Computational Linguistics and Chinese Language Processing
Vol. 19, No. 4, December 2014, pp. 29-46

學術論文簡介的自動文步分析與寫作提示

Automatic Move Analysis of Research Articles for Assisting Writing

黃冠誠*、吳鑑城*、許湘翎*、顏孜曦*、張俊盛*

Guan-Cheng Huang, Jian-Cheng Wu, Hsiang-Ling Hsu,

Tzu-Hsi Yen, and Jason S. Chang

摘要

學術論文是一種特殊的文體，有外顯、制式化的結構，如「簡介」、「相關文獻」、「方法」、「結果」、「討論」等。在各節中，又透過所謂文步的隱藏性修辭結構，有條不紊地呈現研究的背景、動機、內容。因此，在學術論文寫作的教學，分析文步扮演了重要的角色。在本論文中，我們提出了一個方法，將所給予的學術論文的每一個句子，標示所隱含的文步（moves），藉以幫助英文非其母語學生，寫作學術論文。我們採取透過常見寫作樣板（common patterns）取得訓練資料的研究路線。而我們的方法涉及擷取常見寫作樣板、標示樣板的文步、產生標示文步的訓練資料、設計分類特徵值、訓練一個文步分類器。在執行時，我們將句子轉換成特徵向量，運用分類器預測句子的文步。我們提出一個雛型系統 WriteAhead，應用分類的句子的資料，提示學習者，如何寫作各種文步的句子。

關鍵詞: 學術英文寫作、電腦輔助語言學習、修辭學、文脈分析

Abstract

Rhetorical moves are a useful framework for analyzing the hidden rhetorical organization in research papers, in teaching academic writing. We propose a

* 國立清華大學資工系
NTHU NLPLAB
Email: {cheng, hsiang, joe}@nlplab.cc; wujc86@gmail.com; jschang@cs.nthu.edu.tw

method for learning to classify the moves of a given set sentences in a academic paper. In our approach, we learn a set of move-specific common patterns, which are characteristic of moves, to help annotate sentences with moves. The method involves using statistical method to find common patterns in a corpus of research papers, assigning the patterns with moves, using patterns to annotate sentences in a corpus, and train a move classifier on the annotated sentences. At run-time, sentences are transformed into feature vectors to predict the given sentences. We present a prototype system, MoveTagger, that applies the method to a corpus of research papers. The proposed method outperforms previous research with a significantly higher accuracy.

Keywords: Academic English Writing, Computer-assisted Language Learning, Rhetoric, Context Analysis

1. 簡介

近年來，英文逐漸變成全世界學術研究最主要的溝通的媒介。而學術英文寫作，也成為非常重要的研究與教學的領域。學者也很重視，如何透過電腦的輔助，幫助一般性的語言學習，甚或特定性的學術寫作。學術寫作包含許多的文章類型，包括學術論文、計畫申請書、回顧與評論文章等（Swales, 1990）。其中，研究論文占有最重要的角色。

在學術論文中，「簡介」是絕大部分論文都有的第一個小節。現今，幾乎沒有學術論文，沒有「摘要」與「簡介」，而直接詳細地描述研究的目的、方法、結果。而且，對寫者和讀者而言，「簡介」在學術論文中都扮演非常重要的角色。一篇好的簡介，要能為整篇論文定調，抓住讀者的興趣，提供論文的扼要資訊。換言之，「簡介」 肩負重大責任——吸引讀者注意，讀完全文。

因此，有一些研究開始分析論文簡介如何達成其溝通的任務。Graetz (1985) 發現論文簡介似乎有共同的「問題—解法」修辭結構，依序包括問題（problem）、方法（solution）、評估(evaluation)、結論（conclusion）等部分。

Swales (1990) 分析大量的論文簡介，歸納出一套修辭的動機與模式：「創造研究空間」（Create A Research Space, CARS）。Swales 認為論文爭取研究得到讀者的認同，有如環境中生物爭取生存空間。為此，大部分作者依循三個修辭的步驟——也就是文步（moves）——來說服讀者。如圖 1 所示，這三個文步包括了「界定研究範圍」、「建立利基」、「佔據利基」。在每一個文步下，又需要描述若干必要或選項的內容。另外，美國國家醫學圖書館，也主張醫學論文作者，應提供分段有標題（labeled sections）的結構化摘要（structured abstract）[1]。

[1]詳見 www.nlm.nih.gov/bsd/policy/structured_abstracts.html

CARS 文步	子文步與資訊內容
文步 I	1. 聲明研究領域的重要性，及/或
界定範圍	2. 聲明研究課題的廣泛性與普及性，及/或
	3. 回顧與評論前人研究
文步 II	1A. 提出與前人不同的聲明，或
建立利基	1B. 指出前人研究的缺口（gap），或
	1C. 提出本論文的研究議題（research question），或
	1D. 說明本研究所根據的典範與傳統
文步 III	1A. 概述本論文的目的，或
佔據利基	1B. 概述本論文的方法
	2. 宣布本論文的主要結果與發現
	3. 指出本論文的結構

圖 1. Swales (1990) 提出的 CARS 模式的文步與資訊內容

目前已經有許多學術寫作教材，透過文步分析來教導英文非母語的學生，如何寫作學術論文（如 Swales & Feak, 2004; Glasman-Deal, 2010）。也有研究者開發軟體系統（例如，Marking Mate: writingtools.xjtlu.edu.cn:8080/mm/markingmate.html），分析學生的作文並自動產生批改的建議與評分。但是很少有系統能夠在學生寫作中，依照文步的推進，適時地提供寫作提示與輔助。直覺上，如果我們能將大量的論文簡介加以處理，自動化分析其中每句的文步，繼而分析特定文步句子的常見片語或句型，我們將可以在寫作的過程，有效地協助學生。

然而，過去所提出的自動文步分析方法，都需費時費工標註大量論文。有鑑於此，我們提出新方法，以降低人工標註的工作量，且標注之資料將運用於訓練統計式分類器，來預測論文簡介中句子的文步，並藉以開發一個線上輔助寫作系統 WriteAhead。在 WriteAhead 的開發過程，我們採用了比 CARS 更簡單的文步分類，如圖 2 所示。用了此一分類方式，除系統較易於自動分類文步外，使用者亦比較容易掌握並使用於寫作過程。

我們期望此一自動文步分析工具，以及 WriteAhead 系統，有助於提升英文非母語者（non-native speakers, NNS）寫作學術論文的能力。在本論文中，我們提出了一套監督式機器學習的方法，能夠自動地學習如何將語料庫內的簡介句子，大略地分類為幾個文步。有了分類的句子之後，我們就可以統計各文步的 N 連詞（ngrams）詞頻。在 WriteAhead 系統，即可參考使用者選擇的文步，以及游標之前的內容，提示單字以及接續片語。

WriteAhead 文步	資訊內容	對應之 CARS 文步
背景（BKG）	領域：重要性、術語定義、缺口	文步 I-1,2,3, 文步 II-1B
	引用與評論前人研究	文步 I-3
本論文（OWN）	目的：輸入、輸出、條件	文步 III-1A, 文步 II-1C
	方法：研究路線、典範、依據、步驟	文步 III-1B, 文步 II-1D
	結果：實作、實驗、評估、結果、發現	文步 III-2
討論（DIS）	比較本論文與前人研究的相同之處 對照本論文與前人研究的相異之處 未來研究方向	文步 II-1A
文本組織（TEX）	提供全文的節大綱（目次表） 提供節內細分子節的大綱 指示圖表（編號） 回顧之前資訊、預告之後資訊	文步 III-3

圖 2. WriteAhead 採用文步與 CARS 模式文步之對照

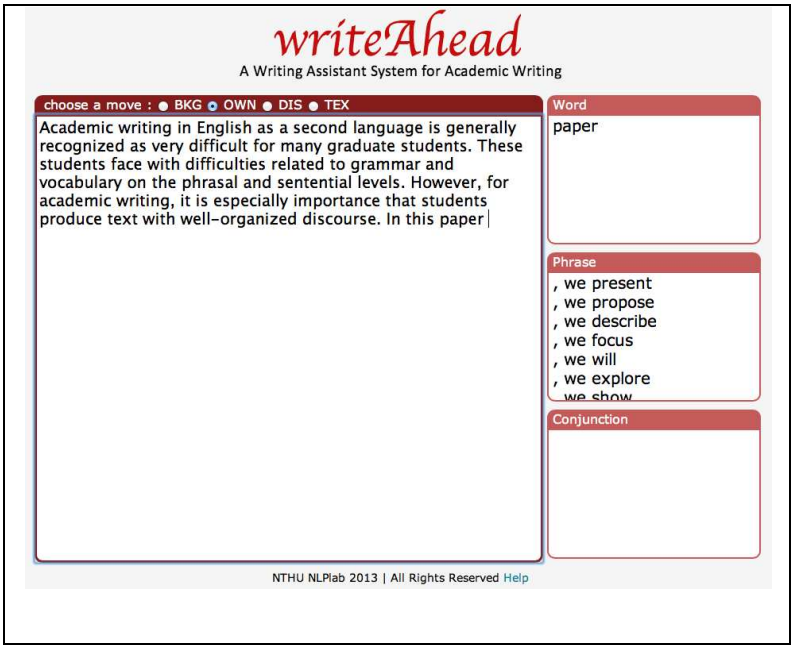

圖 3. WriteAhead 系統操作範例

圖 3 顯示 WriteAhead 系統的操作實例。在圖中，使用者已經介紹了研究背景（BKG 文步），接著使用者選擇了「本論文文步」（OWN），繼而輸入“In this paper” 等字。根據這些資訊，WriteAhead 顯示了適合此一脈絡的提示如下，作為繼續寫作的參考：

, we present	**, we describe**	**, we explore**
, we propose	**, we will**	**, we show**

WriteAhead 能夠提供與排列這些提示，是因為 WriteAhead 透過大量的論文原始資料以及少量的人工標示，學習如何辨識 OWN 文步的句子，並進而統計這些句子內的常見片語及其頻率。我們將在第三節詳述 WriteAhead 所運用的文步分類器的訓練過程。

本論文接下來的部分，安排如下。我們在下一節回顧相關的研究。接著，我們描述如何學習自動將論文簡介句子標註文步（第三節）。我們繼而描述如何將所提出的方法，實際製作成一個考慮文步類別進行寫作提示的雛形系統，以及相關的實驗設定、評估指標、以及實驗結果（第四節）。最後，我們指出未來研究方向，並作結論（第五節)。

2. 相關文獻

學術英文研究與教學（English for Academic Purpose）為相當重要的研究領域。近年來，學者對於研究計劃書，以及學術會議與期刊論文，都有深入的研究（Connor & Mauranen, 1999; Swales & Feak, 2004）。這些研究通常針對論文逐句逐段進行人為分析，經過歸納後，提出一套論文修辭的分析架構。在本研究中，我們則針對學術論文的「簡介」這一個部分，提出一套自動化的結構分析方法，並開發一套能夠讓學生一面寫作，一面獲得寫作提示的電腦輔助寫作系統。我們也討論如何在句子中，擷取能反應修辭結構的特徵，以有助於產生訓練資料，將句子歸類。

許多學者都指出，在表面上以及小節分段上，研究論文大致上有共通的簡單結構——IMRD 結構，即簡介（introduction）、方法（method）、結果（results）、討論（discussion）。也有學者進一步闡述 IMRD 的修辭結構，就像上下寬大，中間狹窄的沙漏：開始時先廣後專（from general to specific），結尾時由專而廣（from specific to general）。Swales (1990) 更為簡介這一個小節，提出了所謂的 CARS 模式（亦即「創造研究的空間」"Create a Research Space"）。CARS 模式歸納了典型的學術論文簡介修辭的動機與模式。CARS 模式提出之後，廣泛地為學者採用作為分析論文「簡介」節的寫作修辭策略 (例如，Cooper, 1985; Hopkins, 1985; Crookes, 1986; Samraj, 2002, 2005)。也有學者沿用 CARS 模式來分析「結果」節（Thompson, 1993），以及「討論」節（如，Hopkins & Dudley-Evans, 1994），以及醫學論文的摘要（Salager-Meyer, 1990, 1991, 1992）。與上述研究不同，我們採用人工督導與機器學習的方式，自動化分類與標註「簡介」節中句子的文步。

在自然語言處理的研究領域，Anthony & Lashkia (2003) 收集了近 700 篇論文摘要，並運用了 CARS 模式，人工標示摘要中每句的文步。之後，再透過機器學習方法，發展出自動文步標示系統 *MOVER*。Anthony 運用 *MOVER* 於學術寫作教學，發現可以幫助學生閱讀、分析、寫作摘要，讓學生有信心地寫出摘要的草稿，突破沒有使用輔助系統時，容易猶豫不決，久久難以下筆的障礙。然而，Anthony 發現 *CARS* 的文步劃分太細，造成 *MOVER* 標示文步的精確度不高。他建議合併相關易混淆文步。如此，可以大幅度提高 *MOVER* 文步分類的正確度，也不至於過於影響 *MOVER* 的效用。我們也將 *CARS* 的 3 大文步共 11 小文步，合併為 4 個文步，以提昇分類正確度，同時也減低使用者的認知負擔。

不同的學術領域的社群有不同文化與溝通的模式。醫學領域的編輯認為摘要應分成有標題的區段，亦即所謂結構化摘要（structural abstracts）。結構化摘要可以讓作者寫出的摘要，資訊完整、流暢易讀（Harley, 2000）。其實這些有標題的一到三句的小段，和文步的觀念是一致的。Shimbo *et al*. (2003) 運用了 MEDLINE 醫學文獻資料庫中標注區段或文步的摘要，開發一套分區檢索的文件資訊檢索系統。該系統運用支撐向量機（Support Vector Machine, SVM），將摘要中的句子劃分為「目的」、「方法」、「結果」」「結論」四種文步。Yamamoto & Takagi (2005) 也開發出類似的 SVM 系統，可將句子分為「背景」在加上以上四類的文步。Hirohata *et al*. (2008) 則是利用 CRF 系列分類器，來標示整個摘要。這些系統通常利用片語、動詞時態、句子位置、前後句特徵，做為分類依據。

近來，學者運用了許多不同的統計式分類的方法，開發文件或文步分類的作法。這些方法包括簡易的貝氏模型（Naïve Bayesian Model, NBM) (Anthony, 2003)，支撐向量機（Support Vector Machines, SVM) (McKnight & Arinivasan, 2003; Shimbo *et al*., 2003; Yamamoto & Takagi, 2005)，隱藏式馬可夫模型（Hidden Markov Model, HMM) (Wu *et al*., 2006; Lin *et al*., 2006)，以及條件式隨機場（Conditional Random Fields, CRFs) (Hirohata *et al*., 2008)。大部分的研究都是針對摘要，只有 Teufel (2000)、Teufel & Moens（2002）這一系列的研究，是針對全篇論文的分析。

和我們最相關的研究，應屬 Teufel（2000）的博士論文研究。Teufel 回顧評論文獻的各種文節分析的架構，並自行提出一套全篇論文的文步分析架構。Teufel （2000）和 Anthony（1993）有相同的意見，主張文步不宜做過度精細的分類，以免人工分類標示時，標示者難以達成共識。她提出兩層式的分類：大分類先分成「背景」、「前人研究」、「本論文」等三個文步。之後，本論文再細分為「目的」、「本論文」、「組織」；而前人研究再細分為「對照」、「立論根據」、「引用前人」，總共七個文步。本研究和 Teufel 的主要區別是，Teufel 利用人工分析，得到一組常見的句型，藉以分析文步，而本研究則透過自動化的語料庫分析得到一組常見的片語與句型。在訓練資料方面，Teufel 依賴對文節的直接標示，而我們透過常見片語與句型，間接標示句子的文步。而我們所採用的分類架構也有所不同，包括了「背景」（含領域、缺口、前人研究），「本論文」（含目的、方法、結果），「討論」（和前人研究的比較與對照），和文節結構（含論文組織、圖表的指示、內容的預告與回顧）等四種文步。

相較於文步分析的文獻中前人的研究，我們提出一套系統，能以較低的人工標示成本，自動學習如何產生訓練資料，進而學習文步的分類。我們並呈現一套實作的系統，其中利用自動化文步標示，輔助英語非母語學生寫作論文的簡介。總體而言，我們利用論文寫作中的常態與常用的表達方式，以及自然語言處理的技術來達到呈現論文修辭現象，並輔助學生寫作的目的。

3. 方法

為了能夠針對學生寫作論文過程中，所想表達的資訊（文步），提供適當的寫作提示，我們需要大量標示文步標籤的句子。人工逐句直接標示文步，無疑地非常費時耗工，絕非最好的作法。比較有潛力省時省力的方法，是先擷取一些論文少量常見句型（例如，“Recently, there have been ...”），透過人工檢視這些句型。決定句型是否大都表達特定的文步（如，背景與重要性）。如果句型有表達特定文步的傾向，就可以保留句型，並標註所屬文步。

最後，再以標示文步的句型來比對句子，產生大量標示文步的句子，以產生統計式文步分類器的訓練資料。

3.1 問題陳述

我們試圖收集大量學術論文，並對其中簡介部分的每個句子都標註修辭文步。之後，我們再利用這些標示資料，開發一個寫作輔助系統。這個系統要能接受學生設定的文步，提供適當的寫作提示。我們觀察學術論文中表達特定修辭文步時，常常用幾個相當特定句型。而一個常見的句型，在一份一萬篇論文的語料庫，出現可達數百次。所以，我們只要能夠擷取與標示這些句型，就可以得到大量的有標示的句子，當做訓練資料，運用於開發出一套文步分類器。運用此分類器，自動標示論文句子文步，我們就可以開發寫作輔助系統。我們現在正式地提出問題陳述。

問題陳述：給定 n 個學術論文「簡介」的句子 $S = s_1, s_2, ..., s_n$，我們的目標將 S 標註上一序列對應的文步標籤 $M = m_1, m_2, ..., m_n$，其中 m_i 為 s_i 的文步類型。為此，我們從 S 計算出常見的 k 個句型 $P = p_1, p_2, ..., p_k$，並人工標註對應文步 $T = t_1, t_2, ..., t_k$，而人工標示句型 p_i 為 t_i 時，必須確認符合 p_i 句型的句子大都表達 t_i 文步的資訊。

在本節的其餘小節，我們將描述我們對此一問題的解決方法。首先，在第 3.2.1 節，我們描述如何從網路收集學術會議與期刊的論文，並擷取其中的「簡介」此一節。接著，我們在第 3.2.2 節描述，如何從簡介中，統計常見的句型，以及人工標記常見句型之文步（第 3.2.3 節），進而產生標示文步之訓練資料（第 3.2.4 節）。最後，我們描述如何在訓練資料上，附加特徵值（第 3.2.5 節），以及訓練統計式機器學習模型（第 3.2.6 節）。

3.2 學習將論文句子標注文步

我們試圖找到一組各種文步的常見句型，藉以產生標示文步句子之訓練資料，以訓練一套文步分類器。我們的訓練過程如圖 4 所示。

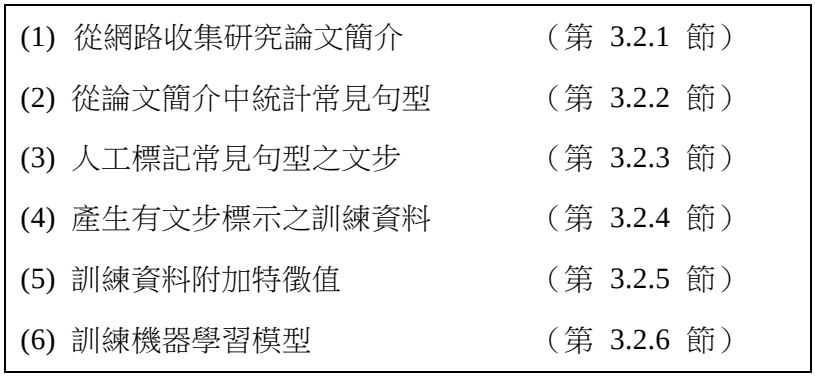

圖 4. 訓練模組的流程

3.2.1 從網路收集學術論文簡介

在訓練過程的第一步，我們收集大量的研究論文，以訓練文步分類器。為此，我們選擇有彙整論文可供直接下載的學會網站，且取得經過 PDF 檔案轉換或光學字元識別（OCR）處理的論文文字檔。然而，通常檔案都未標明節資訊。我們利用簡單規則，大致上辨識出節標題，並擷取論文「簡介」的部份。

3.2.2 擷取簡介常見句型

在訓練的第二步，我們利用現有的句子分割程式，將前一步驟取得的論文簡介，分割成一句一句。然後，再逐句進行切割詞彙（tokenization）、標示詞性（part of speech tagging）與基底片語（base phrases 或 chunks）擷取的預處理作業。

由於專有名詞（如作者名）以及數字（例如年度，或節、圖表編號）變化性大，以及名詞（如 method, approach 等）之前，常有各式的形容詞（如 new, novel）。這些現象都會導致句型發散，不易歸類成常見句型。為了有效歸納常見句型，對於句子內的詞彙，我們做以下的處理：

- 專有名詞、數字詞替換為其詞性標籤（即 NE, CD）
- 名詞片語、動詞片語，去除修飾語的部份，只留下中心語
- 複數名詞替換為單數名詞
- 不同時態的動詞替換為原形動詞

例如，我們會將原始的句子 (1) 替換為 (2) 之後，擷取 N 連詞（ngram）。除了考慮 N 連詞頻率，我們也計算相鄰詞語詞之間的相互資訊（mutual information），篩選所得的常見句型與片語，大都有修辭的功能，而且直覺上對寫作很有幫助的多字詞語（multiword expressions）或短詞串（lexical bundles）。

(1) **Researchers** have successfully **applied** ANN techniques **across** abroad **spectrum of** problem **domains** .

(2) **researcher apply technique across spectrum of domain** .

3.2.3 人工標記常見句型之文步

在訓練的第三步驟，我們挑選一些高頻且文步特性明顯的片語並手動地標記上文步。在此階段，我們將文步分為背景（BKG）、本論文（OWN）、討論（DIS）、文本（TEX）四種類型。 BKG 部分描述領域、課題、缺口、文獻，OWN 部分描述本論文之方法、結果，DIS 部分討論本論文與前人之優劣異同，TEX 部分描述全文或節的目的與組織。 表 1 顯示標了文步的片語範例，以及標籤的簡單定義。所以這個階段的標註對象是處理過後的片語。人工標註的過程中，很難控制標註的品質，因此標註者之間的一致性，需經反覆的核對，調解有衝突的標記 。

表 1. 有文步標記之句型範例

文步	句型	解釋
TEX	in section , we review work	文本：描述全文或節的目的與組織
BKG	research support in part by NE	背景：描述領域、課題、缺口、文獻
DIS	it be important to note that	討論：討論本論文與前人之優劣異同
TEX	rest of paper structure as follow	
OWN	in paper , we propose approach	本文：描述本論文之方法、結果
BKG	follow NE (CD) ,	

3.2.4 產生有文步標示之訓練資料

在訓練的第四步驟，我們利用有標記的句型去匹配大量論文簡介句子，並將句型的文步標註到句子上面。匹配的原則是愈長的句型愈優先。我們利用句型來產生大量有標記文步的句子，用以做為之後模組的訓練資料。表 2 為匹配成功的句子的範例。這個階段的標註範圍是單句。

表 2.句型對應句子的範例

文步	句型	匹配句子
TEX	in section , we review work	In the next section, we will first review some related works.
BKG	in year , there be	In recent years, there has been a rapid growth of interest in the sociological study of childhood.
OWN	in paper , we propose approach	In this paper, we propose a novel unsupervised approach to query segmentation, an important task in Web search.

3.2.5 附加訓練資料之特徵值

在訓練的第五階段，我們要附加特徵值到訓練資料以用來訓練標記文步模型。我們從句子中所抽出 N 連詞特徵值。表 3 為 N 連詞特徵值的例子。為了讓特徵值更能反應文步，我們也加入詞類、語意分類（Word class）的特徵值。我們利用 Teufel（1999）中人工編輯的一組學術論文的分類詞彙。表 4 為我們所使用的 語意分類（Word class）的特徵值。

表 3. 輸入句 "In this paper , we will describe a method …"的N 連詞特徵值

N-gram	Features
Surface unigram	in this paper we will describe a method
Surface bigram	in_this this_paper paper_, ,_we we_will will_describe describe_a a_method
Lemma unigram	in this paper we will describe a method
Lemma bigram	in_this this_paper paper_, ,_we we_will will_describe describe_a a_method
Chunk head unigram	in paper we describe method
Chunk head bigram	in_paper paper_, ,_we we_describe describe_method

表 4. 分類詞類集範例

詞類名稱	詞性	詞彙
AFFECT	v	afford, believe, decide, feel, hope, imagine, regard, trust, think
COMPARISON	v	compare, compete, evaluate, test
TEXT	n	paragraph, section, subsection, chapter

3.2.6 訓練機器學習模型

目前有許多機器學習方法可以處理分類的問題。基本的監督式的方法需要正確的分類資訊，非監督式方法則不需要有正確答案。在本研究中，我們採用監督式訓練方法，但是我們並不直接人工標註正確答案。我們透過標註少量句型，間接地自動產生大量的標記句子，作為監督式機器學習方法所需的訓練資料，並使用最大熵模型（Maximum Entropy, ME）來訓練文步分類器。

訓練完成後，我們就運用此一分類器，將語料庫內所有的論文句子，加以分類，標註上適當的文步。之後，我們就可以運用這些附有文步標籤的句子，來統計各種文步的常見 N 連詞。之後，WriteAhead 系統在輔助寫作時，將參照使用者設定的文步，並根據輸入的內容，查詢適當的片語提供給學習者參考。

4. 實驗與結果

我們設計 WriteAhead 的初衷，是為了提示使用者接著可以寫的數個字詞，以輔助學習者寫作學術論文的「簡介」。因此，我們擷取經過審查、編輯的程序，發表的學術論文，來實作我們提出的方法，以及開發寫作輔助系統。本節中，我們描述模組訓練的實驗設定（第 4.1 節），以及初步實驗的效能評估與結果（第 4.2 節）。

4.1 實驗設定

我們從密西根大學的計算語言學及資訊檢索組（Computational Linguistics And Information Retrieval Group, CLAIR) 設計維護的計算語言學會（Association for Computational Linguistic）會議與期刊論文典藏網站 *ACL Anthology Network*（AAN, clair.eecs.umich.edu/aan），我們擷取 AAN 學會的會議與期刊論文，共四萬多篇的論文的文字檔案。 這些檔案主要是由 PDF 格式的檔案，透過轉檔（類似於 OCR 辨識）所得到的文字檔案。因此，這些檔案有著各式的雜訊，像是殘留的換行連字符號、單字辨識錯誤等。 我們透過設計及分析規則，設定簡單的條件，辨識出節的標題， 並挑選了標示很清楚的論文將近一萬篇。之後，我們根據標題的編號，標題的內容，抽取「簡介」部分來做為研究的訓練資料，以及系統開發的資料。

表 5. 有匹配句型之句子文步分布情形

文步	句數
BKG	3,333
OWN	7,199
DIS	1,572
TEX	5,687
總計	17,791

我們逐篇處理這一萬篇論文簡介。我們利用 Python/NLTK[2] 的分割英文句子、詞彙的工具，將一篇篇論文分割成句子，再將句子分割成詞彙與標點（tokens）。有了句子與詞彙後，我們接著使用 Genia Tagger[3] 標註詞性與基底片語（base phrase 或 chunks）。之後，當所有的緒論單字都被斷詞和標記詞性以及區塊後，我們利用統計方法獲得若干的句型。我們人工的挑選了五百個句型後，手動濾掉文步特性不明顯得的片語並把剩下的句型都標上文步，剩下近約四百個有文步標記的句型。我們在利用這些標記過的句型去匹配一萬篇的論文簡介。我們得到大約一萬八千個句子，其文步的分佈如表 5 所示。再將標記好的句子附加上特徵值 N-gram、詞語分類後，讓 ME 模組做訓練，獲得文步標註模組。

[2] http://www.nltk.org

[3] http://www.nactem.ac.uk/GENIA/tagger/

我們藉由訓練所得的文步標註模組，對一萬篇簡介中的每一句進行文步標註。最後我們統計各種文步中的 N 連詞資訊，我們繼而將一萬多篇簡介內的句子，逐句做文步的分類，運用於 WriteAhead 寫作輔助系統。

4.2 評估與討論

如前所述，WriteAhead 的設計目標是輔助學習者寫作學術論文的「簡介」，所以應該評估各種寫作情境下，使用者覺得 WriteAhead 的提示，是否有助於寫作出更好的「簡介」。然而，一般而言，凡是涉及使用者的評估都是非常困難。退而求其次，我們目前僅針對文步分類器部分，評估其分類正確性。由於論文的文步是依序推移，所以我們針對「簡介」 的整個節，來評估文步的標註是否正確。

表 6. 總共 50 篇簡介之句子標示文步與預測文步與預測正確率

文步	標示句數	預測句數	正確句數	精確率
BKG	621	470	402	.86
OWN	238	259	144	.56
DIS	312	461	241	.52
TEX	117	98	75	.76
總計	1,288	1,288	862	.67

為了達成能自動的為論文簡介句子標註文步此一目標，我們從 *ACL Anthology Network* 中隨機挑選五十篇論文簡介的句子，做為我們文步標註模組的評估資料。表 6 顯示評估的結果。整體的文步預測正確率 67%，還有改善的空間。就個別的文步來看，背景文步 （ BKG）的正確率達 86% 而文脈文步（TEX）達 76%，這可能是因為背景、文脈文步兩者都有比較固定的表達方式。相對的，本論文（OWN）、討論（ DIS）兩種文步的精確率僅僅略高於 50%，這當然是因為表達的方式比較分歧，不易透過常見句型來加以掌握，未來可能還需要發掘比較有效的特徵值。

個別句子的分類正確率並不高，這可能歸咎於幾個原因。首先，標註資料太少，而且標註的正確性也不是非常理想。另外，表達同一類的文步，用字遣詞的差異性很大，很難用有限的資料來掌握，相反地字詞也有不小的詞彙語意歧義。

雖然個別句子的分類正確性不理想，我們觀察統計後的各分類之高頻 N 連詞還算合理。受限於時間，我們尚未評估 WriteAhead 運用各分類高頻 N 連詞，對於提示使用者的效果。不過我們認為，高頻 N 連詞的精確率可能遠高於文步標示的精確率。

本論文所使用的分類器是 Maximal Entropy ，未來也將考慮採用 SVM 或是 CRFs 。本論文所提出的方法，是基於跨領域的論文修辭研究，應該不會受不同學術專業領域特殊性的影響。但是，個別領域表達的方式在用字遣詞仍然有不小的差異，受限於資料，本系統應該對非資訊領域（例如文學、管理學、教育學）的適用性應該不是很理想，需要另外蒐集資料，依照學科建置不同的系統。

5. 結論

對於如何改善我們所提出的系統，我們預見許多可能的未來研究方向。例如，可以運用既有的自然語言處理技術，擷取更具效果的特徵值，來提升文步分類的正確率。例如，我們可以自動產生寫作文體之分類詞彙群。並且，根據分類詞彙群，擷取詞群式的常見樣板（class-based patterns），用來幫助分類的正確性，以及提供富含資訊的寫作提示。另外一個有潛力的研究方向，是讓使用者在另一個文字框，輸入母語（如中文、日文）草稿，而系統參考這些母語草稿，來調整提示的英文句型與片語。另外，我們也可以讓使用者選取部分沒有把握的 2-5 個字，系統提示正確或錯誤的機率，以及其他可以替換的表達方式。

總而言之，我們介紹了一套方法，能處理所搜集到的學術論文，將每一個句子標示上適當的文步（move），並統計各類文步的常見片語，藉以幫助英文非其母語學生，寫作學術論文。 我們的方法涉及擷取常見寫作句型、標示句型的文步、產生大量已標示文步的句子以及特徵值，作為訓練資料來開發文步分類器。我們藉由此一分類器，預測句子的文步。我們提出一個雛型系統 WriteAhead，應用分類的句子與常見片語的資料，提示學習者，如何寫作各種文步的句子。

致謝詞

本研究承蒙科技部補助研究經費，計畫標號 NSC 100-2511-S-007 -005 -MY3。

參考文獻

Anthony, L., & Lashkia, G. V. (2003). Mover: A machine learning tool to assist in the reading and writing of technical papers. *IEEE Trans. Prof. Commun.*, *46*, 185-193.

Connor, U., & Mauranen, A. (1999). *Linguistic Analysis of Grant Proposals: European Union Research Grants*.

Della Pietra, S., Della Pietra, V., Lafferty, J., Technol, R., & Brook, S. (1997). Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 19*(4), 380-393.

Edmundson, H. P. (1969). New Methods in Automatic Extracting. *Journal of the Association for Computing: Machinery, 16*(2), 264-285.

Graetz, N. (1985). Teaching EFL students to extract structural information from abstracts. In Jan M. Ulijn and Anthony K. Pugh, editors, R*eading for Professional Purposed: Methods and Materials in Teaching Languages*, pages 123-135. Acco, Leuven, Belgium.

Hirohata, K., Okazaki, N., Ananiadou, S., Ishizuka, M., & Biocentre, M. I. (2008). *Identifying Sections in Scientific Abstracts using Conditional Random Fields*.

Lin, J., Karakos, D., Demner-Fushman, D., & Khudanpur, S. (2006). Generative Content Models for Structural Analysis of Medical Abstracts. In *Proceedings of th HLT/NAACL 2006 Workshop on Biomedical Natural Language Processing (BioNLP'06)*, 65-72.

McKnight, L., & Srinivasan, P. (2003). Categorization of sentence types in medical abstracts. In *AMIA Annual Symposium Proceedings* (Vol. 2003, p. 440). American Medical Informatics Association.

Ruch, P., Boyer, C., Chichester, C., Tbahriti, I., Geissbühler, A., Fabry, P., ... & Veuthey, A. L. (2007). Using argumentation to extract key sentences from biomedical bstracts. *International journal of medical informatics*, *76*(2), 195-200.

Shimbo, M., Yamasaki, T., & Matsumoto, Y. (2003). *Using sectioning information for text retrieval: a case study with the MEDLINE abstracts.*

Swales, J.M. (1990). Genre analysis: English in Academic and Research Settings. *Cambridge University Press.*

Teufel, S. (1999). Argumentative Zoning: Information Extraction from Scientific Text. *PhD thesis, University of Edinburgh*.

Teufel, S., & Moens, M. (2002). Summarizing Scientific Articles: Experiments with Relevance and Rhetorical Status. *Computational Linguistics*, *28*(4), 409-445.

Wu, J. C., Chang, Y. C., Liou, H. C., & Chang, J. S. (2006). *Computational analysis of move structures in academic abstracts*.

Yamamoto, Y., & Takagi, T. (2005). A sentence classification system for multi-document summarization in the biomedical domain. In *Proceedings of International Workshop on Biomedical Data Engineering*, 90-95.

附錄 A　整合常見句型的寫作樣板

我們擷取常見句型標示文步之後，發現許多句型很類似，只有少數的幾個字變動。我們可將這些句型聚集起來，歸納整合成為正規式樣板（regular expression patterns）。這些樣板避免羅列許多句型的不便，一目了然——既代表了寫作的常態，也呈現了各種變化。運用在教學上讓學生學習很有效果，寫作時也容易加以模仿、改寫。

例如，從附錄 B 中我們可以看到下面左邊這些和時間有關的句型。經過觀察與歸納，我們可以得到下面右邊的樣板及其變化型：

recently , al (CD)
currently , there be
at present ,
over year ,
over decade ,
recently , method
in year ,
there be work
to date ,
in decade ,
currently ,
traditionally ,
recently ,

```
7. In RECENT YEARS, ______ .
   |  |      |      |
   |  |      |      |
   |  |      |
   |  |      | years, decades ---------
   |  |                               |
   |  | recent, the past, the next ----| recently, nowadays, to date,
   |                                  | traditionally
   | in, over, during -----------------
```

變化句型

```
 8. ______  ______  ______  IN RECENT YEARS .

 9. RECENT YEARS have witness ______  ______  ______ .

10. ______  ______  ______ IN RECENT YEARS .

11. RECENT YEARS have witnessed ______  ______  ______ .

12. PREV-WORK has VERBed

13. it has been VERBed that ______  ______  ______ .
                  |
               known, observed, recognized, shown
```

附錄 B 各種文步的常見句型

B.1 背景文步

follow NE (CD) ,
NE (CD) show that
NE (CD) demonstrate that
NE (CD) propose model
it be , however ,
there be , however ,
to knowledge , there be
to good of knowledge ,
in case , however ,
NE (CD) present
NE (CD) describe
however , in case ,
to knowledge , this be
collection comprise CD
in practice , however ,
recognition (NE) be
NE (CD) propose
as matter of fact ,
on hand , approach
currently , there be
this , however ,
first of all ,
however , for language
approach , however ,
research support by NE
however , there be
however , while
study show that
difficulty be that
currently , system
there be also
most of method
challenge be that
recently , model
however , they
at present ,
in general ,
it know that
as alternative ,
over year ,
this be important
much of work
over decade ,
however , if
however , unlike
recently , method
in year ,
it observe that
they show that
there be work
however , when
to date ,
most of system
to knowledge ,
this be task
it recognize that
however , since
in decade ,
however , study
however , approach
unfortunately ,
difficulty be
problem with
challenge be
they describe
currently ,
traditionally ,
in year
while approach
unlike method
recently ,
recently

B.2 「本論文」文步

in paper , we propose approach
in work , we focus on
in paper , we report on
in paper , we show that
in paper , we present approach
in paper , we present mothed
in paper , we present system
in particular , we show that
in paper , we focus on
in study , we focus on
in paper , we show how
in paper , we describe system
in paper , we propose
focus of paper be on
goal of research be to
aim of paper be to
in paper , we explore
in paper , we introduce
in paper , we use
purpose of paper be to
in paper , we consider
in paper , we describe
in paper , we address
in work we focus on
in work , we use
in paper , we study
in paper , we propose
goal of work be to
in paper , we investigate
goal of paper be to
goal in paper be to
in paper we describe
in study , we
paper address problem of
result show that model
in work , we
result show that method
to address problem ,
result show that approach
in paper we present
focus of paper be
we propose that
in study ,
paper focus on
we demonstrate that
in paper we
paper describe system
purpose be to
therefore , we
solution be to
idea be to
we start with
we hypothesize that
aim be to
in paper ,
we argue that
hypothesis be that
goal be
motivation for
in study
in paper
in work
paper present
purpose of
focus be
aim of
paper describe
we demonstrate
paper provide
we evaluate

method
in paper , we argue that
in paper , we propose model
in paper we focus on
in paper , we present
in paper we show that
in paper we describe system
work present in paper
in paper , we
we also show that
paper propose method for
in paper we discuss
in paper we investigate
in paper we propose
to achieve goal ,
in paper , i
thus , method
finally , result
experiment show that
work focus on
goal be to
claim be that
result indicate that
therefore , method
in work ,
evaluation show that
result show that
we evaluate approach
we show that

B.3「討論」文步

it be important to note that
this be due to fact that
contribution of paper be as follow
however , we believe that
advantage of approach be that
contribution of work be :
in order to do this
view express endorse by sponsor
as it turn out ,
reason for this be that
it be worth note that
contribution of paper be :
to overcome problem ,
for example , name
in particular , it
in contrast , model
it be obvious that
it turn out that
contribution of paper be
reason for this be
to knowledge , work
we also show how
in contrast , system
first , it
as result of
contribution be :
by contrast ,
in comparison ,
for reason ,
in practice ,
reason be that
specifically , it
this be problem
this lead to
as consequence ,
that be why
intuition be that
analysis show that
this mean that
we believe that
in principle ,
on contrary ,
example show that
difference be that
in short ,
we then discuss
unlike NE ,
it note that
among them ,
in sum ,
this be because
we note that
this suggest that
contribution be
advantage of
observation be
we believe
although approac

B.4「組織」文步

in section , we review work
remainder of paper organize as follow
in CD , we describe model
rest of paper structure as follow
in CD , we present model
remainder of paper structure as follow
rest of paper organise as follow
part of paper organize as follow
in CD , we present approach
we discuss result in CD
in CD we describe how
paper structure as follow :
in remainder of paper ,
in CD we discuss work
we discuss work in CD
next , in CD ,
in CD we present experiment
finally , CD conclude paper
section present and discuss result
CD present result of experiment
finally , we draw conclusion
in CD we present
in CD we discuss
in what follow ,
result show in CD
finally , CD present
article organize as follow
finally CD conclude paper
in section CD ,
paper organise as follow
in rest of paper
finally , in CD
work discuss in CD
discussion present in CD
paper organize as follow
in section that
we then present
CD describe model
CD present method
CD describe result
CD discuss result
CD review work
CD describe method
CD show result
plan of paper
finally , we
CD describe system
CD present result
CD present work

remainder of paper organise as follow
in CD , we describe system
rest of paper organize as follow
outline of paper be as follow
paper organize as follow : CD
structure of paper be as follow
in CD , we describe method
paper organize as follow : in
finally , we conclude in CD
in CD , we describe corpus
in CD , we review work
organization of paper be as follow
finally , CD present conclusion
finally , in CD ,
in CD we present result
for example , CD show
in section of paper ,
paper organize as follow :
in CD we show that
we conclude paper in CD
in rest of paper ,
finally , we present result
in section , we describe
CD show example of
finally , CD conclude
as we see ,
CD give overview of
result report in CD
result present in CD
as we show ,
paper proceed as follow
we conclude in CD
result discuss in CD
approach describe in CD
in CD we introduce
paper structure as follow
in CD we describe
conclusion draw in CD
result give in CD
after that ,
CD present evaluation
structure of paper
CD conclude paper
CD report result
CD describe algorithm
CD present algorithm
CD present experiment
CD introduce model
CD introduce method
CD present model
CD show example
CD describe how
CD describe experiment
CD describe setup
in section ,
CD show how
CD describe work
CD describe approach
CD give result
CD discuss work
in section
CD describe
CD introduce
CD conclude
CD show
CD detail
CD explain
CD present
CD discuss