# 結合語言模型與網路知識源於列印前檢查

# Print Pickets Combined Language Models and Knowledge Resources in Web

黃昱睿　Yu-Jui Huang
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0990435@mail.ncyu.edu.tw

顏明祺　Ming-chin Yen
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0942795@mail.ncyu.edu.tw

吳冠輝　Guan-Huei Wu
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0962551@mail.ncyu.edu.tw

王耀毅　Yao-Yi Wang
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
s0962576@mail.ncyu.edu.tw

葉瑞峰　Jui-Feng Yeh
國立嘉義大學資訊工程學系
Department of Computer Science and Information Engineering
National Chia-Yi University
ralph@mail.ncyu.edu.tw

## 摘要

人們印製文件時常會疏忽了錯字而在印製完畢後才發現內容有拼字錯誤，在這種情況下往往需要重新印製一份，但這不僅浪費紙張、墨水、印表機的電力等等資源也需花費額外的時間導致降低了工作效率。假如我們能在列印前先發現文件內容的拼字錯誤並及時阻止印表機列印，則可解決我們前面所提及的問題產生。因此，本研究使用語言模型來進行比對之外還另外增加了網路搜尋的新功能搭配一起進行檢查，以便改善拼字錯誤造成列印資源的浪費。

## Abstract

People often find out spelling errors after printing the documents, therefore they need to re-print the contents of the documents. However, it is a waste of paper, ink, printer, power, other resources and so on. Even more they would need to spend extra time without efficient. This paper addresses reducing spelling errors before printing. The language model used in this research to compare and increase a new feature additionally that is a function of Internet search. Combining these research manners, this paper expect to achieve the goals of confirming, improving the spelling error, and reducing the waste of resources.



關鍵詞：拼字錯誤，語言模型，網路搜尋

Keywords: Spelling errors, Language model, Web search


## 一、緒論

隨著科技不斷的進步，人們的環保意識也隨之高漲，「永續發展」和「綠色運算」也逐漸被人們所重視並探討其如何實現，而在這個電子 3C 產品充斥的時代，我們每天日常生活中所消耗的電力是非常可觀的。

表一、台灣電力公司統計每戶家庭平均用電費用

| 台灣電力公司統計每戶家庭平均用電費用 | | | | |
|---|---|---|---|---|
| 統計資料(年度) | 2005 | 2006 | 2007 | 2008 |
| 家庭每月平均用電量(度) | 328 | 322 | 319 | 308 |
| 家庭每月平均電費支出(元) | 823 | 826 | 824 | 796 |

根據表一統計[1]，台灣家庭在 2005 年時的平均用電量是 328 度，每月平均用電支出則是 823 元；而在 2008 年時的平均用電量是 308 度，每月平均用電支出則是 796 元，由上表得知台灣家庭每年的平均用電量都有在逐漸減少的趨勢，可看出國人對於節能減碳觀念的重視。此外，電費支出的費用也隨著用電量減少而逐漸下滑。如此一來，即使台灣電力公司的電價每年都有小幅調漲，藉由減少用電量來降低用電支出仍有不錯的成效。儘管如此，雖然每個人的用電量都有減少，但是一大群人累積起來的大量消耗電力也間接的對環境造成了傷害。因此，如何避免無謂的電力浪費便是我們所要解決的主要課題之一。

現今觀察市面上，可以發現有「綠色印表機」產品，其大部分主要原理是使用了再生紙和對環境無害的大豆油墨等自然原料，經由減少對環境的破壞以達到地球永續的目的。縱使我們所使用的是對地球再無害的原料，列印錯誤所累積的損失對地球而言仍是一種傷害和資源的浪費，因爲使用者輸入錯誤所造成的列印錯誤是可以免除的。然而，目前電腦的周邊設備像是印表機等，其輸出設備常需要一些必要的耗材，如列印所需要

的碳粉、墨水匣、紙張等，如何降低其耗材的浪費便是我們所需要探討的問題。

首先，假如能減少紙張的用量，那勢必可以減少樹木的砍伐並維護目前森林綠地的地表覆蓋面積，間接的也能涵養土壤中的水分和維護空氣的品質。再來是墨水匣、碳粉的部分，如能夠有效的減少用量自然可以省碳且降低其他包裝材料的濫用。電能部分，一旦要使用印表機勢必要消耗一定量的電力，如果免除了因使用者輸入錯誤而需再重新印製的電力，相信可以減少部分碳的排放量。因此目前的印表機皆加入數種減紙少碳的構想於其中，像是列印時可選擇一張多頁、縮小列印、雙面列印或根據使用者的喜好減少紙張周邊的空白以增加可列印的面積，這些設計對於減紙少碳都具有一定的功效，假如能加入內容分析這項技術。相信對於節能減碳會有加分甚至是加乘的效果出現。

內容的分析我們將自然語言處理技術應用於列印前的檢查，排除因使用者的打字錯誤而導致所列印的文件無效造成資源的浪費。可以避免浪費紙張、墨水匣等，更能縮短列印時所花費的時間並提高文件的可靠度。綜合以上各點，我們決定透過文件內容分析這方面切入。一般我們會有列印錯誤的問題，有很大一部分的原因是列印前沒有確實的檢查內容，造成爲數可觀的資源浪費，如果我們能在列印前將列印文件周詳的檢查一遍，經判斷正確無誤後再行列印的動作，相信將會避免掉許多不要的資源耗損，不但可以避免墨水、油墨、紙張的浪費。更可以縮短時間，提升工作效率。

## 二、相關研究

與印表機驅動程式相關之項目主要可以分成微軟的 WDK 開發環境、印表機驅動系統、視窗程式設計。WDK 開發環境可經由安裝 Windows 驅動程式套件-Microsoft Windows Driver Kit (WDK)且研讀 WDK 相關文件並利用其驅動程式套件進行開發，建立驅動程式[2]。而印表機驅動系統可運用前者 WDK 驅動程式套件進行客製化動作，建立所需要的印表機驅動程式。最後視窗程式設計是使用 JAVA 撰寫一個介面視窗，讓使用者在進行列印時，假如發現文章有錯誤可以出現一個提醒的視窗來提示使用者文章中哪裡有錯誤，並且詢問使用者是否要修正或是忽略錯誤直接進行列印的動作，而不僅僅是顯示文章有誤而沒有從使用者的使用習慣著想。

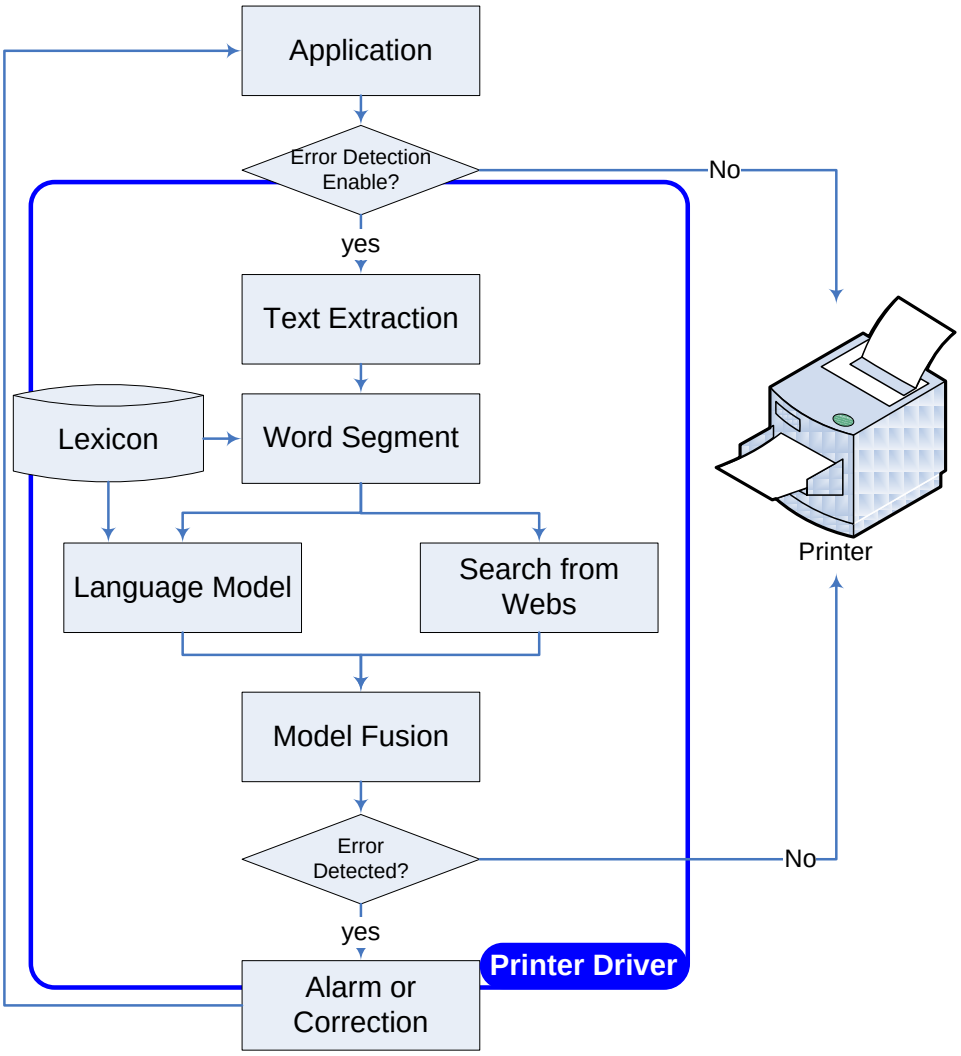
現今一般拼字錯誤檢查系統主要是先將文章裡的文字進行擷取並經過斷詞後，再將經過斷詞的詞彙進行分析，而這些文字擷取與斷詞的方法主要可以分成圖文分離技術(Graphic/Text Detection ASIC)、斷詞系統(Chinese Knowledge and Information Processing, CKIP)、自然語言理解(Natural Language Understanding)、語意擷取(Semantic Retrieval)、本體知識分析(Ontology Analysis)。由工業技術研究院所研發的圖文分離技術[3]，可讓圖文兩者混合的列印文件在圖形、文字中都能表現出原有的效果，簡言之，其爲辨識出「文字」和「圖形」的動作，讓圖形中的文字可以與圖分離。中央研究院的斷詞系統 [4]，其系統包含了大量的詞彙庫和附加詞類、詞類頻率等資料，可用其龐大的詞彙庫來輔助列印文件中文字的切割，以確保斷詞後字詞的正確性。在自然語言方面，一般通指可以隨著文化演化的語言，像是英語、漢語、日語都可算是自然語言，而同樣的字詞在不同的句子當中又會有不同的意思。舉例來說:「我把水果拿給動物們，因爲它們餓了」和「我把水果拿給動物們，因爲它們熟透了」兩句有相同的句子結構，但這兩句子中的「它們」分別代表動物和水果，假如我們不了解動物和水果的屬性就無法做區分判斷。另一

方面，自然語言也會隨著時間不斷的演變，再加上句法、表達上的彈性以及永無止盡的例外和變動也使得自然語言成爲一個不容易判斷的語言類別[5]。語意擷取的技術是將自然語言本中識別出其特定的詞彙、主題或關鍵字，將文件中的原始資料轉爲核心資訊，進一步的與其相關的主題內容作對應，例如人、事、時、地、物等。因此，藉由此項技術，可以依照其相關的主題進行解讀自然語言文本的動作，將原始的文件資料轉換成核心的資訊，以供程式或機器做進一步的使用[6]。本體知識分析主要是研究各名詞「存在」的問題，即探討哪些名詞是實際存在的，哪些名詞只是代表一種觀念，舉例來說:「社團」代表一群人擁有相同理念或性質的集合體，「幾何」表示一種特殊知識的集合。討論此相關問題並分析在各種不同領域的名詞，並描述各領域中實際的特性即爲本體(Ontology) [7]。

在語言模型與網路搜尋的技術，可分爲自動分析網路內容、N-gram 語言模型、文句自動產生、智慧型輸入法、文句自動分析。近年來大陸針對互聯網成功研製了互聯網信息採集分析系統[8]，可針對特定的網絡內容進行採集、跟蹤、預警、監控，目前在國內多個權威機構已經獲得實際應用，取代了國外同類系統，被應用單位評價爲是「融合了最新的人工智能、信息檢索、文本挖掘和互聯網技術的研究成果」，在實際應用中取得了良好的效果。N-gram 語言模型是一種文字斷詞方法，分成 Tri-gram、Bi-gram 和 Uni-gram。Tri-gram 是以 3 個字爲一組的方式來進行文字切割、Bi-gram 是以 2 個字爲一組的方式來進行文字切割而 Uni-gram 則以 1 個字爲一組的方式來進行文字檢查。近年來由美國麻省理工學院蘇波和芭茲萊發表的自動產生維基百科文章的研究[9]，他們事先收集維基百科中有關美國電影明星以及疾病的文章。然後利用這些資料來訓練電腦自動分析文章結構、擷取各段落資料、選擇文章主題文句以及透過文句組合來撰寫文稿。並且透過學界常用的自動化評估方法，得出依照此方式讓電腦自動生成文章的資訊與品質，非常接近真人寫出的文章。智慧型輸入法(網際慧智)是指接受使用者輸入注音符號與音韻，再依據上下文、使用者以往的輸入習慣，依據機率多寡或特定的挑選策略，列出並過濾使用者期望的字詞，例如自然輸入法、倚天忘形輸入法，或是微軟的新注音輸入法等。在網路上收集詞彙資料庫是件很容易的事，但單純根據辭彙頻率，並無法推得操作者的真正需求。因爲辭彙頻率是眾人的累計，對於個人的效果並不明顯。個人除了有用詞習慣、還有專業領域、輸入時期等因素會影響到同音詞的判斷。因此一個拼音注音型的輸入法，無法單憑一個很大的詞庫就搞定一切，爲了提高同音詞的正確率，互動學習是智慧型輸入法的關鍵。透過中研院的斷詞系統(CKIP)與 YAHOO 的斷章取義 API 皆可提供使用者進行自動化的文字語意分析與處理，將文章裡的文字進行切割，並且也能將詞性列出。

## 三、系統介紹

本研究之系統流程圖如下圖一所示：




圖一、系統流程圖

根據上述之系統流程圖，我們將原本的列印驅動程式(Printer Driver)加上拼字檢查模組，完成印表機驅動程式的加值功能。因此，列印文件時若使用者選擇不進行文件偵錯功能(Error Detection Enable)，那程式會立即將文件列印出來，假使選擇進行列印偵錯的動作，將會進行以下的檢查步驟：

(一)文字抽取 (Text Extraction)

首先我們要對列印的文件進行文字正確性的判斷，所以我們只抽取文字部分來做處理。然而文件中的內容格式可能包含了各種的文字大小與各類型的文字格式，因此，我們需要將要列印的文件抽取出純文字(Plain Text)以便進行後續處理的動作。這邊使用 Zan Image Printer 這個工具來幫助我們對一份含有圖文的文件進行文字抽取的動作，最後輸出一個只有文字的文件檔[10]

(二)文字切割 (Word Segmentation)

我們將文件中抽取出來後的純文字進行切割的動作，由於這些純文字內容可能包含許多詞彙或專有名詞，所以要對如何這些名詞如何做正確切割的動作，是一件相當龐大的工作。於是本模組利用中研院斷詞系統(CKIP)來進行輔助，我們利用其龐大的詞彙庫來完成更精確的斷詞，如：我把水果拿給動物們，經過斷詞後的結果可以得到，我、把、水果、拿給、動物、們，使我們在斷詞部分的處理準確性提高。

(三)語言模型(Language Model)

將切割完畢後的文字進行第一次的拼字檢查，我們使用中文十億詞語料庫(Chinese Gigaword Corpus)訓練出的語言模型進行與詞彙庫的交叉比對檢查，利用了 n-gram 中的 tri-gram、bi-gram、uni-gram 一層層的進行分析檢查[11]，假如文件中有新的詞彙或無法判讀的字詞出現，那我們將會使用網路搜尋(Web search)進行檢查，更進一步的判讀字詞的正確性以求更精確的拼字矯正。本研究將 N 值提高，精確度會提升，可是效果不明顯，且效能會由於搜索字串增加而降低，因此本研究只使用到 tri-gram。

(四)網路搜尋(Web Search)

如果使用語言模型做拼字檢查但出現無法解讀的詞彙時，將透過網路搜尋引擎進行判斷，向 Google AJAX Search API 提出要求並利用其所回傳之 JSON 編碼結果，其結果包含了搜尋字串的網站、網址、搜尋筆數等，我們取其中的搜尋筆數作判斷，如果這個詞的搜尋筆數相當高，也就能間接證明這個詞是存在的是有意義的，從另外一個角度來看也等同於將網路變身爲一個超大型資料庫，可以不斷的更新詞彙增加詞彙的數目。

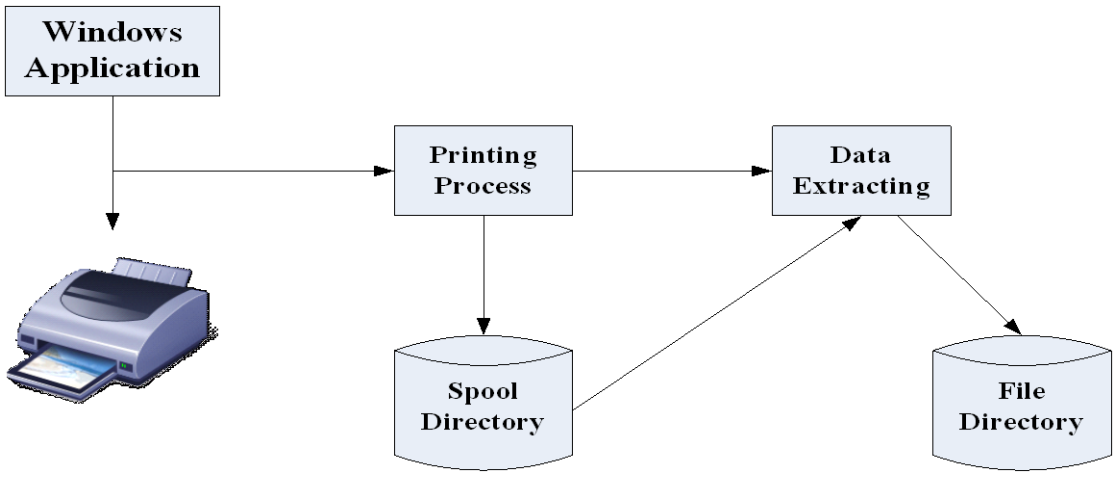
舉例來說：像「阿凡達」這個詞在幾年前沒有出現，在一般的詞彙字典中更是找無此字，是一個不存在直到最近幾年才創出來的新詞彙，不過利用網路來搜尋的特性，如果能找到相關資訊且搜尋筆數不少的話，那也就可以幾乎肯定「阿凡達」這個詞是存在且有意義的[12][13]，用這樣的方法可大幅提升拼字檢查的正確性，並與一般的拼字檢查相比有了較低的錯誤率，而印出來的文件錯誤率降低，效率自然提高，也能使資源能有效的利用不會浪費無謂的資源，進一步的達到節能減碳、地球永續的理念。

(五)模型融合(Model Fusion)

在此步驟我們將上述的語言模型(Language Model)和網路搜尋(Search From Web)以線性組合方式進行整合，並根據兩者最後的綜合計分和結果來進行錯誤偵測，假如判別文件無誤，則將列印訊號傳至印表機進行列印的動作，相反的假如判別文件有誤，則出現提醒(Alarm)或進行修正(Correction)，直到文件全面檢查無誤就可以將文件正確的印製完成。

經過以上的五個步驟，我們就可進行文件的拼字檢查的動作，透過這些動作將文件中有誤的字詞挑選出來，提醒使用者作修改的動作，進而達到減少錯誤列印發生的機率。

綜合以上所述，我們可以將列印程序簡化成如下圖二所示：




圖二、印表機列印程序示意圖

當列印的動作產生時，其中會有三個執行程序(Process)被啓動，包括使用者接觸使用的視窗應用程式(Windows Application)、列印處理程序(Printing Process)以及資料擷取程序(Data Extracting)。在資料擷取上可以著力在由列印處理程序(Printing Process)與經過列印後暫存列印緩衝區的 Spool Directory 中的檔案兩部分。但在 Spool Directory 中的檔案必須是經過轉譯成印表機語言像是 Printer Command Language (PCL)、ECL 的檔案，若從 Spool Directory 處理勢必引入許多來自印表機的語言的問題。因此本研究優先由列印處理程序中將文字擷取出來，這樣的處理程序會比在列印緩衝區的資料存取速度來的較快，可靠度(Reliability)也較高。而資料擷取出來後可以直接進行處理或者暫存至另一個緩衝區(File Directory)。

## 四、實驗結果與討論

(一)評估標準

精準度(Precision)和召回率(Recall)常常用來衡量一個系統的效能，尤其是在資訊檢索或資料探勘的領域中，我們常常會想知道當我們在檢索或搜尋系統中進行一個查詢(query)時，其回傳的結果是不是使用者需要的，還有其回傳的效率好不好等問題。因此計算出 Precision 和 Recall 這兩個值便可以解決前面所提的問題，運用量化的方式將數值呈現出來，便能使我們能更清楚的了解其系統的效能。而 Precision 和 Recall 其定義如下[14]

1. 精準度(Precision)定義：Relevant Documents Retrieved / Total Retrieved Documents
   (1) Relevant Documents Retrieved：系統判爲錯誤中實際錯誤的個數
   (2) Total Retrieved Documents：系統判爲錯誤的個數

2. 召回率(Recall)定義：Relevant Documents Retrieved / Total Relevant Documents
   (1) Relevant Documents Retrieved：系統判爲錯誤中實際錯誤的個數

(2) Total Relevant Documents：文章中實際錯誤的個數

(二) Precision 與 Recall 的定義象限表說明

表二、Precision 和 Recall 之象限表

| | 相關 | 不相關 |
|---|---|---|
| **回傳** | **Tp (true positive)** | **Fp (false positive)** |
| **未回傳** | **Fn (false negative)** | **Tn (true negative)** |

Tp：代表回傳(系統判斷爲錯誤)文字與query(實際錯誤文字)相關，系統判斷正確並回傳

Fp：代表回傳(系統判斷爲錯誤)文字與 query(實際錯誤文字)無關，系統判斷錯誤並回傳

Fn：代表回傳(系統判斷爲錯誤)文字與 query(實際錯誤文字)相關，系統判斷錯誤並無回傳

Tn：代表回傳(系統判斷爲錯誤)文字與 query(實際錯誤文字)無關，系統判斷正確並無回傳

精準度與召回率的公式如下：

1. 精準度(Precision)：Tp / Tp + Fp
2. 召回率(Recall)：Tp / Tp + Fn

(三)實驗說明

在我們的實驗中總共使用三種拼字檢查模式，經由這三種模式的比較可以得知本研究的拼字錯誤檢查系統應該使用哪一種檢查模式才可以得到較好的效率及偵錯能力，下列針對該三種模式做一個概略說明：

1. 網路+語言模型(tri-gram+bi-gram)：此模式運用了語言模型的 tri-gram 和 bi-gram，而 uni-gram 因爲是經由 CKIP 斷詞後之結果，大致上都是屬於有意義之詞彙，因此在這邊不列入考慮當中。除此之外，還搭配網路搜尋的功能進行檢查，而這裡 tri-gram 的網路搜尋的判斷筆數設定爲 1 筆，假如查詢筆數結果大於 1 筆則將該詞彙視爲正確；而 bi-gram 的網路搜尋判斷筆數設定爲 10 筆，假如查詢筆數結果大於 10 筆則將該詞彙視爲正確。
2. 網路+語言模型(bi-gram)：此模式運用了語言模型的 bi-gram，並搭配網路搜尋功能，bi-gram 的網路搜尋判斷筆數設定爲 10 筆。
3. 語言模型(tri-gram+bi-gram)：此模式運用了語言模型的 tri-gram 和 bi-gram，此外此模式無搭配網路搜尋的功能。

本研究所使用的語言模型是由「十億詞中文語料庫」（Chinese Gigaword Corpus）訓練而成的，其內容分別來自台灣與大陸，1990-2002 年的完整通訊社新聞語料。

本研究的實驗文件是從網路上抓取新聞稿，總共分爲 10 大類，分別是政治、科技、教育、旅遊、社會、生活、財經、國際、運動和影劇，每一類分別有 20 篇，實驗新聞稿總計 200 篇，並且將每篇文件內容手動設定 5 個錯誤字，經由我們的拼字錯誤檢查系統後，將其檢查結果分別依照上述 Precision 和 Recall 的公式計算出表三和表四的數據。表五爲根據 Precision 和 Recall 計算出的 F1 值，表中的生活類別之網路+語言模型得出最高的 F1 值，因爲此類別的測試不只有語言模型還有加入網路搜尋的支援，所以得出的結果最高；而財經類別之語言模型的部分，由於只有語言模型的檢查，所以財經新聞中可能有許多未知的字詞，所以才會得出最差的結果。
注意：每一格所代表的數值爲 10 大類別，各 20 篇新聞稿共同實驗後所得的平均值。

表三、精準度(Precision)實驗數據

| Precision (%) | 政治 | 科技 | 教育 | 旅遊 | 社會 | 生活 | 財經 | 國際 | 運動 | 影劇 |
|---|---|---|---|---|---|---|---|---|---|---|
| **網路+語言模型 (tri-gram+bi-gram)** | **63.70** | **72.70** | **65.09** | **57.71** | **55.77** | **60.91** | **47.55** | **47.02** | **9.95** | **59.01** |
| **網路+語言模型 (bi-gram)** | **43.06** | **50.61** | **41.18** | **42.52** | **33.00** | **37.07** | **30.58** | **37.93** | **9.37** | **5.99** |
| **語言模型 (tri-gram+bi-gram)** | **6.94** | **7.75** | **7.34** | **8.22** | **8.08** | **2.60** | **3.36** | **7.88** | **3.95** | **4.05** |

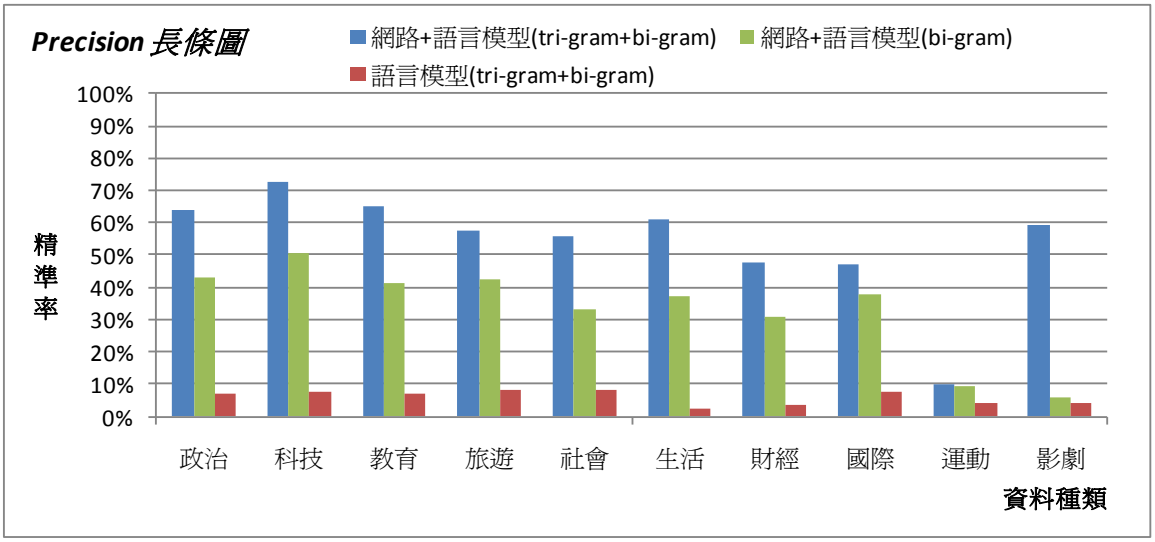
表四、召回率(Recall)實驗數據

| Recall (%) | 政治 | 科技 | 教育 | 旅遊 | 社會 | 生活 | 財經 | 國際 | 運動 | 影劇 |
|---|---|---|---|---|---|---|---|---|---|---|
| **網路+語言模型 (tri-gram+bi-gram)** | **61.00** | **52.00** | **59.00** | **63.00** | **65.00** | **68.00** | **66.00** | **63.00** | **58.00** | **68.00** |
| **網路+語言模型 (bi-gram)** | **61.00** | **52.00** | **62.00** | **65.00** | **64.00** | **66.00** | **67.00** | **64.00** | **66.00** | **65.00** |
| **語言模型 (tri-gram+bi-gram)** | **96.00** | **96.00** | **97.00** | **94.00** | **95.00** | **51.00** | **54.00** | **61.00** | **69.00** | **66.00** |

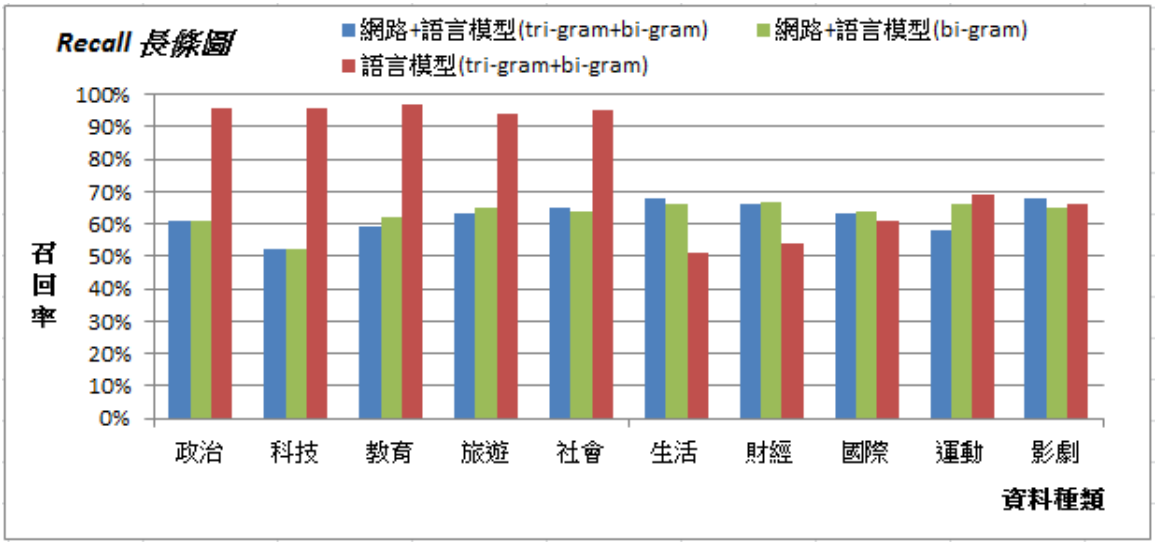
表五、F1 值(F1-score)實驗數據

| F1-Measure (%) | 政治 | 科技 | 教育 | 旅遊 | 社會 | 生活 | 財經 | 國際 | 運動 | 影劇 |
|---|---|---|---|---|---|---|---|---|---|---|
| **網路+語言模型 (tri-gram+bi-gram)** | **62.32** | **60.63** | **61.90** | **60.24** | **60.03** | **64.26** | **55.28** | **53.85** | **16.99** | **63.19** |
| **網路+語言模型 (bi-gram)** | **50.48** | **51.30** | **49.49** | **51.41** | **43.55** | **47.47** | **41.99** | **47.63** | **16.41** | **10.97** |
| **語言模型 (tri-gram+bi-gram)** | **12.94** | **14.34** | **13.65** | **15.12** | **14.89** | **4.95** | **6.33** | **13.96** | **7.47** | **7.63** |

圖五和圖六則是將表三和表四分別做成長條圖，依照下面 2 張圖所示可得知網路+語言模型(tri-gram+bi-gram)其 Precision 和 Recall 的平均值比其他兩者的 Precision 和 Recall 平均還要高，因此我們可以知道網路+語言模型(tri-gram+bi-gram)在文字檢查上擁有較高的效能表現。




圖五、精準度(Precision)長條圖




圖六、召回率(Recall)長條圖

## 五、結論與未來研究方向

本研究經過不斷的測試後發現中研院的 CKIP 系統會將錯字分別斷詞爲單一字詞，而這些單一的字詞經過語言模型與網路搜尋後幾乎會判爲正確，這也使的整個模組的檢測會疏忽這一類的錯字，因此我們將斷詞結果長度爲一的詞進行不同的組合，以期填補這類的缺失，使的整個模組更加完善，辨錯率提高並落實完善檢查的功能。

在本研究中加入網路搜尋是在拼字檢查模組中應用網路搜尋技術，將檢查的誤判率大幅減低，並且藉由網路上龐大的資料作爲拼字檢查的輔助，將網路資源當作一個大型的動態資料庫，其資料能隨著時間不斷的更新，除了涵蓋過去所有的詞彙還包含了列印當下所有的流行詞語或新創的詞彙等，使得本拼字檢查模組的正確率可以高於其他一般的拼字檢查。以及增加人性化提示訊息，常見的拼字檢查(如 Microsoft word)，在文件內容有錯誤時，雖然在該部分字詞會有特殊顏色的標記，但不會在列印時提醒使用者有錯誤，而列印了錯誤的文件，造成列印資源的浪費。而本模組在列印的時候如發現文件有錯誤，即會跳出一個提醒的視窗警告使用者需要更正，同時也可以讓使用者自行選擇是否要進行修正或是忽略錯誤直接將文件列印出來，以此模式進行文件檢查的最後一步把關。

本研究的可行性相當高，虛擬列表機的建立可以利用原有的 windows driver 改寫後作爲檢查用途，知識經濟可以分爲文字擷取與錯誤判斷兩部分。利用中研院的 CKIP 系統依照詞性種類進行斷詞，而後將斷詞後得到的結果透過模組內的語言模型(Language Model)和網路搜尋(Search From Web)來校正。語言模型中使用 N-gram 並分成三種方式來進行檢查，分別爲三字爲一組的 Tri-gram、二字爲一組的 Bi-gram 以及一字爲一組的 Uni-gram。網路搜尋使用 Google AJAX Search API，依照其傳回搜尋字串的網站內容、網址、搜尋筆數等進行判別。將以上各種功能結合，並給予人性化的操作介面讓使用者

可以容易的使用本模組，達到列印前文件內容校正的目標。

本研究可帶來環保性或節能減碳效益，如：避免紙張浪費，印刷的主要原材料是紙張和油墨，而其中紙張就占印刷總成本的 60%～70%[15]，而造成印刷錯誤的主要原因是文件內容有誤，因此減少文字錯誤是爲重要課題之一。以及減少油墨消耗，根據估計[16]台灣有超過六十萬台雷射印表機，每年約使用 300 萬支以上的碳粉匣，若能增加文字正確性，將避免重複列印而大幅降低碳粉夾的消耗量。還有可以節約能源，印刷業的能源消費(CO2 排放)占工業部門 0.13 %[17]，在總工業部門中排名 22，若能避免文字錯誤而減少重複列印，將可有效節約能源。

本研究在拼字檢查模組上結合網路搜尋技術，在檢查拼字正確的效果明顯高於只使用語言模型，但正確的拼字不代表在特定情境下是正確的用字，未來擬將情境納入拼字檢查，來改善這方面的問題。

## 致謝

本研究係國科會研究計劃「應用語言模型與網路知識源之列印拼字錯誤偵測驅動模組之開發(NSC100-2622-E-415-001-CC3)」之成果，承蒙. 國科會提供經費上的支援，特此申謝。

## 參考文獻

[1] 台灣電力公司電價查詢

http://www.rod.idv.tw/fastfood/electricity0001.html

[2] ACULIST’SBLOG Available: http://miraculist.blogspot.com/2009/08/windows-driver-kit-wdk.html

[3] Industrial Technology Research Institute – Graphice/Text Detection ASIC. Available: http://www.itri.org.tw/chi/tech-transfer/04.asp?RootNodeId=040&NodeId=041&id=2353

[4] Academia Sinica – Chinese Knowledge and Information Processing. Available: http://ckip.iis.sinica.edu.tw/CKIP/publication.htm

[5] Bates, M. (1995). Models of natural language understanding. Proceedings of the National Academy of Sciences of the United States of America, Vol. 92, No. 22 (Oct. 24, 1995), pp.9977–9982.

[6] R. K. Srihari, W. Li, C. Niu and T. Cornell,"InfoXtract: A Customizable Intermediate Level Information Extraction Engine",Journal of Natural Language Engineering, Cambridge U. Press , 14(1), 2008, pp.33-69.

[7] Eldred, Michael, Social Ontology: Recasting Political Philosophy Through a Phenomenology of Whoness ontos, Frankfurt 2008 xiv + 688 pp.ISBN 978-3-938793-78-7

[8] 互聯網內容的海量信息自動分析技術

http://www.ilib2.com/A-cstaID~CG2008435623.html
[9] 科學人雜誌網站-語言研究維基化
http://sa.ylib.com/circus/circusshow.asp?FDocNo=1448&CL=95
[10] Zan Image Printer. Available:
http://www.zan1011.com/index.htm
[11] Christopher D. Manning, Hinrich Schütze, Foundations of Statistical Natural Language Processing, MIT Press: 1999. ISBN 0-262-13360-1
[12] Jaime Teevan, Eytan Adar, Rosie Jones, Michael Potts.History repeats itself: Repeat Queries in Yahoo's query logs. Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '06). 2005: pp. 703-704.
[13] Amanda Spink, Dietmar Wolfram, Major B. J. Jansen, Tefko Saracevic. Searching the web: The public and their queries. Journal of the American Society for Information Science and Technology. 2001,52(3): 226-234.
[14] Precision and Recall. Available:
http://en.wikipedia.org/wiki/Precision_and_recall
[15] 如何減少商業輪轉印刷機的紙張消耗
http://www.hkprinters.org/news/news.asp?sub_id=560
[16] 優彩環保再生碳粉匣
http://www.ucolor.org.tw/
[17] 台灣綠色生產力基金會-節能服務網
http://www.ecct.org.tw/print/52_3.htm