漢語歷時語料庫中官爵名的識別與提取*

熊丹[1]，徐建[1]，陸勤[1]，羅鳳珠[2]

[1]香港理工大學電子計算學系

{csdxiong, csjxu, csluqin}@comp.polyu.edu.hk

[2]台灣元智大學中國語文學系

gefjulo@saturn.yzu.edu.tw

摘　要: 官爵名以不同的形式出現在歷代各種體裁的文字記錄中，並富含歷史內涵和現實意義。本研究將官爵名視作一種重要的命名實體，使用人工完成命名實體標註的四部明清小說作為訓練語料，通過機器學習算法自動識別明清之前和之後不同時代、不同體裁文本中的官爵名，包括先秦典籍《左傳》和《國語》、唐詩、及現代新聞《人民日報》。實驗結果顯示，經過訓練的系統識別先秦典籍和唐詩中官爵名的效果較好，相對而言，對現代新聞的識別準確率則有待改進，但對其他使用計算機處理歷時語料的工作還是具有一定的參考價值。本研究的結果不僅能夠幫助改進計算機對歷時語料的處理，還有助於促進歷代官爵名總體概貌及其發展脈絡的研究。

關鍵詞: 命名實體識別，官爵名，歷時語料庫

Recognition and Extraction of Titles in Chinese Diachronic Corpora

Dan Xiong[1], Jian Xu[1], Qin Lu[1], Fengju Lo[2]

[1]Department of Computing, The Hong Kong Polytechnic University

{csdxiong, csjxu, csluqin}@comp.polyu.edu.hk

[2]Department of Chinese Linguistics & Literature, Yuan Ze University, Taiwan

gefjulo@saturn.yzu.edu.tw

Abstract: Titles in this paper refer to the names of official position (官職名) and titles of nobility or honor (爵位封號). They can be found in various kinds of written records in different periods and have great historical significance. This paper introduced a machine learning system to recognize the titles in diachronic corpora using annotated data. A tagged corpus of four classical novels produced in the Ming and Qing dynasties is used to train the system. The system is then used to automatically recognize and extract the titles in pre-Qin classics, Tang-dynasty poems, and modern Chinese news. Experimental results show that the system can achieve relatively good results in recognizing the titles in the pre-Qin classics and Tang-dynasty poems. This attempt to extract the titles in diachronic corpora is able to provide reference to other researches related to recognition of named entities in texts of different styles. This work is a first attempt to improve the performance of automatic title recognition in diachronic corpora. It should be a helpful tool for the studies on the evolution of titles throughout the Chinese history.

Keywords: named entity recognition, titles, diachronic Chinese corpus

1　前言

官爵名是名號稱謂的重要內容，散落在歷代不同形式的文字記錄中，並承載著重要的歷史信息。因此，在自然語言處理及其應用中，有效識別和提取官爵名具有重要意義。從古至今，中國歷代官爵名隨著政治的變遷、經濟文化的發展而不斷變化，但在不同的時代之間也存在一定程度的沿襲. 本研究嘗試從資訊處理的角度對不同時代、不同體裁的歷時語料庫中的官爵名進行研究，首先使用已完成分詞和命名實體標註的四部明清章回小說[1]作為訓練語料，然後通過機器學習算法自動識別早期時代的先秦典籍（《左傳》和《國語》）、唐代詩歌、以及現代紀實性新聞《人民日報》，測試系統識別和提取不同時代、不同體裁文本中官爵名的性能。本研究的成果可對其他使用計算機處理歷時語料的工作提供參考，實驗結果還可用於探索官爵名在不同時代、不同體裁的文字資料中的沿襲和演變。

*本文承九十八年度蔣經國國際學術交流基金會“歷代語言知識庫建置計畫”（RG013-D-09）部分資助。

2 研究概況

無論在任何體裁的文本中，官爵名都傳遞了重要的信息、富含歷史內涵及現實意義。因此，從自然語言處理的角度出發，有必要將官爵名作為一種命名實體來研究。然而，歷代官爵系統的複雜性和官爵名在文本中使用的多樣性增加了識別和提取這類名稱的難度。首先，歷代官制系統十分繁複，隨著政權的更替和時代的變遷，官爵名數量變得非常龐大，當代影響較大的一些歷代官制辭典就收錄了多至兩萬多條官爵名[2-5]。其次，和其他詞彙一樣，官爵名也隨著歷史的發展而消長，有些名稱雖使用至今，但詞義發生了變化，例如“博士”，在古代用作官職名，而現代詞義改變較大。另外，綜觀各時代的文本，無論是紀實性的史籍還是創作性的文學作品，都不僅涉及當前時代的官爵名，往往還會引經據典、引用更早期時代的名稱，例如，即使現代紀實新聞《人民日報》也曾引用“宰相”。因此，在使用計算機處理自然語言時，無論是哪一時代、哪種體裁的文本，對官爵名的有效識別都有助於計算機對語言的正確處理和理解。

目前的中文信息處理中，官爵名未得到足夠的重視，基於歷代語料庫對官爵名的研究更是鮮見。國內通用的現代漢語語料庫加工規範[6]將官爵名作為普通名詞（/n）處理，如“省長/n”、“財政部/nt 部長/n”。台灣中央研究院建立的近代漢語標記語料庫[7-8]中，將官爵名標註為普通名詞（Na），例如“李(Nb)御史(Na)”；其建立的上古漢語標記語料庫[7, 9]中，將官爵名歸於有生名詞（NA1），例如“內史(NA1)[+attr]過(NB1)[+prop]”。

我們在對歷代語料進行研究和標註的過程中發現，如果簡單地將官爵名處理為普通名詞，將會丟失一些重要的信息。官爵名可單獨使用，例如“御史”；也可和人名連用，如“李御史”；還可以和稱呼連用、以示尊重，如“御史大人”。無論以哪種形式出現，在文本中都承擔命名實體的功能，並常常用於指代。本工作將官爵名視作一類重要的命名實體，使用已完成分詞和命名實體標註的明清章回小說作為訓練語料。在官爵名自動識別方面，需要確定官爵名的邊界，為此本研究使用了序列標註算法[10]，加入了上下文的文本特徵及命名實體的類別標籤，然後通過實驗測試系統識別不同時期、不同體裁文本中官爵名的有效性。

3 官爵名的分類和標註

鑒於官爵名在各類文本中的重要意義，本研究將其視作一類重要的命名實體，標註為“/nu”，並基於對歷代語料庫的分析，將官爵名分為以下兩類：

- 官職名：包括政府、軍事機構中經過任命的官銜名，不包括“教授”、“主管”等頭銜名，此類頭銜名在對稱謂的研究中另行描述[11]，此處不再贅述。官職名標註為“/nu1”，如“尚書/nu1”。
- 爵位封號：包括由帝王根據血緣親疏、功勞等授予的爵銜、尊號，含對皇室和宗室女子的封號。爵位封號標註為“/nu2”，如“郡王/nu2”、“貴妃/nu2”。

如果官爵名和機構名、地名等連用，爲了保留這些信息的關聯，使用“[ ]”將整個複合名稱總括，其內部的機構名、地名加相應標識符，例如“[戶部/nt 尚書]/nu1”（戶部為機構名）、“[臨淄/ns2# 侯]/nu2”（臨淄為地名）。但在系統訓練時，只將“尚書”、“侯”作為官爵名，“戶部”、“臨淄”作為其上下文特徵。

由於本工作是從自然語言處理的角度對命名實體進行研究，因此研究對象僅包括具有特指意义的官爵名，即用於指稱且根據上下文能判斷其指稱對象的名稱，如“司徒/nu1 王允/nr3#”，此處的官職名“司徒”特指“王允”。泛指的官爵名視作普通名詞，如“出/了/一/個/郎中/缺”中的“郎中”雖然是官職名，但並不用於指稱，因此不標註為官爵名，作為普通名詞。

4 算法設計

白話語體文的明清章回小說是文言文向現代文過渡時期的產物，因此，本研究將明清時期的小說《三國演義》、《水滸傳》、《金瓶梅》和《紅樓夢》用作訓練語料，該小說已完成分詞和人名、稱謂、官爵名、地名、建築名、機構名等命名實體標註[1]，未進行詞性標註。在系統訓練後，使用序列標註算法自動識別先秦史籍《左傳》和《國語》、唐詩、現代新聞《人民日報》中的官爵名，測試系統識別不同時代、不同體裁的文字資料中官爵名的性能。

本研究將官爵名的自動提取當做是序列標註的問題，即確定官爵名的開始和結束位置，為此

我們採用了條件隨機場模型（CRFs）[10]。該模型定義為：將長度為 T 的輸入語句定義為 $\boldsymbol{x} = x_1, x_2, \ldots, x_T$，$\boldsymbol{y}$是與$\boldsymbol{x}$對應的長度為$T$的標註序列，其中第$j$個字符的標籤為$y_j$，$1 \le j \le T$，並且$y_j \in${B-GZ, I-GZ, B-JF, I-JF, O}。其中 B-GZ 代表官職名的開始，I-GZ 表示官職名的延續，O 表示該字符不屬於官職和爵位的部分。由於 CRFs 模型中能夠加入不同類型的特徵，本工作使用了 6 類特徵，在第 2 類特徵中，字符的實體類別標籤包括 nr（人名）、 na（稱謂）、ns（地名）、nt（機構名）、nv（建築名）和 O（不屬於任何實體類別）。在自動提取官爵名的過程中，CRFs 模型能夠有效利用字符特徵及上下文特徵，以“布視之，乃司徒王允也”為例，特徵設計樣例如表 1 所示。

表 1　特徵設計樣例

	特徵類別	解釋	樣例
1	C_n (n $= -2, -1, 0, 1, 2$)	當前字符以及前、後兩個字符	C_{-2} =乃　C_1 =王 C_{-1} =司　C_2 = 允 C_0 = 徒
2	T_n (n $= -2, -1, 1, 2$)	前、後兩個字符的命名實體類別	T_{-2} =O　T_1 =nr3 T_{-1} =O　T_2 = nr3
3	$C_n\, C_{n+1}$ (n $= -2, -1, 0, 1$)	bigram	$C_{-2}\, C_{-1}$ = 乃司　$C_0\, C_1$ = 徒王 $C_{-1}\, C_0$ = 司徒　$C_1\, C_2$ = 王允
4	$C_{-1}\, C_1$	前一個字符和後一個字符的組合	$C_{-1}\, C_1$ = 司王
5	$Punc(C_0)$	當前字符是否為標點符號	$Punc(C_0 = \text{徒}) = False$
6	$Num(C_0)$	當前字符是否為數字	$Number\ (C_0 = \text{徒}) =$ False

在第 1 類和第 2 類特徵中，本文設計的窗口大小為 2，此外還測試了窗口大小 n 為 3 和 4 的情況。由於 CRFs 算法使用了大量的特徵，因此我們需要估計的參數數目也相當龐大。爲了解決這一問題，採用了 SampleRank 算法[12]。之前使用 Gibbs 採樣方法估計參數，在每一輪算法迭代中，所有參數都要更新一次，但是 SampleRank 算法通過對比模型估計和實際估計，如果兩者對聯繫樣本產生不一致的估算結果時開始更新參數，這降低了參數更新的次數，並且使得學習到的參數更能代表樣本的實際分佈。

5　實驗結果及數據分析

5.1 訓練語料和測試語料信息

本工作所用的訓練語料為四部明清小說，而測試語料則囊括三個不同時代、不同體裁的語料集，包括先秦文言文、唐代韻文、及代表現代新聞的《人民日報》。由於官爵名的上下文中常出現人名、地名等命名實體，爲了有效利用這些特征，在使用系統自動識別前，參照明清小說命名實體標註規範[1]對先秦典籍和唐詩中的人名、稱謂、地名、機構名、建築名進行了標註。表 2 顯示了這些語料集的詳細規模，表 3 則列出了訓練語料和測試語料中的官爵名數量。

表 2　訓練語料和測試語料信息

訓練語料				測試語料			
時代	語料集	文本類型	規模（萬字）	時代	語料集	文本類型	詳細信息
明清	三國演義	章回小說	48.25	先秦	左傳	史籍	全文，19.88 萬字
	水滸傳		75.92		國語		全文，7.04 萬字
	金瓶梅		65.24	唐朝	詩歌	古詩	李白、杜甫、韓愈含官爵封尊謚名之 126 首詩之 142 句
	紅樓夢		72.95	現代	人民日報	新聞	1998 年 1 月全文，160.58 萬字
說明：以上字數均不含標點符號。							

作爲訓練語料的四部明清小說包含了大量的官場描述，涵蓋的官爵名既包括歷史真實名稱，也包括小說虛構的名稱，且使用靈活多樣，因此能幫助提高系統性能。眾所周知，詩歌與散文、小說等語體文在結構和句法上都有很大差異，爲了驗證實驗效果，我們選取了 126 首唐詩中包含官爵封尊謚名的 142 句用於測試。本文採用的現代新聞測試語料為北京大學現有的《人民日報》標註語料庫[6]中的 1998 年 1 月全文，該語料庫將官爵名作為普通名詞處理，因此無法獲取其中的官爵名數據。

表 3　訓練語料和測試語料中的官爵名數量

訓練語料						測試語料					
時代	語料集	官職名		爵位封號		時代	語料集	官職名		爵位封號	
		類數[1]	次數[2]	類數	次數			類數	次數	類數	次數
明清	三國演義	198	2535	39	848	先秦	左傳	119	844	12	2433
	水滸傳	124	3045	27	121		國語	49	166	7	165
	金瓶梅	132	1671	13	36	唐朝	詩歌	36	67	20	55
	紅樓夢	53	202	59	741	現代	人民日報	-	-	-	-

[1] 類數：對語料集中的官爵名去除重複，即相同的官爵名只計一次，體現其類型。

[2] 次數：對語料集中的官爵名不去除重複，即對每個官爵名均進行統計，體現其頻率。

5.2 實驗結果

在系統評價方面，使用了準確率（P）、召回率（R）以及 F 值（F）來衡量官爵名自動提取系統的性能。在模型構建以及參數學習方面，本研究使用了 Factorie 工具[13]訓練 CRFs 模型，迭代次數設置為 20。由於上下文特徵會影響官爵名的識別性能，本研究通過改變特徵窗口的大小驗證其對官爵名識別的效果、以及對不同時代文本和不同文體的影響。特徵窗口設置為 2（即當前字符及前後兩個字符）、3、4，分別對各測試語料進行了實驗。爲了檢測算法的功效，我們還列出了訓練語料和測試語料中人工標註的官爵名（即標準答案）的重合率，其計算公式為 $\frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$,其中的 S_1和S_2分別為兩個對比語料中官爵名的集合。代表現代新聞的測試語料《人民日報》規模較大，該語料庫中將官爵名作為普通名詞處理，現階段由於時間和人力的限制，無法對其中的全部官爵名進行標註，因此也就沒有官爵名的重合率和召回率，僅能對系統自動識別的名稱進行人工檢查，計算出其準確率。表 4 展示了總體實驗結果，P、R 和 F 取 F 值達到最高值時的數據。

表 4　總體實驗結果

測試語料		官職				爵位封號			
時代	文本	P	R	F	重合率	P	R	F	重合率
先秦	左傳	68.44%	22.87%	34.28 (窗口 4)	2.16%	79.77%	34.69%	48.35 (窗口 2)	8.18%
	國語	54.05%	24.10%	33.33 (窗口 3)	1.81%	46.11%	50.30%	48.12 (窗口 2)	6.54%
唐朝	唐詩	94.44%	25.37%	40.00 (窗口 2)	5.30%	42.86%	21.82%	28.92 (窗口 2)	5.83%
現代	人民日報	21.89%	-	-	-	6.52%	-	-	-

表 4 的結果顯示，總體上，在訓練語料和測試語料官爵名重合率非常低的情況下，系統對先秦史籍和唐詩中官爵名的自動識別效果可謂相當好。其中一個原因是，先秦史籍和唐詩中的官爵名均為歷史上的真實名稱，不似虛構名稱靈活多變，更有利於系統將其自動識別出來，例如《左傳》和《國語》中準確率最高的官職名均為“司空”，唐詩中準確率最高的為“太守”，這些都是歷史上常見的真實名稱。

以明清小說作為訓練語料來自動識別現代漢語中的官爵名效果則較差，僅具有一定的參

考價值。這主要因為現代官制體系發生了巨大變化，封建社會官制系統下設置的很多官職名已經消亡，且在很多國家已沒有封爵制。即使有些詞彙仍使用至今，其詞義也發生了明顯變化，不再用作官職名，如“博士”、“大夫”。即便如此，系統仍然能夠自動識別現代漢語中的一些官爵名，主要包括：1）沿襲至今且詞義未發生大變化的官爵名，這類名稱數量較少，例如“參謀”、“王后”（當代一些國家仍存在）；2）詞彙仍用作官職名、但詞義已發生變化，系統仍然能自動識別，例如《三國演義》中有“書記”一職，《人民日報》中多次出現“（省委）書記”，雖兩者意義不同，但系統仍能將它們識別出來；3）現代漢語中引用典故，如《人民日報》中引用的“中郎將太史慈”。

由於本研究的對象為命名實體，因此，在用作訓練語料的明清小說中，將不用於指代的官爵名視作普通名詞，未加命名實體標註。由於訓練上下文未做詞性標註，沒有相關特徵信息，因此在測試時，系統較難做到將用於指代的官爵名和普通名詞區別開，這在很大程度上影響了其性能。例如“司徒具徒，司空視途”，此處的“司徒”、“司空”都是官職名但不具指稱功能，而系統進行了標註，從而影響了總體測試結果。

圖 1.1 至圖 1.3 分別顯示了將特徵窗口設置為 2、3、4 之後，《左傳》、《國語》和唐詩的測試結果 F 值的變化情況。數據顯示，在大約 10 輪迭代之後，F 值都趨於穩定。圖上的數字顯示了每一類數據的最高值。

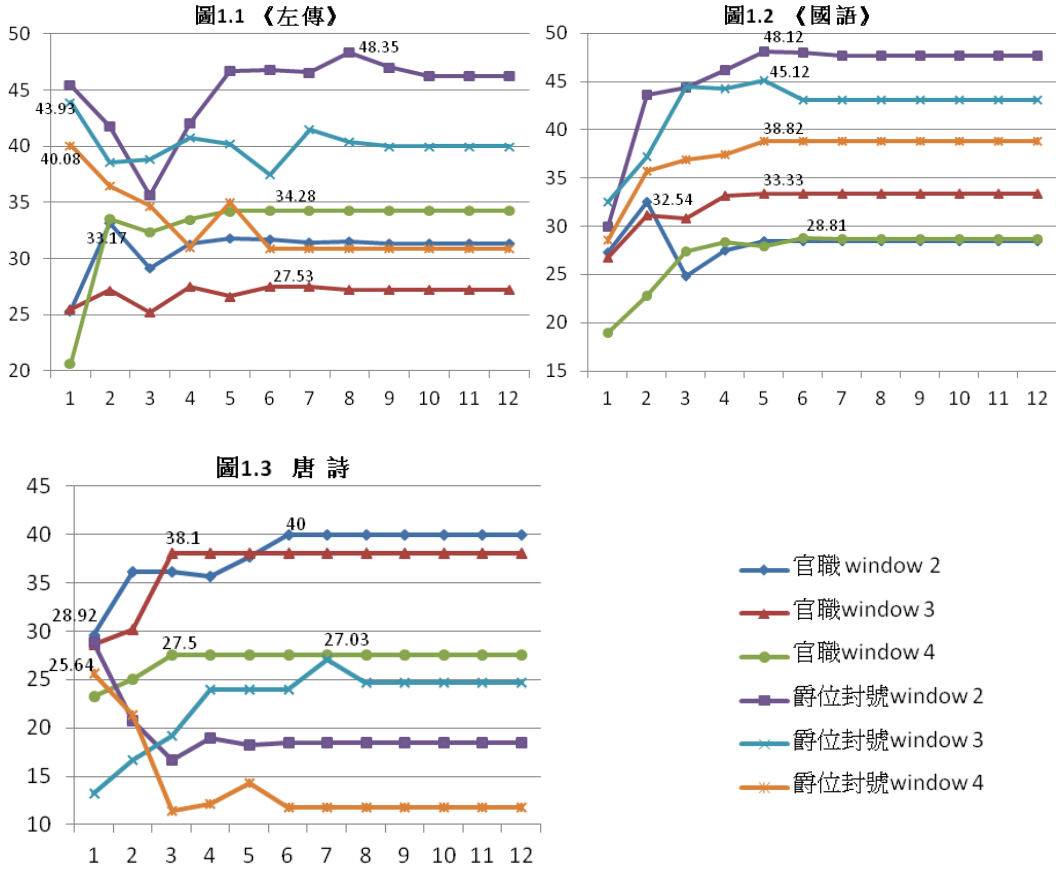

圖 1　先秦史籍、唐詩 F 值變化趨勢

圖 1.1 和圖 1.2 的數據顯示，在兩部先秦典籍中，系統自動識別爵位封號的效果比官職名更好，主要因為先秦史籍中的爵位基本為“公、侯、伯、子、男”五等，規律較明顯。兩部先秦典籍的爵位封號 F 值都在窗口設置為 2 時達到最高，主要因為這兩部史籍中的人物包括很多諸侯國國主，使用爵位指稱時，前面常包含國名，如“[齊/ns1 侯]/nu2”（指代齊國國主）。相對爵位封號，兩部先秦史籍中官職名的識別效果較差，最主要的原因是這兩部史籍和訓練語料中官職名的重合率僅約 2%。《左傳》、《國語》中的官職名 F 值分別在窗口設置為 4、3 時達到最高，這意味著，自動識別《左傳》中的官爵名需要使用更多的上下文特徵，主要因為《左傳》以記事為主，相比以記言為主的《國語》，其上下文中存在人名、地名等命名實體的情況更普遍，更多

地利用這些信息有利於自動識別。先秦史籍中的通假字、古今字和異體字也對自動識別的效果有一定影響，例如《左傳》中準確率最低的名稱為“大史”，該官職名後來寫作“太史”。

在重合率低且文體差異非常大的情況下，圖 1.3 所示的唐詩的測試結果可謂很好。除了因唐朝的年代介於明清和先秦之間外，另一個重要原因是，詩中採用官職名指代歷史人物時，常常和人名連用，例如測試語料中準確率最高的名稱“太守”、“太傅”，均和姓連用。相比先秦典籍，詩歌的結構更加緊湊，因此，唐詩中的官職名和爵位封號 F 值都在窗口設置為 2 時達到最高。由於句子長度的限制及格律的要求，詩中經常使用簡稱，這是影響系統自動識別的重要因素，例如“早晚報平津”，此處“平津”是“平津侯”的簡稱，指代漢公孫弘。

6 结语

從先秦至當代中國，官制系統不斷變化和發展，致使官爵名數量龐大、形式多樣。然而，各代之間也存在一定程度的沿襲。由於官爵名以不同的形式滲透在歷代各種體裁的文字資料中，在使用計算機處理歷代文本、尤其是歷史典籍時，如果不能正確識別官爵名，則會成為計算機理解語言的障礙。因此，本研究基於官爵名作為命名實體進行分類和標註的語料庫，嘗試通過機器學習的方法使計算機識別歷代不同形式文本中的官爵名。本文首先使用完成分詞和命名實體標註、但未做詞性標註的明清小說作為訓練語料，然後使用序列標註算法自動識別明清之前和之後不同時代、不同體裁文本中的官爵名，包括先秦典籍《左傳》和《國語》、唐詩、及現代新聞《人民日報》。在訓練語料和測試語料官爵名重合率非常低、且訓練語料中缺乏詞性標註的情況下，系統能夠較好地自動識別先秦典籍和唐詩中的官爵名，相對而言，對現代新聞的識別準確率則有待改進，但也對其他使用計算機處理歷時語料的工作具有一定參考價值。此項工作涉及的語料跨越從先秦至現代多個時代，涵蓋小說、史籍、詩歌、新聞四種體裁，是使用計算機對歷代不同體裁文本中的官爵名進行研究的一次探索。本研究的成果除了能夠幫助改進計算機對歷時語料庫的處理之外，也有助於對斷代和歷代官爵名的研究，例如，通過比較異同、考證其沿革了解歷代官爵名的總體概貌及發展脈絡。在後續的工作中，可以對訓練語料增加詞性標註，以進一步完善語料加工，使其提供更完整的特徵信息用於機器學習。

参 考 文 献

[1] Xiong D, Lu Q, Lo FJ, Shi DX, Chiu TS. Specification for Segmentation and Named Entity Annotation of Chinese Classics in the Ming and Qing Dynasties. In: Chinese Lexical Semantics (CLSW2012 Revised Selected Papers), Lecture Notes in Computer Science, Volume 7717. Berlin, Heidelberg: Springer, 2013. 280~293.

[2] 徐連達.中國歷代官制辭典.合肥:安徽教育出版社,1991.

[3] 賀旭志.中國歷代職官辭典.長春:吉林文史出版社,1991.

[4] 俞鹿年.中國官制大辭典.哈爾濱:黑龍江人民出版社,1992.

[5] 張政烺,呂宗力.中國歷代官制大辭典.北京:北京出版社,1994.

[6] 俞士汶,段慧明,朱學鋒,孫斌,常寶寶.北大語料庫加工規範:切分•詞性標注•注音. Journal of Chinese Language and Computing, 2003, 13(2):121~158.

[7] 魏培泉,譚樸森,劉承慧,黃居仁,孫朝奮.建構一個以共時與歷時語言研究為導向的歷史語料庫. Computational Linguistics and Chinese Language Processing, 1997, 2(1):131~145.

[8] 中央研究院近代漢語語料庫: http://app.sinica.edu.tw/cgi-bin/kiwi/pkiwi/kiwi.sh

[9] 中央研究院上古漢語語料庫: http://app.sinica.edu.tw/cgi-bin/kiwi/akiwi/kiwi.sh

[10] Lafferty J, McCallum A, Pereira F. Conditional Random Field: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. 2001. 282~289.

[11] 熊丹,陸勤,羅鳳珠,石定栩,趙天成.基於語料庫的明清小說人名與稱謂研究.中文信息學報,待刊.

[12] Wick M, Rohanimanesh K, Culotta A, McCallum A. Samplerank: Learning preferences from atomic gradients. In: Neural Information Processing Systems (NIPS), Workshop on Advances in Ranking. 2009.

[13] McCallum A, Schultz K, Singh S. Factorie: probabilistic programming via imperatively defined factor graphs. In: Advances in Neural Information Processing Systems 22. 2009. 1249~1257.