基于短語分析和動態語義搭配的拼音漢字智能轉換*

張小衡　陸發興
(香港理工大學中文及雙語學系)

摘要：關于拼音漢字智能轉換的研究大都致力于整句分析轉換。這樣做難度很大，且在漢字鍵盤輸入的應用中，修改錯誤的漢字也不方便。本文提出一個新途徑：基於短語分析，并採用動態產生的語義搭配模扳。音字轉換從左至右，漸進處理，使同音詞歧義得以在最小的前后文語境內解決。

關鍵詞：短語分析 語義搭配 拼音漢字智能轉換

Intelligent Chinese Pinyin-Character Conversion Based on Phrase Analysis and Dynamic Semantic Collocation

Zhang Xiaoheng, Lu Faxing
(Dept. of Chinese & Bilingual Studies, Hong Kong Polytechnic University)

Abstract: Research on Chinese pinyin-character conversion has been concentrated on whole sentence analysis and whole sentence conversion, which is extremely difficult and has the side-effect of inconvenience for error correction in Chinese character input. This paper introduces a new method: conversion based on hierarchical phrase analysis and word collocation templates dynamically generated for corpus matching. Conversion of a pinyin word string is carried out from left to right step by step and homophone word ambiguity is tackled within minimally sufficient contexts.

Key words: phrase analysis, semantic collocation, Chinese Pinyin-character conversion

一、引言

學術界對拼音-漢字智能轉換的研究已走過了十年的歲月，成果不少，但尚未形成實用性較強的產品，漢字輸入的市場仍被非智能軟件所佔領。仔細觀察一些代表性系統[1,2,3,4]，不難看出兩個特點：(1) 整句分析整句轉換，力圖全自動化；否則 (2)由用戶在同音詞中手工選取。前者困難極大，因漢語語法研究仍不夠完善，編製通用的語句分析程序不太實際。而後者又回到問題解決前的狀態，很少有智能可言。

居于句子和單詞之間的語言單位是短語。短語結構通常比句子簡單，又能提供比單詞豐富得多的上下文信息。有關短語的語言學研究也比較成熟。加之，短語靈活性很大，可以僅由一兩個詞組成，也可以是整句。這說明短語分析法是很有發展前途的，可由簡到繁，

* 香港理工大學資助課題。

循序漸進。如在基於短語語法分析的基礎上，同時考慮詞語的語義搭配，效果會更佳。本文將沿著這一新途徑探討拼音漢字智能轉換。

基于短語分析的拼音-漢字轉換還有重要的心理學依據。人類在語言交際中往往聽到少數幾個詞後就能理解這組詞所表達的意思并作音字轉換，否則聽演講作筆記和同聲翻譯等活動都是不可能的了。其實，語言學家曾通過實驗證明，在即時對話中，有一些稱為思想單位(idea unit)的基本組合元素，大概六個英文詞長或兩秒之久[11]。這種單位想必是以短語的形式出現的。因此拼音漢字轉換並非一定得整句進行。

二、中文的基本短語簡介

短語，也叫詞組，是能自由運用的詞的組合。短語一般是由三條結構原則建立起來的[5]，即：雙成份性，層次性與向心性。所謂雙成份性是指每一個短語都是由兩個成分構成。所以長的短語可以切分成短的短語，逐步二分，直到構成短語的詞。反過來說，詞可以兩兩組成短語，短的短語兩兩組成長的短語，直至成句。這就體現了短語的層次性。短語的雙成份性和層次性意味著，短語(包括成句的短語)一般可以分解為一棵二叉樹，其根是該短語，葉是單詞，中間節點為各級“部件”短語。短語的向心性是指不同層次上的短語各有其中心單詞或中心短語，中心詞語決定短語的基本語法語義性質。

根據其結構的緊密度，可將短語分為三類[5]：(1)成語，如“事半功倍”，“狐假虎威”，(2)名稱性短語，如“清華大學”，“香港特別行政區”，“魯迅先生”，和(3)自由短語，如“偉大的祖國”，“學習中文”。成語性短語結構固定，數目有限，可按單詞處理。名稱性短語屬專有名詞，主要包括人名，地名和機構名。根據國家規範[7]，我們要求待轉換成漢字的用拼音表示的專有名詞中各部件詞的首字母應大寫，如“Qinghua Daxue(清華大學)”，“Luxun Xiansheng(魯迅先生)”。自由短語是最為普通的短語，內部結構鬆散，可按一定的模式置換。如“讀書”可以換城“寫書”，“吃飯”可以換成“吃菜”。本文主要關心的是自由短語。自由短語可進一步分為[5,6]:

1) 動詞短語：是以動詞為中心語的短語，包括述賓型短語(也叫動賓短語，如“上大學”，“是老師”)、述補型短語(如“打破”，“看清楚”)、狀動型短語(如“快跑”，“感慨地說”)、兼語短語(如“請他來”)和聯動短語(如“出去打球”)。

2) 名詞短語：具有名詞功能的短語，包括定名型短語(如“廣州人”，“藍天”，“寫的書”，“兩個蘋果”)、名位型短語(如，“門前”，“河邊”)，同位短語(如“首都北京”，廠長李明)和“的”字短語(如“吃的”(動詞+的)，“銅的”(名)，“大的”(形))。

3) 形容詞短語：以形容詞為中心語，包括狀形短語(如“不漂亮”，“很好”)和形補型短語(如:“好極了”，“好得很”)。

4) 主謂短語：常充當句子，由主語(常由名詞短語承擔)加謂語(常由動詞短語或形容詞短語承擔)組成，如“春天來了”，“空氣新鮮”。

5) 介詞短語：介詞加名詞短語組成，如“從北京”，“把他”，“對于那件事”。

6) 并列短語：上述各種短語的組織結構還可以是并列的。如兩個或兩個以上的名詞或名詞短語可構成并列型的名詞短語。如“老師和學生”，“正確的答案和錯誤的答案”。

為方便計算機信息處理，我們規定在句子中獨立充當某種成分的動詞，名詞等實詞可單獨構成短語。 這與國際慣例是一致的。

三、關于面嚮中文輸入的短語分析的一些思考

智能拼音漢字轉換的主要途徑是通過語句分析獲取上下文信息，并以此為限制條件去解決中文詞的同音歧義問題。這種分析與一般自然語言理解中的分析不完全一樣。首先，它處理的是一個拼音詞串，分析的深度比較淺，旨在確定與拼音串相對應的正確漢字串，而不需進一步理解該漢字串所表達的意思。此外，一般自然語言處理中的分析往往是對一個已輸入的語言單位做處理， 可自左向右，或自右向左。 由于漢語短語大多為“後中心”結構，因此自右向左可能還方便些。但對于漢字輸入，這就意味著要等用戶輸入完整個短語甚至整個句子後才能分析轉換成漢字。如有轉換不妥，則用戶需(長距離)回移光標來改錯。因此，面向中文輸入的拼音短語分析宜遵循人們的書寫習慣，從左至右，由小的語言單位到大的語言單位，隨時轉換。類似于秘書聽口授寫書信。

此外，在輸入漢語時，我們可要求用戶分詞輸入，但不能要求劃分短語。因為詞是語言思維的基本單位，分詞連寫又是拼音語句的書寫規範[7]，而短語的概念對一般人來說卻比較陌生。因此，面向中文輸入的短語分析首先應有識別各種短語的能力。

在詞語分析過程中，要特別注意相鄰兩個詞或短語之間的關係。在語法上，兩個相鄰詞或短語有兩種基本關係:

1) 這兩個語言單位直接組成某種短語。這時不僅要滿足短語語法規則的要求，還要受到語義搭配的限制。如:可以說“讀書”，“吃飯”，不能說“讀飯”，“吃書”。

2) 兩者不直接構成短語，如“放心 從這裡(經過)”，“我們 經常(工作)”。

對于情形 1)，應該及時合成短語，存蓄有關信息，並利用新短語所提供的種種限制條件，排除拼音-漢字轉換中的歧義對應。對于情形 2)，應該暫時將兩個語言單位分開記錄，等包含他們的公共短語出現時才合成。

四、系統設計

4.1 系統程序的基本算法:

```
Begin Program
/* 要求用戶分詞連寫逐句輸入拼音(可帶調、不帶調或部分帶調)*/
Repeat
    (自左至右)讀新句的第一個拼音詞，從拼音-漢字詞典中找到相應的(同音)
    漢字詞；
    Repeat
        輸出新短語/詞中的無重碼漢字詞，取代用戶輸入的相應拼音詞；
        新短語/詞入棧；
```

```
            讀下一個拼音詞；
            while 能與(位于棧頂的)前驅短語/詞合成短語，
                  彈出前驅詞語，合成新短語;
            endwhile ;
        Until 讀到“。”，“！”和“？”等句末標點符號。
    Until 讀到“。。”
    End  Program
```

程序逐句處理拼音輸入。用戶以兩個句號“。。”退出拼音-漢字轉換系統。 每句的分析是自左至右，遞進處理，將拼音詞隨時轉換成漢字詞。程序用堆棧的技術來模擬短語的二分性和層次性。系統中最重要的工作是判斷當前詞或短語是否能同前驅(棧頂)語言單位合成新的短語。這要用到語法語義的詞語搭配要求，主要依據有：

1) 短語規則(如本文第二節所述)，說明各種短語的可能結構和不可能結構。如：及物動詞後緊接名詞短語構成述賓短語，不及物動詞後不能緊接名詞短語
2) 搭配詞典，列出各詞的語法語義搭配(如 [8])
3) 查閱拼音-漢字短語對照表
4) 使用標點符號和關聯詞等詞組邊界標記
5) 使用短語模板

4.2 動態模板法

以往的短語模板一般是人工歸納的[9]，在系統中是靜態的。然而我們的設想是讓系統根據具體的同音歧義問題和語料匹配情形動態產生和變動模板。為此系統中應有一個語料庫和一個多層次的詞語語義分類表(樹形結構)，如圖 1 所示：

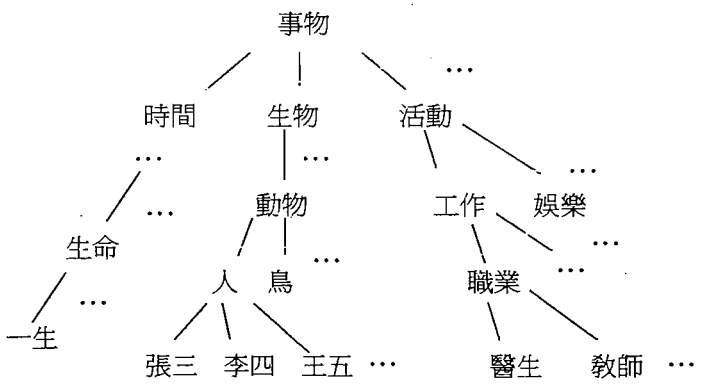

圖 1 ：名詞語義分類示意圖

在計算機中，可用一組表示成員-集體關系的謂詞事實語句來表示，如：

IS-A(醫生, 職業);

IS-A(張三, 人);

IS-A(人, 動物).

除了語義分類樹之外，系統中還應有語法語義類別信息足夠詳細的詞典，其實語義分類樹也可在詞典中實現。如果所用語料庫是作了相應的語法語義標注的，那就更好了。

簡單地說，動態模板法的基本工作原理是：對于一個含同音歧義的拼音詞串，根據語義分類樹和詞典信息，為該拼音詞串的每種可能的漢字詞解釋產生一個最低層的模板，然后去語料庫中匹配。如有成功的，則按其中匹配成功次數最多的模板的相應的解釋漢字作轉換。否則，再根據語義分類樹和詞典，稍微放寬要求，修改模板，然后再去語料庫中匹配。如此重復，直到某個模板匹配成功或模板間無區別。后種情況是模板法不能解決的問題，原因可能是所處理的拼音詞串不是一個短語，或含有語法錯誤。

五、實例

例 1: 短語分析法基本算法的應用

如在“南極星”中文文書系統中輸入拼音“ yinggai nuli de xuexi 。”，則會得到“應該奴隸的學習”。這是單純按“高頻優先”的原則處理同音歧義，逐詞轉換得來的。現在，讓我們看看根據上節介紹的基本算法設計，這個拼音-漢字轉換問題是如何解決的。

當用戶輸入第一個拼音詞“ yinggai ”時， 拼音-漢字轉換系統讀取該詞，查詞典得知，該拼音串對應漢字詞“應該”，無同音重碼(根據“南極星”)。故輸出漢字詞“應該”，取代拼音詞“ yinggai ”。詞“應該”(連同從詞典中查取的有關信息)入棧。 接著，系統讀入下一個詞“ nuli ”，對應詞典中的“奴隸”和“努力”。因為“應該”是個助動詞，根據語法規則它后面一般只能接動詞或動詞短語，不能緊接名詞，因此優先選擇候選詞“努力”。由于“努力”有后繼詞，在這里當副詞用，同后面的詞語直接構成短語。這就意味著“努力”和其前驅詞“應該”不直接構成短語，故根據算法，輸出新詞“努力”，然后“努力”的有關信息進棧。這時的情形如圖 2(a)所示。

	堆棧	轉換現狀
(a)	[應該][努力]	應該 努力 de xuexi 。
(b)	[應該][ADVP,努力,地]	應該 努力 地 xuexi 。
(c)	[VP,應該,[VP,[ADVP,努力,地],學習]]	應該 努力 地 學習。

圖 2 ：“ yinggai nuli de xuexi 。”的拼音-漢字轉換過程

接著讀入“ de ”，對應結構助詞“的”、“得”和“地”。因為這三個同音詞中只有“地”能同前驅副詞“努力”構成短語(屬狀語短語(ADVP))，故“努力”出棧，同“地”合成短語“[ADVP,努力,地]”。該短語不同前驅詞“應該”直接合成短語，故輸出漢字詞“地”，新短語入棧，如圖 2(b)所示。

最後讀“xuexi”。對應“學習”和“血洗”。由于“血洗”只能當及物動詞，后面必須跟名詞短語，而現在後面只有句號，所以這個詞不可能。因為“學習”既可當及物動詞，又可當不及物動詞。因此“xuexi”只能對應“學習”，而且是不及物動詞。“學習”能同前驅短語“努力地”構成動詞短語“[VP,[ADVP,努力,地],學習]”。故合并成新短語。該新短語又同前驅詞“應該”構成更高層的動詞短語“[VP,應該,[VP,[ADVP,努力,地],學習]]”。最后，輸出“學習”，新短語進棧，如圖2(c)所示。

例 2: 動態模板法的應用

在一般系統中輸入“yisheng Zhangsan”，很可能會得到“一生張三”。因為同音詞“一生”比“醫生”在語料庫中出現的頻率較高。但這不是一個合法的詞組。按照我們的動態模板方法，系統首先查詞典確定拼音“yisheng Zhangsan”的兩個可能漢字解釋：“一生張三”和“醫生張三”，并查看語料庫中有無它們的實例，結果前者沒有，後者也可能沒有(如有的話問題就解決了)。接著，根據語義分類樹(圖1)，用“張三”的最小語義類別代入，生成模板“一生<人名>”和“醫生<人名>”，並用它們去語料庫中匹配。結果第一個模板仍匹配失敗，如第二個摸板匹配上“醫生李四”，則說明正確的答案應是“醫生張三”。如第二個模板也匹配不上，則可根據語義分類樹進一步放寬要求，產生新模板，如“<生命><人名>”和“<職業稱呼><人名>”。結果第一模板仍不成功。而第二模板很可能成功，例如匹配上語料庫中的“教授王五”等。萬一不成，還可根據更上一層的語義詞類產生覆蓋面更廣的新模板。可見這種方法自動化程度高，靈活性大，而且可以做到有的放矢，在最小的語義場中解決歧義。

例 3. 非相鄰詞語匹配

待轉換的拼音詞串：“jiu3 suo3 Beijing Shi daxue。”

本例我們只簡單解釋其拼音-漢字轉換過程。先看前兩個拼音詞。由于“jiu3(九，酒，久等)”和“suo3(所，鎖，索等)”兩組同音詞中，可形成數量詞組“九所”，該詞組進棧。對應于“Beijing”的詞有“北京”、“背景”和“北進”；對應于“Shi”的有“是”、“事”和“市”等。但由于這兩個拼音詞的首字母大寫，表示一個專有名詞，因此它們對應的漢字應該是“北京 市”。由于根據漢語語法，數量詞組，尤其是表示復數的數量詞組(如：九所)，不能修飾專有名詞，因此，“北京 市”進棧。最后讀入“daxue”，對應“大學”和“大雪”。語法上這兩個詞都可同前驅詞組“北京市”組成短語，即“北京市大學”和“北京市大雪”。但這兩個候選短語中只有前者能同數量詞組“九所”構成更大的詞組。因此，“daxue”，應該對應“大學”。結果，系統將“jiu3 suo3 Beijing Shi daxue。”成功轉換成“九所 北京 市 大學”。這里有一點要特別引起大家注意的，就是我們的方法能夠根據短語的層次性，通過非相鄰(甚至遠距離)的詞語搭配(如：“九所 …… daxue”)，解決鄰接搭配(如：“北京市 daxue”)無法解決的同音歧義問題(如：“大雪/大學”)。

六. 結論與討論

關于拼音漢字智能轉換的研究大都致力于整句分析轉換，這樣難度很大，修改漢字輸

入錯誤也不方便。本文介紹了一種新的方法，它的主要特點有：

- 基於層次性的短語分析，使轉換歧義得以在最小範圍內解決。
- 由同音歧義問題本身來驅動，動態產生和修改短語模板，提高智能和自動化程度。
- 由左至右，實時漸進轉換，與用戶寫作習慣一致，提高方便度。
- 考慮詞語的相鄰搭配和非相鄰搭配。
- 語料庫和語言規則相結合，分析結果可用于其它應用領域。

這種方法還可能有助于解決一些其它拼音漢字智能轉換系統難以解決的問題。 如，據報道[2,3]，基于規則的拼音-漢字轉換系統無法區分語法語義同類的同音詞，如：其他-其它，第一縣-第一線，制定—制訂，僅僅—緊緊，等等。其中有一些是可以按我們的方法來解決的， 例如：假設用戶輸入"qita xuesheng"，分別對應"其他/其它"和"學生"。如系統能在語料庫中找到與"qita xuesheng"對應的短語(應該是"其他學生")，則問題就解決了。如不能，根據我們的設計，系統會通過語義分類樹，放寬"xuesheng"的語義要求，產生模板"qita <人>"，去語料庫中匹配，這時很可能會成功，例如匹配上"其他教師"。于是計算機便可作出判斷：拼音"qita xuesheng"對應"其他學生"，而不是"其它學生"。

對于本文提出的音詞轉換新方法，我們僅僅作了些初步探索，有許多不完善的地方。編程工作剛起步，采用 C++和 FoxPro，目前正在建立詞典。這些方法還可同其它方法結合，例如作為高頻先見法或基于語料庫的統計法[10]的前處理或后處理。另一個需要深入研究的問題是：為進一步提高拼音-漢字轉換的效率，在做短語分析時，不僅要考慮前驅詞語和當前詞語的關系，還應同時考慮后繼詞語所提供的信息。為此，可借鑒 Marcus 的句法分析法[12]。此外，由于短語分析難免出錯， 因此還應考慮回溯推理。

參考文獻

[1] 俞士汶，中文輸入中語法分析技術的應用，《中文信息學報》，1988，第 2 卷第 3 期。

[2] 王曉龍，拼音語句漢字輸入系統 InSun，《中文信息學報》，1993，第 2 期。

[3] 成華，尹寶林，一個拼音漢字自動轉換系統的設計與實現，《計算語言學進展與應用》，清華大學出版社，1995，pp340-345。

[4] 萬建城，語音代碼--漢字智能轉換研究，《中文信息學報》，第 8 卷第 2 期，1994。

[5] 胡樹鮮，《現代漢語語法理論初探》，中國人民大學出版社，1990。

[6] 範曉，《漢語的短語》，商務印書館，北京，1991。

[7] 國家教委語委，漢語拼音正詞法基本規則，《語言文字規範手冊》，語文出版社，1993，pp293-307。

[8] 張壽康、林杏光主編，《現代漢語實詞搭配詞典》，商務印書館，1992。

[9] 許文廉、陳克健，[國音]智慧型輸入系統的語義分析[脈絡會意法]，《計算語言學研究與應用》，北京語言學院出版社，1993，pp338-343。

[10] 吳軍，王作英，郭進，王政賢，一種基于語言理解的輸入方法---智能拼音輸入法，《中文信息學報》，1996 年第 2 期。

[11] Chafe, W. Integration and involvement in speaking, writing, and oral literature, In D. Tannen(ed), *Spoken and Written Language*, ABLEX, 1982。

[12] Marcus, M., *A Theory of Syntactic Recognition for Natural Language*, MIT Press, 1980.