The 2016 Conference on Computational Linguistics and Speech Processing
ROCLING 2016, pp. 103-114

以語言模型評估學習者文句修改前後之流暢度
Using language model to assess the fluency of learners sentences edited by teachers

蒲冠穎　Guan-Ying Pu,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
sweetmilk425@gmail.com
*shwu@cyut.edu.tw (contact author)

摘要

因應自動化作文教學系統之需求，我們開發了一個偵測學生作文句子的通順度的系統，此系統基於語言模型（language model）方法，結合新聞語料及國中生作文語料訓練而成[1]，　　此句子通順度的偵測系統，我們蒐集了 339 句國中生所寫出來的句子
339

，

，　　，

關鍵詞：中文，作文，語言模型，N 元語言模型，句子流暢度

一、簡介

隨著科技的發展，現在 3C 產品可說是非常的普遍，也因為如此現在非常多的孩子從小就接觸電腦、手機、平板等 3C 產品，使得現在學生更有可能以電腦作為寫作文的工具。雖然教育政策將作文納入考試評分項目，使得學生跟家長再度重視寫作能力，但是受限於教學時數，可以練習寫作的時間實在是不足以將那些寫作能力較弱的學生作有效提升。因此我們認為未來可以藉由自動化的作文教學系統幫助學生在家自學作文。而我們所開發作文教學之句子流暢度偵測系統，經由系統回傳的診斷結果，幫助學生提升詞句組合的理解能力以寫出較順暢的句子，藉此提升他們作文的分數。本系統依賴 N-gram 的語言模型[1]，其特色是計算字詞間組合的機率，機率越高字詞組合的正確性就越高句子也就越順暢，然而語言模型其效果相當依賴大型的訓練語料，這是語言模型仍待克服的問題，而且如果訓練語料的性質跟要測試的文章性質越不相關，效果就會越差，因此語料庫需要根據測試文章做改變。

系統需要知道如何判斷出一篇作文是好的，藉此才能幫助學生寫好作文，國中基測作文評分主要分為四個面向:立意取材、結構組織、遣詞造句、錯別字、格式及標點符號等四項核心技巧。這四項面向是依照作文構成過程所需要的元素所決定，這些作文評分範疇不容易被變更。以下是作文評分為六種不同等級的說明(如表一 [2])，本系統是針對四個範疇中的遣詞造句的句子流暢度作為研究目標。

本文主要的研究在於偵測句子的流暢度，正確的判斷句子是否通順，系統設計用於解決一般性問題，隨著訓練集增加，可增強對句子的判斷。雖然現在系統只算是一個起步，在未來此系統將整合到電腦作文自動評分系統，在學校時一名老師需要面對多名學生，學生難以得到即時的評價，而作文自動評分系統能全天不間斷地提供服務，提供可以隨時學習的機會。自動化作文系統是由多個診斷模組所構成(本文中的系統為其中一個診斷模組)，作文會經由這些分散是診斷模組分別診斷個面向的優缺點，之後產生一份可擴展的診斷清單，此清單整合各面向的診斷結果，產生評分模組以及雷達圖。當作文經由「錯別字、格式與標點符號」、「立意取材」、「遣詞造句」以及「結構組織」等診斷模組產生個別的診斷結果後，再來就可以給作文評定等級。根據作文在四個面向的表現，機器學習程式可訓練出穩定的分類器，將作文分為零到六級分，並產生對應四面向強弱的雷達圖，接著作文評語的部分則是依各個面向產生的診斷結果，合併在一起呈現特徵細節。但是要讓電腦可以詳細地呈現各細節特徵，這需要搭配自然語言處理工具以及語言資源才能做到，最基礎的前處理動作就是文章斷詞以及標註詞性(POS tagging)，然後再依照各個模組的需求來增加處理的知識。本實驗我們使用新聞語料所建的語言模型跟作文語料的語言模型，將 339 句國中生所寫出來的句子 339

ˌ §

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

表一、國中生基本學力測驗作文測驗評分規準[2]

(一)、立意取材

主要是評量是否能切合文章的主題並選擇適合的素材，以表達主題意念。

(二)、 §

結構組織的基本要求是意念的前後一致，也就是首尾要連貫，以及結構要勻稱。結構是文章的「骨架」。結構組織的好文章才能成形，不然只是一堆句子，成不了一篇文章。

(三)、 §

在之前我們做過初步的分析，我們分析了一百份國中生的作文，其結果顯示，如果作文很少使用修飾詞的作文，評量結果大概會落在三到四級分。我們使用了國中三年級的國文課本裡的詞彙以及國中作文語料庫裡面的詞彙，這些詞彙符合國中生的使用程度也不會出現艱澀以極少用詞彙的情況。但是雖然同為國中三年級的國文課本，在各版本中教材仍然有難易度上的差異，通常三級分以下的作文所使用到的詞彙程度都停留在國二以下。因此分類國中國文課本詞彙的等級是具有意義的。

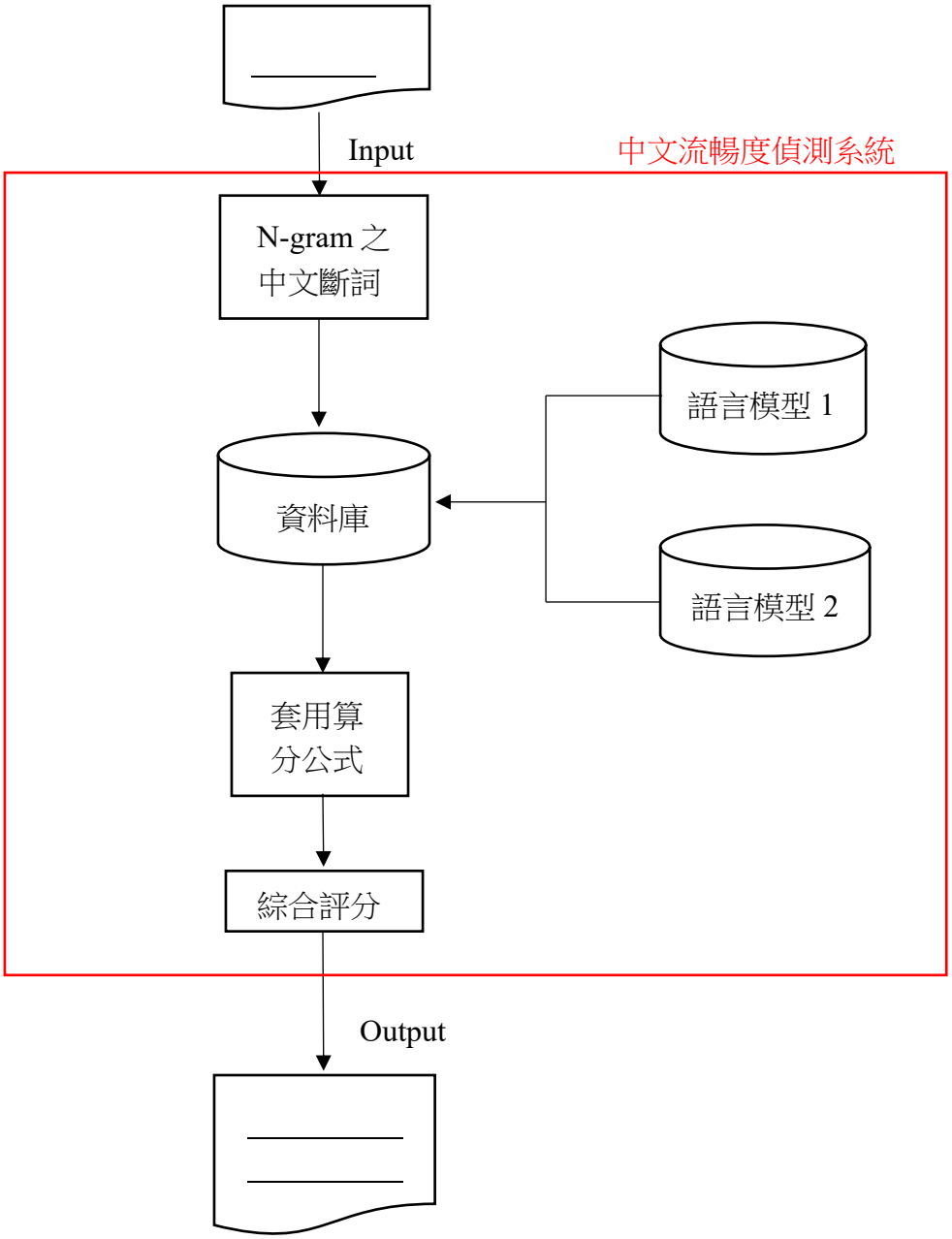
(四)、 ˛ §

錯別字方面，部分系統可以藉由正確作文的語料庫來找尋並比對新作文的錯別字，如此我們可以偵測出錯別字。格式方面，作文長度通常會影響作文的等級，作文的分段也會影響到評分的結果，依據修改國中作文的專家判斷，一級分的作文大多是一到三行寫成一段或兩段；行數在四到七行之間，不包含抄錄題目引導的句子，有段落數有兩段的大概是兩級分；如果只有三段，行數在七到十二行，大部分最高是三級分；最少寫到四段，最多不超過六段通常有四級分以上的分數；而內文是空白的就是零級分。文章如果用錯符號，會容易引起誤解。正確的使用標點符號將文章斷句，不只讀起來通順、文意明確，也有強化語意的功效。

二、系統架構

圖一是中文句子流暢度偵測系統運作的流程圖，首先將要偵測的資料輸入到系統裡，系統會自動計算分數，之後分數若高於一定門檻值，系統將會提示這可能是不通順的句子。系統效能評估的部分，我們把測試結果讓中文叫流利的人進行檢閱，接著計算 Recall 與 Precision 來評估系統能力。我們分析實驗結果，並觀察特殊案例，包含系統誤判為不通順或錯放不通順的句子進一步分析錯誤原因，根據系統的缺失來提出改善系統方法，希望未來系統更新版本能改善實驗結果以及提升效能。語音辨識[8][9]、資訊檢索[3][4]、文件分類、手寫辨識以及機器翻譯[5][6]等，這些都屬於自然語言處理(Natural Language Processing, NLP)的領域。其中語言模型(Language Model, LM)是自然語言處理重要的技術之一[7]，語言模型可以統計並記錄大量語料庫的詞頻及機率，其特性是可依據已訓練的資料，也就是過去統計並記錄過的字，預測下一個字出現的機率，藉此計算一個句子的機率，機率越大就表示此句子越常出現，也表示句子越通順；反之，機率越低，表示這句子很少出現這種寫法，除非是創新的句子，不然極有可能是不通順的句

子，由上述可知訓練語料的性質越接近所測的文章，其效果越好，因此語料庫需要跟著做改變。



圖一、中文作文流暢度偵測系統運作的流程圖

語言模型會因使用方法的不同而有所改變，例如:混合式(mixing)語言模型，其特點是混合使用多種不同類型的語言模型來改善中文斷詞的效果[10]，而本實驗中分別使用新聞語料庫和國中生作文語料庫所建立的語言模型。

(一)、N-gram 語言模型

語言模型是大量語料庫經過訓練、斷詞以及計算詞頻等建立而成的統計資料集，資料集中每個單字或詞的計算方式是使用最大似然法則(Maximum Likelihood Estimation，MLE)[11]來計算每個字詞出現的頻率並藉此計算機率，如以下公式 1:

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

其中 C 表示某個字 W 出現的頻率。

一個句子由 n 個字所組成，所以一整個句子的機率就可以計算，其公式如下:

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

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。

我們假設詞彙的機率為獨立的條件下，根據[12]可以得知依據條件機率句子的計算可定義如公式(3):

$$P(W_1^n) = P(W_1)\, P(W_2|W_1)\, P(W_3|W_1^2) * \ldots * P(W_n|W_1^{n-1}) = P(W_1) \prod_{k=2}^{n} P(W_k|W_1^{k-1}) \quad (3)$$

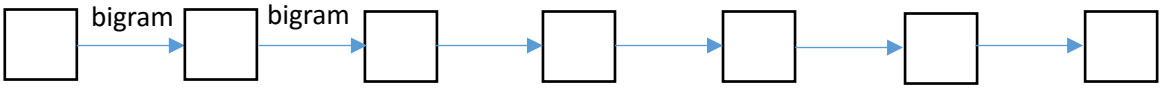
由於無法從過去的語料中來做無限字的預測，所以將公式(3)改成公式(4):

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

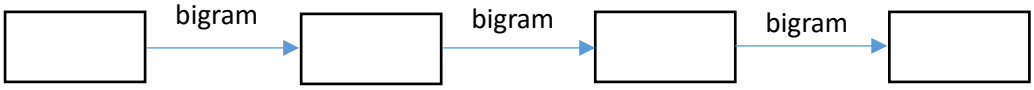
表示假設要預測第 n 個出現的機率，可根據(n-1)個字出現的機率來做預測。N 是指給定的一段文本中 N 個項的序列，當 N=2 時，稱為 bigram，如公式(5):

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

本實驗中的語言模型是採取 bigram 以及 unigram 兩模式，以及 bigram 和 unigram 兩模式先經過 CKIP 系統[13]斷詞後所建立，以下我們用兩個示意圖(圖 2 和圖 3)，以 bigram 模式建立語言模型為例，來講解有無先做斷詞的差異，首先簡單了解一下 bigram 語言模型，bigram 語言模型就是在統計完語料後，在記錄詞會中每一個字出現的條件下，下一個字接在此字後面的機率，用圖 2 的示意圖來說明，以圖 2 中” 今”與” 天”為例，在” 今”出現的情況下，推測” 天”出現的機率，依此類推，” 天”與” 的”； ” 天”與” 氣”也都是 bigram，也因為中文字中出現兩字的組合比例較高，因此我們實驗使用 bigram。再來講解是否先斷詞的差異，如上述所說，bigram 是依上一個字來推測目前的字的機率，假設沒先做斷詞就會如圖 2 示意圖中，舉例的句子含有 7 個字，就需要做(7-1)次的 bigram，但如果如圖 3 示意圖所表示的，在做 bigram 前先做斷詞，就會以” 詞”為單位做 bigram，用” 天氣”與” 真好”為例，依” 天氣”出現的條件下，推測” 真好”出現的機率，如此我們可以清楚了解是否先做斷詞的差別。藉由上述所說的方式，也就能從語言模型中推算出一個句子的機率。



圖二、無先做斷詞的 bigram 示意圖



圖三、先做斷詞的 bigram 示意圖

熵(Entropy)是自然語言中的重要概念，他是很重要的評估標準之一。信息是相當廣泛的概念，很難用簡單的定義將其完全準確的把握。然而，對於任何一個機率分布，可定義一個稱為熵的量，其被定義為下列公式(6):

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

公式中隨機變數 X 涵蓋的範圍包含可預測的 T 集合(例如:字母、字詞或部分語音)，因為 P(X)值極小，為了避免 H(X)太小，所以我們實際使用則套用改寫過的公式(7):

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

上述兩公式所用的$P(X)$跟$P'(X)$都是由 MLE 所計算出來的機率值。

接著定義複雜度(Perplexity)，複雜度是一種衡量 NLP 中語言模型好壞的指標，其定義如下列公式(8):

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

實際計算時套用改寫的公式(9):

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

W 表示句子的單字數，除以 W 目的是避免當句子越長時機率越低的情況發生。複雜度的值越低表是句子中字詞組合的機率越高，也就是表示句子越通順。N-gram 語言模型還有缺點必須克服，就是語言模型不夠龐大時，無法涵蓋所有可能的字詞組合，也就是資料量稀疏的問題，即表示有些字詞組合沒有被訓練到，使查詢頻率時有機率是零的問題發生，而導致無法正確算分的情況。因此為了解決此問題我們還須使用平滑(Smoothing)的方法來改善機率為零的例外情況。

(三)、$f$moothing

平滑($f$moothing)法可分為模式結合的方法[14]與折扣的方法，模型結合是利用內插法與補插法，例如:使用 bigram 無效時，就使用 unigram；而折扣的發法是調整機率，就是將機率較高者的值分給機率為零者。本實驗室使用 Interpolated Kneser-Ney smoothing 的方法。其公式如下公式(10):

$$P_{\text{interpolated}}(\text{W}|W_{i-1}W_{i-2})=\lambda P_{\text{trigram}}(\text{W}|W_{i-1}W_{i-2})+(1-\lambda)\,[\,\iota\,P_{\text{bigram}}(\text{W}|W_i)+(1-\iota\,)\,P_{\text{unigram}}(W)] \quad (10)$$

由於本實驗是使用 bigram 語言模型，因此將公式改寫成(11):

$$P_{\text{interpolated}}(\text{W}|W_{i-1})=\iota\,P_{\text{bigram}}(\text{W}|W_i)+(1-\iota\,)\,P_{\text{unigram}}(W) \quad (11)$$

由於本系統使用混合式語言模型概念，語言模型由新聞語料加上作文語料建立索引檔，其大小為 303MB，此語言模型中新聞語料占了大多數，因此我們又另外建一個純粹只含國中生作文的語言模型，其建立的索引檔大小為 7.21MB，前者語料庫是沒有經過斷詞的，後者則是有先經過斷詞處理，因為使用兩種語言模型，所以計算時必須採用加權計分的方式，其公式如(12):

$$PPL=(1-\alpha)PPL_1+\alpha PPL_2 \quad (12)$$

PPL 就是複雜度(Perplexity)，PPL_1是語言模型 1 計算的結果，PPL_2是後來的語言模型 2 計算的結果，α介於 0 到 1 之間，α為可以調整的，隨著測試資料不同以及使用者不同設定而改變，使系統能調整不同語言模型所產生的偵測結果來提高準確度。

三、實驗內容

我們使用所開發的線上作文模擬考系統(畫面如圖四)，讓國中學生線上寫作文，然後我們請國文老師做批改，其批改畫面如圖五，線上作文模擬考系統會將老師所批改作文的錯誤依照類型(如圖六)存到資料庫裡，之後我們將這些收集來的資料用在自動化作文系統的研究上。本次實驗我們使用批改作文中錯誤類型屬於句子優化的句子(如圖七)，這些句子我們將其分為優化前句子(也就是學生所寫原本的句子)跟優化後句子(經老師批改的句子)，兩者分別是圖七中 Wrong 和 Correct 欄位裡的句子，這次蒐集句子優化的有 339 個句子(優化前跟優化後各 339 句)。我們將優化前跟優化後的句子先做斷詞後分別經由兩種語料(新聞語料與作文語料)的語言模型計算出句子的分數，其結果如圖八跟圖九，圖中的兩個分數分別表示使用新聞語料語言模型和使用作文語料語言模型的分數，之後我們比較優化前跟優化後句子的分數。

圖四、線上作文模擬考系統首頁畫面

圖五、教師端批改畫面

TypeId ▲ 1	Type
1	錯別字
2	成語使用不當
3	詞語使用不當
9	多餘的片段
11	句子優化
21	詞序不當
22	用詞不當
23	缺字
24	冗詞
25	標點符號使用錯誤
29	連接詞使用不當
32	冗詞、句型使用錯誤
33	標點符號使用錯誤、詞序不當、冗詞、錯字、缺字
34	標點符號使用錯誤、詞序不當

圖六、作文錯誤類型

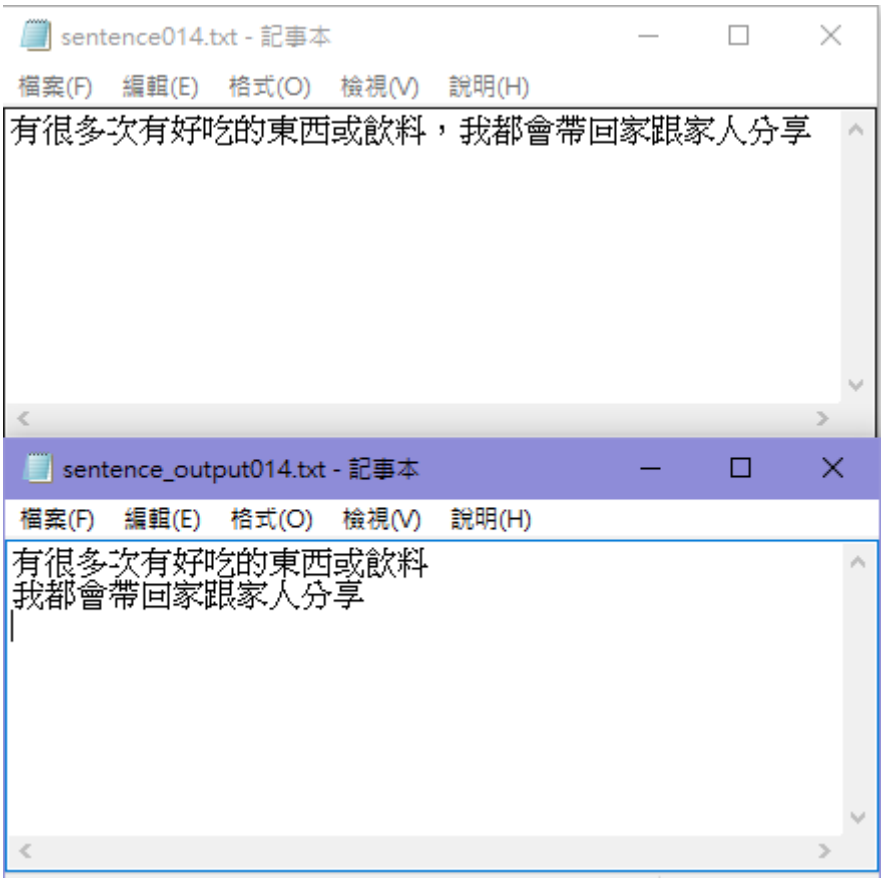
Wrong	Correct	TypeId
於是我們便開始不斷的練習	之後我們便開始不斷的練習	11
所有人之中每個人一生之中一定會有一個令自己滿意的作品，這一幅花費我非常多的心思	每個人一生中，一定會有一個令自己滿意的作品。這一個花費我非常多的心思的作品，	11
我從小就喜歡文房四寶，從小就開始學習書法。	而就我而言，我從小就開始學習寫書法	11
書法穩定了我的性情	書法可以陶冶我的性情	11
在紙上揮霍時，勾、挑、豎，	當我在紙上盡情揮灑時，勾、挑、豎，手中的毛筆	11
學久了，功力也比以前更加純練	等到我學習有一段時間後，筆法也比以往熟練時，就	11
成為我的動力，鞭策我更加打拚	是我繼續堅持的動力，也鞭策我更加地拚命	11
這種作品送去參展，預料之內的落選，只得到優選的成績。	如預期所想的，拿這種作品去參展，最後只得了優選的成績。	11
一股不甘心的心情	這時一股不甘心的心情就此	11
這也是我有史以來最滿意的一篇	因為這篇作品是我有史以來的佳作	11
會考後	而等到會考過後，	11
不要氣餒	都不要氣餒和放棄	11
其實那不是	但其實這不是	11
努力真值得	努力付出是值得的	11

圖七、錯誤類型為句子優化的資料表

圖八、優化前句子計算分數的結果

圖九、 優化過後句子計算分數的結果

優化前後的句子經由新聞語料的語言模型以及作文語料的語言模型跑出來的結果，經比較後發現，使用新聞語料語言模型其結果，優化後分數變好的有 173 個句子，優化後分數變差的有 166 個句子；使用作文語料語言模型其結果，優化後分數變好的有 138 個句子，優化後分數變差的有 201 個句子。根據上述所發現的數據可知道語料庫的大小對結果判斷的影響，由於新聞語料所建的語言模型比作文語料所建的還要大，所以其結果比較好。但是可以發現依結果來看即使新聞語料比較大，可效果並沒有很理想。除了語料庫大小與文章前面所提的性質關係外，我們發現句子資料的長度也影響到分數，這邊所指的句子資料長度的意思不單單只是指句子字數太長，而是依這句子中有幾個標點符號來看。在句子計算分數之前，我們除了做斷詞的處理之外，在斷詞之前我們會做一個叫做切句子的處理，所謂切句子是指將句子中遇到逗號、分號、句號、問號以及驚嘆號去除並做換行的動作，依圖十來舉例，上方是還未切句子的句子，其句子間有一個逗號存在，則將逗號去除並且換行變成下方顯示的那樣。因此在計算分數時，這類句子資料較長的就會分成多個句子做計算，我們將所切的句子計算出來的分數做相乘的動作，所以句子被切越多所得出來的分數值越高。再來講解句子資料長度對於優化前後句子比較分數的影響，以下我們用表二來講解，從表二可以看出，優化後的句子所包含的標點符號變多了，這也表示切句子的數量跟著變多了，所以可以看出優化後的分數結果沒有優化前那麼高。 (表二顯示的分數，文章前面有提到我們是用複雜度(公式 9)來看句子是否通順，所以分數要越低才是越好的。)



圖十、優化前的句子跟其切句子後的結果

優化前句子	優化後句子	優化前分數	優化後分數
§	每個人一生中，一定會有一個令自己滿意的作品。這一個花費我非常多的心思的作品，	**1.93E+04**	3.25E+06
有很多次有好吃的東西或飲料，我都會帶回家跟家人分享	於是在那之後，有好多次得到好吃的東西或飲料時，我都會帶回家跟家人分享	**3.35E+04**	1.00E+07
人生中要成大事老天必先苦其心志	而在人生中，如果要成就大事，必先苦其心志	**1.78E+03**	5.81E+07

表二、句子資料長度對於優化前後分數的影響

四、結論

經過本次實驗可以得知未來系統要改善的地方，首先就是語料庫數量的提升，以增加語言模型的涵蓋範圍，目前的語料庫實在太小，很多字詞的組合沒有出現過導致不管是優化前或優化後的句子分數被拉得太高，我們原本預定分數值以不超過 100 才算是句子是通順的，如表三跟表四所顯示，表三可以看到分數要小於 100 才算是正確的判斷，但表四可以看出，即使句子長度不長，分數還是過高，其原因還是出在於語言模型規模的不足。再來就是解決因切句子的數量太多而造成分數相乘太大問題，此問題的解決方式還有待以後藉由多測試資料來尋找更好的解決方法。林耀等[15]，使用多種的機械學習方法的組合來進行中文的情緒分析，或許在未來我們可以使用他提出的方式來改善我們的系統。

句子	新聞語料模型分數	純作文語料模型分數
我才在五歲時	60.79	41.05
也又自己才能決定	61.7	40.28
甚至是有些人希望自己當個太空人	59.98	52.75

表三、正確判斷的句子範例

優化前句子	新聞語料模型分數	純作文語料模型分數
有運動會時	1954.89	165.96
舉辦運動會時	828.75	565.09
也忘不了我曾看過的	167.32	924.70
想起我忘不了曾看過的	299.19	925.74

表四、因語料庫太小導致分數被拉得太高的句子範例

句子流暢度是針對遣詞造句的範疇，為目前作文評分的四個面向之一。開發另外三個面向同樣需要自種自然語言處理的功能將是我們未來研究的方向。需要每個面向的系統準確度等判斷都達到標準，自動化作文系統才有實現的可能。

參考文獻

[1] Po-Lin Chen, Shih-Hung Wu, “Analyzing Learners 'Writing Fluency Based on Language Model” *Association for Computational Linguistics and Chinese Language Processing ROCLING 2015,*pp:218-232.

[2] 國中教育會考推動工作委員會,“評分規準表” http://cap.ntnu.edu.tw/exam_3_1.html, 2015.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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

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

[11] 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.

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

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

[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] Yiou Lin, Hang Lei, Jia Wu and Xiaoyu Li, “An Empirical Study on Sentiment Classification of Chinese Review using Word Embedding”, *29th Pacic Asia Conference on Language,Information and Computation: Posters*, 2015, pp: 258 – 266.