電子郵件輔助寫作

Email Writing Assistant System

張俊盛 Jason S. Chang

楊謦瑜 Ching-Yu Yang

彭冠復 Guan-Fu Peng

國立清華大學資訊工程學系 Department of Computer Science, National Tsing Hua University

elmon@nlplab.cc, Jason@nlplab.cc, chingyu@nlplab.cc

摘要

在本論文中，我們介紹一種在書寫特定目的之郵件時提供精準寫作建議的方法。在我們方法中，電子郵件先依照意圖分門別類，然後從每個類別的電子郵件中，擷取常用片語，並且講類似的片語，聚集成為一組一組的聚叢。執行時，系統檢視使用者輸入，比對信件類別的片語聚叢，以提供寫作建議。我們將此一方法，加以實際製作，成為一套電子郵件寫作建議系統 *EmailDr* 。

Abstract

We introduce a method for learning to provide writing suggestions for writing an email for a specific purpose. In our approach, emails are divided into categories aimed at finding common phrases for making suggestions. The method involves automatically extracting common phrases for every email category, and automatically clustering phrases for more effective suggestions. At run-time, the system accepts user's input and matches the input with common phrases for the specific purpose, to offer suggestions for what to write next. We have implemented the propose method and present a prototype email suggestion system, *EmailDr*.

關鍵詞：自然語言處理, 智慧寫作, 電子郵件

Keywords: NLP, Smart Compose, Email

一、緒論

隨著科技的日新月異，人們愈加依賴電子郵件做為通訊的方式，因此支援電子郵件寫作的工具也愈來愈多，如 Grammarly (https://www.grammarly.com/)、Gmail (https://www.gmail.com) 、Whitesmoke (http://www.whitesmoke.com/) 等等。目前既存的電子郵件寫作工具，多半協助使用者，改正拼字與文法錯誤，並且提供改善寫作風格建議，這些功能都與自然語言處理的技術息息相關，可以做到相當好的效果，然而如何寫作某一特定性質的書信，更是使用者迫切需要的功能，目前這方面還有相當改進的空間。

近年，Gmail 已經開始在使用者書寫郵件時，建議使用者往下可以使的片語，然而這些建議並未考慮使用者書信的類別。對於使用者寫作的特定類別的書信，未能直接提供對應的寫作建議。在網路上，有許多網站提供不同目的的書信範本，給使用者參考。然而使用者不容易一面寫作，一面查考書信範本的資訊。此外，在使用者尚未開始寫作時，Gmail 無法提供寫作建議。

因此本論文提出方法，利用既有的書信類型與範本，統計分析常用詞語，在使用者寫作時，比對使用者已經輸入的用詞，建議適當的常用片語，讓使用者可以寫出適當的書信。

舉例來說，寫作一封邀請信時，當使用者輸入 *"I would like to"* 時，最好的建議，可能是 *"I would like to invite you to . . ."*。直覺上，為了提供適當的寫作建議，我們可以分析出現在「邀請類」的多篇書信範本的高頻的片語。如此一來，我們可以提供比較正確的寫作預測，顯示比較適合的片語。

我們提出一套新的電子郵件寫作輔助系統 *EmailDr*，可以自動地學習如何接受使用者寫作的未完成句（例如 I would like to)，來預測特定類別書信的接續片語(例如 *I would like to invite you to* ），如圖一所示。*EmailDr* 藉由分析 Enron Email Dataset 中的資料，並計算個別類別書信的高頻片語，已得到適合提供給使用者參考的片語。此外，此一系統也可以透過互動的方式，協助學習者，檢視不同類別之書信常用詞彙與片語，提升學習者的寫作能力。

二、相關研究

近年來為了資訊交流愈加方便，拼字文法校正、建議用詞等等在寫作輔助系統上的功能，是自然語言處理中熱門的研究領域。在拼字文法校正，Grammarly 以及 Linggle 在這方面做了許多研究，而建議用詞的功能則在 EmailPro 以及 Gmail 之中有被提供，EmailPro 是在以不考慮類別下，提供使用者建議用詞，而 Gmail 的功能較偏向精準用詞，所以當建議用詞為高成功率的搭配詞，才會被顯示給使用者作為利用，因此在本論文中，會針對電子郵件的建議用詞上，提出將類別納入考量的方法。

學者針對建議詞彙的生成，也提出不同的方法在這過程之中。[1]先將文法模式取出後，利用 Macmillan English Dictionary 的資料，去做建議詞彙的優先順序。[2] 則是利用機器學習，針對來信的內容去做分析，藉此提供回信的建議用詞。

本文我們是參考 Smadja 演算法去生成搭配詞，設計如何過濾出較佳的詞彙束方式。搭配詞的篩選通常是利用統計的方式從語料庫中選取候選的詞彙，再判斷這些候選的搭配詞何者是合理的。計算兩個單詞相隔的距離在資料中出現的次數，可有效的協助篩選良好的詞彙束

與我們最相關的研究，[3] 也是利用 Smadja 演算法，這個演算法有兩個假設：搭配詞的出現頻率遠高於非搭配詞，以及搭配詞出現的次數在詞與詞之間的距離上的分布有峰值。會先計算任意兩個單字在特定距離的時候出現的次數此為 特定距離時的 skip bigram，進而選出次數最高的 skip bigram 作為最終的搭配詞，而與[3]在過程中的差別是預先篩選搭配詞，利用被預期的搭配詞排序先後，以及實際的搭配詞排序先後，相除得到比例去得到它認為較好的搭配詞，再進行詞彙束的篩選。但我們的文章是直接針對詞彙束做篩選，以 NLTK 中的停用詞作為功能詞，先留下有功能詞的資料，再利用我們生成的搭配詞進行查詢以提取我們要的詞彙束。此外，我們也利用 Linggle 的相似詞功能，去擴增資料，藉以提升輸出的效果。

相對於以往的郵件寫作輔助系統，在本文中會提出一套寫作輔助系統，利用已經標記好的片語整理成 N-連詞(N-gram)，將搭配詞與現有的樣板資訊相互比對，篩選出效果較佳的詞彙束，套用至系統上，幫助使用者書寫郵件。

三、研究方法

本論文的目標為開發一個電子郵件輔助寫作系統，提供不同類別的建議用語。研究方法我們依下列階段進行，分別為（一）資料前處理，（二）生成搭配詞，（三）篩選良好的詞語束，（四）建立預測模型。

（一）資料前處理:

利用 spaCy 套件解析 WriteExpress (https://www.writeexpress.com/) 資料的句子，得到對應的詞性 (part-of-speech tag)、名詞片語 (noun chunk) 等資訊，再以名詞片語為 token 單位來切 Ngram (n 從 3~5)。之後，再根據 Ngram 中的每個 token 詞性來找出相對應的 grammar pattern (這邊的 grammar pattern 會做 lemmatization)，詳細步驟如下：

1. 辨別句子之詞性(Part-of-Speech Tag)

句子 (e.g., "My husband and I will be delighted to be part of the celebration.") 經由 spaCy 產生資訊，例如詞性為 [DET NOUN CCONJ PRON VERB AUX ADJ PART AUX NOUN ADP DET NOUN PUNCT]，名詞片語為 [My husband, I, part, the celebration]

2. 產生 Ngram 資料

將句子，以三個到五個詞為單位，產生 Ngram，以上述句子為例，切出來的 Ngram 有[My husband and, My husband and I, My husband and I will, ...]等等組合。

3. 將 Ngram 轉為針對每組 Ngram 搭配名詞片語的資訊，再轉換成 grammar pattern 形式 (e.g.,“My husband and I will“轉換為“My n. and SOMEONE v.” 以及 “My husband and SOMEONE v.”)

我們共有 63 種信件類別，其中有 2 類資料短缺，實作時只能排除。這個步驟後產生 4,132,282 種 Ngram，共有 3069634 個 grammar pattern 分布在 61 類中，共有 818,169 種不同的 grammar pattern。

（二）生成搭配詞

研究搭配詞時通常只研究 base word 和 collocate 兩個單詞，可以記錄成(w,w1)，比如 listen 和 music 的 collocation 會表示成 (listen, music)，省略中間的單詞。這種類似 bigram 的形式，稱為 skip bigram。藉由統計 Enron Email Dataset 的 skip-

gram 等資訊，並利用 Smadja 演算法篩選出合理的搭配詞。結果可以用來過濾掉不適合或不完整的 Ngram (使用 詞彙束 Lexical Bundle 的概念，因此在第三步我們將過濾完的 Ngram 稱為 Lexical Bundle)。詳細的步驟如下（見表一）：

1. 找出 window size = 5 之內的所有 skip-bigram。
2. 計算每個候選搭配詞的頻率、分佈等資訊。
3. 保留符合 Smadja’ s Algorithm 篩選標準的搭配詞。

left token 左字元	right token 右字元	distance	count
invite	you	1	2087
invite	to	2	2081
invite	attend	3	210
thank	you	1	5861
thank	for	2	3655
thank	your	3	1208
total collocations		969750	

表一、Smadja 演算法生成的搭配詞

（三）生成詞彙束

利用第二步所產生的搭配詞來過濾第一部產生的 Ngram。因為使用到 lexical bundle 的概念，因此我們稱過濾出的乾淨資料為 lexical bundle。lexical bundle 的概念提到，一組 lexical bundle 通常都擁有 function word (這裡我們使用 nltk 的 stopword 作為 function word)，詳細步驟如下（可見表二）：

1. 先留下有 function word 的 Ngram，接著找出每個 Ngram 的 skip-bigram。
2. 用第二步產生的搭配詞表查詢，只留下「至少有一個 skip-bigram 為搭配詞表中出現過」的 Ngram。

未過濾的 Ngram	過濾後的 Lexical Bundle
Mr. and Mrs. John Doe	accept your kind invitation to
accept your kind invitation to	to brunch at

to brunch at	the twentieth of
of October at	
the twentieth of	

表二、詞彙束過濾前後的結果

（四）分析同義詞

由於雖然有些 pattern 有不同的用字，用法實際上卻是很相近的（如 “I am glad to” 及 “I am happy to” ），我們利用查詢 Linggle（Linggle.com） 得到相近詞來擴展 pattern，讓輸入更容易被對應到 pattern 。

1. 計算出信件樣板之 關鍵詞

利用 Chi-square test ，從 WriteExpress data 得到關鍵詞，在這裡我們取分數最高的前 10 名。

2. 藉由 Linggle 查詢相似詞

利用 Linggle API 查詢，得到各個配對出現在例句的次數，從而得知哪些配對較常一起出現、可能有相近的意思。例如我們從上一步得到三個關鍵詞 “pleased” 、 “honored” 及 “sorry” ，透過 Linggle API 查詢 “pleased/honored/sorry and pleased/honored/sorry” ，便能知道 “pleased” 及 “honored” 較常一起被提及，因此較有可能具有相近的詞義。在這裡，我們取出次數前三名的配對，以及第四名後次數超過 10000 的配對列為相似詞。

3. 擴展片語

將片語中的單字，以其同義詞取代，產生另一組片語，藉此可以擴增片語，形成片語的聚叢。

（五）統計式預測模型

我們以 lexical bundle 為 knowledge base 來跟使用者輸入做比對，最主要是比對輸入的最後兩個 token 和 knowledge base 的每個 pattern，最後產生相對應的建議。詳細步驟如下：

1. 輸入句(通常是不完整句)先使用 spaCy 斷詞，再取出最後兩個 token。

2. 首先比較 pattern 的前兩個 token 和輸入句的最後兩個 token ，相同則 hit。

3. 若沒有，則比較輸入的最後一個 token 和 pattern 的第一個 token，相同則 hit。

4. 若沒有，則比較輸入的倒數第二個 token 和 pattern 的第一個 token，相同則 hit。

5. 從 hit 中按照頻率取出前五名成為給使用者的建議。

四、實驗

（一)、資料集

透過網路爬蟲收集 WriteExpress(https://www.writeexpress.com/) 所提供的各項類別的樣板，並依照信件樣本、例句、搭配用語作為分類，另一個則是 Enron Email Dataset 共五十萬筆的信件內容，這兩項資料皆以 3-grams、4-grams，以及 5-grams 的方式將文字做切割，再利用這些資料進行分類的預測模型訓練。

（二）實驗模型與結果

1. 利用範例信件進行評估準確率

我們使用 《How To Say It》一書中的範例信件來評估系統，書中共有 50 類書信，其中有 29 類與我們的系統相同，因此我們取這些類別中的每篇範例信件來進行評估。評估方式如下：

$$\sum_{i=1}^{w} \frac{((N_i - n_i + 1)/N_i)}{\text{總字數 w（去除標點符號)}}$$

輸入為 token[: i]，系統輸出共輸出 N_i 筆 pattern，其中第 n_i 名 pattern 的 Ngram 與 token[i-2 : i+1] 或 token[i-1 : i+1] 相同，i 從 1 ~ w，w 為範例信件長度。我們共做了 156 篇測資，去除標點符號及不在 Training corpus 內的字 (out of vocabulary) 後共 12950 字。由於測資的部分並沒有完全對應到 WriteExpress 上所提供的類別，所以單以類別作為評估標準，而測試結果準確率最高的是 response 類別，最低的是 application，下表（見表三）顯示前三高的類別與最後三名的類別。我們認為由於訓練的資料不夠豐富，以及分析測資內容時，有特別針對某些詞語做歸類，會影響由 WriteExpress 資料所統計出的常用搭配詞，進而導致資料不平均，推測藉由其他資料來源，以及擴增特定用語的數量，可以將準確率不佳的類別做程度上的改善。

類別	準確率
response	0.611

love	0.607
appreciation	0.566
refusal	0.422
introduction	0.416
application	0.387

表三、信件類別與準確率

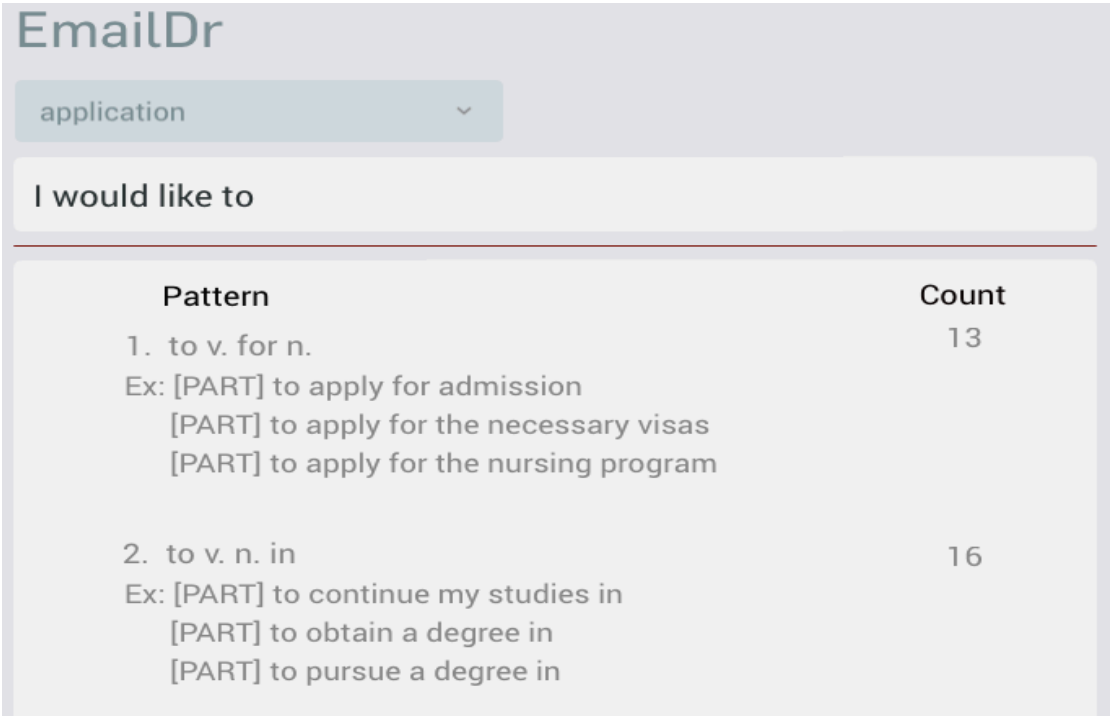
2. EmailDr 與其他系統的比較

(1)EmailDr 與 EmailPro

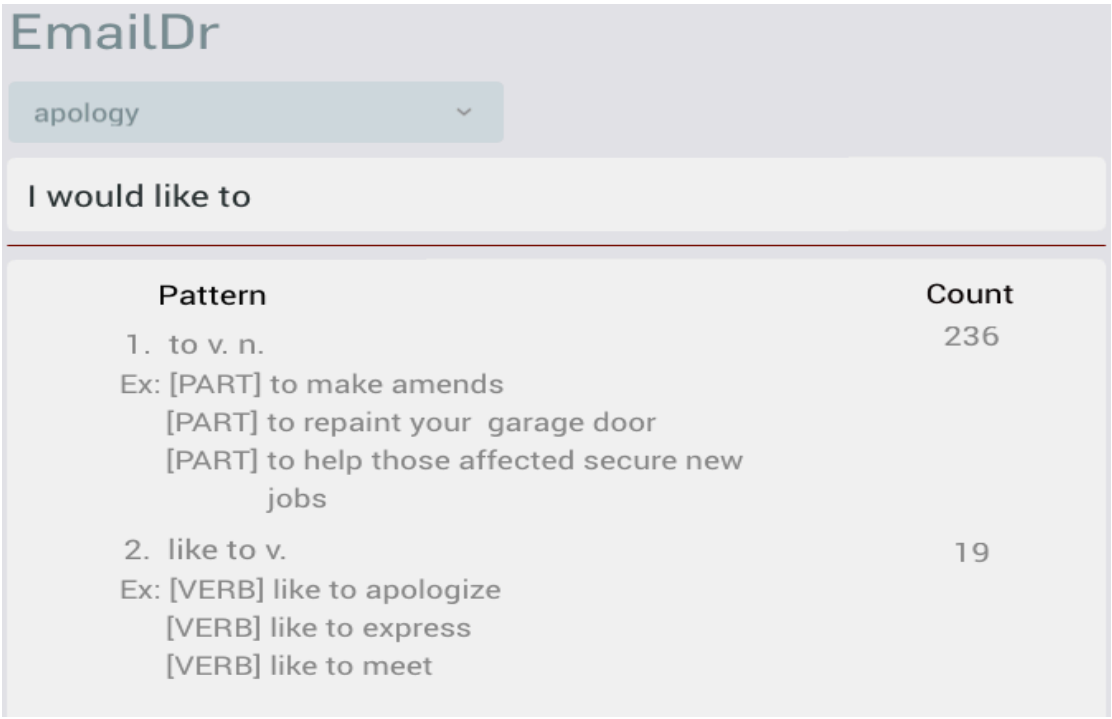
我們利用 EmailDr 選出類別後所提供的資訊（可參見圖二、圖三)，以及圖四，能夠發覺在有類別的選擇之下，可以比較精準而快速的提供建議用語，相反的 EmailPro 呈現的資訊較為繁雜（可見圖四），容易造成查詢者使用上的困擾。

(2)EmailDr 與 Gmail

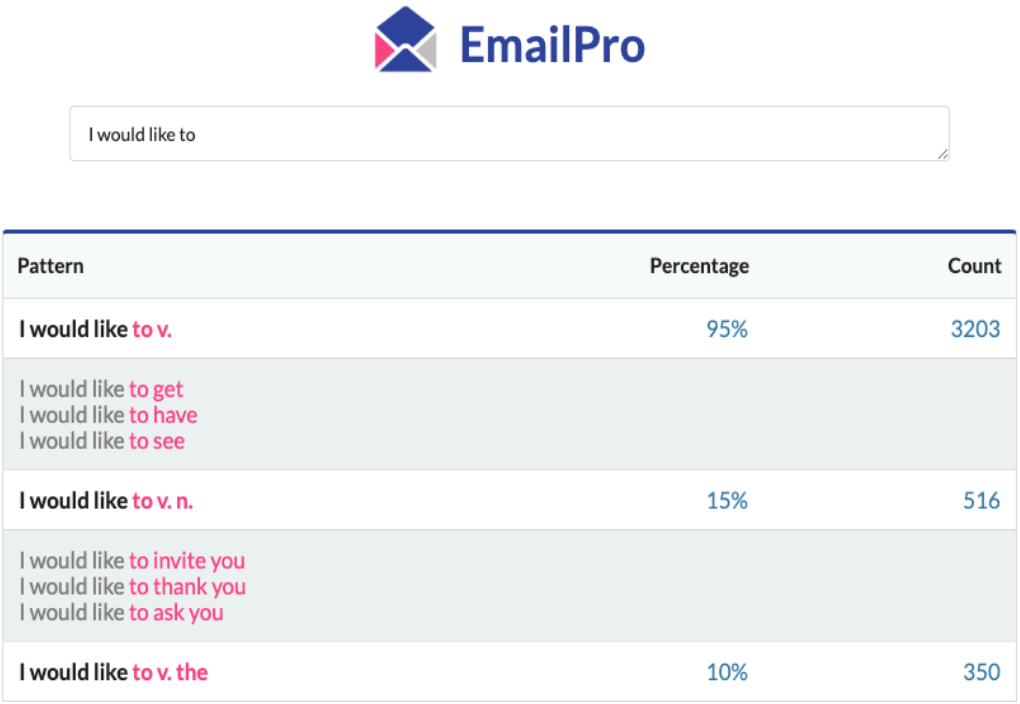
以一封申請信的 *"I would like to..."* 為例，在我們的系統中，選擇類別後，能夠在開頭輸入少量資訊（可參見圖二），便提供該類別的相關建議用語，而 Gmail 可能無法立即地提供資訊（可參見圖五）。除此之外，我們所提供的資訊可能稍多，較適合作為使用者教學以及學習使用，Gmail 則是比較偏向高成功率的搭配詞出現時，才給予使用者建議，使其能降低其錯誤率。



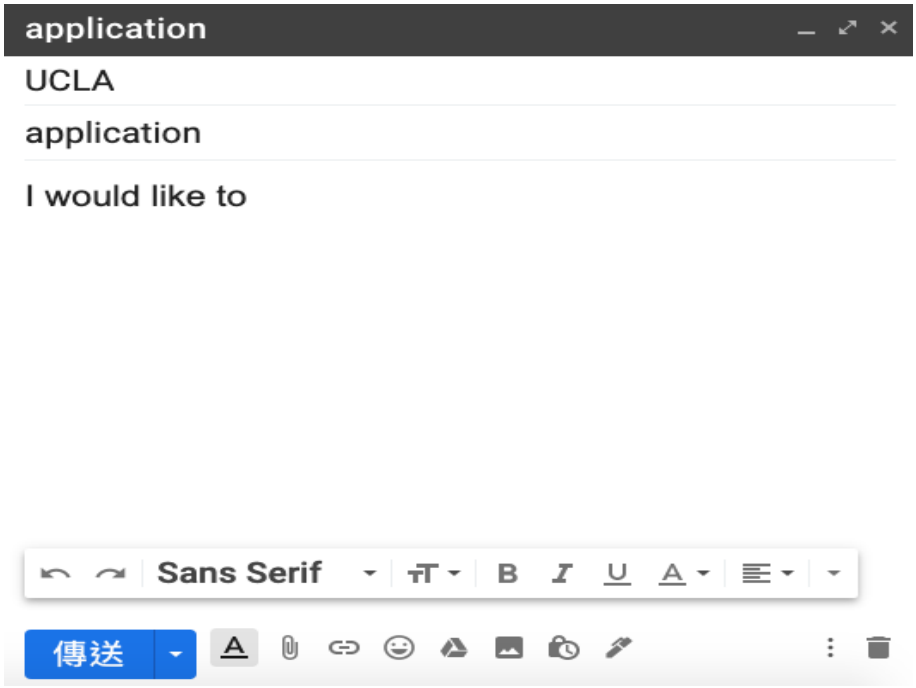
圖二、EmailDr 選取 Application 類別所提供的建議用語



圖三、EmailDr 選取 Apology 的類別所提供的建議用語



圖四、EmailPro 呈現的建議用語



圖五、在 Gmail 輸入未完成句所呈現的情形

五、結論

本篇論文建立一個電子郵件輔助寫作系統，從資料收集、分析處理、統計頻率，至產生供使用者利用的建議用語。實驗結果顯示，我們的系統在參考類別後所提供的建議用語具有不錯的效果。此外也在少量詞彙資訊輸入時，即時提供建議，即使有些建議可能無法貼近使用者的需求，但還是盡可能達到輔助寫作的作用。

未來，我們將會持續收集與標記信件的樣板資料，以及利用機器學習的方式使大量的郵件內容可以被自動分類成各個類別。此外，加強信件開頭語的建議用詞，讓使用者能夠更便捷完成一封信件的書寫。

六、參考文獻

[1] Jim Chang, JS Chang, “WriteAhead2: Mining Lexical Grammar Patterns for Assisted Writing”, Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 106-110. 2015

[2] Hsu, C. L., Ju, H. H., Wu, Y. H., Peng, H. C., Chen, J. J., Chang, J., & Chang, J.. Computer Assisted English Email Writing System. Proceedings of the

International Conference on Computational Linguistics and Intelligent Text Processing, CICLing, 2017

[3] Benjamin N Lee, Gagan Bansal, Yuan Cao, Shuyuan Zhang, Justin Lu, Jackie Tsay, Yinan Wang, Andrew M. Dai, Zhifeng Chen, Timothy Sohn, Yonghui Wu," Gmail Smart Compose: Real-Time Assisted Writing." Knowledge Discovery and Data Mining, KDD, 2019

[4] Email Pattern dataset source web site: https://www.writeexpress.com/