The 2015 Conference on Computational Linguistics and Speech Processing
ROCLING 2015, pp. 87-99

基於貝氏定理自動分析語料庫與標定文步 *

A Bayesian approach to determine move tags in corpus

張瓊文 Chiung-Wen Chang 、徐嘉連 Jia-Lien Hsu[1]
私立輔仁大學資訊工程學系
Department of Computer Science and Information Engineering,
Fu Jen Catholic University, Taiwan (R.O.C.)

張俊盛 Jason S. Chang
國立清華大學資訊工程學系
Department of Computer Science,
National Tsing Hua University, Taiwan (R.O.C.)

摘要

利用科技幫助語言學習，是一個重要的研究議題，英文是現今人們主要的溝通語言，對於非英語體系的國家，學習英語 (從聽力、閱讀到寫作) 是一件困難的事情。尤其在寫作方面，由於英文文法跟中文文法上的差異，導致在學習英文寫作時，常常會將組成句子的架構搞混，使得在學習寫作有較大的困難。

英文學術論文寫作，不同於一般文章寫作，通常有明確的架構與段落，如「簡介」、「相關文獻」、「方法」、「結果」等，此結構稱為「文步」。此外，學術論文寫作與一般寫作有些許的不同，在寫作的用詞上就有些差異，因此，為了幫助需要寫學術論文的同學們，我們參考學術論文的文步架構，設計文步分類器訓練語言模組，擷取在特定文步使用的字詞。

在語言處理方面，學者們依照文步架構，提出自動化分析，但是在訓練語言模組中通常需要大量人工標註資料，為了降低人工標註的部分，我們將專家整理歸納的詞彙，透過機器學習與迭代 (bootstraping) 的方法達到學習效果，再利用訓練過的語言模型，預測文章句子當中的文步。

在本研究中，我們提出一套系統，以貝氏方法 (Bayesian approach) 做語言文步分析，此系統分為兩部分，一為訓練階段 (Training phase)，另為測試階段 (Testing phase)。在訓練階段中，透過大量的文本 (Corpus) 建立學習模型，採用專門蒐集學術論文簡介的語料集 (CiteSeerX) 與初始規則 (Initial pattern) 做為分析的依據，利用貝氏方法判斷語料庫中每篇簡介裡的句子所屬的文步 (move)，當句子被標定完文步之後，利用迭代的方法更新貝氏模型，達到學習效果。而在測試模型中，將訓練階段得到的結果，給予一篇新的簡介，一樣透過貝氏方法，預測文步，經過測試階段，我們得到文步預測精確率為 56%。

關鍵詞：學術英文寫作、輔助寫作、文步分析

Abstract

English of Academic Writing (EAW) is essential to the research community for sharing knowledge. Research documents using EAW, especially the abstract and introduction, may

*此研究由科技部資助，編號為：MOST-103-2511-S-007-002-MY3
[1]通訊作者：徐嘉連 Jia-Lien Hsu (E-mail: alien@csie.fju.edu.tw)

follow a simple and succinct picture of the organizational patterns, called *move*. This paper introduces a method for computational analysis of move structures, the Background-Purpose-Method-Result-Conclusion in this paper, in abstracts and introductions of research documents, instead of manually time-consuming and labor-intensive analysis process. In our approach, sentences in a given abstract and introduction are automatically analyzed and labeled with a specific move (i.e., B-P-M-R-C in this paper) to reveal various rhetorical functions. As a result, it is expected that the automatic analytical tool for move structures will facilitate non-native speakers or novice writers to be aware of appropriate move structures and internalize relevant knowledge to improve their writing.

In this paper, we propose a Bayesian approach to determine move tags for research articles. The approach consists of two phases, training phase and testing phase. In the training phase, we build a Bayesian model based on a couples of given *initial patterns* and the corpus, a subset of CiteSeerX. In the beginning, the priori probability of Bayesian model solely relies on initial patterns. Subsequently, with respect to the corpus, we process each document one by one: extract features, determine tags, and update the Bayesian model iteratively. In the testing phase, we compare our results with tags which are manually assigned by the experts. In our experiments, the promising accuracy of the proposed approach reaches 56%.

Keyword: Academic English Writing, Assisted Writing, Move Tag Analysis

一、緒論

自然語言處理是近幾年學術所關心的議題，在科技尚未發展以前，語言處理幾乎靠人力檢查與校正拼字與文法錯誤，但是靠人力，則會產生人為的失誤，意思是指並非人工檢查就表示寫作的用詞與語法正確，所以採用機器學習來替代人工的方式，相較於機器學習，人工校正或是處理文字相對花費較多時間。

英文是在學術上主要溝通的語言，所以非英語體系的國家，對於英文寫作這一部分相較之下，發生文法與拼字的錯誤率會明顯提高，因此在資訊發達的世代，學術機構開始收集寫作資料，譬如：英文檢定考的作文 (ETS)、學生寫的作文資料集 (CLEC) 與維基百科的編輯紀錄等等，有這些語料集 (Corpus)，學者們開始從事多方面的語言處理與分析研究。利用語料庫，分析英語的用法 (搭配詞、文法)，運用統計，找出大部分人們所使用的句法，嘗試著從數據當中找到理論，藉此幫助學習，以及提升寫作上的效率。

在學術論文中，簡介此一章節，通常會描述：問題的背景、主要目的、解決方法、結果與結論，此修詞結構的組成稱之為「文步」。在過去的研究中 [1–3]，針對論文簡介定義出四個文步，包括：問題 (Problem)、方法 (Solution)、評估 (Evaluation) 與結論 (Conclusion) 等部分。美國國家標準協會 (American National Standard Institute, ANSI) [4]，審核並規範寫作的文步結構為目的 (Problem)、方法 (Method)、結果 (Result) 與結論 (Conclusion)。Swales [5] 定義在論文寫作依循的三大文步修辭結構 (Creating a Research Space, CARS)，包括：為建立研究領域 (Establishing a research territory)、建立利基 (Establishing a niche)、占領利基 (Occupying the niche)，並在每一個文步修辭結構之下定義細節，藉此幫助描述文章內容。

特別針對學術英文寫作，Glasman-Deal [6] 提出寫作上文步模組，包括：介紹、方法、結果、討論。Weissberg & Buker [7] 定義學術論文寫作文步為 BPMRC，即背景 (Background, B)、目的 (Purpose, P)、方法 (Method, M)、結果 (Result, R)、討論 (Conclusion, C)。

在本篇論文使用 Weissberg & Buker 提出文章的文步架構 (背景、目的、方法、結果、結論)，利用大量的學術論文資料 (CiteSeerX) 與少量初始規則，訓練貝氏模型 (Bayesian approach)，學習如何判別句子所屬的文步。

為了得知訓練完畢的貝氏模型所提供文步的精確度，則利用單一篇新的學術論文簡介，透過貝氏分類器進行文步標定，最終由人為判別文步的正確性。

本論文接著的部分會先探討相關研究 (Section 2)，進而描敘分類器自動學習標定文步的過程 (Section 3)，與實驗設計、結果 (Section 4)。最後，討論未來的研究方向與結論 (Section 5)。

二、相關研究

隨著資訊發展，為了讓資訊交流快速，關於自然語言處理為相當重要的研究領域，在純文字的應用包括機器翻譯、拼字校正、資料檢索等等。近年來學者對於學術論文或是期刊，有進一步的研究 (Swales & Feak, 2004)。主要針對論文的段落與句子進行人為的分析研究，經過歸納之後提出關於論文修辭的架構規則-「文步」。在本研究中，則是針對論文的「簡介」這一個章節做分析，提出自動化分析論文文步結構的方法。

大部分論文簡介有著簡單文步結構–IMRD [8]，即為介紹 (Introduction)、方法 (Method)、結果 (Result)、討論 (Discussion)，許多學者也定義出不同的論文文步結構，例如 Swales [5] 為簡介此小節提出 CARS(Creating a Research Space) 模組，CARS 主要為 3 大文步並細分為 11 文步，使得許多學者使用 CARS 模組探討寫作上的修辭方法，Weissberg & Buker [7] 整理出 BPMRC 文步結構，即背景 (Background)、目的 (Purpose)、方法 (Method)、結果 (Result)、結論 (Conclusion)，為學者與作者提供研究方向與寫作建議。

近幾年來，有許多學者採用不同機器學習的方式訓練文步分類器，例如 Teufel & Moens [9] 利用簡易的貝氏分類器 (Naive Bayesian Model, NBM) 透過修辭的狀態與關聯針對論文全文進行文步分類。Ling [10] 提出隱馬可夫模型 (Hidden Markov Model, HMM) 利用統計機率去做文步標註，Wu & Jason S. [11] 提出一套系統 (CARE)，利用 HMM 標記文步。Shimbo [3] 透過 MEDLINE，提出一套系統，讓使用者可以搜尋簡介特定的文步，此系統利用支撐向量機 (Suport Vector Machines, SVM)，系統將簡介分為四個部分，為目的、方法、結果、結論，每個句子可以利用位置找出上下文，作為判別文步的依據。Yamamoto & Takagi [12] 將簡介中的句子分為背景、目的、方法、結果、結論，訓練線性 SVM 找出動詞時態與相對的句子位置當作分類依據，進行文步標註。

在本文當中，所採用的機器學習演算法為貝氏定理 (Bayesian)，貝氏定理在自然語言處理上常被用於統計式翻譯 (Statistical Machine Translation, SMT)，在條件機率理論上，預測原文被翻譯為譯文的方式，去做機器訓練 (Jia Xu, 2008) [13]，利用大量的論文資訊，運用貝氏定理採取半監督式分析法，預測一篇簡介的句子，屬於何種文步來做討論。

與本文最相關的研究，為 Guan-Cheng Huang [14] 的論文研究，主要的區別為所採用的分類架構有所不同，Guan-Cheng Huang 提出：背景（領域、缺口、前人研究）、本論文 (目的、方法、結果）、討論（和前人研究的比較與對照）與文節結構（論文組織、圖表的指示、內容的預告與回顧）等四種文步，而本篇所採用的文步為五種 (背景、目的、方法、結果、結論)，在應用上，訓練文步分類器的演算法有些差別，本文是採用貝氏分類 (Bayesian) 而 Guan-Cheng Huang 提出最大熵模型 (Maximum Entropy, ME)，差別在於貝氏在運算的一開始需要先驗機率條件，依據先驗條件推理出文步機率，而最大熵模型則不需先驗條件，所以會平均分佈，不傾向於任何文步，但在訓練過程中接觸到其他訊息，則會調整文步的機率分佈。

相對於前人研究文步分析的文獻，在本文當中提出一套自動學習系統，利用專家已經歸納的文步片語整理成 N-連詞 (n-gram)，以降低人工標示的成本，在訓練的過程中，利用文步特徵，自動將句子標示，使得系統可以分類文步並從中擴充字詞，利用自動化文步標示而得到的字詞，套用到英文輔助寫作系統，幫助學生寫學術論文。

三、方法

為了提供使用者在寫學術論文時，在不同章節 (文步) 可以使用較正確的字詞，我們必須擁有大量已經被標註的文步字詞來做寫作上的提示，而人工自行標註字詞的文步需花費大量的時間，因此，我們採取專家整理過的字詞透過自動學習的方法，省去人工標註所需花的時間，我們將問題定義如下。

我們將句子經過 Genia Tagger 斷字之後，採用三種特徵訓練出貝氏模型 (OW, BF, BPC)，以迭代 (bootstrapping) 的方法擴增貝氏模型，計算之後，將一篇文章當中的句子，單獨觀察一種文步，找出在此文步機率最高的句子進行文步標註，避免一篇文章當中只有一種文步的情形發生，將已被標定文步的句子分為 N-連詞 ($S = \{ng_1, ng_2, ... \}$) 回饋到初始表，藉以達到訓練的效果。

在測試階段，則會選取一篇新的文章簡介，透過訓練完畢的模型結果，評估文步標註的精確率。

在此章節，敘述我們所使用的演算法，包含貝氏定理所需要的先驗機率與文步的挑選，並

問題陳述
給定：以學術文章組成的語料集 (*Corpus*) 與初始訓練規則 (Initial pattern)
我們先計算一個初始模型
給定：一篇學術文章 D ($D \in Corpus$)
目標：為單一篇文章 ($D = \{S_1, S_2, ...\}$) 中每一句子 S_i
判定句子文步
標上文步標籤 (move-tag = $\{B, P, M, R, C\}$)
同時，新增或更新規則

介紹系統架構圖與模組訓練的過程。

3.1 系統架構

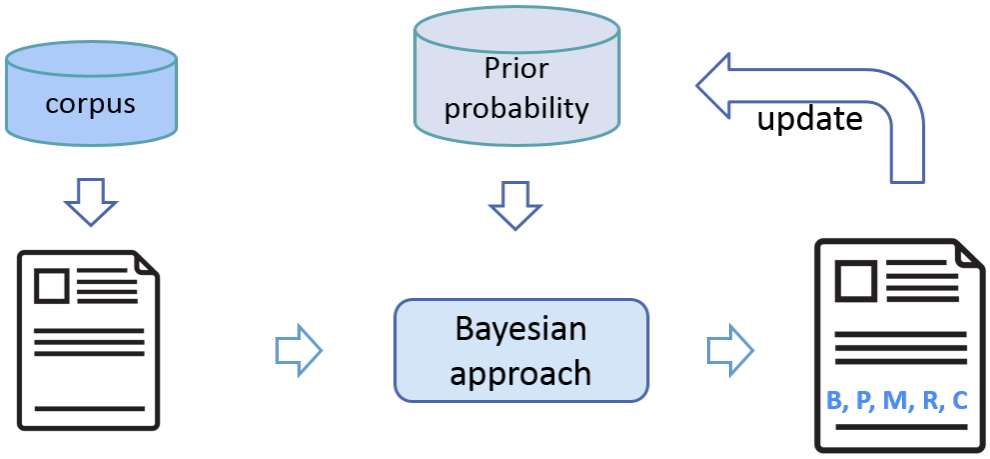

圖 1: 系統架構 (System architecture)

參考系統架構圖 (圖 1)，分為兩大部分: 一部分是利用 CiteSeerX 語料庫，訓練貝氏分類器，另一部分則是使用訓練好的貝氏分類器預測新文章的句子。

在訓練階段，從 Glasman-Deal 此書當中依照文步所提供的資訊，擷取 155 句 N-連詞當作初始文步訓練規則，而在資料方面則是取專門收集學術論文簡介的語料庫 (CiteSeerX)，須先將語料庫逐篇經過 Genia Tagger 進行斷字處理，利用程式將簡介中的句子分割成一句一句，然後再把句子依照 Genia Tagger 提供的詞性標註 (Part of Speech, POS)、字根還原 (Base form) 與語意區塊 (chunk) 做預先處理，將處理過後的句子分為 N-連詞，依照初始文步當作依據，經過分析將語料庫所提供句子字詞進行文步標註，並回饋到初始表當中，經過反覆訓練的過程，擴增已被標記的 N-連詞當作下一次計算的依據。

在測試階段，選取新的一篇文章，一樣使用 Genia Tagger 進行預先處理，將訓練階段得到大量被標記文步的 N-連詞，當作先驗資訊，測試文章經過計算之後，逐句所得的文步標籤是否正確，進而得知方法的效率。

3.2 特徵選取

本文中，挑選 BPMRC 此文步架構當作句子分類類別，從語料庫逐篇處理句子，若篇幅當中的句數少於五句，則會忽略不做處理。

而文章中的句子 (S) 會經過 Genia tagger 處理每個字詞 (W)，Genia tagger 會提供字詞的一些特徵，例如詞性標記 (Part-of-Speech, POS)、意元集組 (Chunk)。例如一篇文章中其中一句 (S_1) 為"*Glyoxysomal citrate synthase in pumpkin is synthesized as a precursor that has one*

cleavable presequence at its N-terminal end." 經過 enia tagger 分析之後 (表 1)，我們依照結果，將句子整理成三種表達方式，分別為

1. 原始資料 (OW: Original word)
   將句子保留原始資料，包含過去式、複數等等，但是捨去部分符號，使得句子只由單字組成，所以原始句子 (S_1)，將會轉成如下：
   "*Glyoxysomal citrate synthase in pumpkin is synthesized as a precursor that has one cleavable presequence at its N-terminal end.*"

2. 字根還原 (BF: Base form)
   將句子包含的單字，還原字根，使得句子被簡化，所以原始句子 (S_1)，將會轉成如下：
   "*Glyoxysomal citrate synthase in pumpkin be synthesize as a precursor that have one cleavable presequence at its N-terminal end.*"

3. 利用意元集合與詞性 (BPC: Base form & POS & Chunk)
   透過 POS 給的規則，將代表數字 (*CD*、*LS*) 或是非英語單字 (*FW*) 的字詞換成標籤 (*one*→*CD*)，而符號 (*SYM*、*$*、*:*) 則是忽略，並考慮 *Chunk*，找出單字的前後屬性是否為一個組合 *Base form* 使得單字可以統一，則句子會轉成如下：
   "*Glyoxysomal citrate synthase in pumpkin be synthesize as a precursor that have CD cleavable presequence at its N-terminal end.*"

表 1: 將 S_1 = "*Glyoxysomal...*" ，經過 Genia Tagger 分析之結果

Original word	Base form	POS	Chunk	Named entity (NE)
Glyoxysomal	Glyoxysomal	JJ	B-NP	B-protein
citrate	citrate	NN	I-NP	I-protein
synthase	synthase	NN	I-NP	O
in	in	IN	B-PP	O
pumpkin	pumpkin	NN	B-NP	O
is	be	VBZ	B-VP	O
synthesized	synthesize	VBN	I-VP	O
as	as	IN	B-PP	O
one	one	CD	B-NP	O
precursor	precursor	NN	I-NP	O
that	that	WDT	B-NP	O
has	have	VBZ	B-VP	O
...	...	...	...	...
. (Period)	. (Period)	.	O	O

3.3 初始規則

針對初始 N-連詞的選用，採用 Glasman-Deal 所撰寫的教科書 [6]，此書歸納出在不同文步上該如何建立一個寫作架構與在文步上所該使用的詞彙，從中挑選，我們將選取出的 N-連詞 (參考表 2，依照文步給予初始值，比如"a basic issue for" 在書中的建議在背景 (B) 當中使用，所以代表此 N-連詞在背景出現次數為 1(表 3)。利用被標註分類的 N-連詞，當作訓練資料，運用貝氏定理計算而自動產生大量標註完的論文句子，將句子分為 N-連詞，回饋於初始值，當作下一次訓練資料，而最後將訓練完的結果，進一步的分析。

所以在實驗當中，選出 155 個詞彙片語作為 N-連詞 (N-gram)，當作初始的特徵參數，其初始句數的分布如表 4。

表 2: 從 Glasman-Deal 撰寫的書 [6] 所擷取出部分的初始規則 (Initial pattern)

Pattern
a basic issue for
approach was developed by
majority of the tests
...
in future it is

表 3: 將初始規則給予出現次數，稱之為 *Count table* ($\mathcal{CT}$)

Pattern	B	P	M	R	C
a basic issue for	1	0	0	0	0
approach was developed by	0	1	0	0	0
majority of the tests	0	0	1	0	0
...	...	...	...	...	...
in future it is	0	0	0	0	1

3.4 貝氏方法

首先，貝氏定理需要有先驗機率，才能計算與分析，所以利用書本在文步架構上推薦的寫法，當作貝氏的特徵參數，再來，將一篇文章的簡介當中的字詞經過處理，使得文章當中的特殊符號不影響單字，利用貝氏定理計算文章中的每個句子去預測為背景、目的、方法、結果、結論的機率，當一篇文章的所有句子都計算完畢，才從所有句子當中是關於背景此文步最大值的句子標註為背景，已被標註的句子則不能重複被標註，當句子都被標註完，則將獲得辨識結果，回饋到一開始的先驗特徵參數。

從 *Corpus* 取一篇簡介 (D_1)，因為我們的分類模型為五個文步，若該篇簡介數句少於五，則忽略該篇文章，而句數超過五句，就定為一篇完整的簡介，而進一步分析。如表 5為擷取的一篇完整篇幅，接著將文章當中的一個句子 (S_1) 分別計算出可能為背景 (B)、目的 (P)、方法 (M)、結果 (R)、結論 (C) 的機率。

3.4.1 計算文步機率

以 S_1 為例，我們將分別計算文步的機率值。

$$P(\textit{move-tag}|S_1) = \frac{P(\textit{move-tag}) \times P(S_1|\textit{move-tag})}{P(S_1)}, \text{ when } \textit{move-tag} \in \{B, P, M, R, C\} \quad (1)$$

句子會計算每個文步的機率，由於每個文步的計算方法都相同，所以在往後的敘述將已背景 (B) 文步做代表。

而本文中給予的先驗特徵參數是給予 N-連詞，因為 S_1 假設為一組獨立 N-連詞所組成 (S_1 is approimated by set of n-grams as follows: $\{ng_1, ng_2, ...\}$) 所以要計算句子的所屬的文步機率，在此將句子劃分為 N-連詞來做運算，例如 S_1 近似為 n 個 N-連詞所組成。

$$S_1 \leftarrow \{ng_1, ng_2, ..., ng_n\} \quad (2)$$

表 4: 初始規則 (Initial pattern) 中，各種文步的次數分佈

move-tag	B	P	M	R	C
次數	25	34	33	41	22

表 5: 範例文章 $D_1 = \{S_1, S_2, ... , S_6\}$

S	Sentence
S_1	Glyoxysomal citrate synthase in pumpkin is synthesized as one ….
S_2	To investigate the role of the presequence in the ….
S_3	Lmmunogold labeling and cell fractionation studies ….
S_4	The chimeric protein was transported to functionally ….
S_5	These observations indicated that the transport of ….
S_6	Site-directed mutagenesis of the conserved amino acids in ….

所以 S_1 的機率定義為

$$P(S_1) \simeq P(ng_1) \times P(ng_2)... \times P(ng_n) \tag{3}$$

而 $P(S_1|B)$ 的條件機率定義為

$$P(S_1|B) \simeq P(ng_1|B) \times P(ng_2|B) \times ...P(ng_n|B) \tag{4}$$

根據上述的定義，將公式 (1)，定義為

$$P(B|S_1) \simeq \frac{P(B) \times P(ng_1|B) \times P(ng_2|B) \times ...}{P(ng_1) \times P(ng_2) \times ...} \tag{5}$$

3.4.2 計算 N-連詞機率

句子經過分割之後，得到其中一段 N-連詞 (ng_1)，例如為"ng_1: glyoxysomal citrate synthase in"，先計算 ng_1 出現的機率。

	n-gram	B	P	M	R	C
ng_1	glyoxysomal citrate synthase in	B_1	P_1	M_1	R_1	C_1
ng_2	a basic issue for	B_2	P_2	M_2	R_2	C_2
…	…	…	…	…	…	…
ng_m	role of the presequence	B_m	P_m	M_m	R_m	C_m

$$P(ng_1) = \frac{B_1 + P_1 + M_1 + R_1 + C_1}{\sum_{i=1}^{m}(B_i + P_i + M_i + R_i + C_i)} \tag{6}$$

則會判斷此 ng_1 是否存在於初始表 (表 3)，若存在於初始表 ($\mathcal{CT}$)，則會計算 N-連詞在該文步 (B) 次數出現的機率。

$$P(ng_1|B) = \frac{B_1}{\sum_{i=1}^{m} B_i} \tag{7}$$

若該 N-連詞不存在於初始表格或是在未曾出現於某文步，比如"a basic issue for" 此 N-連詞不曾出現於結論 (C)，則會給予極小值 ($\delta = 10^{-8}$) 當作機率。將每個 N-連詞，對照初始規則表 ($\mathcal{CT}$)，因此 S_1 透過運算則會得近似所屬的文步機率值。而各文步的機率，則是依照文步次數做為依據。

由於句子組成單字的多寡，會影響計算上的公平性，所以我們將結果正規化。

$$normalized\,(P(B|S_1)) = \frac{P(B|S_1)}{\#\ of\ n\text{-}gram\ in\ S_1} \tag{8}$$

3.4.3 文步標定

在文步標定上，在本文中，先將一篇文章 (D_1) 中所有句子的各文步機率計算完畢，才逐句標定文步。

而每個文步要標定幾個句子，則是依據給定的比例去做計算，由於句數不能為小數，所以取四捨五入的方法。

表 6: 一篇文章中 (D_1)，文步句數算法。

move-tag	文章句數比例	句數
B	$0.15 \times 6 = 0.9$	1
P	$0.20 \times 6 = 1.2$	1
M	$0.30 \times 6 = 1.8$	2
C	$0.15 \times 6 = 0.9$	1
R	$6 - (1 + 1 + 2 + 1)$	1

為了避免某一文步造成多數制 (Majority rule) 結果，我們根據文步在 *Corpus* 內文寫的比例多寡，依序標定文步 (以表 6為例，先標定 B 往後順序為 $C \rightarrow P \rightarrow R \rightarrow M$)。由於我們是由一篇文章判定句子文步，依照比例，我們先標定為 B 的句子。

$$\begin{aligned} &\because B_2 = max\{B_1, B_2, ..., B_6\} \\ &\therefore S_2 \leftarrow B \end{aligned} \tag{9}$$

若該句 (S_2) 已經被標上標籤 (B)，則將句子移除序列中，經由表 6計算，文章當中為 B 的內容為一句，則換標定下一個文步 C。

表 7: 經過第一次文步標定

Sentence	B	P	M	R	C
S_1	B_1	P_1	M_1	R_1	C_1
~~S_2~~	~~B_1~~	~~P_1~~	~~M_1~~	~~R_1~~	~~C_1~~
S_3	B_3	P_3	M_3	R_3	C_3
S_4	B_4	P_4	M_4	R_4	C_4
S_5	B_5	P_5	M_5	R_5	C_5
S_6	B_6	P_6	M_6	R_6	C_6

反覆標定過程，將文章當中的句子標上文步。

$$\begin{aligned} &\because C_6 = max\{C_1, C_3, ..., C_6\} \\ &\therefore S_6 \leftarrow C \end{aligned} \tag{10}$$

表 8: 經過第二次標定

Sentence	B	P	M	R	C
S_1	B_1	P_1	M_1	R_1	C_1
~~S_2~~	~~B_2~~	~~P_2~~	~~M_2~~	~~R_2~~	~~C_2~~
S_3	B_3	P_3	M_3	R_3	C_3
S_4	B_4	P_4	M_4	R_4	C_4
S_5	B_5	P_5	M_5	R_5	C_5
~~S_6~~	~~B_6~~	~~P_6~~	~~M_6~~	~~R_6~~	~~C_6~~

當一篇文章當中所包含的句子都已經標上文步，而我們也會根據結果，更新 $\mathcal{CT}$ 相對應的規則 (表 3)。假設 S_1 被標定為 B，而句子當中包含一個 N-連詞"ng: glyoxysomal citrate synthase in"，不存在 $\mathcal{CT}$，依照句子被標定的文步，新增 ng 至 $\mathcal{CT}$ 中並在 B 給予初始次數 (表 9)。

表 9: 新增規則

pattern (4-gram)	B	P	M	R	C
glyoxysomal citrate synthase in	1	0	0	0	0
...	...	...	...	...	...

若 ng 存在於 $\mathcal{CT}$ 中，則會依照 S_1 被標定的結果，更新 ng 在該文步出現的次數 (表 10)。

請注意，在這步驟中，我們僅是將前一步驟中、用貝氏方法判定的文步結果（沒有人為介入判定），加回 $\mathcal{CT}$ 中。我們並不立即判定所標定的文步是否為正確，而是以不斷迭代（iterative）的方式，利用貝氏方法，來抓住訓練資料（training data）的特性。

表 10: 更新規則

pattern (4-gram)	B	P	M	R	C
glyoxysomal citrate synthase in	$B_i + 1$	0	0	0	0
...	...	...	...	...	

四、實驗

在本節中，我們將討論實驗設定與結果討論。

4.1 語料庫

本文針對輔助英文學術論文寫作，因此我們採用專門收集發表過的學術論文語料集 (CiteSeerX)[1]。CiteSeerX 是一個關於文獻的搜尋引擎，在 1997 年，由美國普林斯頓大學開發 CiteSeer，建立一個數位圖書館，由於 CiteSeer 只能收集公開的文件，使得所收集文章領域有限，為了克服侷限性，針對系統架構重新定向 (CiteSeerX)，於 2007 年採用機器學習的方法，自動辨識網路上存在的論文，然後依照索引標示文章，透過引文的影響，連接每篇文章。

CiteSeerX 總共擁有 138 萬多篇的文獻，主要的內容為科學領域（包含資工和生醫領域），而這些資料來源通常為 PDF 格式，經過自動辨識轉檔成文字，因此語料庫裏頭包含許多換行連字符號、特殊符號等雜訊，所以在使用資料之前，我們透過文字處理，將冗餘的符號或是日期格式捨去，進而得到較完善的一篇論文。

4.2 實驗設定

參考系統架構圖 (圖 1)，我們利用初始規則 (Initial pattern)，分析語料庫 (CiteSeerX) 提供的文章，逐篇訓練語言模組，每當經過一千篇訓練的語言模組，則會測試精確度，在本論文當中，取兩萬篇當訓練資料。

在測試階段，事先從語料庫隨機提出 20 篇尚未經過訓練的文章，經過專家逐句標註文步，我們透過四個專家針對此 20 篇 (共 185 句) 逐句給予文步標籤，挑出其中三個人以上給予句子的標籤相同來評估資料的準確性 (三人以上相同句數共 142 句)。

我們將 20 篇文章進行測試，將句子標上標籤。之後定義如下精確率，依照 142 句正確答案，找出標上正確文步的句數。

$$Accuracy = \frac{\#\ of\ sentences\ with\ correct\ move\text{-}tag}{142}. \tag{11}$$

4.3 實驗結果

本文利用 CiteSeerX 提供的資料，計算每經過一千篇的訓練後，則會增加多少 N-連詞的先驗規則，因資料經過 Genia Tagger 處理之後會提供資料原始字詞 (Original Word)、字根還原 (Based form)、詞性標記 (POS) 等資訊，則訓練方法給的文字資料為此三種方式，透過運算得到的結果。

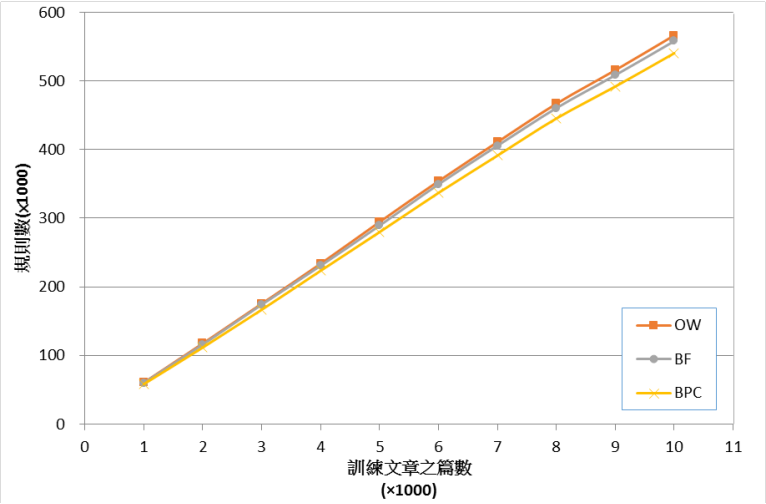

圖 2: 經由訓練，增加的規則數

每經過一千篇的訓練，則 $\mathcal{CT}$ 會增加約六萬的 N-連詞規則，經過測試，並沒有發現收斂的現象，可能是因為文章當中有過多特殊字詞，或者是因為我們設定的 N-連詞太過於長，導致

[1]CiteSeerx: http://citeseerx.ist.psu.edu/about/site

組合過多。

表 11: 專家標註文步的句數 (# of sentence with correct tags)

move-tag	B	P	M	R	C	Total Sentence
# of sentence	27	10	21	60	24	142

在測試階段，為能了解資料經過訓練的篇數是否影響文步標籤的精確率，所以評估每經過一千篇訓練的 $\mathcal{CT}$ 表，預測句子文步的標註是否正確。

首先評估原始資料經過逐篇訓練而得到的 $\mathcal{CT}$ 資料表，所預測句子文步的精確率。

由圖 3得知評估的結果，可以觀察句子文步標籤的精確率，發現 $\mathcal{CT}$ 資料表每經過一千篇的訓練，得到的結果逐漸改善。

由於 BF 做出的實驗結果與 OW 相似，所以在此只顯現精確率的結果。

再者，評估文章經過詞性標記與意元集組處理的句子所訓練的 $\mathcal{CT}$ 資料表，預測句子文步的精確率。

		結果比較				
		B	P	M	R	C
篇數	1000	9	3	9	26	9
	2000	9	4	10	27	9
	3000	9	3	12	27	10
	4000	9	3	11	28	10
	5000	9	3	11	29	10
	20000	10	4	11	30	10

圖 3: 關於 OW 標定句子文步資訊

		結果比較				
		B	P	M	R	C
篇數	1000	9	4	9	31	9
	2000	9	5	11	32	9
	3000	10	5	11	32	12
	4000	9	4	11	31	12
	5000	9	4	12	31	12
	20000	10	5	14	38	13

圖 4: 關於 BPC 標定句子文步資訊

由圖 4得知評估結果，相對於原始資料的精確率略為提高，原因為將句子簡單化，避免意思相同的 N-連詞，因為一些數字或是非英語單字而降低文步的特徵計算。

4.4 討論

整體而言，文步預測的正確率為 56%，尚有進步的空間。在訓練規則少的情況下 (155 個規則)，能達到一半的準確率，對於此情形，我們保持樂觀的態度。

而在統計規則新增圖表中，對於規則數持續增加的問題，我們設定三種特徵選取字詞，將句子的字詞變得較抽象，例如不在乎時態與複數、替換符號或數字等等，想藉此將規則的數量減少，但並未達到預期的效果 (表 15)。

我們採取三種特徵訓練得到的結果，由於原始資料 (OW) 與字根還原 (BF) 此兩種特徵所得到的精確率，沒有預期的差距，而在詞性替換 (BPC) 的特徵下，相對於 OW 與 BF，文步的精確率有明顯提升，可能因為提供的規則表達方式比較簡易，在提供計算時的文步特徵較明顯。

表 12: 利用原始資料 (OW) 所得的精確率 (Accuracy)

篇數	1,000	2,000	3,000	4,000	5,000	20,000
Accuracy	39.43%	41.54%	42.95%	42.95%	43.66%	45.77%

表 13: 利用字根還原 (BF) 所得的精確率 (Accuracy)

篇數	1,000	2,000	3,000	4,000	5,000
Accuracy	40.14%	42.25%	44.36%	45.07%	45.77%

表 14: 利用詞性標記與意元集組處理文章 (BPC) 所得的精確率 (Accuracy)

篇數	1,000	2,000	3,000	4,000	5,000	20,000
Accuracy	43.66%	46.47%	49.29%	47.18%	47.88%	56.33%

五、結論

我們設計一套語言訓練方法，專門處理學術論文，逐篇逐句標上適合的文步標籤，將收集到的 N-連詞進而整理，以幫助學生寫作學術論文。我們所使用的方法，是利用專家提供在特定文步常使用的字詞，藉以透過語料庫進行分析並擷取句子的特徵，產生大量已標註的 N-連詞，將得到的訓練資料，應用到文步分類器。

在未來研究中，我們將擷取辨識度高的文步特徵，提升文步辨識的準確率。例如計算 N-連詞之間的相似度，找出屬於文步的句型，找出特殊單字出現的頻率，增強文步屬性的特徵，希望能在學術論文寫作上提供更好的幫助。同時，並考量文步的順序與位置，來調整貝氏規則。在實驗方面，考慮 N-連詞中，不同的 N 值，也將用更多的訓練資料，並與其他的分類方法（例如：SVM, ME）比較。

References

[1] N. Graetz, "Teaching EFL students to extract structural information from abstracts," in *Readings for Professional Purposes: Methods and Materials in Teaching Languages*, J. M. Kline and A. K. Pugh, Eds., 1985, pp. 225 – 335.

[2] F. Salager-Meyer, "Discoursal flaws in medical english abstracts: A genre analysis per research-and text-type," *Text – Interdisciplinary Journal for the Study of Discourse*, vol. 10, no. 4, pp. 365–384, 1990.

[3] M. Shimbo, T. Yamasaki, and Y. Matsumoto, "Using sectioning information for text retrieval: a case study with the medline abstracts," in *Proceedings of Second International Workshop on Active Mining (AM'03)*, 2003.

[4] American National Standards Institute, *American national standard for writing abstracts*, ser. Z39-14. Bethesda, Maryland, USA: NISO Press, 1997.

[5] J. Swales, *Genre analysis: English in academic and research settings.* Cambridge University Press, 1990.

[6] H. Glasman-Deal, *Science research writing: For non-native speakers of English.* Imperial College Press, 2009.

[7] R. Weissberg and S. Buker, *Writing up research.* Englewood Cliffs, NJ, USA: Prentice Hall, 1990.

[8] P. M. Martín, "A genre analysis of english and spanish research paper abstracts in experimental social sciences," *English for Specific Purposes*, vol. 22, no. 1, pp. 25 – 43, 2003.

表 15: 經訓練增加規則的 $\mathcal{CT}$ 表

篇數	1,000	2,000	3,000	4,000	5,000
OW	60,579	117,533	175,520	233,771	294,589
BF	60,111	116,105	174,136	230,741	289,236
BPC	58,354	111,703	166,751	223,436	279,805

[9] S. Teufel and M. Moens, "Summarizing scientific articles: Experiments with relevance and rhetorical status," *Comput. Linguist.*, vol. 28, no. 4, pp. 409–445, 2002. [Online]. Available: http://dx.doi.org/10.1162/089120102762671936

[10] Z.-H. Ling, Y.-J. Wu, Y.-P. Wang, L. Qin, and R.-H. Wang, "USTC system for Blizzard Challenge 2006: an improved hmm-based speech synthesis method," in *Proceedings of Blizzard Challenge Workshop*, 2006.

[11] J.-C. Wu, Y.-C. Chang, H.-C. Liou, and J. S. Chang, "Computational analysis of move structures in academic abstracts," in *Proceedings of the COLING/ACL on Interactive Presentation Sessions*, ser. COLING-ACL '06. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 41–44.

[12] Y. Yamamoto and T. Takagi, "A sentence classification system for multi biomedical literature summarization," in *Proceedings of 21st International Conference on Data Engineering Workshops*. IEEE, 2005, pp. 1163–1163.

[13] J. Xu, J. Gao, K. Toutanova, and H. Ney, "Bayesian semi-supervised chinese word segmentation for statistical machine translation," in *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 1017–1024. [Online]. Available: http://dl.acm.org/citation.cfm?id=1599081.1599209

[14] 黃冠誠, 吳鑑城, 許湘翎, 顏孜曦, and 張俊盛, "學術論文簡介的自動文步分析與寫作提示," *International Journal of Computational Linguistics Chinese Language Processing*, vol. 19, no. 4, pp. 29 – 46, Dec. 2014.