee, J., (2007). Detecting Erroneous Sentences Using Automatically Mined Sequential Patterns. *In Proceeding of the 45th annual meeting of the Association of Computational Linguistics*, 81-88.

Sun, G., Cong, G., Liu, X., Lin, C.-Y., & Zhou, M. (2007). Mining Sequential Patterns and Tree Patterns to Detect Erroneous Sentences. *Proceedings of the 22nd national conference on Artificial intelligence - Volume 1*. 925-930.

Shi,Y., & Zhou, L. (2005). Error Detection Using Linguistic Features. *Proceeding of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing,* 41-48.

Levenshtein, VI. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady,* 10(8), 707-710.

Lü, Y., & Zhou, M. (2004). Collocation translation acquisition using monolingual corpora. *In Proceeding of Association for Computational Linguistics*.

Leacock, C., Gamon, M., & Brockett, C.(2009). User Input and Interactions on Microsoft Research ESL Assistant. *Proceeding of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications,* 73-81.

Levenshtein. (n.d.). Retrieved from the Levenshtein Wiki: http://en.wikipedia.org/wiki/Levenshtein_distance

Computational Linguistics and Chinese Language Processing
Vol. 15, No. 3-4, September/December 2010, pp. 237-262

學術會議資訊之擷取及其應用

Information Extraction for Academic Conference and It's Application

陳光華[*]

Kuang-hua Chen

摘要

網際網路已成爲學術訊息傳播的主要管道，本研究關注擷取網際網路上學術研究人員關心的學術會議訊息，提供會議主題、時間、空間等訊息，企望減輕研究人員蒐集與管理會議資訊的負擔，進而提升學術研究出版的效率。本研究首先提出一套學術會議資訊檢索與擷取的自動程序，並藉由實驗確認其可行性，實驗結果顯示文件分類績效 F1 measure 超過 80%；具名實體擷取績效 Recall 超過86%，F1 measure超過70%。繼而實際開發學術會議檢索與擷取系統平台，提供文件檢索、資訊擷取、分類瀏覽、行事曆等功能，整合研究人員的學術活動與日常行程安排，展示前述學術會議資訊檢索與擷取程序的實用性。

關鍵詞：學術資訊、資訊擷取、資訊檢索、具名實體

Abstract

Internet has become a major channel for academic information dissemination in recent years. As a matter of fact, academic information, e.g., "call for papers", "call for proposals", "advances of research", etc., is crucial for researchers, since they have to publish research outputs and capture new research trends. This study focuses on extraction of academic conference information including topics, temporal information, spatial information, etc. Hope to reduce overhead of searching and managing conference information for researchers and improve

[*] 國立臺灣大學圖書資訊學系
Department of Library and Information Science, National Taiwan University
E-mail: khchen@ntu.edu.tw

efficiency of publication of research outputs. An automatic procedure for conference information retrieval and extraction is proposed firstly. A sequence of experiments is carried out. The experimental results show the feasibility of the proposed procedure. The F1 measure for text classification is over 80%; F1 measure and Recall for extraction of named entities are over 86% and 70%, respectively. A system platform for academic conference information retrieval and extraction is implemented to demonstrate the practicality. This system features functionalities of document retrieval, named entities extraction, faceted browsing, and calendar with a fusion of academic activities and daily life for researchers.

Keywords: Academic Information, Information Extraction, Information Retrieval, Named Entities

1. 緒論

在全球化的趨勢之下，大學的學術評價更加受到前所未有的重視，有各式各樣以全球大學為標的之學術評鑑報告陸續公告周知，如上海交通大學（ARWU, 2010）與英國 Quacquarelli Symonds（QS, 2010）所做的世界大學排名。此外，Thomson Reuters 公司的 SCI、SSCI、A&HCI 等資料庫，以及 Journal Citation Report（JCR），提供的統計數據，往往成為各國評鑑國內大學學術成果的計量指標。在這種激烈的學術競爭環境之下，且學術競爭力被視為國家競爭力的一環，大學教授莫不兢兢業業地、努力地從事學術研究。學術研究人員掌握學術會議資訊的即時性與確實性，對於其研究工作的進展與研究成果的發表，是非常重要的。本研究在這樣的背景下，研發學術會議資訊檢索與擷取系統，希望能夠有效地由充斥浮濫資訊的網際網路，擷取相關的學術會議資訊。

學術研究人員的學術活動是非常多元的，學術資源服務的類型眾多，本研究將著重於以資訊擷取為基礎的學術會議資訊的檢索與擷取。研究人員的學術活動中很重要的一項便是「學術研究的出版」，學術的出版有兩個主要的方向，一個是學術會議，另一則是學術期刊。會議的 Call For Paper 有時間的期限，而期刊 Special Issue 的 Call For Submission 也有時間的期限，協助研究人員掌握這些重要的訊息，自動地由網路擷取學術會議的時間訊息、空間訊息、與主題訊息，協助研究人員管理時間與空間訊息，將有很大的助益。若能進一步搭配「行事曆（calendar）」的功能，對於研究人員而言更是事半功倍的。換言之，一般行事曆功能僅提供使用者新增資訊、更新資訊、刪除資訊，為了搭配學術研究的出版，行事曆必須有更進階的功能，能夠依據使用者的 profile 搜尋 Call For Paper 與 Call For Submission，填入行事曆，並依據使用者的設定，提供警示（alert）的服務。

研討會通知或會議論文投稿須知，一般是透過既有的郵寄目錄發送，或是以網頁文件的形式發佈，也因此訊息傳播的目標通常局限於特定族群及研究機構。即使使用者自行利用網頁搜尋工具在網際網路上查找，所取得的資訊可能不完整，或是已錯過參與的時機。若要提供即時的且整合的研討會相關資訊，蒐集網際網路上與研討會通知相關網

頁的自動機制，是重要的一環。

一般在網路上大量蒐集網頁的方式，通常利用網頁擷取機器人（web crawler）到處拜訪網站並擷取所有網頁內容。由於 Web Crawler 的建置困難度較高，維護與效能控管也較爲複雜，不當的設計常會佔據網路頻寬資源，或導致被網站封鎖而無法擷取內容。因此另有一種方式，並不採用傳統的 web crawler 而是修改網頁擷取機制，以適當的關鍵字與網頁搜尋引擎的整合來蒐集網頁。

目標式網頁擷取（focused crawling）是一種蒐集研討會通知資訊的方式。有別於一般 Web Crawler 漫無目的地抓取所有的網頁，Focused Crawling 會先過濾與主題無關的內容，也就是會應用一組特定主題的關鍵詞，用以訓練並建立文件分類機制，再由此分類機制引導 crawler 擷取與主題相關的網頁。（Chakrabarti, van den Berg, & Dom, 1999）另外還可以將 Focused Crawling 稍加變化，依據一組系統已經記載的研討會議網站清單，反向地蒐集相關網頁文件，這種網頁資料蒐集的替代方案被稱爲反向式網頁擷取（backward crawling）。（Brennhaug, 2005）這種網頁蒐集機制首先以主題關鍵字，透過搜尋引擎取得相關網頁的網址及網頁內容，以建構候選相關文件集。再接續利用搜尋引擎的反向連結查詢功能（back link query），一併蒐集連結到候選文件的網頁。又考量到這種由反向連查詢所得的網頁也有可能再連結到其他研討會議網頁，所以再繼續以正向連結（forward crawling）擷取該網頁中的其他 URL，以發掘潛在的相關網頁。此程序將會一直重覆執行直到重覆的次數達到預設的門檻。

若以蒐集研討會議徵稿通告的相關資訊來檢視網頁自動擷取機制，無論是正向或反向擷取，都會面臨下列兩項議題：(1)網路上傳播的研討會會議資訊經常更新，例如投稿截止日期的延期、會議地點資訊的更新、或是新加入的 workshop 議程等等，而所蒐集的研討會會議資訊必需能夠即時反應各項更新資訊。(2)目前雖然將「研討會議通知資訊」定義爲與研討會議相關的訊息通知網頁，但網頁內容通常包含許多與研討會無關的各種式樣各種規格的其他資訊，例如文字或影音廣告，網站目錄選項，或其他網站連結等，這也造成在擷取網頁機制建置時，文件相關程度判斷的問題。

本研究基於前述的背景，運用網頁搜尋技術，以及資訊檢索與擷取技術，發展一套學術會議資訊檢索與擷取的自動程序，並實際建構系統平台，以服務學術研究人員。本文的結構如下：文獻探討一節將說明資訊擷取的技術，運用於學術會議檢索的情形，相關資訊服務系統的現況；學術會議資訊蒐集一節討論由網際網路蒐集學術會議資訊的方法，以及過濾不相關資訊與雜訊的作法；資訊擷取模型之訓練與建置一節探討學術會議資訊擷取模型的訓練與建立；系統實作與功能一節討論系統實作的方法，以及各項功能；最後則是簡短的結論。

2. 文獻探討

學術會議資訊之檢索屬於資訊檢索的應用研究，其中牽涉的研究議題眾多，至少有具名實體的辨識（named entities identification）、分群歸類（clustering and classification）、

文件檢索（text retrieval）。然而，若要建置完整的應用系統，則牽涉更多的技術，如時間與空間資訊的搭配，各種 API 應用元件的整合。本研究嘗試建構學術會議資訊檢索與擷取系統，首先探討資訊檢索與擷取技術的現況，以及現有檢索系統的發展。限於篇幅，本文並不嘗試進行全面而完整的相關文獻的探討。

學術會議資訊文件含有許多具名實體，包括會議名稱、會議時間、會議地點、會議主題、截稿日期等等，已有許多學術論文探討這個研究課題，訊息理解會議（Message Understanding Conference，簡稱 MUC）是第一個將具名實體的辨識視爲一項檢索研究的評量項目，企圖推動資訊檢索研究社群，投注研究能量，發展更新的技術，提昇具名實體辨識的績效。（MUC, 2001）訊息理解會議認爲不僅僅需要辨識重要的實體，還必須確認實體之間的關係（relationship），MUC-6 則明確地規範三個層次的資訊擷取的研究議題：具名實體之辨識、照應詞之解析、樣版資訊之建構。照應詞之解析是串連具名實體及其對應的照應詞（如代名詞）；腳本樣版則是依照預先訂定的樣版，由文件中擷取相關的資訊填入樣版的欄位。（Grishman & Sundheim, 1996）

雖然具名實體辨識的研究很早就開始了，但是學術會議資訊擷取的研究則是比較不受到許多研究者的關注。Lazarinis（1998）提出應該應用資訊擷取技術進行論文徵稿通告（call for paper，簡稱 CFP）的檢索，有別於傳統上僅以文件檢索技術檢索 CFP。Lazarinis 發現這種作法在固定 Recall 的情形下，可以提昇 45%-60%的 Precision，這項研究確認應將學術會議資訊的檢索，視爲資訊擷取的問題，而非單純的文件檢索的問題。

Schneider（2005）應用 Conditional Random Fields（CRF）模型，擷取 CFP 的重要訊息，Schneider 特別關注文件版面特徵（layout features）的貢獻，發現版面特徵可以提昇約 30%的 F1 分數（F1 measure）。因爲，Schneider 的研究關注於各項特徵的效益，使用的測試資料僅有 263 篇乾淨無雜訊的 CFP，而避開真實文件各種複雜的情況，因此很難建構一個實際可行的資訊服務系統。

目前亦有許多學術組織，建構了 Conference Calendar 的相關網頁，希望有利於會議資訊的流通，但是這種資訊彙整形式的網頁，僅提供瀏覽的功能，沒有進階檢索功能，使用者仍須耗費相當的精力，才能瀏覽相關的會議資訊。另外，尚有功能比較好的類似系統，例如 WikiCFP 與 EventSeer 等 CFP 資訊共享服務系統，但是提供的多爲電腦科學相關學術領域的學術會議資訊。WikiCFP（http://www.wikicfp.com/）是使用 Wiki 建構的 CFP 共享系統，資訊來源是依賴使用者提供相關會議資訊；EventSeer（http://eventseer.net/）是一個 Web 2.0 的網站，企圖建構一個電腦科學研究的社群網站，除了允許登錄使用者自由發佈學術資訊外，另外運用 Robot 主動搜集網際網路上的 CPF 資訊。

Takada（2008）建構的 ConfShare 資訊服務系統，透過瀏覽器提供學術會議資訊檢索的服務。Takada 認爲研究者爲了參加學術會議學習最新的研究成果，或發表本身的研究成果，都需要蒐集學術會議的相關資訊。蒐集資訊的工作是參加會議不可缺乏的，但也造成研究者不小的負擔。ConfShare 以使用者（亦即研究者）的角度，提供與學術會議相關資訊的各種服務，希望能夠減輕前述研究者的額外負擔。

Xin, Li, Tang, and Luo（2008）使用 Constrained Hierarchical CRF（CHCRF）標註學術會議官方網站的網頁以及屬性，企圖建構一個學術會議的行事曆系統。Xin 等人關注的是學術會議的官方網站而非 CFP，然而官方網站成立的時間通常都很晚，不像 CFP 的快速與即時，而且，官方網站的資料是透過下達會議名稱與時間，由 Google 檢索而得，這樣的假設並非很合理，因爲，類似的系統應該是藉由學術研究的主題取得學術會議資訊，而非藉由特定的會議名稱或是舉辦時間。

本研究企圖建構的學術會議資訊檢索與擷取系統（Academic Conference Information Retrieval and Extraction System，ACIRES），較接近於 Takada（2008）的 ConfShare 系統，但是在功能面仍有差異，使用的技術亦不相同，涵蓋的學科主題範疇亦有很大的差異。下文將說明本研究的資訊的蒐集、處理、模型的訓練、以及系統的實作 。

3. 學術會議資訊蒐集

學術會議資訊的檢索與擷取，當然需要被檢索的標的物，必須有一套機制蒐集網路上的論文徵稿通告，作爲系統開發前，資訊擷取模型訓練之用；系統開發完成，正式運轉時，亦需要這套機制持續蒐集論文徵稿通告，以服務學術研究人員以及一般的使用者。

爲了有效地蒐集相關的學術論文徵稿通告，本研究採用目標式網頁擷取（focused crawling）的概念，先以學門分類表做爲各學科主題的查詢關鍵字，利用網頁搜尋引擎蒐集所需之論文徵稿通告。我們採用澳洲與紐西蘭標準研究分類表（Australian and New Zealand Standard Research Classification，簡稱 ANZSRC）爲主（Pink & Bascand, 2008），再整合 Wikipedia 提供的學術領域列表以補充新興學科。由於論文徵稿通告不一定會標示所屬學科領域，以學門分類名稱爲查詢關鍵詞所蒐集的論文徵稿通告，可能無法涵蓋各學科領域所有重要的研討會資訊。因此，可再進一步分析第一批搜集的論文徵稿通告的研究議題相關詞彙，整合到學科主題關鍵詞列表，形成所謂的 bootstrapped crawling，讓學術會議資訊的蒐集更爲廣泛且完整。表 1 依字母順序，簡要列出部分之主題關鍵詞。

利用前述的主題關鍵詞，透過 Google 搜尋引擎，分別取得查詢結果前五十筆最相關的網頁，再接續依相關網頁的內容執行一次正向連結查詢（forward link query），一併收錄該五十筆網頁中超連結所指到的網頁。透過網頁搜尋引擎，可一次性地蒐集大量的相關網頁，但無法掌控網頁提供的會議資訊是否已過期。再考量研討會資訊的提供，必須符合即時性與時效性，因此再進一步利用網頁快訊服務（Google Alert），補充最新的研討會資訊。

網頁快訊服務就是當新的網頁發佈於網際網路時，網頁搜尋引擎比較該新網頁與使用者預設的 profile 的相關度，若是在搜尋結果的前 20 名內，就會立即以電子郵件通知快訊訂閱客戶。利用此服務特性，將前述的學科主題關鍵詞，做爲取得快訊的搜尋詞彙，即時取得最新發佈的網頁文件。對於以網頁快訊服務取得的相關網頁，本研究也會進一步執行一次正向連結查詢。

無論是從網頁搜尋引擎或是網頁快訊服務蒐集而得的網路資訊，必定會有重覆的情形，因此在蒐集網頁時，必須初步過濾重覆的網頁。以網頁搜尋引擎取得的相關網頁，由於是同一時間取得的網頁內容，因此不需考量網頁更新的因素，直接比對網址過濾重覆者。以網頁快訊服務取得的新網頁，若網址與現有文件相同，則必須考量網頁更新因素，先比對兩筆網頁的上次更新時間，再保留更新時間較近的網頁。若無法取得網頁的上次更新時間，則保留由網頁快訊服務取得的網頁。

由於從網頁搜尋引擎及網頁快訊服務廣泛蒐集的網頁數量龐大，大量的文件中可能包含與研討會論文徵稿通告無關的網頁，爲了提升學術會議資訊自動標註的準確度，必須篩選無關的網頁文件。本研究運用文件自動分類技術，可以迅速處理大量文件，避免繁瑣且冗長的人工分類作業，我們採用開放程式碼 Rainbow Classifier 自動過濾非會議徵稿通告的網頁文件。（McCallum, 1996）由於 Rainbow Classifier 需要一組已分類的文件做爲分類模型所需的訓練文件，此訓練文件將利用人工分類的方式產生，該人工分類的作業一併整合至人工標註輔助系統，讓標註人員可同時並行訓練文件分類與文件內容標註工作。

表1. 部分主題關鍵詞

abnormal psychology accompanying accounting scholarship acoustic engineering acoustics acting actuarial science adapted physical education admiralty law advertising aerobiology aeronautical engineering aerospace engineering aesthetics affine geometry african studies agricultural economics agricultural education agricultural engineering agrology agronomy air force studies algebraic computation algebraic geometry algebraic number theory algebraic topology american history american politics american studies analytical chemistry ancient egyptian religion ancient history animal communications animal science animation anthropology of technology apiculture appalachian studies applied psychology approximation theory aquaculture architectural engineering archival science art education art history artillery arts administration asian american studies asian studies associative algebra astrobiology astronomy astrophysics atheism and humanism atomic, molecular, and optical physics australian literature automotive systems engineering beekeeping behavioral geography behavioural economics behavioural science bilingual education biochemistry bioeconomics biogeography bioinformatics biological psychology biology biomechanical engineering biomedical engineering biophysics black studies or african american studies botany business administration business english business ethics calligraphy campaigning canadian literature canadian studies canon law cardiology cardiothoracic surgery cartography category theory cell biology celtic studies chamber music chemical engineering cheminformatics chemistry education chicano studies child welfare children geographies chinese history chinese studies or sinology choreography christianity chronobiology church music civics civil procedure classical archaeology classics climatology coastal geography cognitive behavioral therapy cognitive psychology cognitive science collective behavior combat engineering communication design communication engineering

4. 資訊擷取模型之訓練與建置

學術會議的論文徵稿通告主要包含會議名稱、會議地點、會議時間、會議主題、會議官方網站、以及各項截止日期或公佈日期等。論文徵稿通告與一般文件最大的差異在於其重要資訊不一定是以完整的語意文句組成，可能利用內容配置及排版以突顯各項資訊。例如，一份論文徵稿通告的會議名稱通常單行置中且前後各有空行，研討會議題以項目符號逐項表列，各項重要期限或公佈日期通常利用表格呈現。除了排版上的特色之外，還可利用特定詞彙判斷是否爲重要通知資訊，例如會議名稱通常會出現 conference、international、annual 等詞彙，submission、notification、deadline 等詞彙則經常伴隨日期出現，另外也可以利用完整的地名詞典擷取會議舉行地點。雖然可利用排版及詞彙兩種特性設計論文徵稿通告的資訊自動擷取機制，但是網路上或電子郵件提供的論文徵稿通告，並沒有一致的文件格式，通知項目也沒有統一的名稱，這都增加資訊判斷的困難度。

本研究應用 Conditional Random Field（CRF）建立自動擷取會議資訊的模組，從會議通告網頁文件，擷取重要的會議資訊欄位（如會議名稱，會議日期，會議地點等）。CRF 爲機器學習式（machine learning-based）演算法，需設定數種資料特徵以訓練模型，因此以學術會議徵稿通告必備的重要資訊項目，作爲資料特徵欄位（如表 2 所示），再使用一部分學術研討會徵稿通告，做爲訓練文件集，先以人工的方式標註特徵欄位，並利用特殊詞典或地名資料庫標示特定詞彙（例如地名、會議專有名詞等），建立 CRF 學習樣版，再經由 CRF 自動學習與測試，調整資訊辨識的準確度，以建置資訊擷取的自動機制。

CRF 是在機率演算的架構之下，針對某種結構組成的文字資料進行分段（segment）或是標註（label）的工作，其文字資料結構包含序列式或是矩陣式等。某些機器學習的演算法必須假設每一個序列資訊都是相互獨立，例如 Hidden Markov Model（HMM），但是真實世界的序列資料並不是由一連串獨立的資訊組成的。CRF 不同於其他機器學習演算法，會考量隨機序列資訊的關聯性，以求整體序列的聯合條件機率，以避免詞彙標註的偏置（bias）問題（Wallach, 2004）。本文並不試圖詳細描述 CRF 的理論與技術，相關說明請參考（Sutton, Rohanimanesh, & McCallum, 2004; Lafferty, McCallum, & Pereira, 2001）。

表 2. 徵稿通告之特徵及對應之標籤

中文名稱	英文名稱	HTML 標籤	標籤範例
會議全名	Conference Name	confname	<confname> Multimedia in Ubiquitous Computing and Security Services</confname>
會議名稱縮寫	Abbreviation of Conference Name	confabbr	<confabbr> MUCASS 2008 </confabbr>
會議地點	Conference Location	confloc	<confloc> Hobart, Australia </confloc>
會議日期	Conference Date	confdate	<confdate> October 14-16, 2008 </confdate>
會議網址	Conference Website	confwebsite	<confwebsite> http://www.sersc.org/MUCASS2008 </confwebsite>
會議主題	Conference Topic	conftopic	<conftopic> Real-time and interactive multimedia applications </conftopic>
報名截止日期	Registration Deadline	registdue	<registdue> Registration - 15th October, 2007 </registdue>
摘要提交截止日期	Abstract Submission Due	abstractdue	<abstractdue> Deadline for abstract 11 June 2008 </abstractdue>
摘要錄取通知日期	Abstract Notification	abstractnotify	<abstractnotify> Acceptance of papers - August 30, 2009 </abstractnotify>
論文提交截止日期	Paper Submission Deadline	submissiondue	<submissiondue>February 15 23, 2009 - Paper submission</submissiondue>
論文錄取通知日期	Author Notification	authornotify	<authornotify> March 23, 2009 - Author notification </authornotify>
論文定稿截止日期	Final Paper Due	finalpaperdue	<finalpaperdue> Camera-ready copies: April 7, 2009 </finalpaperdue>
海報論文截止日期	Poster Paper Due	posterdue	<posterdue> Poster Paper Submission Deadline May 15, 2008 </posterdue>
專題提案截止日期	Workshop Proposals Due	workshopdue	<workshopdue> workshop submissions due : Sunday, 2 Mar 2008 </workshopdue>
教學提案截止日期	Tutorial Proposals Due	tutorialdue	<tutorialdue> Tutorial Proposals: June 30, 2003 </tutorialdue>
博士生論壇投稿截止日期	Doctoral Consortium Due	doctoraldue	<doctoraldue> Doctoral consortium submissions due: 6 Apr 2008 </doctoraldue>

整體工作流程如圖 1 所示，包含文件前置處理、分類模型的訓練、CRF 模型的訓練三項工作。文件前置處理包含去除文件雜訊、標註學術會議資訊、Tokenization 與詞彙特性標示。

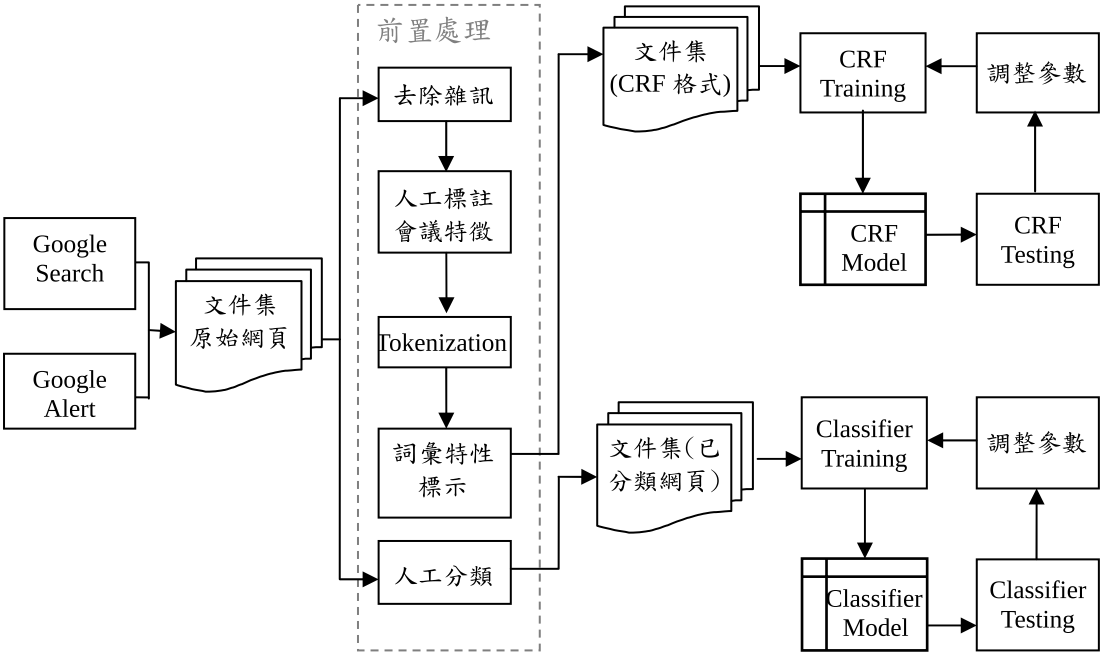


圖 1. 學術會議資訊檢索與擷取自動模型之建置流程

4.1 文件前置處理

4.1.1 去除文件雜訊

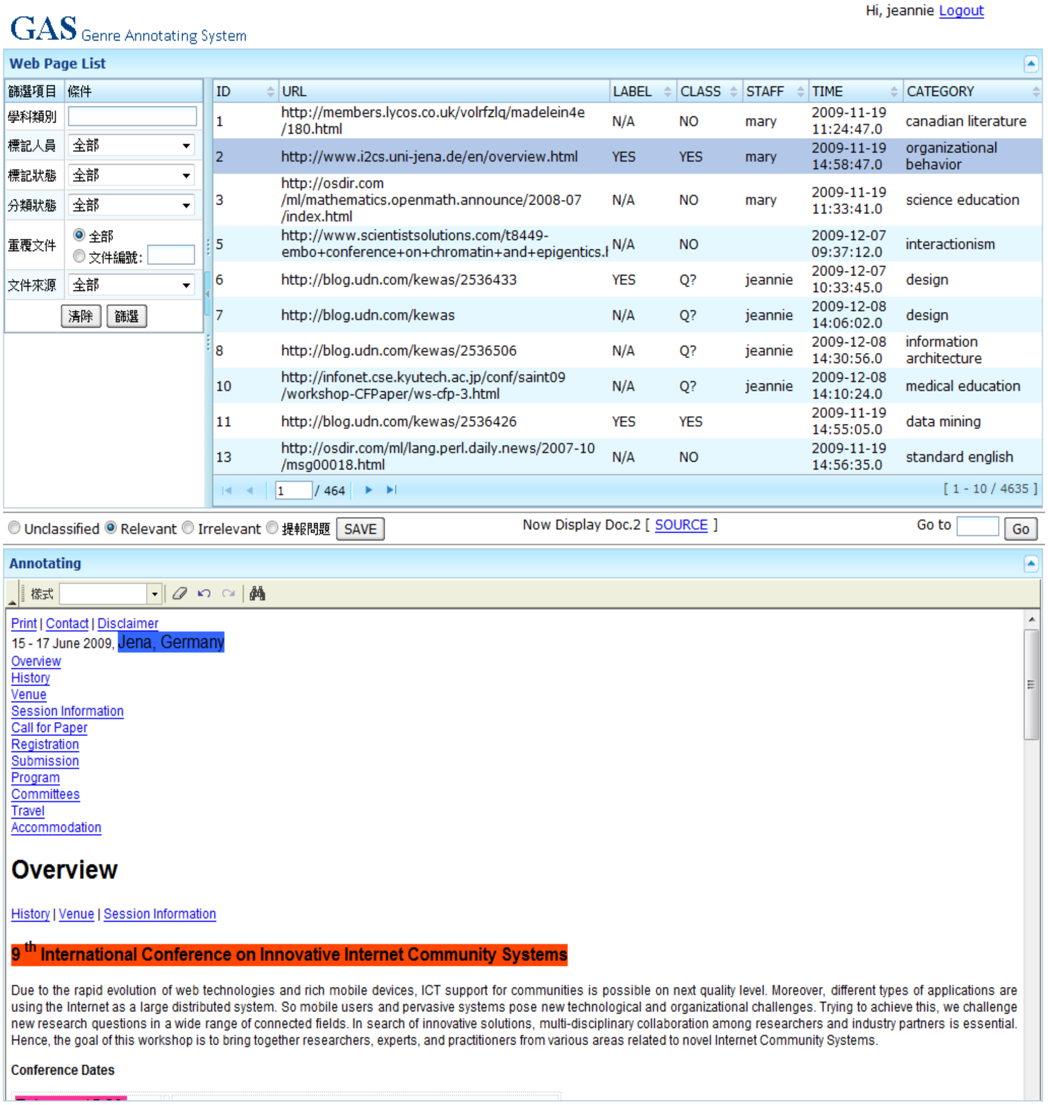
由於由網際網路蒐集的文件，通常爲 html 的網頁，包含許多各式各樣的資訊，除了該網頁的主要內容之外，尙有網頁相互連結的資訊，以及網站外部的延伸資訊。有些網頁的作者爲讓網頁更吸引使用者瀏覽，採用了動態網頁或是多媒體的呈現模式，增加處理網頁內容工作的複雜度。無論在資訊擷取的訓練階段或是正式的應用上，過多與會議資料無關的雜訊將會影響資訊欄位判斷的精確度，因此必須先去除與網頁內容主體無關的雜訊，包含廣告，圖片，網站目錄，視覺特效相關程式段落等等。

4.1.2 標註學術會議資訊

建構自動文件分類機制以及自動資訊擷取模型，需要大量的訓練資料，本研究另外建置類別標註系統（Genre Annotating System，GAS），整合內容標註與文件分類二大功能，以求內容特徵標註與文件分類標註的一致性與效率。GAS 以瀏覽器爲系統平台，爲典型的 Web-Based Application，主要功能分成三部分：候選文件瀏覽、文件分類標註，以及內容特徵標註。圖 2 爲本研究建構之類別標註系統的操作畫面。

1. 候選文件瀏覽區

圖 2 右上方的功能區塊爲候選文件瀏覽區。如前文所述，候選文件是以學門分類表的學科名稱爲關鍵字，經由 Google Search 及 Google Alert 於網路上蒐集與會議論文徵稿通告相關的網頁文件集合，經由去除雜訊處理之後，自動載入 GAS 系統。標註人員登入 GAS 後，系統會於候選文件瀏覽區展示由該人員負責標註之文件清單，標註人員也可以利用左方的查詢功能篩選網頁文件，清單上同時標示每份候選文件的標註狀態及記錄。



Hi, jeannie Logout

GAS Genre Annotating System

Web Page List

篩選項目	條件
學科類別	
標記人員	全部
標記狀態	全部
分類狀態	全部
重覆文件	● 全部 ○ 文件編號：
文件來源	全部

清除 篩選

ID	URL	LABEL	CLASS	STAFF	TIME	CATEGORY
1	http://members.lycos.co.uk/volrfzlq/madelein4e /180.html	N/A	NO	mary	2009-11-19 11:24:47.0	canadian literature
2	http://www.i2cs.uni-jena.de/en/overview.html	YES	YES	mary	2009-11-19 14:58:47.0	organizational behavior
3	http://osdir.com /ml/mathematics.openmath.announce/2008-07 /index.html	N/A	NO	mary	2009-11-19 11:33:41.0	science education
5	http://www.scientistsolutions.com/t8449-embo+conference+on+chromatin+and+epigentics.l	N/A	NO		2009-12-07 09:37:12.0	interactionism
6	http://blog.udn.com/kewas/2536433	YES	Q?	jeannie	2009-12-07 10:33:45.0	design
7	http://blog.udn.com/kewas	N/A	Q?	jeannie	2009-12-08 14:06:02.0	design
8	http://blog.udn.com/kewas/2536506	N/A	Q?	jeannie	2009-12-08 14:30:56.0	information architecture
10	http://infonet.cse.kyutech.ac.jp/conf/saint09 /workshop-CFPaper/ws-cfp-3.html	N/A	Q?	jeannie	2009-12-08 14:10:24.0	medical education
11	http://blog.udn.com/kewas/2536426	YES	YES		2009-11-19 14:55:05.0	data mining
13	http://osdir.com/ml/lang.perl.daily.news/2007-10 /msg00018.html	N/A	NO		2009-11-19 14:56:35.0	standard english

1 / 464 [1 - 10 / 4635]

○ Unclassified ● Relevant ○ Irrelevant ○ 提報問題 SAVE Now Display Doc.2 [SOURCE] Go to Go

Annotating

樣式

Print | Contact | Disclaimer

15 - 17 June 2009, Jena, Germany

Overview
History
Venue
Session Information
Call for Paper
Registration
Submission
Program
Committees
Travel
Accommodation

Overview

History | Venue | Session Information

9 th International Conference on Innovative Internet Community Systems

Due to the rapid evolution of web technologies and rich mobile devices, ICT support for communities is possible on next quality level. Moreover, different types of applications are using the Internet as a large distributed system. So mobile users and pervasive systems pose new technological and organizational challenges. Trying to achieve this, we challenge new research questions in a wide range of connected fields. In search of innovative solutions, multi-disciplinary collaboration among researchers and industry partners is essential. Hence, the goal of this workshop is to bring together researchers, experts, and practitioners from various areas related to novel Internet Community Systems.

Conference Dates

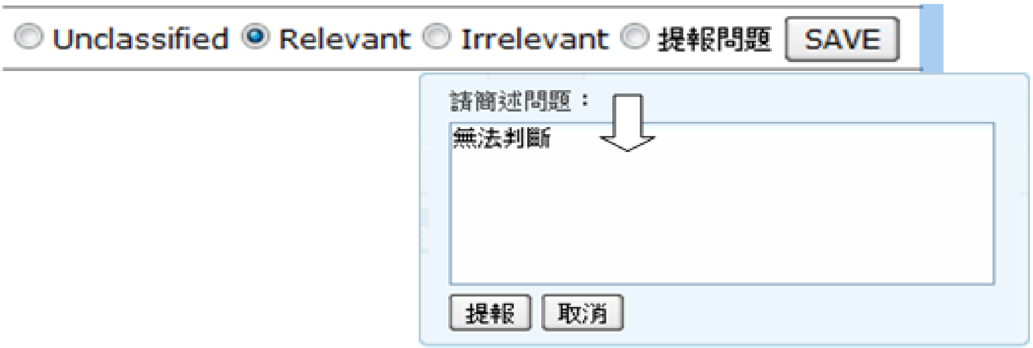
圖 2. GAS -功能畫面

2. 文件分類標註區

 文件分類標註區位於圖 2 系統功能畫面中間的狹長矩形區塊。候選網頁文件主要分成相關與不相關兩類，所謂的相關與不相關，是以該網頁文件是否與會議論文徵稿通告相關與否，作爲判斷的依據。但是，考量有些網頁文件內容資訊太複雜而無法斷定，也可以暫時不將該網頁歸類，且可以註記無法歸類的原因，作爲後續文件分類例外處理的參考，如圖 3 所示。標註人員從內容特徵標註區可檢視網頁文件，判斷該文件內容是否是會議論文徵稿通告，若確定是會議論文徵稿通告，才需要進一步針對文件內容標註各項會議資訊。

3. 內容特徵標註區

 內容特徵標註區位於圖 2 的 GAS 系統功能畫面的下方功能區塊。選取候選文件瀏覽區的任一筆資料，系統會將該網頁文件全文載入內容特徵標註區，內容特徵標註區係以 HTML 模式呈現網頁文件內容。內容特徵標註區上方的功能列，除了提供「復原動作」、「重覆動作」、「去除 HTML 標籤」、及「字串查詢」等功能按鈕之外，最重要的功能是「樣式」的下拉式選單，此樣式選單列出所有本研究採用的會議資訊特徵，標註人員於網頁內容中框選特徵資訊後，再選取對應的會議資訊特徵樣式，標註之後，所選取的特徵資訊會以特定的 HTML 標籤標示。例如會議名稱在 HTML 原始碼中標示爲<confname>會議名稱</confname>，本研究考量的會議資訊特徵與對應的 HTML 標籤請再次參見表 2。



○ Unclassified ● Relevant ○ Irrelevant ○ 提報問題

請簡述問題：

無法判斷

圖 3. GAS -文件分類標註區

4. Tokenization 與詞彙特性標示

 CRF 需切割序列性資料爲一連串 Token 後，並賦予各 Token 適當的詞性標示，再依每個 Token 的特徵向量，計算各 Token 之間的條件機率，以做爲建構詞彙辨識模型的依據。因此去除雜訊後的網頁內容，要再抽取非 HTML 標籤的字串，將字串以單一詞彙或標點符號爲單位，切割成更小的片段爲 Token，針對每一個 Token，進一步做一般詞性標示及專門詞性標示。一般詞性標示包含標點符號，大小寫，數字，日期型態等識別。專門詞性則包括地名，會議資訊經常使用專門詞彙，例如 conference、congress、 association、annual、national 等，本研究採用 GeoNames 地名資料庫為地名辨視依據，並整理會議資訊經常使用的專門詞彙，用以比對並標示相關詞彙，如表 3 所示。

表 3. 會議資訊使用之專門詞彙列表

專門詞彙類別	詞彙項目
機構名稱	Center, centre, college, department, institute, school, univ., university
組織名稱	Association, consortium, council, group, society
事件名稱	Colloquium, conf., conference, congress, convention, forum, meeting, round, roundtable, seminar, summit, symposium, table, track, workshop
時間屬性名稱	Annual, autumn, biannual, biennial, European, fall, int., interdisciplinary, international, joint, national, special, spring, summer, winter

4.2 分類模型的訓練

文件分類的目的是為了預先過濾並非論文徵稿通告的文件，以降低內容自動標註時的負擔。當系統運轉後，大量的網路文件進入系統時，必須先判斷是否為論文徵稿通告的相關文件，然後再透過內容特徵擷取功能，擷取所需要的會議資訊。由於目前有許多的開放程式碼可供使用，以開發文件分類的功能模組，本研究使用 McCallum（1996）的 Bow Library，開發統計學習為本的文件自動分類功能模組，用以過濾由網路取得的會議通告文件，Rainbow 則是基於 Bow 的應用程式，可由 http://www.cs.cmu.edu/~mccallum/bow/rainbow/取得。基本上，Rainbow 是利用已知類別的文件，統計分析各文件特徵並建立分類模型，再依此分類模型對新文件進行自動分類。在人工標註輔助系統所產生的相關文件集與不相關文件集，是收錄原始網頁文件，而不是已被人工標註特徵項目的新網頁內容，因為本研究的會議資訊自動擷取系統，是先過濾非會議通告網頁，才進行資訊擷取程序，因此文件自動分類功能模組，是以原始網頁做為訓練文件。我們進行大量的訓練與測試，使用 k-Nearest Neighbor（kNN）、Naive Bayes（NB）、Support Vector Machine（SVM）三種分類模式，隨機抽取文件進行 20 次的實驗，使用訓練文件與測試文件比例分別為（7:3）、（5:5）、（3:7），觀察分類績效的變動情形，以決定系統使用的分類模型。分類結果的優劣是以 Recall（求全率）與 Precision（求準率）評量，可以進一步將兩項指標結合為單一的 F1 指標，計算方式說明如下。每一篇文件皆已有正確的分類標記，在每一次的分類實驗，分類模型會為每一篇自動賦予其分類標記，可能與正確的分類標記一樣，或是不一樣，因此有四種可能性，如表 3 所示。

依據表 4 可以計算 Recall (R)、Precision (P)、以及 F1 Measure。

$$\mathrm{P}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FP}_i}, \qquad \mathrm{R}_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i}, \qquad \mathrm{F1} = \frac{2\mathrm{P}_i\mathrm{R}_i}{\mathrm{P}_i + \mathrm{R}_i}$$

表 4. 分類結果列聯表

Category *i*		Expert Assignment	
		TRUE	FALSE
System Judgment	TRUE	TP_i	FP_i
	FALSE	FN_i	TN_i

因爲進行了 20 次實驗，可以計算 Micro Recall、Micro Precision、Marco Recall、Macro Precision，以及對應的 Micro F1 Measure 與 Macro F1 Measure，以觀察每次實驗的變異情形，計算方式如下所示，其中 n 代表實驗次數。

$$\mathrm{P_{micro}} = \frac{\sum_{i=1}^{n} \mathrm{TP_i}}{\sum_{i=1}^{n}(\mathrm{TP_i} + \mathrm{FP_i})}, \qquad \mathrm{R_{micro}} = \frac{\sum_{i=1}^{n} \mathrm{TP_i}}{\sum_{i=1}^{n}(\mathrm{TP_i} + \mathrm{FN_i})}$$

$$\mathrm{P_{macro}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FP_i}}, \qquad \mathrm{R_{macro}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathrm{TP_i}}{\mathrm{TP_i} + \mathrm{FN_i}}$$

實驗結果如表 5 所示，Outside Test 意指測試資料與訓練資料不同，Inside Test 意指測試資料與訓練資料相同。Inside Test 的結果一定會比 Outside Test 的結果好，如果 Outside Test 的結果很接近於 Inside Test，代表分類模型的適應性很好；訓練資料越多，涵蓋面越廣，分類結果也越好。

實驗結果顯示，SVM 模型的表現最好，Naive Bayes 次之，而 kNN 最差。SVM 在 Inside Test 與 Outside Test 的表現差異最小，而 Naive Bayes 變動的幅度很大，代表 SVM 模型對於未知資料的解釋性很強。除此之外，無論是何種模型，$\mathrm{F_{micro}}$ 與 $\mathrm{F_{macro}}$ 的表現相當，代表每一次實驗結果的變異性很小。值得注意的是，本研究是採用 Recall-Oriented 的作法，調整系統參數，進行文件的自動分類，原因是希望能夠儘量取得會議相關的文件，因此較著重於 Recall。依據前述實驗的結果，本研究發展的系統將採用 SVM 模型，自動分類大量的網路文件，判定是否爲 CFP 文件後，再進一步擷取文件中的會議資訊。

4.3 CRF模型的訓練

本研究使用 CRF 模型建構會議資訊擷取的自動程序，由於目前也已有許多現成的開放程式碼可供使用，決定採用 Kudo（2010）開發的 CRF++套件，以擷取會議論文徵稿通告的特徵資訊，CRF++可由 http://crfpp.sourceforge.net/取得。吾人可以使用 CRF++開發文件自動分詞（segmenting）或內容特徵標註（labeling）等序列性資料的應用系統。CRF++宣稱使用者可以自訂資料特徵，而且計算速度快，僅使用少量的記憶體。由於 CRF++使用特定文件格式，必須將文件內容切割成一連串的 Token，以表格的形式陳列每一個 Token 的詞彙特性、版面特性以及會議資訊等特徵，無論訓練文件或是測試文件，都必須依循此特定格式編排。

表 5. 分類結果績效比較

方法	訓練：測試	Inside/ Outside	P_{micro}	P_{macro}	R_{micro}	R_{macro}	$F1_{micro}$	$F1_{macro}$
SVM	**70%**：30%	Outside Test	75.30	75.34	92.07	92.07	82.84	82.87
		Inside Test	77.94	78.31	92.70	92.70	84.68	84.90
	50%：50%	Outside Test	74.19	74.21	90.36	90.36	81.48	81.49
		Inside Test	76.07	77.09	92.14	92.14	83.34	83.94
	30%：70%	Outside Test	72.90	72.93	89.10	89.10	80.19	80.21
		Inside Test	74.83	76.08	92.85	92.85	82.87	83.63
Naive Bayes	**70%**：30%	Outside Test	78.00	78.07	62.63	62.63	69.48	69.50
		Inside Test	75.29	75.50	95.30	95.30	84.12	84.25
	50%：50%	Outside Test	76.31	76.40	63.02	63.02	69.03	69.07
		Inside Test	75.28	75.59	94.18	94.18	83.68	83.87
	30%：70%	Outside Test	69.76	69.85	95.37	95.37	80.58	80.64
		Inside Test	74.84	75.51	96.33	96.33	84.23	84.66
kNN	**70%**：30%	Outside Test	66.97	69.32	58.67	58.67	62.54	63.55
		Inside Test	56.88	57.39	94.73	94.73	71.08	71.48
	50%：50%	Outside Test	65.74	67.77	61.82	61.82	63.72	64.66
		Inside Test	56.14	56.54	95.70	95.70	70.77	71.09
	30%：70%	Outside Test	63.51	67.03	58.67	58.67	60.99	62.57
		Inside Test	57.98	59.23	91.42	91.42	70.96	71.89

完成人工標註的網頁文件轉換成此特定格式後，將其中四分之三的文件做爲訓練文件集，四分之一做爲測試文件集。透過 CRF 以訓練文件的 Token 特性，演算並建構自動標註模型，再使用測試文件測試自動標註之效果，並依測試結果調校運算參數或調整會議資訊特徵人工標註規則，以提升自動標註模型的績效。CRF 的實驗結果如表 6 所示，由於希望加強 Recall，以儘可能地擷取相關的 Entities，以避免遺漏會議資訊，因此表 6 顯示 Recall 相對較高。對於可能造成的誤判，再應用許多 Heuristic Rules 過濾不適當或是錯誤的訊息，這些 Heuristic Rules 可分爲下列五種型式：

- 序列規則（Sequence Rule）：考量時間資訊的序列性。
- 詞彙規則（Term Rule）：考量特定的詞彙。
- 位置規則（Location Rule）：考量具名實體的相對位置。
- 格式規則（Format Rule）：考量時間資訊的格式。
- 相似規則（Similarity Rule）：考量具名實體的相似性。

表 6. 具名實體的擷取

Documents / System	True Entities	False Entities
Positive Entities	1632	1079
Negative Entities	261	2785

Recall (R) = 1632/(1632+261)=86.21%; Precision (P) = 1632/(1632+1079)= 60.20%
F1 measure (F1) = (2*P*R)/(P+R)=70.89%

5. 系統實作與功能

爲了實作本研究提出的學術資訊自動擷取的機制，並提供學術會議資訊之應用服務，我們建構學術會議資訊檢索與擷取系統平台（Academic Conference Information Retrieval & Extraction System，簡稱 ACIRES）。ACIRES 由後端資訊處理系統與前端使用者系統構成，兩者皆爲自動化與即時性之服務，系統架構如圖 4 所示。後端系統蒐集網路上的學術會議資訊網頁、過濾非相關網頁、擷取會議資訊、並進而建立文件索引，前端系統是與使用者互動的入口，使用後端系統建構之索引資料，提供使用者各項服務，並與 Google Calendar 聯繫，建構個人行事曆。以下分別介紹後端資訊處理系統以及前端使用者系統的各項功能。

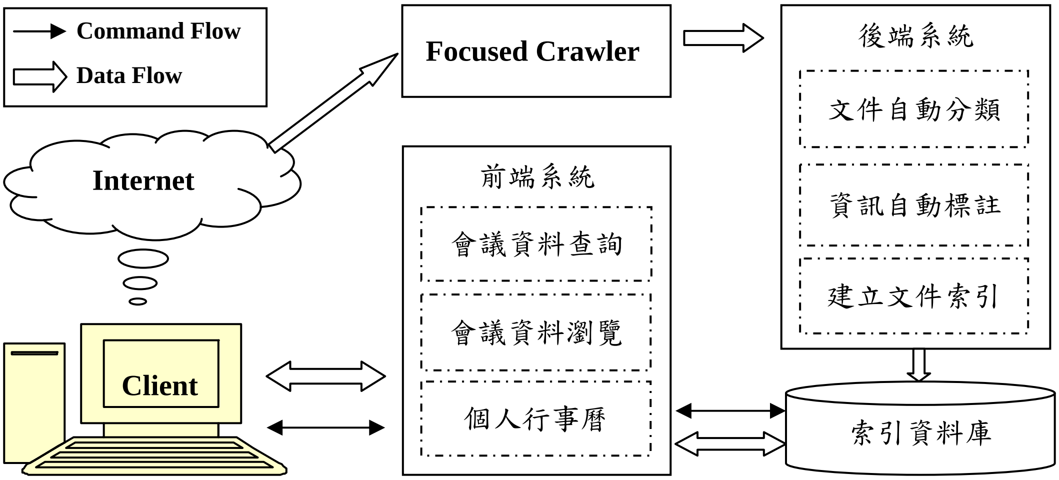


圖 4. ACIRES 整體系統架構

5.1 後端資訊處理系統

後端資訊處理系統主要的工作爲文件自動分類、資訊自動標註、以及建立文件索引，請參考圖5。後端系統使用Google Alert蒐集網路上可能的學術會議資訊、過濾無關的內容、擷取會議各項時間與地點資訊、建置文件索引資料，分別說明如下。

5.1.1 文件自動分類

ACIRES 持續以 Google Alert 快訊服務，以本研究整理的學科主題關鍵字，訂閱各主題相關網頁通知，取得最新的學術會議資訊，保持資料的即時性與時效性。由 Google Alert 蒐集而得的網頁，先經由 Rainbow Classifier 的文件分類模型，自動過濾非相關網頁。再經過去除雜訊的程序，刪除廣告，動態網頁程式等與會議資訊無關的內容。

5.1.2 資訊自動標註

已去除雜訊的網頁，進一步轉製成特定格式，以本研究建置的 CRF 資訊擷取模型，自動標註網頁中的會議資訊特徵。系統解析完成標註的文件，一一擷取各項特徵項目，再針對不同資料格式進一步處理，例如統一日期格式、轉換 HTML 特殊字元等。另外，有些網頁可能包含一個以上的學術會議資訊，因此同一份文件所擷取的項目會有重覆出現的狀況，例如有兩個會議時間、有三個會議地點等。系統則依文件排版的先後順序關係，將特徵項目分組爲多筆會議資料。

5.1.3 建立文件索引

透過自動資訊擷取所取得的各項會議資訊，以及研討會通知網頁中未被擷取的其他相關資訊，都需進一步整合爲容易查找的資料集合，以提供快速且簡便的檢索及瀏覽服務。

ACIRES 採用 Lucene 檢索系統整合所蒐集與整理的會議資料。（Apache Software Foundation, 2010）Lucene 為完整的資訊檢索系統，提供全文資料及欄位資料的索引建立與資料查詢功能。ACIRES 取用已去除雜訊的網頁內容建立全文索引。每一筆會議資料是由一份網頁全文及多個自動擷取的特徵項目所組成，這些特徵項目也是建立索引資料庫時，各學術會議資料的欄位索引項目。

5.2 前端使用者系統

如前文所述，前端系統乃是支援使用者各項功能的入口，其架構如圖 6 所示，各項功能可分為兩大模組：1) 會議資料搜尋；2) 個人行事曆。會議資料搜尋為了滿足使用者檢視資料的不同需求，實際提供了包括基本檢索、進階檢索、分類瀏覽、時間瀏覽、地點瀏覽等功能；個人行事曆則是提供行事曆的管理功能。圖 7 為前端使用者系統的入口首頁，分為時間資訊畫面、檢索功能畫面、分類瀏覽畫面、檢索結果畫面，下文簡要說明各項功能。

5.2.1 查詢學術會議資訊

系統提供基本的全文檢索功能，以及可指定欄位的進階檢索功能。當使用者進行關鍵字檢索時，系統查找研討會通告中含有查詢關鍵字的文件，依序列出查詢結果。使用者亦可進一步利用不同欄位間的布林邏輯進行進階檢索，查找更精確的會議資料。使用者點選進階檢索的鏈結，系統展現進階檢索的功能畫面，使用者可使用"AND"、"OR"、"NOT" 組合不同欄位，進階檢索提供的檢索欄位，包含所有會議資訊特徵項目，請參見圖 8。

5.2.2 檢視詳細會議資訊

查詢結果清單的每筆會議資訊包含會議名稱、會議日期、會議地點以及查詢關鍵字在文件中出現的片段。使用者可點選每筆會議資訊的[Detail]按鈕，檢視更詳細的資料。[Detail]視窗分為二部分，上方是本系統摘錄的會議基本訊息，下方式系統儲存的會議通告文件，使用者也可以進一步在詳細資料視窗點選原始網頁位址，進入該學術會議官方網站取得進一步資訊，請參見圖 9。

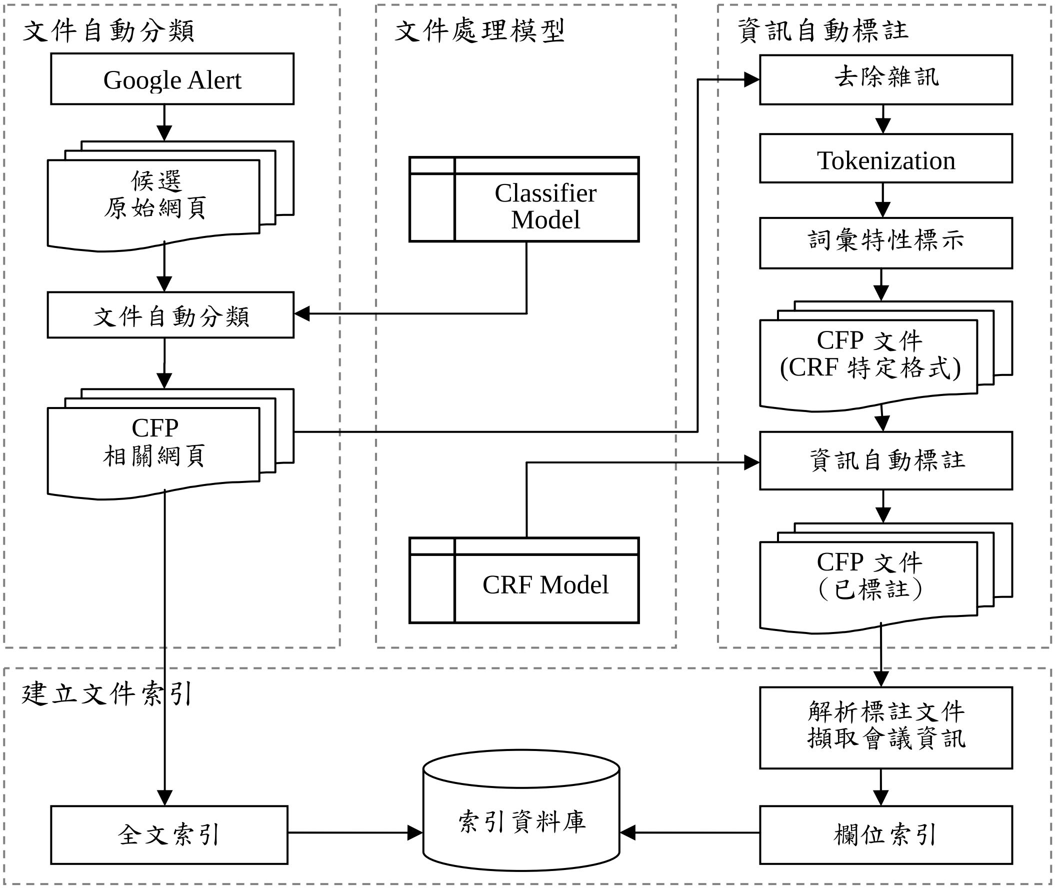


圖 5. ACIRES 系統架構：後端資訊處理系統

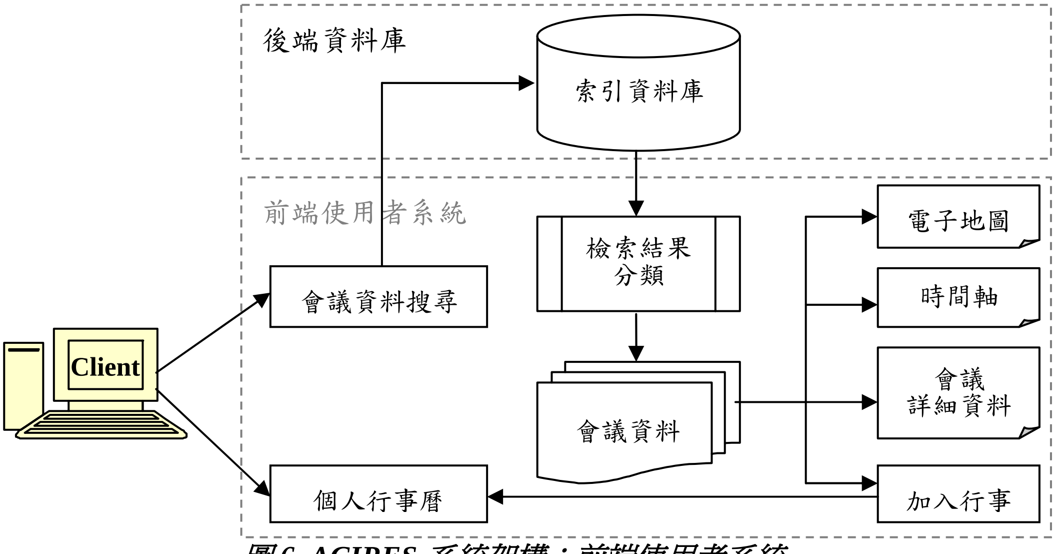
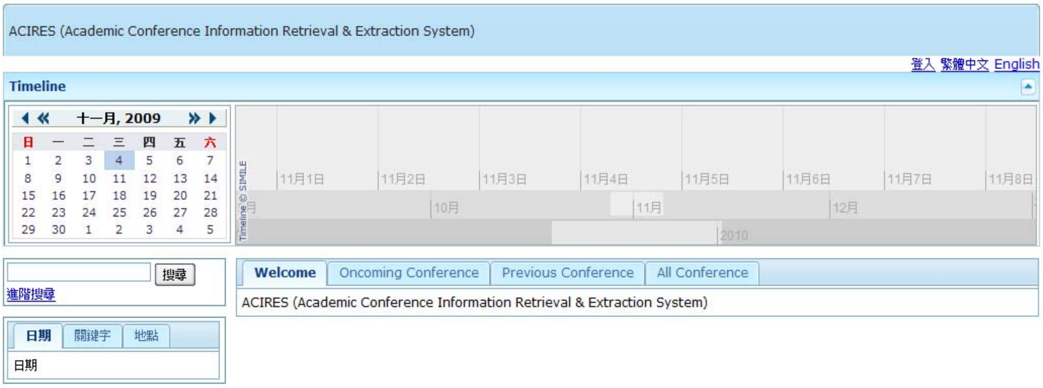


圖 6. ACIRES 系統架構：前端使用者系統



ACIRES (Academic Conference Information Retrieval & Extraction System)

登入 繁體中文 English

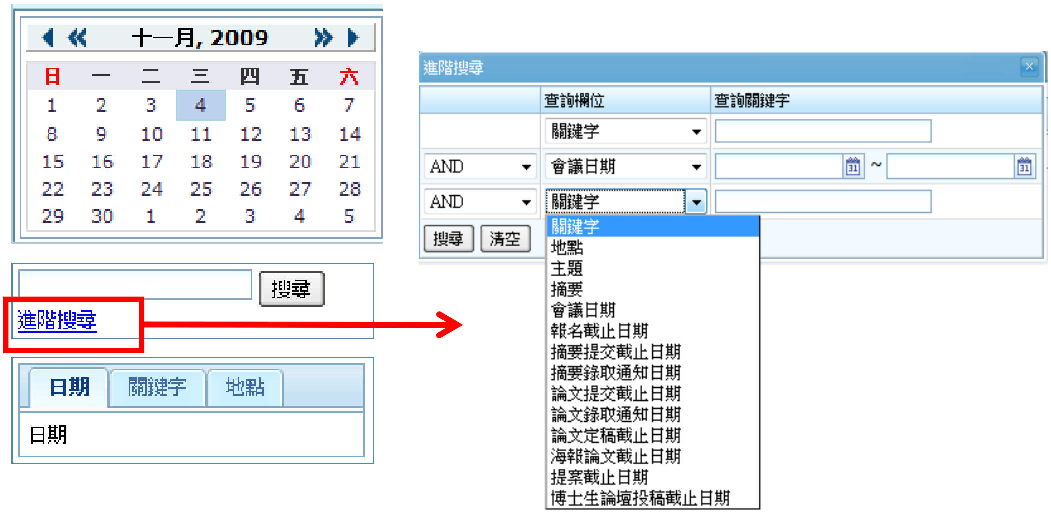
Timeline

日	一	二	三	四	五	六
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

ACIRES (Academic Conference Information Retrieval & Extraction System)

日期

圖 7. ACIRES 首頁



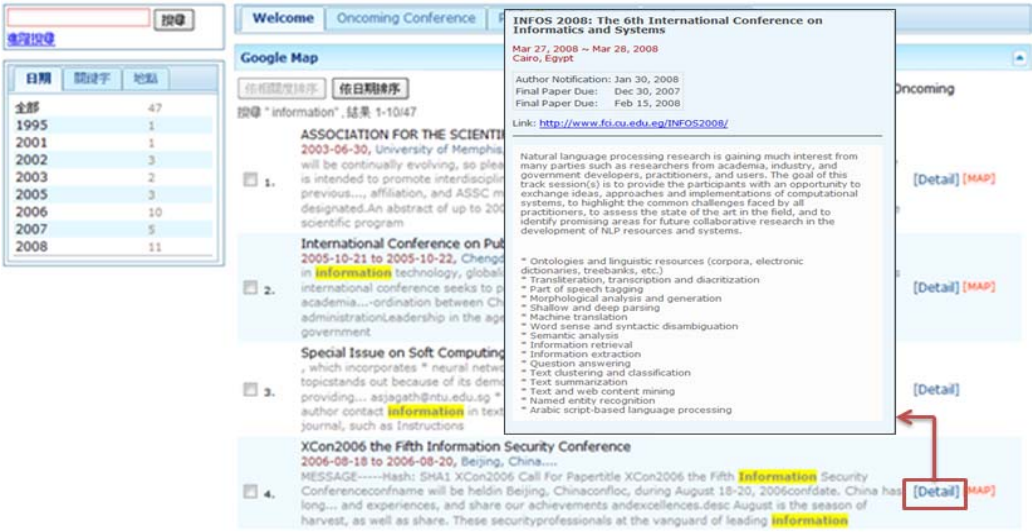
日	一	二	三	四	五	六
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

日期

	查詢欄位	查詢關鍵字
	關鍵字	
AND	會議日期	~
AND	關鍵字	

關鍵字
地點
主題
摘要
會議日期
報名截止日期
摘要提交截止日期
摘要錄取通知日期
論文提交截止日期
論文錄取通知日期
論文定稿截止日期
海報論文截止日期
提案截止日期
博士生論壇投稿截止日期

圖 8. 進階檢索



日期	
全部	47
1995	1
2001	1
2002	3
2003	2
2005	3
2006	10
2007	5
2008	11

搜尋 " information" , 結果 1-10/47

□ 1. ASSOCIATION FOR THE SCIENTI
2003-06-30, University of Memphis,
will be continually evolving, so plea
is intended to promote interdisciplin
previous..., affiliation, and ASSC m
designated.An abstract of up to 200
scientific program

□ 2. International Conference on Pub
2005-10-21 to 2005-10-22, Chengd
in information technology, globali
international conference seeks to p
academia...-ordination between Ch
administrationLeadership in the age
government

□ 3. Special Issue on Soft Computing
, which incorporates " neural netwo
topicstands out because of its demo
providing... asjagath@ntu.edu.sg "
author contact information in text
journal, such as Instructions

□ 4. XCon2006 the Fifth Information Security Conference
2006-08-18 to 2006-08-20, Beijing, China....
MESSAGE-----Hash: SHA1 XCon2006 Call For Papertitle XCon2006 the Fifth Information Security Conferenceconfname will be heldin Beijing, Chinaconfloc, during August 18-20, 2006confdate. China has long... and experiences, and share our achievements andexcellences.desc August is the season of harvest, as well as share. These securityprofessionals at the vanguard of leading information

INFOS 2008: The 6th International Conference on Informatics and Systems

Mar 27, 2008 ~ Mar 28, 2008
Cairo, Egypt

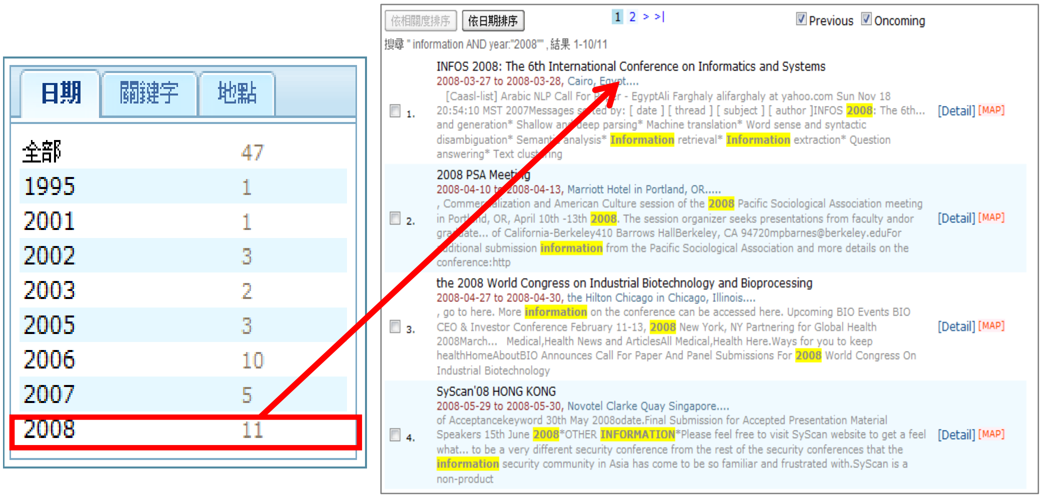
Author Notification:	Jan 30, 2008
Final Paper Due:	Dec 30, 2007
Final Paper Due:	Feb 15, 2008

Link: http://www.fci.cu.edu.eg/INFOS2008/

Natural language processing research is gaining much interest from many parties such as researchers from academia, industry, and government developers, practitioners, and users. The goal of this track session(s) is to provide the participants with an opportunity to exchange ideas, approaches and implementations of computational systems, to highlight the common challenges faced by all practitioners, to assess the state of the art in the field, and to identify promising areas for future collaborative research in the development of NLP resources and systems.

* Ontologies and linguistic resources (corpora, electronic dictionaries, treebanks, etc.)
* Transliteration, transcription and diacritization
* Part of speech tagging
* Morphological analysis and generation
* Shallow and deep parsing
* Machine translation
* Word sense and syntactic disambiguation
* Semantic analysis
* Information retrieval
* Information extraction
* Question answering
* Text clustering and classification
* Text summarization
* Text and web content mining
* Named entity recognition
* Arabic script-based language processing

圖 9. 查詢結果清單及會議詳細資料



日期 | 關鍵字 | 地點

全部	47
1995	1
2001	1
2002	3
2003	2
2005	3
2006	10
2007	5
2008	11

依相關度排序 | 依日期排序 | 1 2 > >| | ☑ Previous ☑ Oncoming

搜尋 " information AND year:"2008"" , 結果 1-10/11

□ 1. INFOS 2008: The 6th International Conference on Informatics and Systems
2008-03-27 to 2008-03-28, Cairo, Egypt....
[Caasl-list] Arabic NLP Call For Paper - EgyptAli Farghaly alifarghaly at yahoo.com Sun Nov 18 20:54:10 MST 2007Messages sorted by: [date] [thread] [subject] [author]INFOS 2008: The 6th... and generation* Shallow and deep parsing* Machine translation* Word sense and syntactic disambiguation* Semantic analysis* Information retrieval* Information extraction* Question answering* Text clustering
[Detail] [MAP]

□ 2. 2008 PSA Meeting
2008-04-10 to 2008-04-13, Marriott Hotel in Portland, OR.....
, Commercialization and American Culture session of the 2008 Pacific Sociological Association meeting in Portland, OR, April 10th -13th 2008. The session organizer seeks presentations from faculty andor graduate... of California-Berkeley410 Barrows HallBerkeley, CA 94720mpbarnes@berkeley.eduFor additional submission information from the Pacific Sociological Association and more details on the conference:http
[Detail] [MAP]

□ 3. the 2008 World Congress on Industrial Biotechnology and Bioprocessing
2008-04-27 to 2008-04-30, the Hilton Chicago in Chicago, Illinois....
, go to here. More information on the conference can be accessed here. Upcoming BIO Events BIO CEO & Investor Conference February 11-13, 2008 New York, NY Partnering for Global Health 2008March... Medical,Health News and ArticlesAll Medical,Health Here.Ways for you to keep healthHomeAboutBIO Announces Call For Paper And Panel Submissions For 2008 World Congress On Industrial Biotechnology
[Detail] [MAP]

□ 4. SyScan'08 HONG KONG
2008-05-29 to 2008-05-30, Novotel Clarke Quay Singapore....
of Acceptancekeyword 30th May 2008odate.Final Submission for Accepted Presentation Material Speakers 15th June 2008*OTHER INFORMATION*Please feel free to visit SyScan website to get a feel what... to be a very different security conference from the rest of the security conferences that the information security community in Asia has come to be so familiar and frustrated with.SyScan is a non-product
[Detail] [MAP]

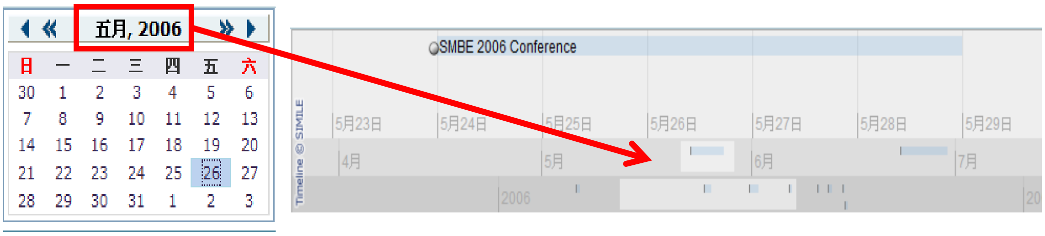
圖 10. 查詢結果分類瀏覽

5.2.3 分類瀏覽查詢結果

過去的檢索系統通常僅僅顯示檢索的結果，本系統則是進一步允許使用者依據會議舉行年份、會議舉行地點、及會議相關議題等不同觀點，更有意義地瀏覽結果。因此，本系統會爲每一次的查詢結果，進行自動分群的工作，讓使用者可進一步縮小檢索範圍，分類瀏覽查詢結果，請參見圖 10。

5.2.4 檢視會議時間資訊

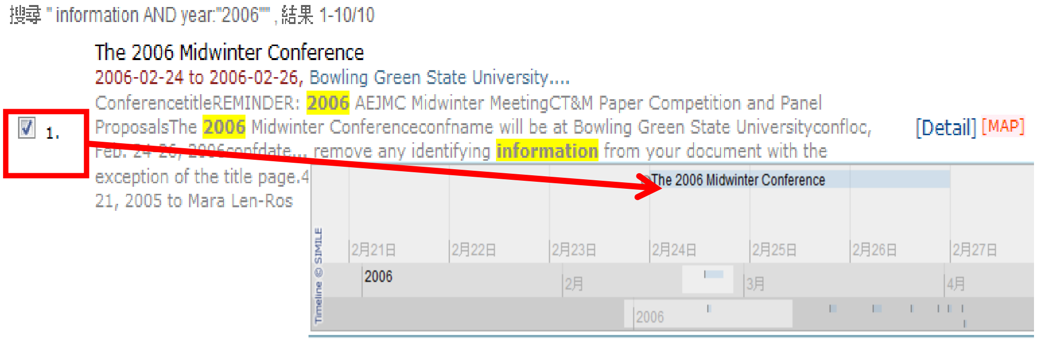
以時序方式展示研討會事件，可以讓使用者更容易安排學術活動，本系統用時間軸移動的概念，表示每個研討會事件的先後順序，讓使用者可以清楚地了解不同時間中的會議舉行狀態。使用者可直接捲動時間軸改變呈現的時間點，或是利用左上角的日曆設定日期，時間軸會即時連動至對應的時間點。當使用者勾選查詢結果清單上任一筆會議資訊，時間軸也會自動捲至該會議舉行的時間點，讓使用者可於時間軸上檢視在同一時間舉行的其他會議。點選時間軸上的事件節點，則可檢視該會議詳細資訊，請參見圖 11、12、13。



五月, 2006

日	一	二	三	四	五	六
30	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31	1	2	3

圖 11. 點選日曆捲動時間軸



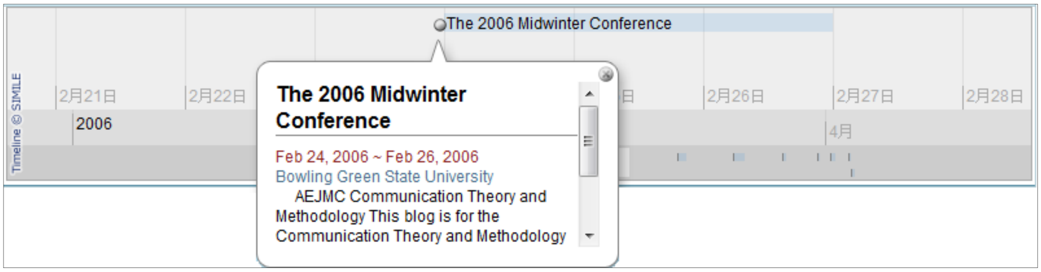
搜尋 " information AND year:"2006"" , 結果 1-10/10

☑ 1.

The 2006 Midwinter Conference
2006-02-24 to 2006-02-26, Bowling Green State University....
ConferencetitleREMINDER: 2006 AEJMC Midwinter MeetingCT&M Paper Competition and Panel ProposalsThe 2006 Midwinter Conferenceconfname will be at Bowling Green State Universityconfloc, Feb. 24-26, 2006confdate... remove any identifying information from your document with the exception of the title page.4 21, 2005 to Mara Len-Ros

[Detail] [MAP]

圖 12. 勾選會議項目自動捲動至對應時間



The 2006 Midwinter Conference

Feb 24, 2006 ~ Feb 26, 2006
Bowling Green State University
AEJMC Communication Theory and Methodology This blog is for the Communication Theory and Methodology

圖 13. 於時間軸上檢視會議資訊

5.2.5 瀏覽會議地點資訊

本系統整合 Google Map 服務，將查詢結果所得之會議舉行地點一一標示在地圖上，請參見圖 14。直接點選任一地點標示，即可檢視對應會議的相關資訊。當使用者勾選任一項會議資料，或點選檢索結果清單的[MAP]按鈕，地圖即自動將該地點放大特寫，請參見圖 15。

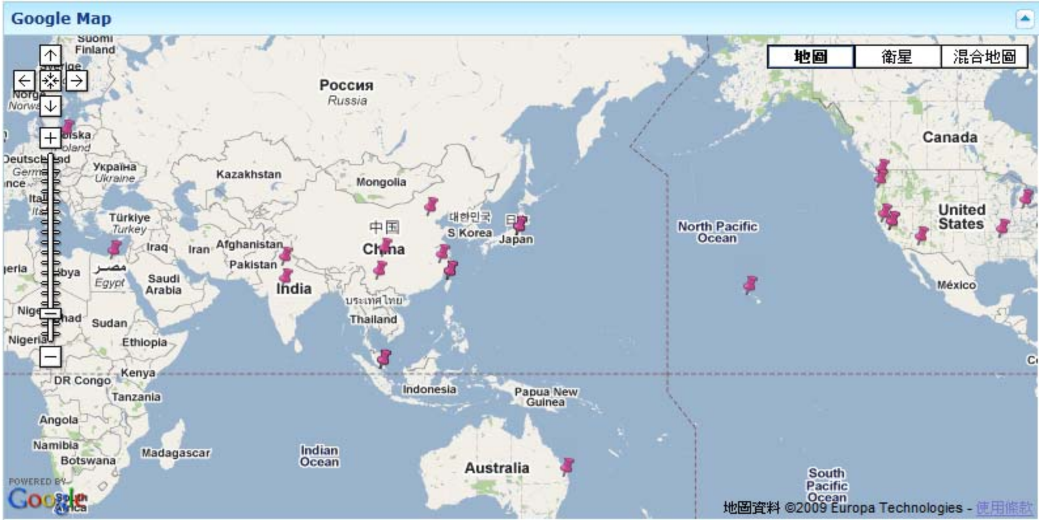
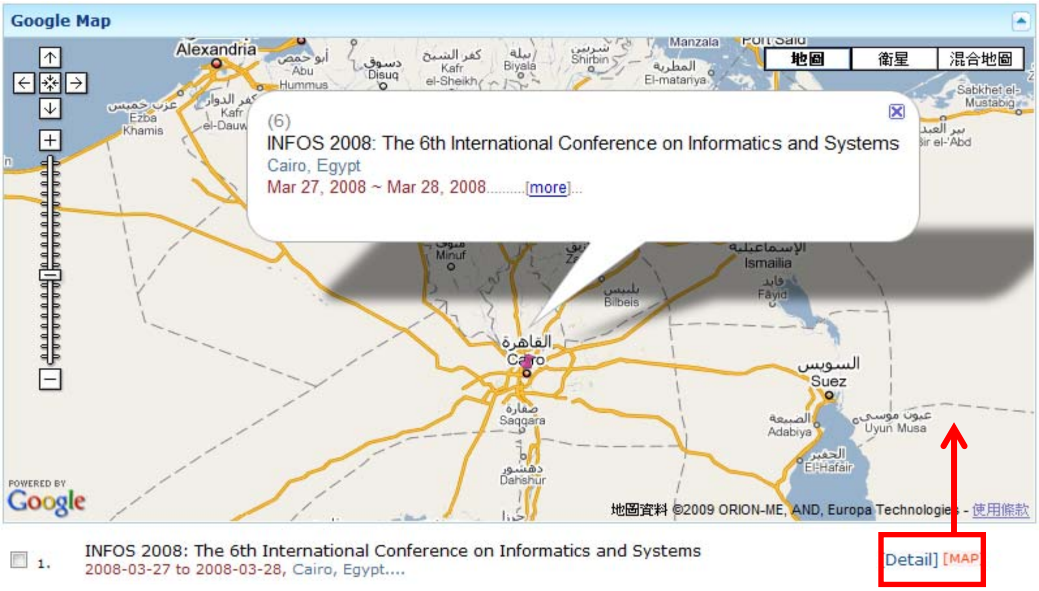


圖 14. Google Map -- Global View



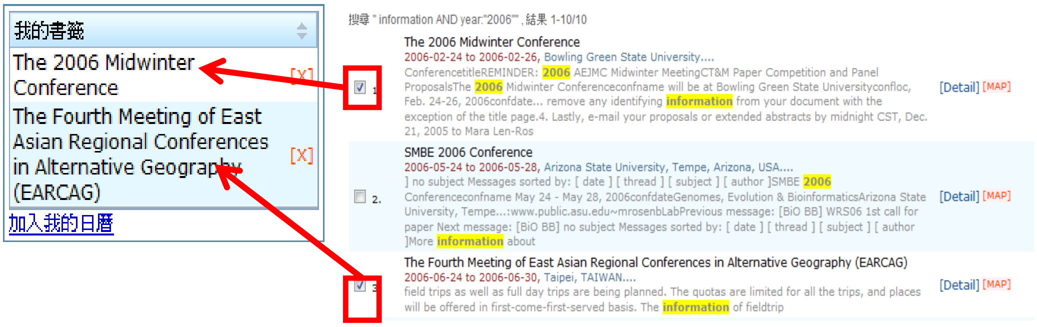
(6)
INFOS 2008: The 6th International Conference on Informatics and Systems
Cairo, Egypt
Mar 27, 2008 ~ Mar 28, 2008..........[more]...

□ 1. INFOS 2008: The 6th International Conference on Informatics and Systems
2008-03-27 to 2008-03-28, Cairo, Egypt.... [Detail] [MAP

圖 15. Google Map -- Single Spot View

5.2.6 個人行事曆

本系統進一步讓使用者儲存並記錄想繼續追蹤的學術會議資訊。由於 Google Calendar 的使用者眾多，也已經有許多應用程式可執行於多種不同的智慧型資訊裝置，使用者可以很方便地使用各種裝置查詢行事曆。因此本研究整合 Google Calendar 個人行事曆服務，讓使用者將學術會議加入行事曆，並可直接在 ACIRES 檢視個人的行事曆內容。使用者於檢索結果清單中勾選感興趣的會議項目後，所勾選項目即加入左方的書籤清單，亦可點選[x]按鈕刪除對應的會議項目，書籤清單可記錄多次檢索結果所勾選的項目，使用者可隨時新增或刪除選取的項目，請參見圖 16。


我的書籤
The 2006 Midwinter Conference [X]
The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG) [X]
加入我的日曆

搜尋 " information AND year:"2006"" , 結果 1-10/10

☑ 1. The 2006 Midwinter Conference
2006-02-24 to 2006-02-26, Bowling Green State University....
ConferencetitleREMINDER: 2006 AEJMC Midwinter MeetingCT&M Paper Competition and Panel ProposalsThe 2006 Midwinter Conferenceconfname will be at Bowling Green State Universityconfloc, Feb. 24-26, 2006confdate... remove any identifying information from your document with the exception of the title page.4. Lastly, e-mail your proposals or extended abstracts by midnight CST, Dec. 21, 2005 to Mara Len-Ros [Detail] [MAP]

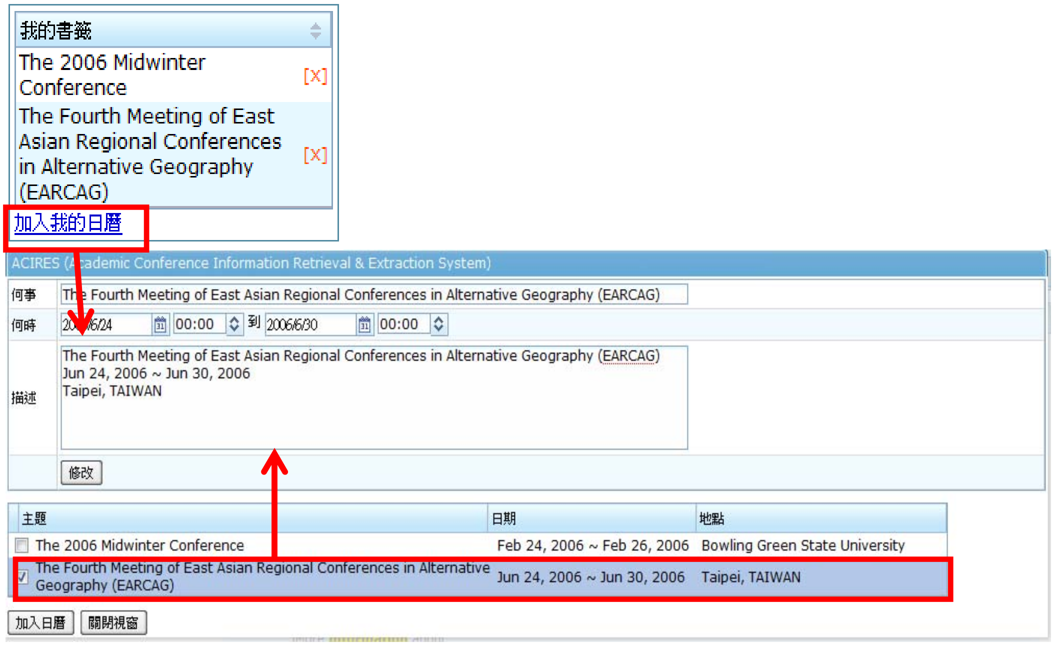
□ 2. SMBE 2006 Conference
2006-05-24 to 2006-05-28, Arizona State University, Tempe, Arizona, USA....
] no subject Messages sorted by: [date] [thread] [subject] [author]SMBE 2006 Conferenceconfname May 24 - May 28, 2006confdateGenomes, Evolution & BioinformaticsArizona State University, Tempe...:www.public.asu.edu~mrosenbLabPrevious message: [BiO BB] WRS06 1st call for paper Next message: [BiO BB] no subject Messages sorted by: [date] [thread] [subject] [author]More information about [Detail] [MAP]

☑ 3. The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG)
2006-06-24 to 2006-06-30, Taipei, TAIWAN....
field trips as well as full day trips are being planned. The quotas are limited for all the trips, and places will be offered in first-come-first-served basis. The information of fieldtrip [Detail] [MAP]

圖 16. 個人行事曆--加入書籤

會議資訊在放在「我的書籤」的清單後，尚未真正進入 Google Calendar，此時使用者可以預覽每筆會議項目資料，並可以編輯修改寫入行事曆的相關說明，請參見圖 17，當一切就緒後，可以點選「加入我的日曆」，相關會議資訊才是真正地寫入 Google Calendar。

當使用者使用 ACIRES 系統時，若已經使用 Google 帳號登入，即可在首頁直接檢視個人的行事曆內容，請參見圖 18 與圖 19。



我的書籤

The 2006 Midwinter Conference [X]

The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG) [X]

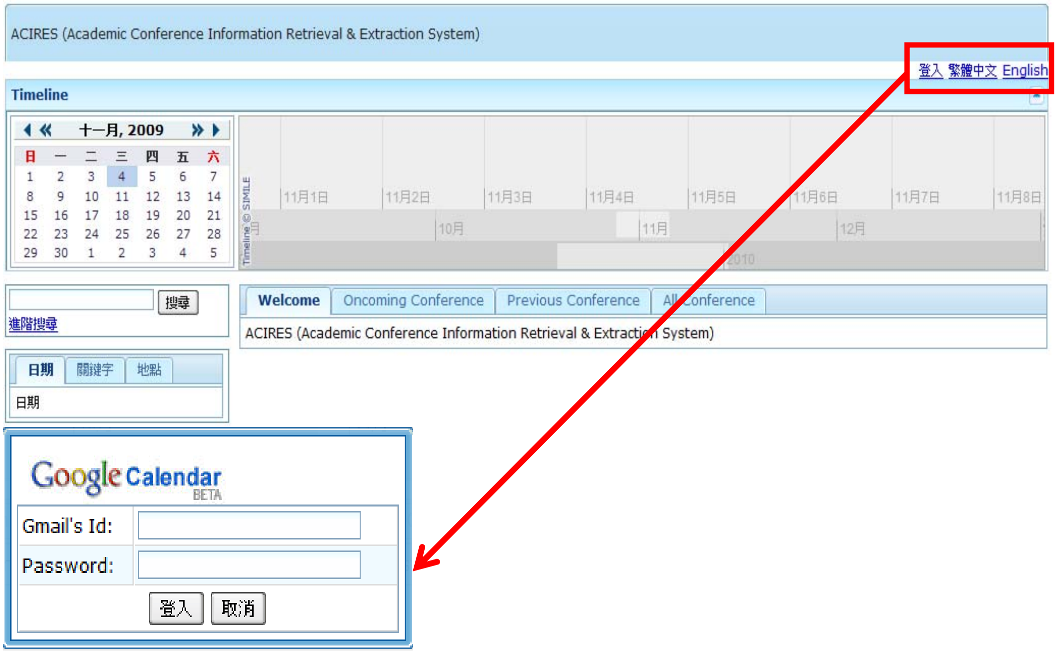
加入我的日曆

ACIRES (Academic Conference Information Retrieval & Extraction System)

何事	The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG)
何時	2006/6/24 00:00 到 2006/6/30 00:00
描述	The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG) Jun 24, 2006 ~ Jun 30, 2006 Taipei, TAIWAN
	修改

主題	日期	地點
□ The 2006 Midwinter Conference	Feb 24, 2006 ~ Feb 26, 2006	Bowling Green State University
☑ The Fourth Meeting of East Asian Regional Conferences in Alternative Geography (EARCAG)	Jun 24, 2006 ~ Jun 30, 2006	Taipei, TAIWAN

圖 17. 個人行事曆--加入行事曆前預覽



ACIRES (Academic Conference Information Retrieval & Extraction System)

日	一	二	三	四	五	六
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	1	2	3	4	5

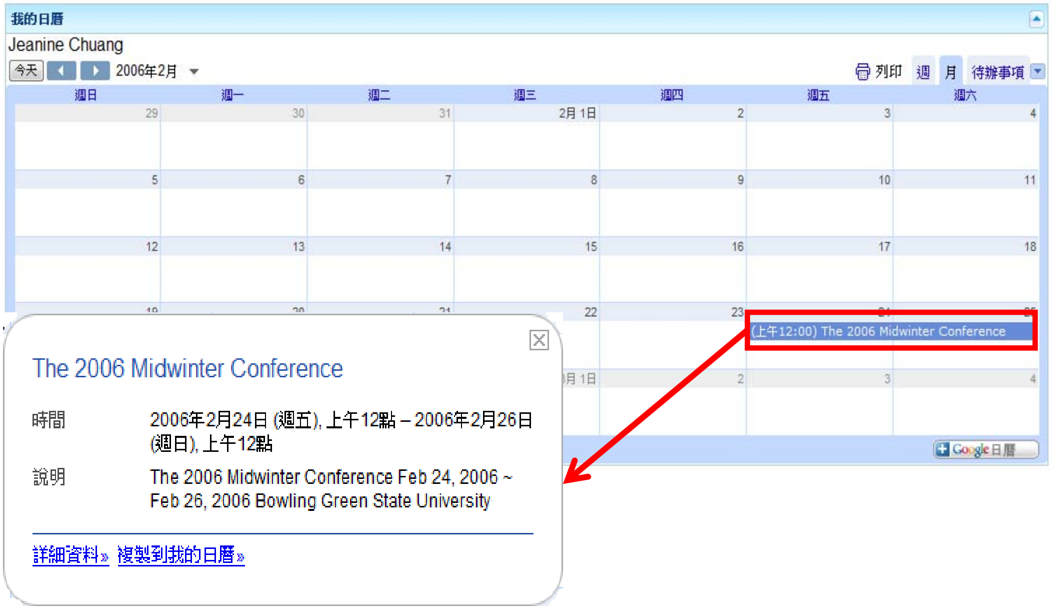
ACIRES (Academic Conference Information Retrieval & Extraction System)

Google Calendar BETA

Gmail's Id:

Password:

圖 18. 以 Google 帳號登入



The 2006 Midwinter Conference

時間 2006年2月24日 (週五), 上午12點 – 2006年2月26日 (週日), 上午12點

說明 The 2006 Midwinter Conference Feb 24, 2006 ~ Feb 26, 2006 Bowling Green State University

詳細資料» 複製到我的日曆»

圖 19. 個人行事曆--總覽

6. 結論

本文提出學術會議資訊檢索的自動處理程序，以因應廣泛的學術會議資訊檢索需求。為了確認處理程序的可行性，本研究首先進行了一系列分類績效與擷取績效的實驗，實驗顯示分類績效 F1 measure 超過 80%；擷取績效 Recall 超過 86%，F1 measure 超過 70%。第二步則是實作學術會議資訊檢索與擷取系統平台。本研究不僅提供傳統的主題檢索功能，考慮使用者的時間考量與空間考量，允許使用者由時間軸線與空間地圖瀏覽學術會議資訊，並應用檢索後分類的策略，讓使用者可以分類瀏覽檢索結果，更有意義地看待學術會議資訊。本研究同時整合了個人行事曆功能，讓學術會議資訊檢索融入研究人員的活動行程，使得前述的系統平台更具實用性。

一個相對完整的資訊系統，除了提供檢索功能外，還應該提供系統內的知識框架，允許使用者應用瀏覽的方式，檢視系統提供的各項資訊，未來本研究提出的學術會議資訊檢索與擷取系統平台將加入這樣的知識框架，以及使用者個人化的功能，例如個人專題資訊選萃（資訊過濾）。

誌謝

本研究部分研究成果獲得國科會專題研究計畫「個人化資訊服務：學術資訊之擷取」的補助，計畫編號為 NSC 96-2413-H-002-018。感謝研究助理莊雅蓁小姐、陳新瑋先生、與陳瑞呈先生的協助。

參考文獻

Apache Software Foundation. (2010). *Apache Lucene - Overview*. Retrieved Oct. 1, 2010, from http://lucene.apache.org/java/docs/index.pdf

ARWU (2010). *Academic Ranking of World Universities - 2010*. Retrieved Oct. 1, 2010, from http://www.arwu.org/

Brennhaug, K. E. (2005). *EventSeer: Testing Different Approaches to Topical Crawling for Call for Paper Announcements*. Unpublished Thesis. Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway. Retrieved Oct. 1, 2010, from http://ntnu.diva-portal.org/smash/get/diva2:348108/FULLTEXT01

Chakrabarti, S., van den Berg, M., & Dom, B. (1999). Focused Crawling: A New Approach to Topic-Specific Web Resource Discovery. In *Proceedings of the 8th International World Wide Web Conference* (Vol. 31, pp. 1623-1640). Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.43.1111&rep=rep1&type=pdf

Grishman, R., & Sundheim, B. (1996). Message Understanding Conference-6: A brief History. In *Proceedings of the 16th International Conference on Computational Linguistics* (pp. 466-471). Retrieved Oct. 1, 2010 from http://www.aclweb.org/anthology/C/C96/C96-1079.pdf

Kudo, T. (2010). *CRF++: Yet Another CRF Toolkit Version 0.54*. Retrieved Jun. 2, 2010 from http://crfpp.sourceforge.net/

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the 18th International Conference on Machine Learning* (pp. 282-289). Retrieved Oct. 1, 2010, from http://www.cis.upenn.edu/~pereira/papers/crf.pdf

Lazarinis, F. (1998). Combining Information Retrieval with Information Extraction for Efficient Retrieval of Calls for Papers. In *Proceedings of IRSG98*. Retrieved Oct. 1, 2010, from http://www.cs.strath.ac.uk/~mdd/research/publications/98lazarinis.pdf

McCallum, A. (1996). *Bow: A Toolkit for Statistical Language Modeling, Text Retrieval, Classification and Clustering*. Retrieved Aug. 4, 2009, from http://www.cs.cmu.edu/~mccallum/bow.

MUC (2001). *Message Understanding Conference Evaluation*. Retrieved Oct. 1, 2010 from http://www-nlpir.nist.gov/related_projects/muc/

Pink, B., & Bascand, G. (2008). *Australian and New Zealand Standard Research Classification (ANZSRC)*. Retrieved Mar. 2, 2010, from http://www.arc.gov.au/pdf/ANZSRC_FOR_codes.pdf

QS (2010). *World University Rankings*. Retrieved Oct. 1, 2010, from http://www.thes.co.uk/worldrankings/

Schneider, K.-M. (2005). An Evaluation of Layout Features for Information Extraction from Calls for Papers. In *Proceedings of Lernen, Wissensentdeckung und Adaptivitat* (pp.

111-116). Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.87.118&rep=rep1&type=pdf

Sutton, C., Rohanimanesh, K., & McCallum, A. (2004). Dynamic Conditional Random Fields: Factorized Probabilistic Models for Labeling and Segmenting Sequence Data. In *Proceedings of the 21st International Conference on Machine Learning*. Retrieved Oct. 1, 2010, from http://www.cs.umass.edu/~mccallum/papers/dcrf-icml04.pdf

Takada, T (2008). ConfShare: A Unified Conference Calendar that Assists Researchers in the Tasks for Attending an Academic Conference. *Journal of Information Processing Society of Japan*, 49(12), 4093-4104.

Wallach, H. M. (2004). *Conditional Random Fields: An Introduction*. Technical Report MS-CIS-04-21, University of Pennsylvania. Retrieved Oct. 1, 2010, from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.64.436&rep=rep1&type=pdf

Xin, X., Li, J., Tang, J., & Kuo, Q. (2008). Academic Conference Homepage Understanding using Constrained Hierarchical Conditional Random Fields. In *Proceedings of the 17th ACM conference on Information and knowledge management* (pp. 1301-1310). Retrieved Oct. 1, 2010, from http://doi.acm.org/10.1145/1458082.1458254

The individuals listed below are reviewers of this journal during the year of 2010. The IJCLCLP Editorial Board extends its gratitude to these volunteers for their important contributions to this publication, to our association, and to the profession.

Guo-Wei Bian

Sin-Horng Chen

Hao-Jan Chen

Zhao-Ming Gao

Moheb Girgis

Jyh-Shing Roger Jang

Lun-Wei Ku

June-Jei Kuo

Cheng-Te Li

Yuan-Fu Liao

Chuan-Jie Lin

Shu-Yen Lin

Khaled Shaalan

Hsiao-Chuan Wang

Mi-Yen Yeh

Ming-Shing Yu

2010 Index
International Journal of Computational Linguistics & Chinese Language Processing
Vol. 15

This index covers all technical items---papers, correspondence, reviews, etc.---that appeared in this periodical during 2010.

The Author Index contains the primary entry for each item, listed under the first author’s name. The primary entry includes the coauthors’ names, the title of paper or other item, and its location, specified by the publication volume, number, and inclusive pages. The Subject Index contains entries describing the item under all appropriate subject headings, plus the first author’s name, the publication volume, number, and inclusive pages.

AUTHOR INDEX

SUBJECT INDEX

The Association for Computational Linguistics and Chinese Language Processing

(new members are welcomed)

Aims：

1. To conduct research in computational linguistics.
2. To promote the utilization and development of computational linguistics.
3. To encourage research in and development of the field of Chinese computational linguistics both domestically and internationally.
4. To maintain contact with international groups who have similar goals and to cultivate academic exchange.

Activities：

1. Holding the Republic of China Computational Linguistics Conference (ROCLING) annually.
2. Facilitating and promoting academic research, seminars, training, discussions, comparative evaluations and other activities related to computational linguistics.
3. Collecting information and materials on recent developments in the field of computational linguistics, domestically and internationally.
4. Publishing pertinent journals, proceedings and newsletters.
5. Setting of the Chinese-language technical terminology and symbols related to computational linguistics.
6. Maintaining contact with international computational linguistics academic organizations.
7. Dealing with various other matters related to the development of computational linguistics.

To Register：

Please send application to:

The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

payment： Credit cards(please fill in the order form), cheque, or money orders.

Annual Fees：

regular/overseas member：	NT$ 1,000 (US$50.-)
group membership：	NT$20,000 (US$1,000.-)

life member：ten times the annual fee for regular/ group/ overseas members

Contact：

Address： The Association for Computational Linguistics and Chinese Language Processing
Institute of Information Science, Academia Sinica
128, Sec. 2, Academy Rd., Nankang, Taipei 11529, Taiwan, R.O.C.

Tel.：886-2-2788-3799 ext. 1502　　Fax：886-2-2788-1638

E-mail: aclclp@hp.iis.sinica.edu.tw　　Web Site: http://www.aclclp.org.tw

Please address all correspondence to Miss Qi Huang, or Miss Abby Ho

The Association for Computational Linguistics and Chinese Language Processing

Membership Application Form

Member ID#： ________________

Name： __________________________________ Date of Birth： _______________

Country of Residence： ____________________ Province/State： ____________________

Passport No.： ___________________________ Sex: _______________________

Education(highest degree obtained)： ____

Work Experience： ____

____

Present Occupation： ____

Address： ____

____

Email Add： ____

Tel. No： ____________________________ Fax No： __________________________

Membership Category：□ Regular Member □ Life Member

Date： _____/____/____（Y-M-D）

Applicant's Signature：

Remarks： Please indicated clearly in which membership category you wish to register, according to the following scale of annual membership dues：

Regular Member ： US$ 50.- （NT$ 1,000）
Life Member ： US$500.-（NT$10,000）

Please feel free to make copies of this application for others to use.

Committee Assessment：

中華民國計算語言學學會

宗旨：

（一） 從事計算語言學之研究
（二） 推行計算語言學之應用與發展
（三） 促進國內外中文計算語言學之研究與發展
（四） 聯繫國際有關組織並推動學術交流

活動項目：

（一）定期舉辦中華民國計算語言學學術會議（Rocling）

（二）舉行有關計算語言學之學術研究講習、訓練、討論、觀摩等活動項目

（三）收集國內外有關計算語言學知識之圖書及最新發展之資料

（四）發行有關之學術刊物，論文集及通訊

（五）研定有關計算語言學專用名稱術語及符號

（六）與國際計算語言學學術機構聯繫交流

（七）其他有關計算語言發展事項

報名方式：

1. 入會申請書：請至本會網頁下載入會申請表，填妥後郵寄或E-mail至本會

2. 繳交會費：劃撥：帳號：19166251，戶名：中華民國計算語言學學會
信用卡：請至本會網頁下載信用卡付款單

年費：

終身會員：	10,000.-	（US$ 500.-）
個人會員：	1,000.-	（US$ 50.-）
學生會員：	500.-	（限國內學生）
團體會員：	20,000.-	（US$ 1,000.-）

連絡處：

地址：台北市115南港區研究院路二段128號 中研院資訊所(轉)
電話：(02) 2788-3799 ext.1502 傳真：(02) 2788-1638
E-mail：aclclp@hp.iis.sinica.edu.tw 網址: http://www.aclclp.org.tw
連絡人：黃琪 小姐、何婉如 小姐

中華民國計算語言學學會
個人會員入會申請書

會員類別	□終身 □個人 □學生	會員編號			（由本會填寫）
姓名		性別		出生日期	年　月　日
				身分證號碼	
現職		學歷			
通訊地址	□□□				
戶籍地址	□□□				
電話		E-Mail			
			申請人:		（簽章）
		中華民國			年　月　日

審查結果:

1. 年費：

   終身會員：　10,000.-

   個人會員：　1,000.-

   學生會員：　500.-（限國內學生）

   團體會員：　20,000.-

2. 連絡處：

   地址：台北市南港區研究院路二段128號 中研院資訊所(轉)

   電話：(02) 2788-3799　ext.1502 傳真：(02) 2788-1638

   E-mail：aclclp@hp.iis.sinica.edu.tw　　網址: http://www.aclclp.org.tw

   連絡人：黃琪 小姐、何婉如 小姐

3. 本表可自行影印

The Association for Computational Linguistics and Chinese Language Processing (ACLCLP)

PAYMENT FORM

Name :____(Please print) Date: ___________________

Please debit my credit card as follows: US$______________________

❑ VISA CARD ❑ MASTER CARD ❑ JCB CARD Issue Bank:_________________________

Card No.: __________- __________ - __________ - __________ Exp. Date:________________

3-digit code: ______________ (on the back card, inside the signature area, the last three digits)

CARD HOLDER SIGNATURE : ____________________________________

Tel.: ___________________________________ E-mail: ___________________________________

Add: ____

PAYMENT FOR

US$ _____ ❑ Computational Linguistics & Chinese Languages Processing (CLCLP)

Quantity Wanted: ______

US$ _____ ❑ Publications:____________________________________

US$ _____ ❑ Text Corpora: __________________________________

US$ _____ ❑ Speech Corpora:_________________________________

US$ _____ ❑ Others: _______________________________________

US$ __________ ❑Life Member Fee ❑ New Member ❑Renew

US$ __________ = Total

Fax : 886-2-2788-1638 or Mail this form to :

ACLCLP
% Institute of Information Science, Academia Sinica
R502, 128, Sec.2, Academia Rd., Nankang, Taipei 115, Taiwan

E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

中華民國計算語言學學會

信用卡付款單

姓名: ________________________________(請以正楷書寫)　日期:：______________

卡別：□ VISA CARD □ MASTER CARD □ JCB CARD　發卡銀行：________________

卡號:__________-_________-_________- ________　有效日期：_________________

卡片後三碼：_____________________（卡片背面簽名欄上數字後三碼）

持卡人簽名：__________________________________(簽名方式請與信用卡背面相同)

通訊地址：____

聯絡電話：_______________________　E-mail：_________________________________

備註：為順利取得信用卡授權，請提供與發卡銀行相同之聯絡資料。

付款內容及金額：

NT$__________□ 中文計算語言學期刊(IJCLCLP)

NT$__________□ 中研院詞庫小組技術報告

NT$__________□ 中文（新聞）語料庫

NT$__________□ 平衡語料庫

NT$__________□ 中文詞庫八萬目

NT$__________□ 中文句結構樹資料庫

NT$__________□ 平衡語料庫詞集及詞頻統計

NT$__________□ 中英雙語詞網

NT$__________□ 中英雙語知識庫

NT$__________□ 語音資料庫______________________________

NT$__________□ 會員年費　□續會　□新會員　□終身會員

NT$__________□ 其他: __________________________________

NT$_______________=　合計

填妥後請傳真至 02-27881638 或郵寄至:
115台北市南港區研究院路2段128號中研院資訊所(轉)中華民國計算語言學學會 收
E-mail: aclclp@hp.iis.sinica.edu.tw
Website: http://www.aclclp.org.tw

Publications of the Association for Computational Linguistics and Chinese Language Processing

			AIR	AIR		
		Surface	(US&EURP)	(ASIA)	VOLUME	AMOUNT
1.	no.92-01, no. 92-04(合訂本) ICG 中的論旨角色與 A Conceptual Structure for Parsing Mandarin -- Its Frame and General Applications--	US$ 9	US$ 19	US$15	_____	_____
2.	no.92-02 V-N 複合名詞討論篇 & 92-03 V-R 複合動詞討論篇	12	21	17	_____	_____
3.	no.93-01 新聞語料庫字頻統計表	8	13	11	_____	_____
4.	no.93-02 新聞語料庫詞頻統計表	18	30	24	_____	_____
5.	no.93-03 新聞常用動詞詞頻與分類	10	15	13	_____	_____
6.	no.93-05 中文詞類分析	10	15	13	_____	_____
7.	no.93-06 現代漢語中的法相詞	5	10	8	_____	_____
8.	no.94-01 中文書面語頻率詞典（新聞語料詞頻統計）	18	30	24	_____	_____
9.	no.94-02 古漢語字頻表	11	16	14	_____	_____
10.	no.95-01 注音檢索現代漢語字頻表	8	13	10	_____	_____
11.	no.95-02/98-04 中央研究院平衡語料庫的內容與說明	3	8	6	_____	_____
12.	no.95-03 訊息爲本的格位語法與其剖析方法	3	8	6	_____	_____
13.	no.96-01 「搜」文解字－中文詞界研究與資訊用分詞標準	8	13	11	_____	_____
14.	no.97-01 古漢語詞頻表（甲）	19	31	25	_____	_____
15.	no.97-02 論語詞頻表	9	14	12	_____	_____
16.	no.98-01 詞頻詞典	18	30	26	_____	_____
17.	no.98-02 Accumulated Word Frequency in CKIP Corpus	15	25	21	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	4	9	7	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	8	13	11	_____	_____
20.	Computational Linguistics & Chinese Languages Processing (One year) (Back issues of *IJCLCLP*: US$ 20 per copy)	---	100	100	_____	_____
21.	Readings in Chinese Language Processing	25	25	21	_____	_____
				TOTAL	_____	_____

10% member discount: ___________Total Due:__________

- **OVERSEAS USE ONLY**
- PAYMENT： □ Credit Card (Preferred)
  □ Money Order or Check payable to "The Association for Computation Linguistics and Chinese Language Processing " or “中華民國計算語言學學會”
- E-mail：aclclp@hp.iis.sinica.edu.tw

Name (please print): ____________________________ Signature: ____________________________

Fax: ____ E-mail: ______________________________

Address：____

中華民國計算語言學學會

相關出版品價格表及訂購單

編號	書目	會　員	非會員	冊數	金額
1.	no.92-01, no. 92-04 (合訂本)　ICG 中的論旨角色 與 A conceptual Structure for Parsing Mandarin--its Frame and General Applications--	NT$ 80	NT$ 100	_____	_____
2.	no.92-02, no. 92-03 (合訂本) V-N 複合名詞討論篇 與V-R 複合動詞討論篇	120	150	_____	_____
3.	no.93-01　新聞語料庫字頻統計表	120	130	_____	_____
4.	no.93-02　新聞語料庫詞頻統計表	360	400	_____	_____
5.	no.93-03　新聞常用動詞詞頻與分類	180	200	_____	_____
6.	no.93-05　中文詞類分析	185	205	_____	_____
7.	no.93-06　現代漢語中的法相詞	40	50	_____	_____
8.	no.94-01　中文書面語頻率詞典（新聞語料詞頻統計）	380	450	_____	_____
9.	no.94-02　古漢語字頻表	180	200	_____	_____
10.	no.95-01　注音檢索現代漢語字頻表	75	85	_____	_____
11.	no.95-02/98-04　中央研究院平衡語料庫的內容與說明	75	85	_____	_____
12.	no.95-03　訊息爲本的格位語法與其剖析方法	75	80	_____	_____
13.	no.96-01　「搜」文解字－中文詞界研究與資訊用分詞標準	110	120	_____	_____
14.	no.97-01　古漢語詞頻表（甲）	400	450	_____	_____
15.	no.97-02　論語詞頻表	90	100	_____	_____
16	no.98-01　詞頻詞典	395	440	_____	_____
17.	no.98-02　Accumulated Word Frequency in CKIP Corpus	340	380	_____	_____
18.	no.98-03 自然語言處理及計算語言學相關術語中英對譯表	90	100	_____	_____
19.	no.02-01 現代漢語口語對話語料庫標註系統說明	75	85	_____	_____
20	論文集 COLING 2002 紙本	100	200	_____	_____
21.	論文集 COLING 2002 光碟片	300	400	_____	_____
22.	論文集 COLING 2002 Workshop 光碟片	300	400	_____	_____
23.	論文集 ISCSLP 2002 光碟片	300	400	_____	_____
24.	交談系統暨語境分析研討會講義（中華民國計算語言學學會1997第四季學術活動）	130	150	_____	_____
25.	中文計算語言學期刊（一年四期）　年份：______（過期期刊每本售價500元）	---	2,500	_____	_____
26.	Readings of Chinese Language Processing	675	675	_____	_____
27.	剖析策略與機器翻譯 1990	150	165	_____	_____
			合　計	_____	_____

※　此價格表僅限國內 (台灣地區) 使用

劃撥帳戶：中華民國計算語言學學會　　劃撥帳號：19166251

聯絡電話：(02) 2788-3799 轉1502

聯絡人：　黃琪 小姐、何婉如 小姐　　E-mail:aclclp@hp.iis.sinica.edu.tw

訂購者：　______________________________ 收據抬頭：______________________________

地　　址：____

電　　話：______________________________ E-mail: ______________________________

Information for Authors

International Journal of Computational Linguistics and Chinese Language Processing (IJCLCLP) invites submission of original research papers in the area of computational linguistics and speech/text processing of natural language. All papers must be written in English or Chinese. Manuscripts submitted must be previously unpublished and cannot be under consideration elsewhere. Submissions should report significant new research results in computational linguistics, speech and language processing or new system implementation involving significant theoretical and/or technological innovation. The submitted papers are divided into the categories of regular papers, short paper, and survey papers. Regular papers are expected to explore a research topic in full details. Short papers can focus on a smaller research issue. And survey papers should cover emerging research trends and have a tutorial or review nature of sufficiently large interest to the Journal audience. There is no strict length limitation on the regular and survey papers. But it is suggested that the manuscript should not exceed 40 double-spaced A4 pages. In contrast, short papers are restricted to no more than 20 double-spaced A4 pages. All contributions will be anonymously reviewed by at least two reviewers.

Copyright：It is the author's responsibility to obtain written permission from both author and publisher to reproduce material which has appeared in another publication. Copies of this permission must also be enclosed with the manuscript. It is the policy of the CLCLP society to own the copyright to all its publications in order to facilitate the appropriate reuse and sharing of their academic content. A signed copy of the IJCLCLP copyright form, which transfers copyright from the authors (or their employers, if they hold the copyright) to the CLCLP society, will be required before the manuscript can be accepted for publication. The papers published by IJCLCLP will be also accessed online via the IJCLCLP official website and the contracted electronic database services.

Style for Manuscripts: The paper should conform to the following instructions.
1. Typescript: Manuscript should be typed double-spaced on standard A4 (or letter-size) white paper using size of 11 points or larger.
2. Title and Author: The first page of the manuscript should consist of the title, the authors' names and institutional affiliations, the abstract, and the corresponding author's address, telephone and fax numbers, and e-mail address. The title of the paper should use normal capitalization. Capitalize only the first words and such other words as the orthography of the language requires beginning with a capital letter. The author's name should appear below the title.
3. Abstracts and keywords: An informative abstract of not more than 250 words, together with 4 to 6 keywords is required. The abstract should not only indicate the scope of the paper but should also summarize the author's conclusions.
4. Headings: Headings for sections should be numbered in Arabic numerals (i.e. 1.,2....) and start form the left-hand margin. Headings for subsections should also be numbered in Arabic numerals (i.e. 1.1. 1.2...).
5. Footnotes: The footnote reference number should be kept to a minimum and indicated in the text with superscript numbers. Footnotes may appear at the end of manuscript
6. Equations and Mathematical Formulas: All equations and mathematical formulas should be typewritten or written clearly in ink. Equations should be numbered serially on the right-hand side by Arabic numerals in parentheses.
7. References: All the citations and references should follow the APA format. The basic form for a reference looks like

Authora, A. A., Authorb, B. B., & Authorc, C. C. (Year). Title of article. *Title of Periodical, volume number*(issue number), pages.

Here shows an example.

Scruton, R. (1996). The eclipse of listening. *The New Criterion, 15*(30), 5-13.

The basic form for a citation looks like (Authora, Authorb, and Authorc, Year). Here shows an example. (Scruton, 1996).
Please visit the following websites for details.
(1) APA Formatting and Style Guide (http://owl.english.purdue.edu/owl/resource/560/01/)
(2) APA Stytle (http://www.apastyle.org/)

No page charges are levied on authors or their institutions.

Final Manuscripts Submission: If a manuscript is accepted for publication, the author will be asked to supply final manuscript in MS Word or PDF files to clp@hp.iis.sinica.edu.tw

Online Submission: http://www.aclclp.org.tw/journal/submit.php

Please visit the IJCLCLP Web page at http://www.aclclp.org.tw/journal/index.php

International Journal of
Computational Linguistics & Chinese Language Processing
vol.15 no.3-4 September/December 2010

Contents

Papers

ISSN: 1027-376X

The Association for Computational Linguistics and Chinese Language Processing