Computational Linguistics and Chinese Language Processing
Vol.7, No.2 , August 2002 , pp. 101-112

基於詞彙語義的百科辭典知識提取實驗

An Experiment on Knowledge Extraction from an Encyclopedia Based on Lexicon Semantics

宋柔*、許勇+

Song Rou, Xu Yong

摘要

本文研究百科辭典釋文信息提取方法，設計了一個基於詞彙語義屬性和關係的形式系統。在對百科辭典的詞目按語義分類的基礎上，對釋文的線性詞串進行簡單的語義屬性匹配，便可提取文本中的簡單知識。在一項百科辭典信息提取的實驗中，這一方法的有效性得到了初步的驗證。

關鍵詞：知識提取，詞彙語義

Abstract

The typical approaches to extracting text knowledge are sentential parsing and pattern matching. Theoretically, text knowledge extraction should be based on complete understanding, so the technology of sentential parsing is used in the field. However, the fragility of systems and highly ambiguous parse results are serious problems. On the other hand, by avoiding thorough parsing, pattern matching becomes highly efficient. However, different expressions of the same information will dramatically increase the number of patterns and nullify the simplicity of the approach.

Parsing in Chinese encounters greater barriers than that in English does. Firstly, Chinese lacks morphology. For example, recognition of base-NP in Chinese is more difficult than that in English because its left boundary is hard to discern.

* 北京語言大學計算機系 Beijing Language and Culture University
E-mail: songrou@blcu.edu.cn
+ 北京工業大學計算機學院 Beijing Polytechnic University
E-mail: hopexy163@163.com

Secondly, there are many stream sentences in Chinese which lack subjects and cause parsing to fail. Finally, in Chinese, the absence of verbs is also pervasive. Sentential parsing centering on verbs, which is used with English, is not always successful with Chinese.

We are engaged in research on knowledge extraction from the Electronic Chinese Great Encyclopedia. Our goal is to extract unstructured knowledge from it and to generate a well-structured database so as to provide information services to users. The pattern-matching approach is adopted.

The experiment was divided into two steps: (1) classifying entries based on lexicon semantics; (2) establishing a formal system based on lexicon semantics and extracting knowledge by means of pattern matching.

Classification of entries is important because in the text of the entries of different categories there are different kinds of patterns expressing knowledge. Our experiment demonstrated that an entry of the encyclopedia can be classified precisely merely according to the characters in the entry and the words in the first sentence of the entry's text. Some specific categories, e.g., organization names and Chinese place names, can be classified satisfactorily merely according to the suffix of the entry, for suffixes are closely related with semantic categories in Chinese.

The formal system designed for knowledge extraction consists of 4 kinds of meta knowledge: concepts, mapping, relations and rules, which reflect lexicon semantic attributes. The present experiment focused on the extraction of knowledge about various areas from the texts regarding administrative places of China (how large is a place or its subdivisions). The results of the experiment show that the design of the formal system is practical. It can accurately and completely denote various expressions of simple knowledge in a Chinese encyclopedia. However, when the focus of knowledge changes, e.g., from administrative areas to habits of animals, it is a labor-intensive task to renew the formal system. Therefore the study of auto or semi-auto generation of this kind of formal system is required.

1. 問題背景

以信息技術爲基礎的在線知識服務是信息產業的發展方向。目前已經出現具有初步實用價值的在線知識服務，其中最具生命力的是問答服務（Q&A），而這一服務的關鍵技術

之一是文本知識的自動提取，它是爲各種各樣的問題提供答案的基礎。

目前，網絡技術的發展使得人們能輕易地獲取幾乎無窮無盡的文本。但由於網絡文本範圍太廣，涉及的語言現象太複雜，全自動、高準確率的信息提取和知識提取困難較大，短期內難以實用。百科辭典是一種受限的文本，知識含量高，知識表述比較規範。無論是從理論的角度看還是從應用的角度看，從百科辭典中自動獲取知識可當作文本知識自動提取的突破口。

文本知識提取的方法主要有兩種：基於語句分析的方法和基於模式的方法[Tsujii, J. 2000]。理論上說，文本知識的提取需要在徹底理解的基礎上進行。因此，句法分析和語義分析技術很自然地使用於這一領域，但它有脆弱性和多歧義問題。基於模式的方法可以避免對語句進行徹底分析，效率較高，但同一信息的不同表達形式會使模式數量大爲膨脹。[Tsujii, J. 2000，Hull, R. *et al.* 1999 and Soderland, S. G. 1996]處理的是英語或日語文本，主要使用基於語句分析的方法。其中[Hull, R. *et al.* 1999]的工作是基於大容量的知識庫，採用部分分析、語義解釋和推理的步驟；[Soderland, S. G. 1996]也是進行句法分析，包括名詞短語分析、同位關係識別（Appositive Recognition）和同指消解（Coreference Analysis）；[Tsujii, J. 2000]本身的工作主要使用語句分析的方法，但吸收了模式方法的優點。

漢語文本的知識提取使用語句分析方法比英文問題更大。首先是因爲漢語缺乏形式標誌。比如基本名詞短語的識別在英語中並不困難，但在漢語中由於難以確定其左邊界而識別率較低。其次，漢語常有缺主語的流水句，會造成句法分析的失敗。此外，英語句子的句法分析和語義分析一般都以動詞爲核心，而相當一部分漢語的句子沒有動詞，如“昌平縣面積 1352 平方公里，人口 43 萬”。如果照搬英語中的做法做句法分析（或淺層句法分析）、找動詞的語義格，其效果不會好。

漢語文本知識提取的工作已發表的並不多。[Gu, F. *et al.* 2001]的工作也是從百科辭典中提取知識，它的結果是一個框架結構的知識庫，可以提供實用的知識服務。但爲了得到這個知識庫，需要先設計一個形式語言，並用它對辭典文本進行人工標注。

本文研究漢語百科辭典的知識提取。我們的目標也是把百科辭典中的無結構的知識提取出來，生成帶結構的數據庫，向用戶直接提供知識服務。這項工作當然只能在一個受限範圍內通過人機結合的方式來完成。但是，我們希望使人的勞動集中於詞彙的語義屬性研究和詞庫中詞彙的語義屬性標注，避免人工標注語料所需的巨大勞動量。由於上述漢語分析中的困難，我們不採用常規的句法語義分析，而嘗試關鍵詞語爲核心的模式匹配的方法，其中關鍵詞語不一定是動詞，但具有信息提示的功能（如“面積”提示其後面有關於面積數量的信息），模式匹配主要依靠詞語的語義屬性。

我們的處理對象是《中國大百科全書》（光盤版），工作步驟是：（1）根據詞目確定題材類別，根據題材類別確定知識提取的目標；（2）建立基於詞彙語義的形式系統，用詞語模式匹配的方法提取知識。本文介紹了相關研究的一些實驗，測試結果證明這一方法是有效的。

2. 百科辭典詞目的分類

2.1 百科辭典詞目按題材分類的試驗

爲了提取知識的方便，首先需要把按領域分卷的《中國大百科全書》中的詞目進行分類。

這裏所說的詞目的類別，不是按專業領域劃分，而是按題材劃分的。比如，人物和概念是不同的題材。《中國大百科全書》美術卷中人物"徐悲鴻"釋文與數學卷中人物"華羅庚"釋文的風格相似，所表達的信息內容的類型十分接近，但與同在美術卷中概念"油畫"的釋文風格和信息內容的類型完全不同。

題材的異同取決於詞目的語義類。所有類別的釋文第一句話總是對於詞目給出一個概括性的說明，指出它的最重要的特徵，如人物的國籍和歷史地位，行政區劃的行政隸屬和政治經濟地位，動物的目科屬種等。第一句話以後，不同語義類的釋文有不同的信息內容。比如，人物的釋文包括人物的生卒時間和地點、生平事蹟、主要成就等，行政區劃的釋文包括該地區的面積、人口、沿革、地形、氣候、經濟、特產、名勝等，動物的釋文包括動物的體形、各部位的形狀大小顏色、分佈區域、生活習性、繁殖方式、與人類的關係等。

從百科辭典知識提取的使用目標出發，我們目前採用的詞目分類系統中的大類是人物、行政區劃、自然地理、動物、植物、機構組織、事件、裝置、其他。之所以採用這樣的分類體系，一是因爲這些類的詞目和釋文有比較明顯的特徵，知識抽取相對容易；二是因爲這些類在整個百科辭典中所占比重較大，詞目較多，有條件使用統計方法進行信息提取。有些大類下面還要分小類，如自然地理類中包括山脈、河流、湖泊、沙漠、島嶼等等，分小類的目的是使同類釋文的信息特徵更加一致。

我們使用現代漢語通用分詞系統 GWPS 的專名識別功能將詞目中的人名、地名（包括行政區劃名、自然地理名、古地名、景點設施名）、機構名挑選出來，實驗對象是美術卷、外國文學卷、世界地理卷、中國地理卷。我們只使用詞目內部的用字信息和釋文第一句話最後兩個詞的信息，識別結果如下：

	實有詞目	識別詞目	遺漏	誤識	準確率	召回率
美術卷人名	935	935	1	1	99.9	99.9
外國文學卷人名	2470	2471	0	1	99.96	100
世界地理卷地名	1153	1154	5	6	99.5	99.6
中國地理卷地名	1498	1500	0	2	100	99.9
美術卷機構名	98	98	0	0	100	100
合計	6154	6158	6	10	99.8	99.9

如果不用釋文信息，只用詞目內部的用字信息，則對機構名和中國地名影響不大，對人名和外國地名來說召回率大大降低。如此時美術卷人名詞目識別結果爲：

實有詞目	識別詞目	遺漏	誤識	準確率	召回率

935	779	166	10	98.7	82.2

原因是有些人名詞目使用的是法號（如“法常”）、綽號（如“泥人張”），有些是 GWPS 不具有識別能力的日本人名（如“奧村土牛”）。人名釋文的首句最後一個詞絕大部分是“身份詞”（“畫家”、“建築師”等），大部分首句前還帶有說明生卒年代的括號，所以利用了釋文首句的信息後，召回率大大提高。

地名詞目中，不同小類的詞目的釋文風格差別仍然很大。比如，行政區劃名釋文的主要信息是行政隸屬關係和政治經濟地位、面積、人口、沿革、地形、氣候、經濟、特產、名勝等，自然地理名的釋文中沒有這些內容。行政區劃名和自然地理名中還需分更小的類，因爲行政區劃中，關於國家的釋文同關於城市的釋文在詳盡程度上很不同，信息內容的類型上也有區別。自然地理名中，關於山脈的要介紹山脈地理分佈、走向、山峰高度、地質歷史等，關於河流的要介紹河流發源地、走向、流域面積、經濟功能等。爲此，我們對於中國地理卷中的地名進行了細分類試驗。對於行政區劃詞目，分爲省、自治區、地區、自治州、市、區、縣（包括自治縣）、鎮，共 8 類；對於自然地理詞目，挑選出江河、湖泊、山嶺、山脈、盆地、沙漠、平原、高原、丘陵、草原、島嶼，共 11 類。我們試驗完全依據詞目後綴進行識別。行政區劃詞目所用後綴和識別結果如下：

類名	省	市	地區	自治州	區	縣	鎮	合計
後綴	省	市	地區	自治州	區*	縣	鎮	
實有	23	385	5	9	22	275	36	755
標識	23	385	5	9	25	275	36	758
誤識	0	0	0	0	3	0	0	3
漏識	0	0	0	0	0	0	0	0
準確率%	100	100	100	100	88	100	100	99.6
召回率%	100	100	100	100	100	100	100	100

注：“區”類要從後綴“區”中去掉後綴“自治區”、“地區”、“風景區”、“風景名勝區”、“自然保護區”、“灌區”。該類的 3 個誤識錯誤是“皖西山區”、“皖南山區”、“神農架林區”。簡單地把“山區”當作後綴從“區”類中去掉是不行的，因爲上海有“寶山區”，北京曾有“燕山區”，等等。

自然地理詞目所用後綴和識別結果如下：

類名	河流	湖泊	山嶺*	山脈	島嶼	盆地	沙漠	平原	高原	草原	丘陵	合計
後綴	江,河*, 溪, 水*	湖, 錯, 池, 海*	山, 峰,	山脈	島*	盆地	沙漠*	平原	高原	草原	丘陵	

			嶺									
實有	144	65	162	19	20	14	5	19	11	2	8	466
標識	137	59	162	19	20	14	5	19	11	2	8	456
誤識	3	1	1	0	0	0	1	0	0	0	0	6
漏識	7	7	1	0	0	0	1	0	0	0	0	16
準確率%	97.81	98.31	99.38	100	100	100	80	100	100	100	100	98.68
召回率%	93.06	89.23	99.38	100	100	100	80	100	100	100	100	96.57

注：以“河”爲最後一個字但不是河流的詞目是“三河”；以“水”爲最後一個字但不是河流的詞目是“中國的地表水”和“中國的地下水”;漏識的河流是以“布”和“曲”爲後綴的西藏地區河流；以“海”爲湖泊詞目的後綴，需要人爲地去掉渤海、黃海、東海和南海，但仍然有一個誤識：“中國的近海”；“山嶺”類的誤識是“中國的火山”，漏識是“神農頂”；“島嶼”類要從後綴“島”中去掉後綴“半島”；漏識的湖泊是“月亮泡”、“大布蘇泡”和以“茶卡”爲後綴的西藏地區鹹水湖；誤識的沙漠是“中國的沙漠”，漏識的沙漠是“毛烏素沙地”。

2.2 關於詞目分類方法的結論

我們的試驗說明，僅根據詞目的用字構成和詞目釋文的首句用詞，就可以對於百科辭典詞目的主要題材類別進行分類，準確率和召回率可達到實用要求。對於某些類別，比如機構名和中國地名，則僅使用詞目後綴就能達到相當好的識別效果，其原因是漢語後綴成分與語義類別緊密相關。

3 百科辭典釋文知識提取實例

3.1 一個基於詞彙語義屬性的形式系統

我們把處理對象限定爲行文規範的百科辭典，目前只提取比較易於形式化的信息。我們的基本思想是：建立起一個基於詞彙語義的屬性和關係的形式系統，其中的屬性和關係同欲提取的信息緊密相關；使用屬性模式匹配的方法在線性詞串中提取信息。

我們首先做的是中國行政地名詞目釋文中面積信息的提取。

大部分面積信息的表述中有“面積”二字，但是在成串的說明中，有省略的情況；“填海”、“種植”等動詞帶數詞和面積量詞表示面積的情況下，有時也不使用“面積”。如關於

香港的釋文中有：

陸地面積 1071.8 平方公里。其中香港島 75.6 平方公里，九龍 11.1 平方公里，“新界”（包括大嶼山島等周圍 230 多座島嶼）975.1 平方公里，另新填土地 9.2 平方公里。

此外，中國行政地名詞目的釋文中，有 4 處“面積”的錯別字：2 處錯成“南積”，1 處錯成“面和”，1 處錯成“面只”。

作爲信息提取的初步研究，我們只考慮出現“面積”二字時的情況。

在 755 個中國行政地名詞目的釋文中，“面積”出現了 1668 次，其中 38 個“大面積”和 2 個“單位面積”用作修飾成分，如“形成眾多的鹽湖和大面積沼澤”和“樹木種類多,單位面積蓄積量高”，其餘 1628 處“面積”確實表達面積信息。

利用面向語言教學研究的文本檢索工具 CCRL 作爲輔助工具,我們用人工分析研究了這些“面積”的上下文。

與“面積”相關且帶有數值的信息可以看成是某些關係：

數量關係。論元爲主體、數值、度量單位。如“海壇島面積 323 平方公里”，“海壇島”爲主體，“323”爲數值，“平方公里”爲度量單位。

比例關係。論元爲分子主體、分母主體、比例數。多比例關係則涉及多個比例主體和多個比例數。如“青海……天然草場面積約占全省土地總面積的 46.39%”，“天然草場”爲分子主體，“全省土地”爲分母主體，“46.39%”爲比例數。

變化數量關係。論元爲主體、擴縮標記、數值、度量單位。如“……城區面積擴大了 15 平方公里”，“城區”爲主體，“擴”爲擴縮標記，“15”爲數值，“平方公里”爲度量單位。

變化比例關係。論元爲主體、擴縮標記、倍數或比值。如“貴州……茶園面積較 50 年代初擴大 20 多倍”，“茶園面積”爲主體，“擴”爲擴縮標記，“20 多倍”爲倍數。

變化數量關係和變化比例關係還應當涉及變化前時間和變化後時間。變化前時間往往顯式地給出，變化後時間有時省略，其實就是百科全書資料收集的時間。如上面最後一例，變化前時間是“50 年代初”，變化後時間爲百科全書資料收集的時間，文中省略。數量關係和比例關係也應當涉及時間，被省略的時間也是百科全書資料收集的時間。這些關係往往帶有修飾成分，如“***約*** 10 公頃”，“***不到*** 30%”，“5 倍***以上***”，“***擴大至*** 23 平方公里”等。這些也應當作爲論元加入到各關係中。

信息提取的任務就是確定這些關係中的論元在文本中所指的內容。其中，數值、比例數、倍數或比值、度量單位、擴縮標記、修飾比較容易確定，因爲它們形式規範，位置比較固定，而且後三者的集合基本上是封閉的。時間論元也有形式標記，包括“世紀”、“年代”、“年”、“月”等，表示朝代或事件的詞語後面加上“初”、“末”、“前”、“後”、“期間”等時間方位詞。確定時間論元的主要困難在於出現位置不固定。我們的策略是從其它論元出現的位置往前看 6 個逗號或句號，找到了時間論元特徵就可以提取出來，找不到就歸結爲省略，即時間論元是百科全書資料收集的時間。

最大的困難在於各種面積主體的確定。爲此，我們從實例中提取了一個基於詞彙語

義屬性的形式系統，它的內容包括 4 類元知識：

概念：

行政區劃 xq，往往是當前詞目本身，也可能是當前詞目所代表的行政單位的上級單位。

詞目替代詞 td，包括“省境”、“全省”、“市境”、“全市”、“區境”、“全區”、“縣境”、“全縣”。

行政區劃的分部 fb，包括“市區”、“城區”、“郊區”、“海域”、“陸域”、“陸地”，還包括方位分部如“東部”、“西北部”。

具有面積屬性的名詞性詞語 mc，包括“草原”、“平原”、“耕地”、“土地”、“陸地”、“森林”、“荒地”、“荒山”、“喀斯特地貌”、“茶園”、“桑園”、“果園”等。（注：“山嶺”、“山脈”、“河流”等不具有面積屬性。）

具有面積屬性的動詞性詞語 dc，包括“種植”、“播種”、“養殖”、“淡水養殖”等。

與具有面積屬性的動詞關聯的名詞性詞語 md，如與種植和播種關聯的有“作物”、“經濟作物”、“糧食作物”，以及具體的作物名稱“水稻”、“小麥”、“棉花”、“茶葉”、“菸草”、“甜菜”、“橡膠”等；與養殖有關的有“魚”、“蝦”等。

具有面積屬性的專名 zm，其類型包括“農場”、“林場”、“風景區”、“自然保護區”以及各種建築物等。

行政區劃類型 xl，包括“省”、“自治區”、“市”、“地區”、“自治州”、“區”、“縣”、“鎮”。

映射：

{td→xq}，由詞目替代詞到詞目本身，如在“江蘇省”釋文中，“全省”映射爲“江蘇省”。

{xq→xl}，由行政區劃名到它本身的行政區劃類型，如由“江蘇省”映射爲“省”。

{xq→xq}，由行政區劃名到它的上級行政區劃名，如由“蘇州市”映射爲“江蘇省”。

{md→dc}，由名詞到與它關聯的具有面積屬性的動詞，如由“棉花”映射爲“種植”。

{mc→mc}，由名詞到它的上級語義名詞，如由“糧食作物”映射爲“作物”。

關係：

數量關係：sl(time, body, number, area-unit, modifier)，即時間、主體、數值、面積單位、修飾成分滿足數量關係。

比例關係：bl(time, body-numerator, body-denominator, ratio, modifier-before, modifier-after)，即時間、分子主體、分母主體、比值、前修飾成分、後修飾成分滿足比例關係。

變化數量關係：bsl(time-before,time-after, body, extend-reduce, number, area-unit, mordify)，即變化前時間、變化後時間、主體、擴縮標記、數值、面積單位、修飾成分滿足變化數量關係。

變化比例關係：bbl(time-before,time-after, body, extend-reduce, ratio, mordify) ，即變化前時間、變化後時間、主體、擴縮標記、比值、修飾成分滿足變化比例關係。

其中面積主體 body 的構成方式爲：

xq [fb[fb]] [(zm | mc | md {md→dc}md)]

式中（ | ）表示選擇，[ ] 表示可有可無。

規則：

規則的作用就是從文本中的適當位置抽取關係中論元所指的內容。規則的形式是：文本模式→關係，其中文本模式列出關係中各論元所指內容在文本中的相對於“面積”的位置。同一個規則中的同一個變元若重複出現，則代表同一個內容。

下面列出一些常用的規則。其中，text-begin 表示篇首，dot-comma 表示句號或逗號，no-area-string 表示一個句串，其中不出現“面積”。string 表示一個句串，其中每個標點句的首詞不帶有 xq、fb、mc、md、zm、time 屬性，而且句串中包含的標點句不超過 6 句。我們這裏所說的句串就是一串標點句，而標點句就是文本中以逗號、句號、分號、嘆號、問號分隔的字串。

text-begin no-area-string dot-comma [總] 面積 [modifier] [爲] number area-unit

→sl(nil, xq, number, area-unit, modifier)

例如：“阿克蘇市”釋文的開始幾句是：

新疆阿克蘇地區轄市和行署駐地,新疆重點墾區。位於塔里木盆地西北部。面積 1.83 萬平方公里,人口 38.13 萬。

匹配規則的條件部分後，得到的數量關係是：

sl(nil，阿克蘇市，1.83 萬，平方公里，nil)

這個關係的 5 個數據“nil”、“阿克蘇市”、“1.83 萬”、“平方公里”、“nil”分別存放在 sl 數據庫的 5 個字段 time、body、number、area-unit、modifier 下，表示在該百科辭典編制時阿克蘇市面積恰爲 1.83 萬平方公里。

dot-comma time string fb 面積 [modifier] [爲] number area-unit

→sl(time, xq td, number, area-unit, modifier)

例如：“安順市”釋文中有：

20 世紀 50 年代以前，城區面積僅 1.4 平方公里，人口 2.4 萬人。

匹配規則的條件部分後，得到的數量關係是：

sl(20 世紀 50 年代以前，安順市城區，1.4，平方公里，僅)

這個關係的 5 個數據“20 世紀 50 年代以前”、“安順市城區”、“1.4”、“平方公里”、“僅”分別存放在 sl 數據庫的 5 個字段 time、body、number、area-unit、modifier 下，表示在 20 世紀 50 年代以前安順市城區面積僅 1.4 平方公里。

dot-comma td string mc 面積 [modifer-before] 占 [td][[總]面積] [的] ratio [modifier-after]

→bl(nil, {td→xq}td mc, {td→xq}td , ratio, modifier-before，modifier-after)

例如："安達市"釋文中有：

市境地形平坦,平均海拔 150 米。草原面積占 51.5%以上，宜發展畜牧。

匹配規則的條件部分後，得到的比例關係是：

bl(nil，安達市市境草原，安達市市境，51.5%，nil，以上)

這個關係的 6 個數據"nil"、"安達市市境草原"、"安達市市境"、"51.5%"、"nil"、"以上"分別存放在 bl 數據庫的 6 個字段 time、body-numerator、body-denominator、ratio、modifier-before，modifier-after 下，表示在該百科辭典編制時安達市市境草原面積占安達市市境面積 51.5%以上。

由於這些被提取出來的信息以關係數據庫的形式存放，所以可以借助數據庫檢索工具來檢索。

4. 測試與討論

我們檢查了中國行政地名詞目按漢語拼音排序 a-d 的 107 個詞目的釋文，這裏面出現"面積"176 次，帶有數量的面積信息 153 條，其中有些是行政區劃本身的面積，有些是行政區劃內部某個分區的面積。使用該系統的規則能夠正確提取信息的 141 條，準確率約爲 92%。其中，上述第 1 條規則使用 94 次，第 2 條規則使用 22 次，全部正確。特別是，107 個詞目釋文中，有 103 個提到了該詞目所代表的行政區劃的面積，它們都出現在靠近篇首的位置，其中 102 條可以用規則將面積信息提取出來，94 條用上述第 1 條規則，8 條用第 2 條規則。發生錯誤的大都是行政區劃內某一部分中某種特定地域的面積，主要問題是面積主體過於複雜。如"安徽省"釋文中有：

皖中丘陵水旱作物過渡區。以水稻、小麥爲主的水旱兼作、一年兩熟區。位於淮河以南、江淮分水嶺—滁河一線以北，土地面積占全省 23.7%，……

該例中，第一句是個小標題，後面幾句是對該標題所涉地區的說明。最後一句中的"面積"的主體是"安徽省皖中丘陵水旱作物過渡區土地"。這一主體的構成方式過於複雜，難以識別。

從這一實驗中可見，

(1) 由概念、映射、關係和規則組成的形式系統可以比較全面準確地表示一些簡單知識在百科辭典文本中的形式，這一個基於語義屬性的形式系統的框架設計是成功的。

(2) 爲了構造這一形式系統需要做大量的人工調查、分析、標注工作。若信息提取的焦點不變而僅僅換掉文本（把中國大百科全書換成其他百科辭典文

本），則由於各種百科辭典中同種知識的表達形式基本上是有限多種，所以增大的人工工作量不會太大，這是這種做法的優越性所在。但當信息提取的焦點改變（比如從提取行政區劃的面積知識轉而要提取動物的生活習性知識）時，人工的投入量仍然會相當大。爲此，必須研究這類形式系統自動（或半自動）生成的方法，這將是我們的下一步工作。

鳴謝

本文得到中國國家自然科學基金（60141001）和國家高技術計劃（2001AA114111）的資助，謹在此致謝。

參考文獻

Tsujii, J., “Generic NLP Technologies: Language, Knowledge and Information Extraction”, *Proc. of ACL2000*, 2000, pp.11-18.

Hull, R., and Gomez, F., “Automatic acquisition of biographic knowledge from encyclopedic texts”, *Expert Systems with Applications* 16(1999), pp.261-270.

Soderland, W. D., Fisher, J. Aseltine, and W. Lehnert, “CRYSTAL: Inducing a Conceptual Dictionary”, *Proc. of the International Joint Conference on Artificial Intelligence*, Montreal, Canada, 1995, pp. 1314-1319.

Gu, F. and Cao, C., “Biological Knowledge Acquisition From the Electronic Encyclopedia of China”, *Proc. of ICYCS’2001*, 2001, pp.1199-1203.