The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 218-232

以語言模型判斷學習者文句流暢度

陳柏霖　Po-Lin Chen, *吳世弘　Shih-Hung Wu
朝陽科技大學資訊工程系
Department of Computer Science and Information Engineering
Chaoyang University of Technology
streetcatsky@gmail.com
*shwu@cyut.edu.tw (contact author)

摘要

因應自動化作文教學系統之需求，我們將開發多種中文自然語言處理功能。本文將以作文句子的通順程度偵測為目標，我們提出基於語言模型（language model）結合國中生作文語料知識庫的方法，並且使用資訊檢索的技術來改善系統效能，開發出第一套針對句子通順程度的偵測系統，能更快更正確偵查學生文章內容不通順的地方。系統分為二個部份：語言模型訓練模組和中文語料擷取測試模組。我們的實驗證明了以語言模型理論為基礎的句子通順度自動偵測系統能夠有效偵測不通順的句子。提供本國學生或外籍學生學習作文時的輔助工具。

關鍵詞：中文，作文，語言模型，N 元語言模型，句子流暢度

一、緒論

由於現代科技以及 3C 產品的普及，使得孩子頻繁的接觸電視、網路、手機…等，因此容易缺乏與人之間互動、溝通以及情感的表達，相對的，學生寫的作文常常是以流水帳交代經過，有的學校甚至不考作文，但隨著教育政策的變動，國中教育會考加入了作文評量的項目，使的作文再度受到學生及家長的重視。可是受限於學校教學時數，作文較弱的學生容易缺少補救的機會。我們認為未來自學作文以及在家練習，可以藉由自動化的作文教學系統輔助。而本系統開發作文教學系統之句子流暢度偵測，經由系統回饋的診斷結果可以讓學生對詞句組合的理解力有所提升，幫助學生寫出較流暢的句子，藉此提高他們的作文分數。系統所依賴的 N-gram 語言模型，它的特性是計算字詞間組合的機率，機率越高的話字詞組合的正確性越高也就是越流暢，而語言模型效果相當依賴大型的訓練語料，這是語言模型然能待克服的缺點，例如資料稀疏(Data sparseness)的問題，可以使用平滑(smoothing)的方法解決；以及跨領域的問題，只要訓練語料的性質越不同於測試的文章，我們所建立語言模型的效果就越差，因此語料庫也要跟著改變。

二、研究動機

要幫助學生寫好的作文首先要讓系統知道如何判斷出一篇是好的作文，國中基測作文的評量主要以四個範疇為主:”基測寫作測驗雖然採用整體性評分方法，但評分的時候仍

然已考慮立意取材、組織結構、遣詞造句、錯別字、格式及標點符號等四項核心技巧為主軸”(陳滿銘 2007，396)。這四個作文評量範疇並不是任意規定的，而是依照作文的構成過程中所需要的元素決定這些評量範疇的。因此這些作文評量範疇不容易被變更。以下說明如何將作文評量為 6 種不同的等級(如表一 [1])，而本系統針對四個面向中-遣詞造句的句子流暢度進行研究。

表一、國中生基本學力測驗作文測驗評分規準[1]

級分	國民中學學生基本學力測驗寫作測驗評分規準一覽表
六級分	六級分的文章是優秀的，這種文章明顯具有下列特徵： ※遣詞造句：能精確使用語詞，並有效運用各種句型使文句流暢。
五級分	五級分的文章在一般水準之上，這種文章明顯具有下列特徵： ※遣詞造句：能正確使用語詞，並運用各種句型使文句通順。
四級分	四級分的文章已達一般水準，這種文章明顯具有下列特徵： ※遣詞造句：能正確使用語詞，文意表達尚稱清楚，但有時會出現冗詞贅句；句型較無變化。
三級分	三級分的文章在表達上是不充分的，這種文章明顯具有下列特徵： ※遣詞造句：用字遣詞不太恰當，或出現錯誤；或冗詞贅句過多。上的錯誤，以致造成理解上的困難。
二級分	二級分的文章在表達上呈現嚴重的問題，這種文章明顯具有下列特徵： ※遣詞造句：遣詞造句常有錯誤。
一級分	一級分的文章在表達上呈現極嚴重的問題，這種文章明顯具有下列特徵 ※遣詞造句：用字遣詞極不恰當，頗多錯誤；或文句支離破碎，難以理解。
零級分	使用詩歌體、完全離題、只抄寫題目或說明、空白卷

(一)、立意取材

這裡主要評量所寫的作文內容是否符合主題，就如蔡英俊(2006，1)提到”立意取材:主要在評量學生是否能切合題旨並選擇合適的素材”。

(二)、結構組織

目前國中作文修改系統少了針對連接詞錯誤的處理:林素珍分析國中作文的錯誤中，結果顯示:”在行文佈局方面所犯錯誤的統計:有 45.3%的作品在文意的承接上不連貫，是比較嚴重的問題”(林素珍 2007，158)。蔡英俊(2006，2)提到”在結構組織上的基本要求，則是意念前後一致(首尾連貫)和結構勻稱。

(三)、遣詞造句

我們初步分析了一百份國中作文的結果顯示，如果作文平鋪直敘很少用修飾詞的作文大概是三到四級分。洪美雀(2013， 279-280)建議學生使用「敘事加描寫」取代「單純敘事」而且將修飾詞分級，如下表二。我們使用國中三年的國文課本裡面的詞彙以及國中作文語料庫裡面的詞彙，這些詞彙不僅府和國中生的使用程度也不會有艱澀以及少用詞彙的發生。

表二、修飾詞分級表(洪美雀 2013， 81)

四級分用語	五級分用語	六級分用語
很累	疲累	疲憊
很吵	吵鬧	喧囂

由於國中三年的各色國文教材仍然有難易度的分別，通常三級分以下的作文使用的詞彙都停留在國二以下，沒有達到國三的等級。所以分類國中國文課本的詞彙等級是具有意義的。文獻中依照各級分作文詞彙的程度不同，相同意思但表達方式不同，依照等級由低到高分類，例如:3 級分:我肖想能考第一名，4 級分:考第一名對我來說真是非分之想，5 級分:我妄想能考到第一名，6 級分:覬覦著第一名寶座的我(洪美雀 2013， 70)，以及 4 級分:遇到，5 級分:相逢，6 級分:邂逅(洪美雀 2013， 81)，我們這裡主要針對冗贅詞的偵測來處理。

(四)、錯別字、格式與標點符號

錯別字部分系統可以藉由正確作文的語料庫來尋找、比對新作文的錯別字，因此我們可以偵測錯別字。將作文格式轉變成電腦可以辨識的格式，國中生作文在書寫時是由上而下，由右至左，此外抄寫標題時要空 4 格，每段前面空兩格。作文長度常常會影響作文的等級就如洪美雀(會考的核心老師)提到:”作文考題的導文後面均出現「文長不限」，「文長不限」是為了怕人批評以字數論文章優劣，不過不要被騙了，學測至少寫六百字，指考至少寫五百二十字，沒有這樣的分量，都不可能得到高分! ”(洪美雀 2013， 102)。作文的分段也會影響到評分，一級分大都是一到三行寫成段或兩段(依據修改國中作文的老師之專家判斷)，行數 4 行以上約 6、7 行以下但不包含抄綠題目引導的句子，且有兩段大概是二級分。如果只有三段大部分最高是三級分，通常是七到十二行。至少寫 4 到 6 段，不要超過 6 段通常是四級分以上。內文空白的話就是零級分。根據林素珍的研究顯示標點符號錯誤的比例很高”32.9%的作品標點使用不當，26.5%的作品斷句不當或誤置標點，兩者占了將近六成的比重自然是不容忽視的” (林素珍 2007，159)。

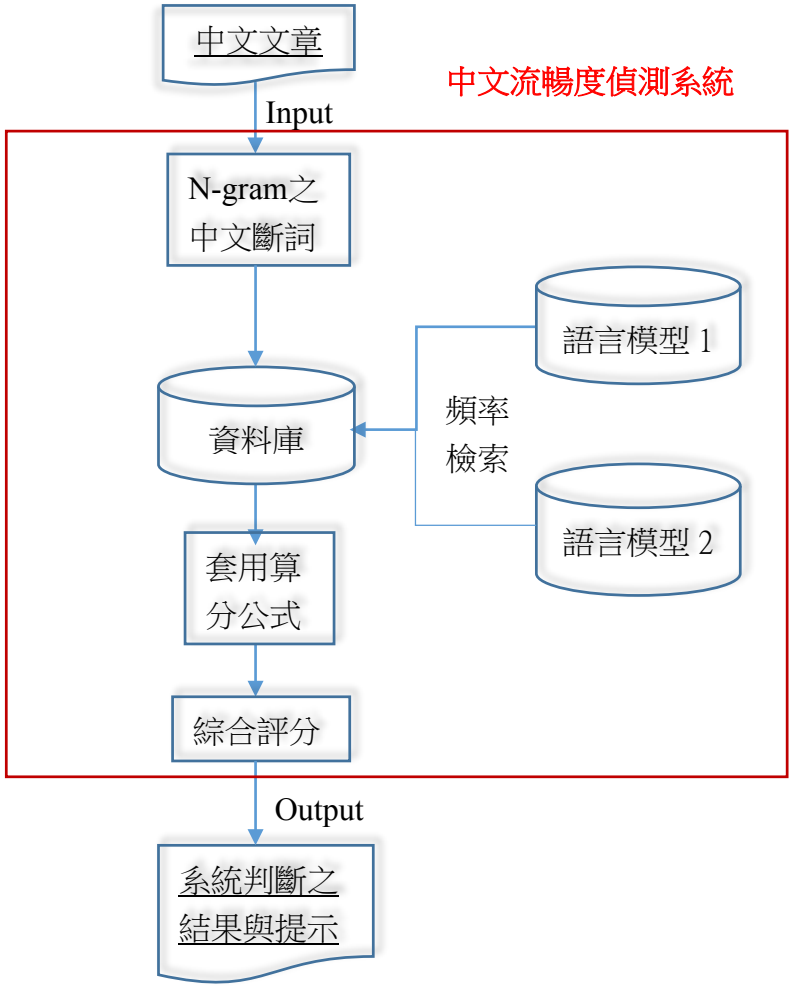
(五)、研究目的

綜合上述的說明，本論文主要的研究在於句子流暢度的偵測，正確判斷句子是否通順，系統設計於可解決一般性問題，可隨著訓練集的增加，而增強對句子的判斷。雖然這只是一小起步，此系統未來將會整合到電腦作文自動評分系統，電腦自動評分系統能 24

小時不間斷地提供服務，隨時提供學習的機會，而且學校的老師一次面對許多的學生，學生難以獲得即時的評價回饋。學生寫的作文經由分散式的診斷模組(本系統為其中一個診斷模組)分別診斷個別面向的優缺點，之後產生一份可擴展的作文診斷清單。這個清單裡整合各面向最後的診斷結果，提供後面評分模組以及雷達圖的產生。在四個面向裡面，「錯別字、格式與標點符號診斷模組」技術上是目前最成熟的，如有明顯的錯誤將均會被診斷出來並且糾正。當作文各個的診斷結果產生後(立意取材診斷模組、遣詞造句診斷模組、結構組織診斷模組等)，我們就可以給作文評定等級。依照作文在四個面向的表現，機器學習程式可以訓練出穩定的分類器，將作文分為零到六級分，產生對應的雷達圖，接著評語依照各別面相診斷的結果產生，然後合併一起呈現特徵細節，但是電腦可以詳細地將各種特，這些知識的檢測需要搭配自然語言處理的工具程式以及語言資源，最基礎的前處理就是斷詞以及標註詞性(POS tagging)，然後依照各模組需求來增加處理的知識。最後，說明實驗結果與評估的分析討論，藉以驗證本論文擷取之效能。

三、研究架構

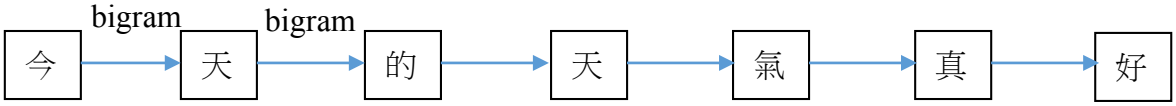
下圖一是中文作文流暢度偵測系統運作的流程圖，首先將測試的資料，包括手工設定的句子以及中文作文輸入到句子流暢度偵測系統，系統會自動偵測計算分數，之後分數若



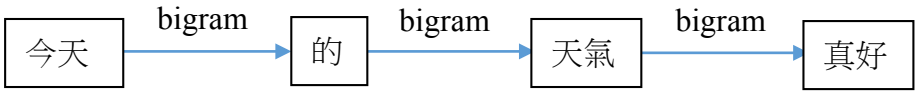
圖一、中文作文流暢度偵測系統運作的流程圖

高於一定的門檻值將會提示這可能是個不通順的句子，評估系統的效能的部分，我們把測試結果經由中文流利人是來進行檢閱，接著使用 Recall 與 Precision 評估系統偵測能力。我們將分析本系統的實驗結果，並且根據系統缺失，來一步一步提出改善的方法，希望未來更新版本的系統，能改善實驗結果、提昇效能，並且觀察特殊案例，包括系統誤判為不流暢的句子以及錯放不流暢的句子，進一步分析錯誤的原因，設法改善系統。自然語言處理(Natural Language Processing, NLP)的領域包含了語音辨識[7][8]、資訊檢索[2][3]、文件分類、手寫辨識以及機器翻譯[4] [5]…等等，而語言模型（Language Model, LM）是自然語言處理重要的技術之一[6]，語言模型統計並且紀錄了大量語料庫的詞頻及機率，它特性就是可以依據過去的訓練資料，也就是曾經出現的字，預測下一個字出現的機率，因此也能藉此計算出一個句子的機率，機率越大代表這句子越常出現，也就是越為通順，反之如果機率越低，代表這句子的寫法很少出現，如果不是創新，極有可能是寫出了不通順的句子，所以語言模型也能應用在中文句子流暢度偵測的方面。語言模型規模相當依賴大型訓練語料，訓練語料的性質越接近測試的文章，所建立的語言模型效果越好，所以語料庫也要跟著改變與適應。

語言模型會基於使用方法的不同而有所改變，例如： mixing 語言模型，使用混合多種不同的語言模型來改善中文斷詞的效果[9]，圖二舉例說明非斷詞 bigram 使用單字建立語言模型，圖三舉例說明斷詞 bigram 先斷詞後建立語言模型。而本實驗中分別使用了新聞語料庫以及國中生作文語料庫來建立語言模型。



圖二、舉例說明非斷詞 bigram 示意圖



圖三、舉例說明斷詞 bigram 示意圖

(一)、N-gram 語言模型

語言模型是由大量語料庫經過訓練、斷詞、計算詞頻等建立而成的統計資料，文集中每個單字詞的計算方式是使用 Maximum Likelihood Estimation (MLE) [10]來計算每個字出現的相對頻率並藉此計算機率，如下式：

$$P(W_n|W_{n-N+1}^{n-1}) = \frac{C(W_{n-N+1}^{n-1}W_n)}{C(W_{n-N+1}^{n-1})} \quad (1)$$

其中 C 代表某個字 W 出現的頻率。

一個句子是由 n 個字所組成，所以一整個句子的機率就可以計算如公式(2)：

$$P(W_1^n) = P(W_1, W_2, \ldots, W_n) \quad (2)$$

其中W_n表示句子中第 n 個字。$P(W_1^n)$ 表示 1 到 n 個字出現的機率值。

我們假設詞彙的機率為獨立的條件之下，根據[11]可以得知句子依據條件機率可定義如公式(3)所示：

$$P(W_1^n) = P(W_1)P(W_2|W_1)P(W_3|W_1^2) * ... * P(W_3|W_1^{n-1}) = P(W_1)\prod_{k=2}^{n} P\left(W_k\middle|W_1^{k-1}\right) \quad (3)$$

公式(3)改成(4)是由於無法從過去的語料中來做無限字的預測：

$$P(W_n|W_1^{n-1}) \approx P(W_n|W_{n-N+1}^{n-1}) \quad (4)$$

代表依據前(n-1)個字出現的機率來預測目前第 n 個字所出現的機率，而所謂的 N-gram 就是當 N=2 時，稱為 bigram，如公式(5)：

$$P(W_n|W_{n-1}) \quad (5)$$

在本實驗中所建立的語言模型採取 bigram 以及 unigram 的模式以及是否先經過 CKIP[12]斷詞，簡單來說 bigram 語言模型就是統計完語料之後，紀錄詞彙中每一個字出現的條件下，下一個字接在此字後面的機率，也因為中文字中兩兩字的組合比例較高，因此我們實驗使用 bigram。如圖二表示，此圖舉例說明，以“今” 與“天”為例，由“今 ”出現的情況下，推測“天”出現的機率，就稱為 bigram，同理“天”與“的”；“天”與“氣”也都是 bigram…依此類推，而如果句子先經由斷詞我們就會以詞為單位，bigram 的情況就會變成以“天氣”與“真好”為例，由“天氣”出現的情況下，推測“真好”出現的機率，也因此我們就能從語言模型中推算出某一個句子的機率。

“Entropy”是很重要的評估標準之一，它也被廣泛的使用在測量資訊上[13]，“Entropy”被定義為下列的式子(6)：

$$\mathrm{H(X)} = -\sum_{x\in T} P(X)\log_2 P(X) \quad (6)$$

其中隨機變數 X 涵蓋的範圍包含可預測的 T 集合(例如字母，字詞或部分的語音)。P(x)、P’(x)都是 MLE 所計算出來的機率值，實際使用時則是套用下列改寫過的公式(7)：

$$\mathrm{H'(X)} = -\sum_{x\in T} \log_{10} P'(X) \quad (7)$$

另外再定義 Perplexity，如下式(8)：

$$\mathrm{Perplexity} = 2^{H} \quad (8)$$

實際計算時亦套用改寫過的公式：

$$\mathrm{Perplexity'} = 10^{H'/W} \quad (9)$$

其中 W 是一個句子的單字數除以 W 的目的是避免當句子越長時機率越低的情況發生。Perplexity 越低代表句子中字詞的組合機率很高，也就是說這個句子是比較多人這樣寫的當然也會較為通順。但是 N-gram 語言模型還是有缺點必需要克服：語言模型在不夠龐大時對無法涵蓋所有可能的字詞組合，也就是資料稀疏的問題，即有些字詞的組合沒有被訓練到，使的在查詢頻率時會有零的問題發生，導致無法正確算分的錯誤狀況。因此為了解決這個問題我們還需使用平滑(smoothing)的方法來改善機率為零的例外情況。

(二)、Smoothing

Smoothing 的方法可分成模式結合的方法[14]以及折扣的方法，模型結合的方式就是利用內插法和補插法，bigram 無效時，使用 unigram；而折扣的方法就是調整機率，將機率較高者把值分配給機率為零者。實驗是用 Interpolated Kneser-Ney smoothing。而

Good-Turing(GT) [14]與 modified Kneser-Ney (mKN) [15]的演算法效果不錯，以下將會簡單介紹 GT 與 KN 的演算法。

1、 Good-Turing Discounting(GT)

Good-Turing 的演算法是調整從"r" (r: 表示出現 r 次的字數)至"r*"，依據它是二項式分布的假設 ，如公式(10)。

$$r^* = (\mathrm{r}+1)\frac{N_{r+1}}{N_r} \quad \mathrm{r} < \mathrm{M} \tag{10}$$

其中 r N 是 N-gram 中出現 r 次的字數，M 是界限值通常都小於 5。特別需要注意的是 r =0 時，代表 N-gram 中出現 0 次的字數：

$r^* = \frac{N_1}{N_0}$

其中 0 N 是表示從未出現過， 因此折扣過後改寫成下式(11)：

$$P_{GT}(W_1 \ldots W_n) = \frac{r^*}{N} \tag{11}$$

Good-Turing 僅適用於 r < 5，而且必須重新標準化以確保機率總和為 1。如此調整過後，原本出現頻率為 0 的字，將會被調整提昇成為小數位數，所以避免了機率為零的而導致無法計算整個句子機率的錯誤情形。

2、 Modified Kneser-Ney discounting (mKN)

Kneser-Ney 利用內插法的方式，例如 trigram 無法計算時，改以用 bigram，若依然無法使用，再改 unigram 的方式，則一定可以找出其出現的機率值，因此可以提供正確的估計值，mKN 的方法是由 Chen 與 Goodman 共同提出。mKN 的 smoothing 方法有 3 個參數 ：D1，D2，和 D3，這三個參數是分別用來對應於 unigram，bigram 與 trigram。 mKN 折扣方法的演算法如下式(12)：

$$P_{mKN}\left(W_i \middle| W_{i-n+1}^{i-1}\right) = \frac{C\left(W_{i-n+1}^{i-1}\right) - D\left(C\left(W_{i-n+1}^{i-1}\right)\right)}{\sum_{wi} C(W_{w-n+1}i)} + \gamma(W_{i-n+1}^{i-1}) P_{mKN}\left(W_i \middle| W_{i-n+2}^{i-1}\right) \tag{12}$$

其中$\mathrm{D(c)} = \begin{cases} 0 \; if \; c=0 \\ D_1 if \; c=1 \\ D_2 \; if \; c=2 \\ D_3 \; if \; c\geq 3 \end{cases}$ $D_1 = 1 - 2\frac{N_1}{N_{1+2N_2}} * \frac{N_2}{N_1}$ $D_2 = 1 - 3\frac{N_1}{N_{1+2N_2}} * \frac{N_3}{N_2}$ $D_3 = 1 - 4\frac{N_1}{N_{1+2N_2}} * \frac{N_4}{N_3}$

3、 Interpolated Kneser-Ney smoothing

Interpolated Kneser-Ney smoothing 其公式(13)：

$$P_{\text{interpolated}}(W|W_{i-1}W_{i-2}) = \lambda P_{\text{trigram}}(W|W_{i-1}W_{i-2}) + (1-\lambda)[\mu P_{bigram}(W|W_i) + (1-\mu)P_{unigram}(W)] \tag{13}$$

本篇實驗也是使用 Interpolated Kneser-Ney smoothing 的公式，但是實驗是 bigram 語言

模型，因此我們將公式改寫成(14):

$$P_{\text{interpolated}}(W|W_{i-1}) = (1-\lambda)[\mu P_{bigram}(W|W_i) + (1-\mu)P_{unigram}(W) \quad (14)$$

雖然此方法較 Modified Kneser-Ney discounting 的方法簡單一點，但是卻較易執行，更重要的是效果略比 Modified Kneser-Ney discounting 還要好。[11]

語言模型是大量語料庫經過訓練所建立的，語料庫與測試的資料彼此間是有所關聯的，性質越相同的語料庫偵測的效果越好，例如受測的資料是以國中生的作文為主，我們就加入以國中生作文到所建立的語言模型中，由於系統使用了混合式語言模型概念，語言模型的規模分別為新聞語料加上作文語料建完索引檔有 303MB，這個語言模型新聞語料佔了大多數因此我們特別又建立了一個只含國中生作文的語言模型建完索引檔有 7.21MB，前者語料庫沒有經過斷詞後者純作文的部分則是先經過 CKIP 斷詞處理，所以必須加權計分，權重計分的公式如(15)：

$$PPL = (1-\alpha)PPL_1 + \alpha PPL_2 \quad (15)$$

其中 PPL 是 Perplexity，PPL1 是語言模型 1 計算的結果，PPL2 是後來的語言模型 2 計算的結果，α 介於 0 到 1 之間，α 是可以調整的，隨著測試資料不同以及使用者設定的不同而改變 α，使系統能調節不同語言模型產生的偵測結果來提高準確度。

四、實驗內容

國中生作文語料庫的部分我們所蒐集的學生作文是來自於某國中七、八年級手寫的考試作文蒐集而成如圖四，並且每篇作文皆由教師訂正過錯別字，最後將這些作文輸入成電腦可處理的 XML 格式如圖五。

圖四、學生作文原文

<doc>
<class>八年四班</class>
<number>11</number>
<title>走出青春的光彩</title>
<score>5</score>
<essay>
此刻，正在考場中振筆疾書的我，感受到教室中有股青春的氣息在空氣中<revise><wrong>瀰漫</wrong><correct>瀰漫</correct></revise>著，裡面有著大家的夢想、有著光明的未來。
正值青春年華的我們，應保有一顆堅強以及積極的心，去面對各個關卡，大家心中熱血<revise><wrong>沸鶩</wrong><correct>沸騰</correct></revise>，正朝著自己的目標努力著，為了自己的未來所奮鬥著。
其實用心體會生活中的微小事物，幫助別人，使自己心情開朗，和同學<revise><wrong>分賞</wrong><correct>分享</correct></revise>快樂的事情，聽見同學們的歡笑聲，也能使自己開心，和爸媽訴說在學校所發生的喜怒哀樂，不只能增進親子間的<revise><wrong>觀係</wrong><correct>關係</correct></revise>，也讓他們的生活多了些趣味性，這樣的生活不也是多彩多姿嗎？這樣的生活不也是許多人所嚮往的嗎？
…
</essay>
</doc>

圖五、學生作文電子檔

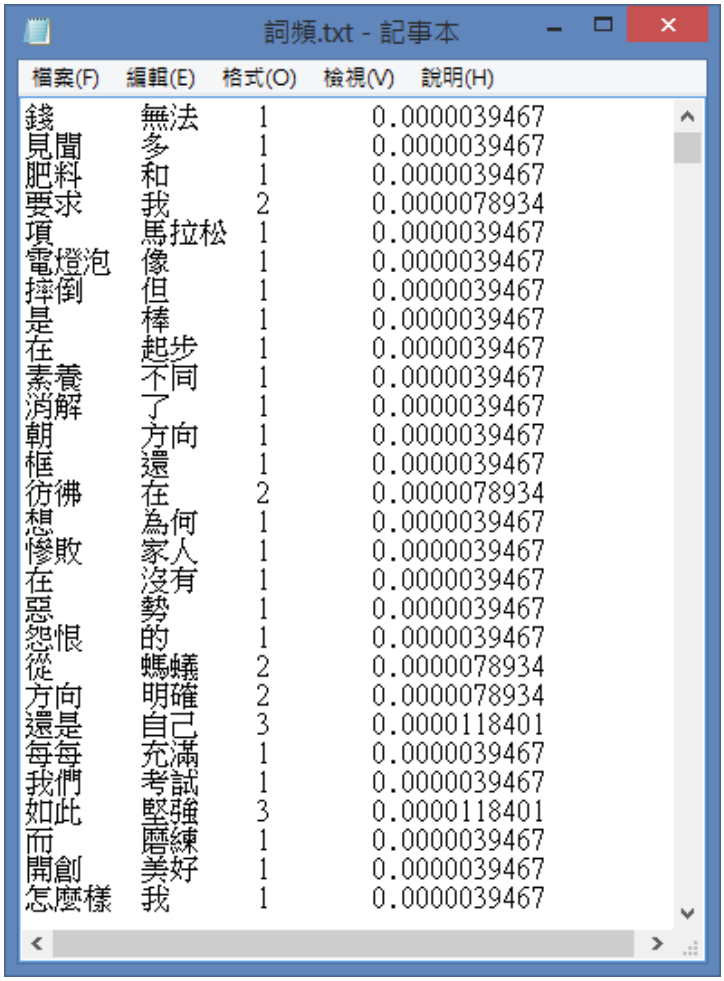
而學生作文電子檔的 Tag 我們定義如下：
<doc></doc>： 文件的貣始與結束，一篇文件包含下列的資訊。
<class></class>： 學生的班級。
<number></number>： 學生的座號。
<title></title>： 作文的標題。
<score></score>： 學生的得到的級分。
<essay></essay>： 學生的文章內容。
： 段落。
<revise></revise>： 老師批改到的錯別字以及老師更正的結果。
<wrong></wrong>： 錯誤字詞。
<correct></correct>： 老師所提供的正確的字詞。

而我們把其中五級分以及六級分一共 833 篇的作文來做斷詞建立語言模型。而訓練過程如下列圖片所示:

此刻
正在 考場 中 振筆 疾書 的 我
感受到 教室 中 有 股 青春 的 氣息 在 空氣 中 瀰漫 瀰漫 著
裡面 有 著 大家 的 夢想 、 有 著 光明 的 未來
正值 青春 年華 的 我們
應 保有 一 顆 堅強 以及 積極 的 心
去 面對 各 個 關卡
大家 心 中 熱血 沸騰
正 朝著 自己 的 目標 努力 著
為了 自己 的 未來 所 奮鬥 著
其實 用心 體會 生活 中 的 微小 事物
幫助 別人
使 自己 心情 開朗
和 同學 分賞 分享 快樂 的 事情
聽見 同學 們 的 歡笑聲
也 能 使 自己 開心
和 爸媽 訴說 在 學校 所 發生 的 喜怒哀樂
不只 能 增進 親子 間 的 觀 係 關係
也 讓 他們 的 生活 多 了 些 趣味性
這樣 的 生活 不 也 是 多彩多姿 嗎
這樣 的 生活 不 也 是 許多 人 所 嚮往 的 嗎

圖六、文件經過去除 Tag、分割句子與 CKIP 斷詞

圖七、統計各字詞詞頻，計算其句首及各詞條件機率，建立成 unigram



圖八、統計各字詞詞頻，計算其條件機率，建立成 bigram

（一）、實驗一測試國中生所寫的作文

我們擷取了跟訓練集同樣的國中生所寫的作文來測試，我們從五級分和六級分作文中一共一百個句子，我們預設如果算出來的加權分數超過 50 分就有可能是不通順的句子，根據測試結果在 90 句 50 分以下的句子有 5 句經由中文流利人事審查為不流暢，10 句 50 分以上句子其中 6 句是判斷正確 4 句為誤判的，整體的效率評估達 89%，而偵測不流暢句子的 Recall 則有 60%而 Precision 54.5%。

表三、實驗一測試結果

實驗一		已知分類		Results			
		不通順	通順	Precision	Recall	F–measure	Accuracy
預測分類	不通順	6	4	60%	54.5%	57.11%	89%
	通順	5	85	94.4%	95.5%	94.99%	

（二）、實驗二測試外國人所寫的作文

我們從 NLP-TEA1[17] Testing Data 收集語料(圖九)，我們依照文件的標籤收集 843 筆正確的句子以及 273 筆含有冗詞的句子，這個實驗主要是測試希望在正確的句子中分數都能很低而在包含冗詞的句子中分數是高得，可是由於這是由外國人所寫的中文句子，或許句子是被歸類在正確的那一類，可是他們用詞的習慣多少還是跟本國人有些差距，所以我這邊預設如果算出來的加權分數超過 100 分就有可能是不通順的句子，根據測試結果在 843 句正確的句子當中有 450 句認為是通順的，在 273 句包含冗詞的句子中有 139 句正確判斷為不通順的，整體的效率評估達 52.77%，而偵測不流暢句子的 Recall 則有 51%而 Precision 26%。

```
<ESSAY title="不能參加朋友找到工作的慶祝會">
<TEXT>
<SENTENCE id="A2-0003-1">我以前知道妳又很聰明又用功</SENTENCE>
</TEXT>
<MISTAKE id="A2-0003-1">
<TYPE>Redundant</TYPE>
<CORRECTION>我以前知道妳又聰明又用功</CORRECTION>
</MISTAKE>
</ESSAY>

<ESSAY title="不能參加朋友找到工作的慶祝會">
<TEXT>
<SENTENCE id="A2-0005-1">我替你很高興</SENTENCE>
<SENTENCE id="A2-0005-2">我們應該找時間可以好好地聊聊</SENTENCE>
</TEXT>
<MISTAKE id="A2-0005-1">
<TYPE>Redundant</TYPE>
<CORRECTION>我替你高興</CORRECTION>
</MISTAKE>
<MISTAKE id="A2-0005-2">
<TYPE>Redundant</TYPE>
<CORRECTION>我們應該找時間好好地聊聊</CORRECTION>
</MISTAKE>
</ESSAY>

<ESSAY title="不能參加朋友找到工作的慶祝會">
<TEXT>
<SENTENCE id="A2-0010-1">最近找到工作很難</SENTENCE>
</TEXT>
<MISTAKE id="A2-0010-1">
```

圖九、語料來源格式

表四、實驗二測試結果

實驗二		已知分類		Results			
		包含冗詞句子	正確的句子	Precision	Recall	F–measure	Accuracy
預測分類	包含冗詞句子	139	393	26%	51%	34.44%	52.77%
	正確的句子	134	450	77%	53.3%	62.99%	

五、結論

在本實驗中我們使用混合式的語言模型來實作中文句子的流暢度偵測系統，經由實驗結果顯示本系統對於不流暢的句子有一定的識別度，未來我們也會繼續做各式各樣的實驗，想辦法改善系統效能讓它能確實應用在作文教學系統上輔助學生學習，表六為正確判斷的句子範例，我們根據下表五分析了幾個錯誤的原因並提出解決的方法:

1. 句子中夾雜英文以及人名，未來我們應該要把英文字母去除以及如果是人名的話應該用其他標籤取代以降低它的誤判。
2. 句子中包含標點符號，未來我們應該也要把頓號以及其他的標點符號去除再來算分應該可以大大提升準確度。
3. 由於句中包含了少見的專有名詞” 釙和鐳”使的分數大大的提高，為啥我們可以考慮建立一個專有名詞辭典來改善這個問題。

表五、誤判句子範例

句子	新聞語料模型	純作文模型	加權分數
例如 google 公司的大老闆布林和佩吉	215394.11	127.08	100760
我曾在這打鬧、嬉戲、悲傷、難過、歡樂、憤怒	200086	70	100078
因而找到新元素釙和鐳	58493.4	140.4	29316

表六、正確判斷的句子範例

句子	新聞語料模型	純作文模型	加權分數
我才在五歲時	60.79	41.05	50.92
也又有自己才能決定	61.7	40.28	50.99
甚至是有些人希望自己當個太空人	59.98	52.75	51.37

表六為正確判斷的句子範例，未來我們會持續實驗改進系統效能，並且增加語言模型的質量。句子流暢度偵測於作文當中遣詞造句的範疇，是目前各級升學考試作文評分四個面向之一。開發另外三個面向同樣需要的各種自然語言處理的功能將是我們未來研究的方向。像是使用邏輯連接詞(例如:因為…所以)以及組織文章的連接詞(例如:首先…之後)的出現率，來評估文章的組織結構。藉由語言學的 Isotopie 理論讓電腦辨識文章是否合乎主題。因為組成文章的句子不會是彼此都沒有相關的句子，Isotopie 可以分析句子是

否相互呼應且具有聚合力。且由於一篇文章的撰寫會根據主題選用同一辭彙場域的詞，所以我們可以利用辭彙的詞場來測試作文是否合乎主題。

參考文獻

[1] 教育部,“評分規準表” http://www.bctest.ntnu.edu.tw/writing.htm, 2015.

[2] Dequan Zheng, Feng Yu, Tiejun, Sheng Li, “Documents Ranking Based on a Hybrid Language Model for Information Retrieval” *IEEE International Conference on Information Acquisition*, Aug. 2006, pp: 279-283.

[3] Fei Song, W. Bruce Croft , “A general Language Model for Information Retrieval”, Proc. of Eighth International Conference on Information and Knowledge Management, 1999, pp: 316-321.

[4] Brown, Peter E; Cocke, John; Della Pietra, Stephen A.; Della Pietra, Vincent J.; Jelinek, Frederick; Lafferty, John D.; Mercer, Robert L.; and Roossin, Paul S., "A statistical approach to machine translation ." Computational Linguistics, Volume 16 , Issue 2, 1990, pp: 79-85.

[5] Jason S Chang, David Yu, Chun-Jun Lee, “Statistical Translation Model or Phrases” In Processing of Computational Linguistics and Chinese Language, Vol. 6, No. 2, August 2001, pp: 43-64.

[6] Ronald Rosenfeld, “Adaptive Statistical Language Modeling: a Maximum Entropy Approach” Ph.D. Thesis Proposal, Carnegie Mellon University, September 1992.

[7] Lalit R. Bahl, Peter F. Brown, Peter V. De Souza, Robert L. Mercer, “ATree-Based Statistical Language Model for Natural Language Speech Recognition”, IEEE Transactions on Acoustics Speech and Signal Processing, Vol. 37, No. 7, July 1989, pp: 1001-1008.

[8] Sergios Theodoridis and Konstantion Koutroumbas, “Pattern Recognition(Third Edition) ”, Academic Press. pp 13-19

[9] Wu, A.-D., and Z.-X. Jiang, "Word Segmentation in Sentence Analysis,"International Conference on Chinese Information Processing, 1998, Beijing,China, pp: 169-180

[10] Slavam. Katz, “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer ”, IEEE Transactions on ACOUSTICS, SPEECH, and SIGNAL PROCESSING, VOL. ASSP-35, NO.3, MARCH 1987, pp 400-401

[11] J. Goodman, "A Bit of Progress in Language Modeling, Extended Version," Microsoft Research, Technical Report MSR-TR-2001-72, 2001.

[12] National Digital Archives Program , “CKIP” http://ckipsvr.iis.sinica.edu.tw/, 2015.

[13] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics, Volume 22, Issue 1, March 1996, pp: 39-71.

[14] S. F. Chen, Joshua Goodman “An Empirical Study of Smoothing Techniques for Language Modeling”, Proc. of the 34th annual meeting on Association for Computational Linguistics ,Santa Cruz, California, 1996, pp:310-318.

[15] J. Goodman, “A bit of Progress in Language Modeling”, Microsoft Research, Aug. 2001.

[16] Adam L. Berger, Vincent J. Della Pietra, Stephen A. Della Pietra, “A maximum entropy approach to natural language processing”, Computational Linguistics, Volume 22, Issue 1, March 1996, pp: 39-71.

[17] Ru-Yng Chang, Chung-Hsien Wu, and Philips Kokoh Prasetyo (2012). Error Diagnosis of Chinese Sentences Using Inductive Learning Algorithm and Decomposition-Based Testing Mechanism. ACM Transactions on Asian Language Information Processing, 11(1), article 3, March 2012.