Computational Linguistics and Chinese Language Processing
Vol. 10, No. 4, December 2005, pp. 533-542

《人民日報》語料庫命名實体分類的研究

The Chinese Named Entity Categorization Based on the People's Daily Corpus

夏迎炬*、于浩*、西野文人*

YingJu Xia, Hao Yu and Fumihito Nishino

摘要

在信息檢索、信息抽取等應用中，命名實体的處理十分重要。本文在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体的細分類進行了深入的研究。提出了命名實体的多級分類并給出了每一級的詳細分類。爲了檢驗該分類体系的實際效果，我們在人民日報語料上進行了初步的標注。并使用常用的基于統計模型的命名實体識別算法在人民日報語料上做了一系列的對比實驗。實驗結果表明：面向機器處理的細分類能有效地提高識別系統的性能并最終有助于信息檢索和抽取。

關鍵字：命名實体、分類、語料庫、自然語言處理

Abstract

Named entity recognition is a very important part of information retrieval and information extraction. Classification is also very important. This paper investigates the sub-classification of named entities from the point of view of information retrieval and information extraction. This paper also presents multi-classification and gives detailed information about each sub-class. We have manually annotated people's daily corpus (1998) and conducted a serial of experiments using the statistical model of named entity recognition. The

* 富士通研究開發中心有限公司，100016 北京市朝陽區霄雲路 26 號鵬潤大廈 B306 室
Internet Application Laboratory, Fujitsu Research & Development Center Co., LTD.
Room B306, Eagle Run Plaza No. 26, Xiao Yun Road, Chao Yang District, Beijing, 100016, P. R. China
E-mail: {yjxia, yu, nisino}@frdc.fujitsu.com

[Received April 12, 2005; Revised June 15, 2005; Accepted August 14, 2005]

experimental results show that the sub-classes presented by this paper can enhance the recognition system's performance and aid information retrieval and information extraction.

Keywords: Named Entity, Classification, Corpus, Natural Language Processing

1. 前言

隨著因特網的飛速發展，網絡信息量呈指數增長。如何使用計算機幫助人們從海量的數據中有效地獲取所需信息是自然語言處理的熱點問題之一。在這一類信息檢索和抽取的應用中，命名實体的自動識別是極爲關鍵的步驟。其主要採用的方法有規則方法、統計方法和統計規則相結合的方法。由于規則的方法需要人工總結、編寫大量的規則并且可移植性差等缺點，人們開始將研究的重點轉移到機器學習方法上[Aberdeen *et.al* 1995; Sekine *et.al* 1998; Borthwick 1999; Sun *et.al* 2002; Bikel *el.al* 1997]。而對于基于機器學習的識別系統而言，語料庫起著至關重要的作用[黃昌寧 2002; 馮志偉 2001]。歷來受到學者們的高度重視，隨著基于統計的自然語言處理研究的不斷深入，對于語料庫的需求也日益強烈。

富士通研究開發中心有限公司與北京大學計算語言學研究所、人民日報信息中心合作，以 1998 年人民日報爲對象，製作了大規模漢語標注語料庫[俞士汶 2000; 段慧明 2002]，并已將上半年部分于 2001 年進行了公開，在很多研究單位得到了使用。在 1998 年全年的人民日報標注語料庫定義中，包括有命名實体的標記，語料庫中的命名實体分爲 4 類[俞士汶 1998]：機構團体名、地名、人名、其他專用名詞。與之對應的是國際上權威的評測會議 Message Understanding Conference 于 1995 年（MUC-6）第一次設立命名實体識別評測專項。MUC（MUC-6）定義了三大類（實体、時間和數位）、七小類（人名、機構名、地名、時間、日期、貨幣和百分比）命名實体。

經過多年的實踐證明，這樣的分類對于應用來說顯得過粗，不能起到很好的定義作用。本文的工作就是在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体細分類的顆粒度進行了深入的研究。提出了命名實体的多級分類体系并給出了每一級的詳細分類定義。力求既要適應語言信息處理與語料庫語言學研究的需要，又要能爲傳統的語言研究提供充足的素材；既要適合計算機的自動處理，又要便于人工校對。在本文的第 2 節將分別介紹對人名、機構名、地名子類分類体系。在第 3 節對命名實体簡稱的分類做了初步探索。最后在第 4 節中給出實驗結果及結論并展望未來的工作。

2. 命名實体細分類

2.1 人名分類

在 1998 年人民日報語料中，人名採用統一的標注模式，從標注上無法區分中國人名和外國人名。這樣的標注語料會給機器學習帶來噪聲。因爲不同國家的人名的內部特徵（主要是人名用字集）存在較大的差別。例如，日本人名的常用山本、太郎、大島、藤田等

字和詞；蘇俄人名常用斯、基、娃等字；歐美人名常用朗、魯、倫、曼、尼等字。所以，將他們分開標注和識別將會提高人名的識別性能。基于這樣的考慮，我們將人名的第一級分爲：中國人名、外國人名和不確定三類。其中不確定類別是爲了減少在機器學習的噪聲而引入的，這樣的分類能儘量保證同一類別中的特徵比較集中。當然在外國人名中，可以嘗試根據其人名用字集分爲不同的國別，比如：日本人名、歐美人名、其他等，這樣可以進一步使特徵集中。

在第一級分類的基礎上，我們注意到：即便是同一大類中的名字（比如中國人名），其構成及用字也有很大的不同，比如單詞型（只有姓）、双詞型（姓名）、三詞型（夫婦双姓等），可以通過這些特徵將人名進一步分類。由此，我們得到了第二級的分類。最后在此分類的基礎上，我們對語料庫中的人名進行了統計，發現有必要按照其構成特徵繼續進行細分類。比如單詞型可以細分爲單字型、二字型、三字型和多字型。需要指出的是分類越細，將導致機器學習的訓練數據越稀疏，這是我們不希望看到的情況，但是考慮到標注的語料庫既要滿足語言信息處理與語料庫語言學研究、又要爲傳統的語言研究提供充足的素材。我們還是決定進行第三級的分類。在機器學習的時候可以根據訓練語料的規模以及特徵的相似性將某些特徵合并在一起，比如將三字型和多字型合并等。

表 1 給出了人名的細分類以及每一分類在語料庫中的出現的頻數。

表 1: 人名分類

總類	一級子類	二級子類	三級子類
人名（48709）	中國人名（39372）	單詞型（2673）	單字型（396）
			二字型（2141）
			三字型（96）
			多字型（40）
		双詞型（36663）	單姓單名（6492）
			單姓双名（29983）
			双姓單名（57）
			双姓双名（105）
			其他（26）
		三詞型（36）	
	外國人名（9227）	單詞型（9208）	二字人名（495）
			三字人名（2824）
			四字人名（2083）
			多字人名（3806）
		双詞型（19）	
	不確定（110）	存疑待查（42）	
		語料錯誤（68）	

2.2 機構名分類

關于機構分類，目前有很多可以借鑒的分類体系。比如按照中華人民共和國民法通則的規定，將組織機構分爲企業、事業單位、社會團体、國家機關四個大類。在此四大類的基礎上又詳細的劃分了若干子類，比如國家機關又可以分爲：國家權力機關、國家權力機關分支機搆、國家行政機關、國家行政機關分支、派出機構、國家司法機關、人民法院、人民法院分支機搆、人民檢察院、人民檢察院分支機搆、政黨機關、政協組織等。但是這樣的分類顯然不符合信息檢索和抽取的需要。在綜合考慮現有的各種分類体系基礎上，我們對語料庫中的機構名進行了詞頻統計。制定了機構團体的二級分類体系。由于中外機構名在用詞等特徵上沒有明顯的差別。而且在下一級的子類的劃分上也基本相同。我們并沒有劃分中國和外國機構名這一級。只是在總類上標明了可以有中國機構和外國機構的區別并給出了其在語料庫中的頻數。

第一級的分類有 23 類，在第一級的基礎上，又對其中的經濟機構類、政治組織類、健康組織、媒体組織、体育機構、教育機構進行了細分類。

其具体的分類和頻數統計見表 2。

表 2: 機構名分類

總類（標注）	一級子類（標注）	二級子類（標注）
機構團体（41141） [中國：32450] [外國：8691]	文學藝術類（art_nt）（2000）	
	經濟機構類（economic_nt）（3410）	銀行（bank_economic_nt）（1020）
		基金會（fund_economic_nt）（296）
		股票名稱（stock_economic_nt）（623）
		其他經濟組織（org_economic_nt）（1477）
	政治組織類（political_nt）（1693）	政治聯盟（org_political_nt）（1465）
		政黨名稱（party_political_nt）（228）
	公司（company_nt）（7960）	
	健康組織（health_nt）（1477）	醫院（hospital_health_nt）（659）
		醫療健康組織，但不是醫院（org_health_nt）（818）
	軍事機構（military_nt）（2255）	
	飯店酒店（hotel_nt）（39）	
	女性組織（woman_nt）（142）	
	媒体組織（media_nt）（1696）	電視媒体（tv_media_nt）（304）
		廣播媒体（radio_media_nt）（183）

		出版社（publishhouse_media_nt）（967）
		電影媒体（movie_media_nt）（183）
	製造業（manufacturer_nt）（1977）	
	研究機構（research_nt）（2156）	
	宗教機構（religion_nt）（73）	
	工會（labourunion_nt）（156）	
	体育機構（sport_nt）（1645）	体育運動隊（team_sport_nt）（783）
		其他体育組織（org_sport_nt）（862）
	教育機構（edu_nt）（4400）	學校（school_edu_nt）（4035）
		其他教育組織（org_edu_nt）（365）
	能源部門（energy_nt）（1276）	
	政府部門（gov_nt）（11133）	
	公安部門（police_nt）（1016）	
	檢查部門（procuratorate_nt）（260）	
	法院（court_nt）（624）	
	律師事務所等（law_nt）（245）	
	海關（ciq_nt）（164）	
	其他組織（otherunion_nt）（2799）	

2.3 地名分類

地名分類的情況與機構名分類情況有些類似，也有很多可以借鑒的現有分類体系。比如全國地名委員會編的《地名信息系統規範》就把地名分爲：國名、首都、省級行政區域駐地、地級市、縣級市等共 76 類別，但這樣的分類更多是從行政區劃上考慮的。對于面向信息檢索和抽取的標注語料庫建設來說，需要綜合考慮信息檢索、機器學習方面的需求。我們最終將地名劃分爲 46 個一級子類。同樣沒有區分中國地名和外國地名，其具体的分類見表 3。

表 3: 地名分類

總類	一級子類
地名（ns）（5226） [中國：3921] [外國：1305]	市場（market_ns）（418）
	賓館（hotel_ns）（266）
	劇院禮堂（theater_ns）（129）
	博物館紀念館（museum_ns）（374）
	機場（airport_ns）（255）
	車站（station_ns）（167）
	公園（park_ns）（298）
	草原（grassland_ns）（27）
	大廈寫字樓（mansion_ns）（126）
	地區（area_ns）（845）
	公路（road_ns）（62）
	山脈（mountain_ns）（31）
	街道（street_ns）（113）
	体育運動場所（sportplace_ns）（275）
	行政區（district_ns）（313）
	農場（farm_ns）（57）
	廣場（plaza_ns）（136）
	平原（plain_ns）（47）
	碼頭（dock_ns）（30）
	開發區（developarea_ns）（166）
	教堂寺廟（religion_ns）（51）
	住宅小區（uptown_ns）（51）
	墓地陵園（grave_ns）（55）
	沙漠（desert_ns）（20）
	書店（bookshop_ns）（17）
	電站（powerplant_ns）（5）
	示範區（demonstratearea_ns）（20）
	流域（drainagearea_ns）（35）
	水文站（waterinfostation_ns）（29）

	活動中心（center_ns）（141）
	基地（fundation_ns）（64）
	海峽海域（seaarea_ns）（27）
	餐飲場所（restaurant_ns）（39）
	藥店（drugstore_ns）（3）
	林場（forestcenter_ns）（17）
	茶場（teafield_ns）（4）
	油田（oilfield_ns）（5）
	礦區（mine_ns）（20）
	音樂廳（musichall_ns）（14）
	出版社（publishhouse_ns）（1）
	政府（gov_ns）（29）
	公司（company_ns）（2）
	植物園（arboretum_ns）（17）
	鐵路（railway_ns）（3）
	其他地名（other_ns）（349）
	會議場所（meetinghall_ns）（49）

3. 簡稱分類

在真實語料中，很多命名實体是以簡稱的形式出現。比如 “老張”、“小李”、“京九鐵路”、“京津高速”、“北大”、“政協”等。在 1998 年人民日報標注語料中對這些簡稱的標注方案是：對人名的簡稱標成“人名（nr）”，對其他的簡稱則同一標成“簡稱（j）”。比如：“老張/nr”、“政協/j”、“中/j 美/j 關係/n”。這樣的標注不但混淆了全稱和簡稱的區別（人名簡稱），而且模糊了簡稱之間的界限（地名簡稱和機構名簡稱）。另外，從語義角度、信息檢索和抽取等應用角度應該是將不同類型的簡稱區別開來并且在語料中給予標注。比如：從面向機器學習的角度，“中美關係”較好的標注結果應該是“中/aloc 美/aloc 關係/n”。（其中 aloc 表示簡稱地名）。

基于這樣的考慮，我們將簡稱首先分爲人名簡稱、地名簡稱、機構名簡稱和其他簡稱等幾類。在這幾類簡稱中，機構名簡稱的情況比較複雜，也最難識別。經在語料中統計分析發現，機構名簡稱大致有如下的幾種形式：連續型：“解放軍第 301 醫院（301 醫院）”，不連續型：“北京大學（北大）”，以及混合型：“東風汽車電子儀表股份有限公司（東風電儀）”。更多的具体例子見表 4。通過這樣的分析，我們將語料庫中的簡稱按表 5 所示進行了細分類。

表 4: 機構名簡稱示例

連續簡寫	上海華聯超市股份有限公司	上海華聯
	上海紫江企業集團股份有限公司	紫江企業
	解放軍第 301 醫院	301 醫院
	北京 25 中學	25 中
不連續簡寫	上海證券交易所	上證/上證所
	北京大學	北大
	電子工業部第六研究所	六所
	武漢鋼鐵集團公司	武鋼
連續簡寫與不連續簡寫混合	東風汽車電子儀表股份有限公司	東風電儀

表 5: 簡稱分類

人名簡稱（APER）(716)	
地名簡稱（ALOC）(24760)	
機構名簡稱（AORG）(6378)	連續型簡稱（AORG_SEQ）
	不連續型簡稱（AORG_DIS）
	混合型簡稱（AORG_OTH）

4. 實驗結果及結論

4.1 實驗結果

爲了驗證細分類帶來的效果，我們在人民日報語料上進行了一系列的實驗。實驗採用的[Wu 2003]所示的基于詞類和詞性類的統計模型來進行命名實体識別，訓練語料爲 1998 年 1~5 月份的人民日報標注語料；測試語料爲 1998 年 6 月份的人民日報語料。我們使用的評測指標有：精確率，召回率，F-值

$$准確率 = \frac{正確識別的實体數}{總的識別實体數} ，$$

$$召回率 = \frac{正確識別的實体數}{總的實体數} ，$$

$$F-值 = \frac{2*召回率*正確率}{召回率+正確率} 。$$

首先，我們比較了第 2、3、4 節介紹的人名、地名、機構名細分類和原分類体系之間的差別。然后我們比較了在新分類体系下，使用與不使用簡稱識別的系統性能。實驗

結果如表 6、表 7 和表 8 所示。表 6 是在原分類体系上進行訓練和測試得到的結果，表 7 是在本文介紹的細分類体系上得到的結果，但是不使用簡稱識別。表 8 是在新分類体系上使用簡稱識別的測試結果。從結果上看，對于人名識別來說，新舊分類体系的差別很大。在新的分類体系下系統的性能得到很大的提高。而對于地名和機構名來說，在新的分類体系下，系統性能也有提高但不如人名那樣顯著。其主要原因是人名的細分類更多的抓住了人名的構詞特徵，比如中外人名的區別、單字型和二字型之間的區別等。這樣的劃分是有利于機器學習的特徵抽取的。而地名和機構名的劃分在構詞規律上區別不大，更多的是語法和語義上的區別，比如地名分類中的“賓館”和“大廈寫字樓”類之間以及機構名分類中的“政府部門”和“公安部門”之間從用字規律上基本相同。其主要的區別是職能上的，這已經屬于語義的範疇了，在目前的水準下，處理語義信息是極其困難的。這也是在新的分類体系下，地名和機構名分類不如人名分類效果那麼明顯的原因。

但從表 7 和表 8 的對比，我們可以看出，簡稱的分類和識別對地名和機構名識別的性能提升很明顯。這是對簡稱進行細分類和識別帶來的效果。當然，由于其中加入了簡稱識別技術。簡稱識別的性能也將對系統產生影響。這樣的對比并不是很嚴格。

表 6: 原分類体系下的系統性能

	准確率	召回率	F 值
人名	92.87	89.33	91.06
地名	93.66	88.78	91.15
機構名	84.36	77.96	81.03

表 7: 新分類体系下的系統性能（無簡稱識別）

	准確率	召回率	F 值
人名	94.06	95.21	94.63
地名	93.87	90.78	92.29
機構名	83.76	82.42	83.08

表 8: 新分類体系下的系統性能（有簡稱識別）

	准確率	召回率	F 值
人名	93.96	96.18	95.06
地名	95.77	96.76	95.26
機構名	89.77	86.72	88.22

4.2 結論和未來的工作

本文的在目前的命名實体分類体系的基礎上，從信息檢索和抽取的角度對命名實体的細分類進行了深入的研究。提出了命名實体的多級分類体系并給出了每一級的詳細分類定義并對命名實体簡稱的分類做了初步的探索。在人民日報 1998 年標注語料上根據新的分

類体系進行標注和測試。實驗結果表明：新的分類体系有助于面向信息檢索和抽取的機器學習。這樣的分類体系可以使自動識別系統的性能得到大幅的提高。

常規的命名實体類別包括三大類（實体類、時間類和數字類）、七小類（人名、機構名、地名、時間、日期、貨幣和百分比）命名實体。對于信息提取、文本挖掘、網絡內容管理等應用來說，這些類別已經不能滿足應用的需求，還有一些命名實体也非常必要。例如："事件類實体"（第一屆中國網球公開賽、第五屆中國國際航空航天博覽會、首屆英國戲劇舞蹈節、中國首屆網絡相聲大賽、澳大利亞文化周）、"著作類實体"（斯德哥爾摩環抱公約、《中國人民銀行金融機構反洗錢規定》、《神雕俠侶》）、"股票名稱"（銀河創新、富龍熱力 、深萬山A）等等。我們需要對新的實体類型進行大範圍的調查，看需要擴充哪些類型的命名實体，分別出現在哪些領域，從而建立起命名實体分類体系和與之配套的識別和標注工具，爲信息提取、文本挖掘、網絡內容管理等應用奠定基礎。

參考文獻

段慧明、松井久仁于、徐國偉、胡國昕、俞士汶， "大規模漢語標注語料庫的製作與使用"， *語言文字應用，2002 年，第 2 期，* pp. 72-77.

馮志偉， "中國語料庫研究的歷史和現狀"， *Proceedings of ICCC2001*, pp.1-24.

黃昌寧、李涓子， "語料庫語言學"， *商務印書館， 2002*, pp. 244-256.

俞士汶、朱學鋒、段慧明， "大規模現代漢語標注語料庫的加工規範"， *多語言信息處理國際會議，2000* , pp. 19-24.

俞士汶等， "現代漢語語法信息詞典詳解"，*清華大學出版社， 1998*, pp. 546.

Aberdeen.J, J. Burger, D. Day, L. Hirschman, P. Robinson, and M. Vilain, "MITRE: Description of the ALEMBIC System Used for MUC-6," In *Proceedings of the Sixth Message Understanding Conference (MUC-6)*, November 1995, pp. 141-155.

Bikel, D.M., S. Miller, R. Schwartz, and R. Weischedel, "Nymble: a high-performance learning name-finder," *In Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., 1997, pp. 194-201.

Borthwick, A., "A Maximum Entropy Approach to Named Entity Recognition," PhD Dissertation, New York University, 1999.

Sekine, S., R. Grishman, and H. Shinou, "A decision tree method for finding and classifying names in Japanese texts," In *Proceedings of the Sixth Workshop on Very Large Corpora*, Canada, 1998, pp.171-178

Sun, J., J. Gao, L. Zhang, M. Zhou, and C. Huang, "Chinese Named Entity Identification Using Class-based Language Model," In *Proceedings of the 19th International Conference on Computational Linguistics*, Taipei, 2002, pp. 967-973.

Wu, Y., J. Zhao, and B. Xu, "Chinese Named Entity Recognition Combining Statistical Model wih Human Knowledge," *ACL 2003 Workshop on Multilingual and Mixed-language Named Entity Recognition*, 2003, Japan.