

“俯下身來”和電腦講話

——對《受限漢語》的反思

北京信息工程學院·周錫令

xlzhou0421@vip.sina.com

2007-4-10

一、前言

（一）在“理解”的基礎上進行自然語言處理

以機器翻譯為代表的電腦自然語言處理已經進行了很多年，有很長一段時間人們認為這種工作可以歸結為機械性質的“字串的代換”工作，因此不需要對所處理的文字內容有所理解。在此看法基礎上進行的工作一開始還有一定的成效，然後就長期停滯不前了。

多年的經驗使得越來越多從事自然語言處理的工作者相信：要突破這一壁壘，使“電腦自然語言處理”工作更上一層樓，必須讓電腦對所處理的文字內容有某種程度的理解。那麼，作為一種機械的電腦在企圖理解自然語言的時候，會遇到什麼樣的問題呢？

（二）俯下身來和電腦講話

大家都知道，說話要看對象。和不同的對象說話，要使用不同的辭彙、句型、語氣……，這樣才能使對方理解自己的意思。把在大學專業課堂上的講話方式拿來講給小學生聽，肯定會失敗。你必須俯下身來，用兒童熟悉的語彙和簡單的句式敘述。同樣地，把用來在“人與人”之間溝通資訊的“自然語言”原封不動地交給既沒有感知周圍世界的“五官”，智力和知識水準又比“人”差遠了的電腦去處理，當然也會“問題百出”。

《受限漢語》就是基於以上認識而提出一個研究課題¹。我在1990—1995期間從事這一課題的研究時，心中一個非常樸實的想法就是：不管電腦怎樣“笨”，我們只要把話語說得儘量簡單明白，終究可以達到電腦可以正確處理的地步。需要解決的問題僅在於，我們自己能否忍受這樣十分繁瑣和水準低下的談話方式。而《受限漢語》課題所研究的內容就是：（1）擬定一套簡單而又完備的“句型”，它應當簡單得讓電腦很容易進行句法分析，但是又足夠“完備”，具有足夠的表達能力；（2）開發一個軟體工具，引導用戶使用這些句型來寫出自己所要表達的意思。

（三）怎樣才算“俯下身來”

¹ . 參見俞士汶、朱學鋒著，〈受限漢語研究的必要性〉，收入王均主編《語文現代化論叢第三輯》，語文出版社，1997年10月，P150-160，亦見於：

http://icl.pku.edu.cn/icl_tr/collected_papers/chinese/collection-3/16-SHXHYBYX.htm

句子、文章都是有內部結構的。但是對電腦說來，他所收到的乃是一維的字元流（character stream）。要從這“一維的字元流”中“看出”裏面的語法結構，不是一件容易的事情。《受限漢語》的目標就是要將允許用戶使用的語句的結構複雜性降到電腦能夠處理的程度。從這個觀點來看，當初的這個出發點也不算錯，但是這並不是問題的全部。

經驗表明，這一做法遇到了兩方面的困難：

首先，自然語言中的符號串所傳遞的資訊只是對語境（對談雙方共用的語境）中所含有的大量豐富資訊的一種補充。因此，“語境中已為雙方所共知的資訊”往往可以在語句中被省略。要是遇到這種情況，無論你把句子的結構限制到怎樣簡單的地步，電腦都會暈頭轉向。

其次，人們在說話時，從來不會先在心中意識到自己要去使用哪一種句型，只是根據自己所要表達的語意方面的要求，一個概念一個概念地說下去。因此，要求他一面遣詞造句，一面還要照顧到電腦的水準去選擇句型來說話，實在是一種令他感到非常彘扭的事情。

二、語言溝通的環境

自然語言只是對語境的補充，當對話雙方都處在一定的環境中，不需要通過語言符號就可以從周圍環境中提取的資訊可以稱之為“語境資訊”。能夠提供“語境資訊”的環境有許多層次：

（一）物理環境

1. 促膝談心：可以說是談話雙方共用物理環境最多的一種情況。周圍的風光，雙方的眼神，面部表情，肢體動作，話語的聲調……都可以傳達很多資訊，甚至是語言符號所難以傳達的微妙的資訊。如果話題所涉及的物件就在雙方可視、可聽聞的範圍內，句子中即使缺少指稱這一物件的成分，也不妨礙對方的理解。例如，一面向視走進公園大門的某人，一面說：“來了，來了” 這個句子中缺省的主語（「某某人」來了？）已經由說話人的目光傳遞給聽話的人了。
2. 視頻對話：就差了一點。這時雙方各自所處的已經是兩個不同的物理環境。但是雙方的面部表情，話語的聲調，甚至肢體動作……還可以互相看見，仍然可以通過它們傳遞一些資訊。
3. 電話交談：又差了一點，面部表情，肢體動作……也看不到了。
4. 書面信件往來：更是退了一步。大家都知道，同樣一句話，配上不同的面部表情和聲調，往往具有不同甚至是相反的含義。面部表情和聲調的缺失常常引起誤會。比如說，一面臉上帶著誠摯的微笑，一面說：“你真行！”，和板著臉說這三個字時的含義完全相反。因此落筆寫信時，在遣詞造句方面就應該比口頭談話要採取更加慎重的態度。這一情況也說明：當前在 MSN 和 QQ 平臺上通過短信交流時，在書寫文字的同時，善於使用那些“表情符號”是如何的重要。

從“促膝談心”到“視頻對話”到“電話交談”到“書面信件來往”，對話雙方共用的物理環境一步一步退化，對於說話人的說話水準就要求越來越高。於是討論比較微妙敏感的話題時，經常聽到人們說：“電話裏講不清楚，我們還是約一個時間當面討論吧！”

由於電腦當前都沒有具備視、聽、聞方面的感官和相應的感知能力，所以人和電腦的對話被限制在類似“書面信件來往”的低水準共用環境中，所以話語的組織需要特別完備和嚴密，才能達到互相正確理解的目的。

（二）知識環境

對話雙方除了“身處其中”的物理環境以外，還廣泛利用雙方共同的知識文化背景。舉例如下：

1. 長期的公共知識背景：公共知識按人群劃分。若以一般成年人作為一個群體，它們的公共知識就是“常識”。若按照人們從事工作的專業領域來劃分群體，那麼他們共有的專業知識也是他們的“公共知識背景”的一部分。
2. 短期的知識背景：如果 A 和 B 討論一部兩人都在前不久看過的電影，或者共同目睹過的某一事件的完整過程，那麼他們所說的話，比起和另一位沒有看過這部電影或者沒有目睹過該事件發展過程的老兄 C 討論時，就可以簡短得多。
3. 前文傳遞的知識背景：A 在和 B 說話的過程中，已經說過並被 B 聽到的話（前文）構成了另一種語境，也就是通常所說的“前後文（context）語境”。

（三）公共知識背景內容的引用

對話雙方有了公共知識背景，談起話來就可以採取“引用”的方式使對話變得快捷、高效和濃縮。在自然語言中存在哪些“引用方式”及其分類是另一個需要單獨研究的專門題目，以下舉二個例子說明：

1. 肢體語言：在“促膝談心”式的對話中，可以用肢體語言來指明使用周圍語境環境裏的某一物件來取代被省略的主語或者話題。例如前面所舉例子中用眼睛注視某一人物，說：“來了，來了”。還可以利用“眨眼”，“諷刺性的冷笑”使所說的話的實際含義和字面上的含義相反，或者用眼神來補足說話語言語意的不足，用眼神來傳遞「只可意會而不可言傳」的語意。
2. 成語和典故：這個過程就好像在電腦程式裏調用副程式庫中的副程式一樣。如果把古代發生過的知名故事看成一個“成語和典故庫”，那麼漢語中的大量成語和典故就是引用這些成語和典故的副程式名。

三、“標注”比“受限”更重要

用不著說，目前的電腦還沒有配備如人腦所擁有那樣廣博的知識庫，即使我們作出巨大的努力，為電腦建立起一個具有相當規模的知識庫，它也不具有像中華文化經常要求的“舉一反三”那樣“靈活而巧妙的運用”知識的能力。

既然如此，我們對電腦說話的時候就只好“俯下身來”。各種編程語言

(programming language) 就是採取這一態度的成果。但是，當今的編程語言和在人和人交流中所產生的自然語言之間的差距太大了，而且這些編程語言主要是用來描述“演算法 (Algorithm)”的。因此，現在我們要討論的是，怎樣在自然語言這一基點上採取措施來“俯就”電腦。

這一點我們可以在生活中找到很好的先例，那就是許多學者和老師在普及中華古典文獻的所採取的措施：

- (一) 形式方面，通過“標點符號”對文章的內部結構進行標注。這是古文中從來沒有過的東西。
- (二) 語意方面，對於因年代久遠以至後人難以捉摸的典故（歷史故事），則在正文中添加注解。比如說，文章中有“他家裏這麼困難，你給他補助一百塊錢，簡直是‘杯水車薪’。”這句話，就在相應注解中敘述‘杯水車薪’所指的這個故事。
- (三) 與白話文比較之下相對十分精煉的古文中，往往省略一些句子成分；在詩詞等韻文中爲了照顧格律的要求而違反通常的習慣，使用倒裝句或縮略句。對此，需要把省略的東西補上去，把倒裝句糾正過來。

對於電腦，也可以採取類似的措施：

- (一) 現代漢語的“標點符號系統”對文章內部結構進行的形式標注非常有效，但是對於智力低下的電腦來說，這一系統的標注能力還不夠，還需要擴展。如何擴展呢？這是一個有待研究的問題。不過有些想法已經初見端倪，比如說：首先希望在“語塊”之間添加分隔符號；其次，能不能把句子的“中心動詞(heading verb)”標注出來？第三，在名詞和動詞容易混淆的地方能不能也加上詞性標注？
- (二) 在注解中敘述相應成語或者典故所指的故事，對於電腦來說還不夠，因爲它即使知道了這個故事，也沒有能力“舉一反三”地把“杯水車薪”理解爲“錢不夠”。所以要擴大詞典，把這些典故與成語都包含進去，直接解釋爲“給的東西太少，解決不了問題。”
- (三) 對於句子中被省略的成分，最好把省略的東西補上去；對於違反通常習慣的倒裝句，要把它糾正過來。
- (四) 簡化句子結構，把結構過於複雜以及太長的句子分解爲幾個短句，這是《受限漢語》課題中考慮過的問題。

四、和“反對意見”的討論

- (一) 如何讓計算機與作者對話

開展“自然語言處理”的目標，本來就是要千方百計的克服困難，提高電腦的智慧，讓電腦自動地來處理自然語言。現在你要“人”“俯下身來”將就電腦，豈能見困難就退？

這裏頭的確有一種不得已的折衷和妥協。要麼你將就把電腦根本無力對付的文章讓它處理得一塌糊塗地交還給你。這樣做的結果往往是把事情

弄得“越幫越忙”。你要是有過審閱英漢機器翻譯生產出來的漢語譯文的經驗，就會知道，看這種天外來客式的文章比起你直接讀英文原文來要頭痛的多！；要麼你花一些力氣，把本來是給“人”看的文章改造得為適合“電腦”看的文章。

這樣想，我們至少開闢了一個新的技術工作面。以“機器翻譯”為例。假定有一位不懂英文的人想要把自己寫的一篇文章翻譯為英文。由於他不懂英文，所以就借助於“中翻英”軟體來做這件工作。又由於他不懂英文，所以對於機器翻譯的結果無法檢查。現在設想放在他面前的“中翻英”軟體有兩種工作狀態：一種是直接把給人看的文章交給它，這樣翻譯出來的英文品質很差。另外一種是請你事先對這篇原來是給人看的文章按照電腦的水準作一些標注和簡化，然後再進行翻譯。在翻譯過程中，計算機遇到拿不准的情況時還可以和你對話，交流。這樣翻譯的結果，品質肯定會高很多。

（二）有功夫和耐心為電腦進行逐篇文章的標注

文章那麼多，誰有功夫和耐心為電腦進行逐篇文章的標注呢？

如果寫作人嫌標注工作麻煩，也可以委託專業人員來做，就像當年的打字員。或許還可以形成一個新的產業或是職業呢？

（三）同的人標注出來的結果會不會有很大差異？

不同的人作出的標注肯定會有差異，所以需要為標注工作制定一個規範，按照這個規範去培訓人才。並且設計出與之對應的軟體工具來幫助規範的執行。

（四）軟體工具能不能也幫助減輕一些標注工作的負擔，甚至實現一定程度的標注自動化？

我們當然可以要求電腦儘量“自動”進行標注，並且把它所作的標注結果顯示給作者看，請他檢查這些標注做得對不對。這就在一定程度上相當於電腦以某種方式把它對作者文章的“理解”回饋給作者看；不對的地方請作者糾正。人與人之間的交流不也是這樣子的嗎？以往的機器翻譯完全省略了這種回饋核對的過程，實際上這是對電腦提出了比對人還要高的要求！

值得指出的是，所有這些交流過程都是在文章作者的“母語”領域內完成的。

（五）如何試驗“標注”的效果？

可不可以先作一些小規模試驗，看看“標注”的效果如何？應當是可以的。

比如說，首先以標注出每個句子中的“中心動詞”為例。看看在“標注”和“不標注”兩種情況下漢英翻譯的品質有多大提高。

另外值得實驗的一個問題就是：“空格”的加入密度。如果像最初提出這個問題時那樣要求每一個詞左右兩邊都留空格，就會覺得太繁瑣，因

為在大多數情況下，“分詞”軟體已經做得不錯，只有在容易混淆的情況下才需要添加空格；而且在漢語中什麼叫“詞”，什麼叫“短語”，什麼叫“語意塊”，往往是連續過渡的，沒有截然的界限。

一種模糊地規定可以這樣表述：當我們拿起稿子念一個長句時，凡是你要歇一歇換口氣的地方，最好就在這裏插入一個空格。如果把這一點納入規範的要求，也可以通過已有的翻譯軟體的前端（句法分析部分）看看“空格”的加入密度怎樣才算合適。

五、課題分解

要是覺得上面敘述的意見有一定的道理，那麼接著就要問：按照這種思路，接下來我們應該怎樣開展工作呢？

（一）標點符號系統的擴展

標注的添加顯然可以做得非常細緻，但是這樣一來，就增加了工作量。做得太粗，電腦又很困難。所以要確定分界線放在那裏為好。

要做好這個工作，不妨首先去尋找那些對於電腦幫助最大，而人添加起來也不怎麼費事的標注。我認為這類標注包括：

1. 在“現代漢語標點符號系統”中添加“半個空格”，用來作為“語塊”或者“容易混淆的片語”之間的分隔符號。
2. 把句子的“中心動詞”標注出來。可以用粗字體，或者特殊的括弧。給電腦看的這種標注在顯示給人看時，樣子可以不同。
3. 在名詞和動詞容易混淆的地方能不能也加上詞性標注？²
4. 對於內部結構過於複雜的句子，利用括弧把其中的層次表現出來。³

（二）幫助“補足缺省成分”的人機交互輔助工具

以上已經說明，人們在相互交流時，常常借助內外兩種“語境”，說話時把許多本來應該說出的成分略去不說，從而達到簡潔，快速的目的。這種現象，在古漢語中特別明顯。一個句子中總是保留最重要的“中心動詞”。而動作的發起者（主語）和動作的物件（賓語）這兩種名詞性成分常常被省略。這些被省略的成分當然都可以通過語境（前後文，當時的情境背景（Situation）……）猜出來。但是要現代人做到這一點，很要絞一些腦汁。參見附錄所舉例子。

（三）擬定句型集合

交給電腦處理的句子在形式上當然也不能太複雜和冗長。原先進行《受限漢語》課題的時候，目標之一就是要尋找一組有足夠表達能力，而對電腦來說又比較容易處理的句型集合。這個任務初看簡單，但是論證起來（你

² . 參見拙作《懂了以後才能看懂的句子》

（http://www.nlp.org.cn/docs/doclist.php?cat_id=24&type=10&start_page=0&start=20&range=20），收於「中文自然語言處理開放平台」（Cnlp Platform）網站（<http://www.nlp.org.cn/>）。

³ . 同註 2。

怎樣證明你所選擇的這一套句型集合是最合理的？）卻不容易。

另外一個困難是怎樣引導“人”按照你所選擇的“句型集合”來寫他的文章？或者怎樣把他按照原來的習慣寫出的文章重新改造為用你所選擇的“句型集合”來寫的文章？

六、結語

我們大家都注意到資訊時代的兩個現象：一是人們使用的電腦越來越多，許多辦公室裏，電腦的數量接近乃至超過工作人員的數量。其二，資訊爆炸的結果，使得人們在工作時需要涉獵的資料越來越多，資料的數量比人們有時間閱讀的數量多好多倍。

怎麼辦呢？唯一的辦法就是讓電腦預先替我們泛讀這些海量的資料，然後從中選擇出比較有價值的，再交給工作人員親自閱讀。

換句話說，以後我們寫文章的時候，我們的心目中不能忽視電腦的存在，因為它們也是我們所寫文章的重要讀者。於是“俯下身來和電腦講話”就成為一種必要的心態。讓電腦對我們有更多的理解，然後它們就會為我們作出更多的回報。

附錄：對古文進行標注的例子

〈曹劌論戰〉，選自《十三經注疏》本《左傳》：

十年春齊師伐我公將戰曹劌請見其鄉人曰肉食者謀之又何間焉劌曰肉食者鄙未能遠謀乃入見問何以戰公曰衣食所安弗敢專也必以分人對曰小惠未徧民弗從也公曰犧牲玉帛弗敢加也必以信對曰小信未孚神弗福也公曰小大之獄雖不能察必以情對曰忠之屬也可以一戰戰則請從公與之乘戰於長勺公將鼓之劌曰未可齊人三鼓劌曰可矣齊師敗績公將馳之劌曰未可下視其轍登軾而望之曰可矣遂逐齊師既克公問其故對曰夫戰勇氣也一鼓作氣再而衰三而竭彼竭我盈故克之夫大國難測也懼有伏焉吾視其轍亂望其旗靡故逐之

粗粗一看這段先秦時期的古文，覺得真是夠為難的。那麼語文老師是怎樣為學生們一步一步來化解這些困難呢？

一、第一步：加上標點符號

十年春，齊師伐我。公將戰。曹劌請見。

其鄉人曰：“肉食者謀之，又何間焉？”

劌曰：“肉食者鄙，未能遠謀。”乃入見。

問：“何以戰？”

公曰：“衣食所安，弗敢專也，必以分人。”

對曰：“小惠未徧，民弗從也。”

公曰：“犧牲玉帛，弗敢加也，必以信。”
 對曰：“小信未孚，神弗福也。”
 公曰：“小大之獄，雖不能察，必以情。”
 對曰：“忠之屬也，可以一戰。戰則請從。”
 公與之乘。戰於長勺。
 公將鼓之，蒯曰：“未可。”齊人三鼓，蒯曰：“可矣。”齊師敗績。
 公將馳之。蒯曰：“未可。”
 下視其轍，登軾而望之，曰：“可矣。”遂逐齊師。
 既克，公問其故。
 對曰：“夫戰，勇氣也，一鼓作氣，再而衰，三而竭，彼竭我盈，故克之。
 夫大國難測也，懼有伏焉。吾視其轍亂，望其旗靡，故逐之。”

分好句子，又加上了這些標點符號以後，感覺順當多了。讓我們進一步按照“每個句子只含有一個中心動作”的習慣來檢查一下（結果如下表）。我們會發覺，大概是因為古代書寫漢字很困難（麻煩，字要一個一個刻在竹簡上），所以古文裏的句子總是寫得很精煉，往往只寫出中心動作（謂語成分）。而動作的發起者或者動作的物件經常不明說，讓讀者自己按照情理去猜度。

原文	讀者頭腦中的活動
十年春，齊師伐我。	三個角色：齊師，魯莊公，曹蒯
公將戰。	魯莊公要和誰打仗？
曹蒯請見。	曹蒯要見誰？
其鄉人曰：“肉食者謀之，又何間焉？”	誰的“鄉人”？
蒯曰：“肉食者鄙，未能遠謀。”	
乃入見。	誰見誰？
問：“何以戰？”	誰問誰？
公曰：“衣食所安，弗敢專也，必以分人。”	
對曰：“小惠未徧，民弗從也。”	誰“對”？
公曰：“犧牲玉帛，弗敢加也，必以信。”	
對曰：“小信未孚，神弗福也。”	誰“對”？
公曰：“小大之獄，雖不能察，必以情。”	
對曰：“忠之屬也，可以一戰。戰則請從。”	誰“對”？
公與之乘。	“之”是誰？
戰於長勺。	誰和誰“戰”？
公將鼓之，	
蒯曰：“未可。”	
齊人三鼓，	
蒯曰：“可矣。”	

齊師敗績。	
公將馳之。	
劇曰：“未可。”	
下視其轍，	誰下？誰的“車轍”？
登軾而望之，	誰“登軾”，望“什麼”？
曰：“可矣。”	誰“曰”？
遂逐齊師。	誰戰勝了誰的軍隊？
既克，公問其故。	
對曰：“夫戰，勇氣也，	誰“對曰”？
一鼓作氣，再而衰，三而竭，	
彼竭我盈，故克之。	
夫大國難測也，懼有伏焉。	
吾視其轍亂，望其旗靡，故逐之。”	

對於這一系列的“空缺”，人們要依靠自己的生活經驗與歷史知識加以填補：在這個例子中，問題比較簡單一點。因為文章的頭三句話就為讀者建立了具有“齊師、魯莊公、曹劇”三個角色的語境，而且在隨後的文章中也沒有添加。所以句子中沒有明顯提及的主語和賓語，只需在這三者中進行選擇就可以了。於是進行以下的第二步。

二、第二步：把省略的語句成分補上去

十年春，齊師伐我【魯國】。【魯莊】公將戰。曹劇請見【魯莊公】。

【曹劇】的鄉人曰：“肉食者{大官們}謀之，【你】又何【必參與其】間焉？”

劇曰：“肉食者鄙，未能遠謀。”乃入見【魯莊公】

【曹劇問魯莊公】：“【憑】何以戰？”

公曰：“衣食所安，弗敢專也，必以分人。”

對曰：“小惠未徧，民弗從也。”

公曰：“犧牲玉帛，弗敢加也，必以信。”

對曰：“小信未孚，神弗福也。”

公曰：“小大之獄，雖不能察，必以情。”

對曰：“忠之屬也，可以一戰。戰則請從。”

【魯莊】公與之【同】乘【一輛車】。【和齊軍】戰於長勺。

【魯莊】公將鼓之【擊鼓進軍】，劇曰：“未可。”

齊人三鼓，劇曰：“可矣。”齊師敗績。

公將馳之【追擊齊軍】。劇曰：“未可。”

【曹劇】下視其【車】轍，登軾而望之，曰：“可矣。”遂逐齊師。

既克，公問其故。

【曹劇】對曰：“夫戰，勇氣也，一鼓作氣，再而衰，三而竭，彼竭我盈，

故克之。

夫大國難測也，【我】懼【他們】有伏焉。

吾視其轍亂，望其旗靡，故逐之。”

小結

從以上的例子可以看出，通常在句子中“中心動詞”總是最重要的。因此在古漢語中，爲了簡短，一個句子中總是保留最重要的“中心動詞”。而動作的發起者（主語）和動作的物件（賓語）這兩種名詞性成分常常被省略。這些被省略的成分當然都可以通過語境（前後文，當時的情境背景……）猜出來。爲了“猜出”這些被省略的成分並且明顯地表達出來，有兩種手段：

- 一、根據角色的性質從語義上來判斷。比如說，上面這篇文章中的第三句話“曹劌請見”是“請見”誰的問題；在三個角色：齊師、魯莊公、曹劌中，只有“魯莊公”最合適。“曹劌請見齊師”和“曹劌請見曹劌”都不合情理。
- 二、根據文本中“距離的遠近”從形式上來判斷。比如第六句：“乃入見。”誰“入”？這裏缺了一個主語。按照“就近繼承”的原則，因爲前一句的主語是“曹劌”，因此可以猜想這一句應該擴充爲“曹劌乃入見”。

讓電腦根據語義來判斷是比較困難的。而按照“就近繼承”的原則從形式上來判斷就比較容易。但是這種判斷不一定總是正確的。

那麼怎麼辦呢？比較合理的出路是採取“對話模式”。對一道題目沒有把握的學生可以把自己給出的答案回饋給老師，請求判斷自己答得對不對？因爲在大多數情況下“就近原則”是成功的，這樣也減少了“人”的工作量。

總而言之，對於這種需要智力程度比較高的填補“省略成分”的工作，電腦是比較困難的，所以凡是需要電腦處理的文章，如果希望獲得比較好的效果，應該事先通過“標注”把這些“坑”填平。而這一項工作可以在人機交互的環境中借助工具來完成。