全語料庫中文詞義標記的初步研究

柯淑津[1]、陳振南[2]、黃居仁[3]
[1]東吳大學資訊科學系
[2]銘傳大學資訊工程系
[3]中央研究院語言研究所

摘要

語料庫對自然語言處理研究佔有相當重要的地位。標示詞義(Sense)或語意(Semantic)的語料資源，在理論語言學上，將提供詞彙語意學研究之豐富材料與基本架構；在計算語言學上，將對自然語言處理的核心工作如：WSD 多重詞義辨析研究、辭典建構等，都將會有關鍵性的突破。

本文提出以詞彙為主，結合搭配訊息與機率模式的方式來處理語意自動標示工作。利用機器學習方法，由電腦蒐集標示知識以進行自動詞義標示工作。我們主要以引導(bootstrap)技巧來進行標示工作，這種技巧的運用需要有高精準度詞義標註的起始種子句。因此，我們先由人工標示種子句子後，再透過高精準度的搭配詞彙來決定目標詞彙的詞義，並且用決定出詞義的句子來擴充訓練語料。然後，再利用機率模式設法標註更多句子。這種技巧經過實驗證實再可接受的應用率(74.9%)下能獲得高正確率(92.6%)。

1. 簡介

語料庫對自然語言處理研究佔有相當重要的地位。尤其是統計式計算語言處理，常需仰賴語料庫所蘊藏的豐富資源，作為計算的依據。隨著電腦科技的進步及數位文獻之普及，語料庫的種類越來越多，內容也越來越豐富。標示訊息越完整的語料庫對研究的幫助越大。有些語料庫只以原始語料呈現，有的則加上詞性、詞義等標示資料。目前，標示詞性的語料庫有很多，至於標示詞義的語料庫不論中英文都很少。標示詞義(Sense)或語意(Semantic)的語料資源，在理論語言學上，將提供詞彙語意學研究之豐富材料與基本架構；在計算語言學上，將對自然語言處理的核心工作如：WSD 多重詞義辨析研究、辭典建構等，都將會有關鍵性的突破。另外，這些標示語料經統計處理後所獲得資訊，可應用於資訊檢索、資訊擷取、文件摘要、自動問答等議題之研究。

大量精確的語意標示資料，可提供多項計算語言相關研究的豐富素材。但是，中文語料庫詞義標記有兩個主要的瓶頸：第一是缺乏涵蓋面廣兼具完整理論基礎的詞義區分詞彙庫。換句話說，就是缺乏可靠的詞義標記依據。第二是，因為缺乏足供自動標記參考的資料，而人工標示需要昂貴成本，造成語料庫標示語意工作的難產。為了克服第一個問題，黃居仁（2003）提出了完整的中文詞義區分理論與操作原則，並已完成了將近 2000 個中文詞義的分析（黃居仁 2004）。為了克服第二個問題，本研究提議設法利用機器學習方法，由電腦蒐集標示知識以進行自動語意標示的工作。我們主要以引導(bootstrap)技巧來進行標示工作，這種技巧能不能夠運用成功常隨著起始種子句的詞義標註是否具有高精準度而改變。過去在雙語詞彙對應的研究上，驗證以詞彙為主的對應方式，可以獲得高精確度的效果(Ker 和 Chang, 1997)。另外，我們發現，就特定詞彙而言，表達相同詞義的句子其目標詞彙的周邊往往有固定的搭配詞彙出現。因此，本文提出以詞彙為主結合搭配訊息與機率模式的方式來處理語意自動標示工作。

2. 周邊詞彙決定詞義

在岐義辨識處理上，出現在目標詞彙周邊的特定搭配往往就可以決定詞彙的詞義，Yarowsky 在 1993 年提出的「one sense per collocation」就是指搭配詞彙可決定詞義(Yarowsky, 1993)。

Gale 等人的研究也證實周邊詞彙確實有助於歧義辨識(Gale et al., 1992)。透過觀察我們發現，就特定詞彙而言，表達相同詞義的句子其目標詞彙的周邊往往有固定的搭配詞彙出現，例如： 針對目標詞彙「電」，在例一及例二中，因為前置詞彙「帶」，因此可判斷「電」的詞義為:「指電子和質子運動所產生的能量」，同樣的情形，在例三的「用」也可決定「電」在此例的詞義應為「指發電機所產生的電能量」。當我們把視窗的範圍放大時，較遠距離的詞彙對於目標詞彙的影響力便隨之減小。

例一：左右兩條線是代表帶電粒子

例二：磁層中帶電粒子及磁場的動態變化、磁暴及副磁暴在磁層的影響。

例三：節流方面，鼓勵離峰用量，避開尖峰用電的負荷量，以價制量達到節約能源目的。

Stevenson 等人提出結合多項資訊將有助於歧義辨識工作的精確度（Stevenson and Wilks, 2001)，因此，我們提出一套結合搭配資訊與機率模式的兩段式詞義標註方法。第一階段我們將視窗設定在極小範圍之內，針對特定詞義收集出現在目標詞彙周邊的搭配詞彙，利用搭配資訊來擴充訓練語料之資料量。第二階段再透過機率模式來計算目標詞彙最可能的詞義。研究過程中所需要處理的事項，包括在訓練階段的標示知識收集以及測試階段的自動標示工作。

2.1 搭配資訊詞義標註方法

在收集搭配資訊的過程中，我們認為一個搭配詞彙需要滿足下列三個條件：第一，搭配詞彙與目標詞彙之間的距離必須很小，而且非常固定，第二，搭配詞彙出現頻率高，第三，搭配詞彙與目標詞彙的詞性應該擁有固定的組合。

首先，我們統計設定距離內所有周邊詞彙出現之次數，取出所有次數大過於設定次數的周邊詞彙，並且經過詞性篩選器過濾後，所留下來的詞彙作為目標詞彙在特定詞義下的搭配詞義。接著，我們自測試語料中，將符合搭配條件的句子，直接標上特定詞義，並且，將這些句子加入訓練語料中以擴充訓練語料之資料量。

2.2 機率模式詞義標註方法

訓練階段

在訓練階段針對一個目標詞彙 *w*，以及 *w* 的特定詞義 *s*, 我們自標示語料中收集所有含目標詞彙 *w*，且其詞義標示為 s 的句子，並計數周邊詞彙 c 出現的次數，記為 count(*c*, *w*, *s*)。對於周邊詞彙 *c* 與目標詞彙 *w*，詞義 *s* 的權重值 Pr(*c*, *w*, *s*)，設定為 count(*c*, *w*, *s*)比對於所有周邊詞彙 *c* 與目標詞彙 w 共同出現之次數，如公式一所示。其中，α 值作為 smoothing 用。

$$\Pr(c \mid w, s) = \frac{count(c, w, s) + a}{\sum_{s' \in senses\ of\ w} (count(c, w, s') + a)} \qquad \text{(公式一)}$$

其中， c: 出現於目標詞彙 w 左右視窗內的周邊詞彙，

s: 目標詞彙 w 在給定句子之詞義。

測試階段

首先，針對任意一個目標詞彙 *w*，自未標示語料 U 中選取所有含詞彙 *w* 的句子作為測試集 *D*。由 *D* 中選取任一句子 *S*，我們計算 *w* 被標示為詞義 *s'*的機率值 score(*s'*, *w*, *S*) ，它的計算方法如公式二所示，為目標詞彙 w 與其所有周邊詞彙 c 在詞義 s' 下的共現機率乘積值。我們以擁有最高可能的詞義 s^1 之權重值 Pr(c | w, s^1)比對於次高詞義 s^2 之權重值 Pr(c | w, s^2) 之比值作為標示信心值 R(w, S)（如公式三），認為 R 值越大的句子，標示結果越正確。我們依 R 值由高至低標示，並將完成標示之句子加入標示集合 T，重新計算收集標示知識的機率值，並重

複上述工作，直至完成所有句子標示工作為止。

$$\text{score}(s', w, S) = \prod_{c \in S} \Pr(c \mid w, s'), \qquad \text{（公式二）}$$

$$\text{R}(w, S) = \frac{score(s^1, w, S)}{score(s^2, w, S)}, \qquad \text{（公式三）}$$

其中， $s^1 = \arg\max_{s \in \text{senses of } w} score(s \mid w, S)$ ， $s^2 = \arg\max_{s \in \text{senses of } w, and\ s \neq s^1} \text{score}(s \mid w, S)$.

3. 實驗設計與結果

為驗證本研究所提方法之效益，我們結合搭配詞彙以及機率模式設計兩階段式的實驗。以下分別介紹參與實驗的資料、實驗方法以及對實驗結果進行討論。在效能部分，本研究採用正確率(Correctness)及應用率(Applicability)進行評估。實驗過程中能標示出語意概念的詞彙個數比對於參與實驗的總詞彙數，稱為應用率，至於，正確率定義為標示結果的正確比率。

3.1 實驗資料

我們的詞義標記實驗以中研院平衡語料庫（1995）的 500 萬詞全部資料作為標記範圍，語料庫中所含句子都已經過斷詞及詞性標示處理。標記所需的詞義區分，以黃居仁 2004 年完成約 2000 個中文詞義分析作為基準（黃居仁，2004）。在訓練階段為降低人工標記詞義所需成本，每個目標詞彙我們最多自語料中隨機選出 100 個句子作為訓練語料，其餘的部分則作為測試語料。

至於標示目標詞，在第一批研究中，我們共選了 134 個詞形，455 個詞義。每個詞形的詞義數，最低為 2(如：信 Na、靈魂 Na 等)，最高為 18（如：收 VC）。

3.2 實驗設計

在實驗部分我們分為兩個階段進行，首先利用搭配詞彙來擴充訓練語料，我們將已由人工標記詞義的句子作為種子句子。接著依前節所述方法，對標註詞義統計出現在視窗內的周邊詞彙訊息，並進行詞性等語言特徵篩選，收集搭配詞彙資訊。然後，自測試語料中選出符合搭配條件的句子，直接依搭配條件標記詞義，並以此方法自動擴充訓練語料規模。

我們將第一階段實驗所得標記句子與原先人工標記句子合併之後作為第二階段實驗的訓練語料，進行周邊詞彙共現機率值的資訊收集。以目標詞彙的前、後兩端各 d 個詞做為視窗（本實驗以前後各 10 個詞彙為視窗大小，即 d＝10），收集視窗內的周邊詞彙出現頻率。再依公式一計算目標詞彙在標註詞義下與周邊詞彙的共現機率值。

最後的測試階段，我們自測試語料中選出含目標詞彙的句子，再依公式二，計算目標詞彙在所屬上下文中為各可能詞義的機率值。我們自測試語料中選出可靠度最高的 r 個句子。然後，將新產生之標註結果，併入訓練語料中，以引導式方式 (bootstrap) 按演算法二將測試語料中的句子，逐一標註詞義。

3.3 實驗結果

為了方便實驗結果評估，我們自測試資料選取 24 個檔案，經人工標記詞義後作為標準答案。表一是這些標準答案檔所含的資料量，包含各詞彙的詞義數以及在訓練語料、測試語料中的句數。總共有 2957 個例數，詞義數最高為 18，最低為 2，平均詞義數為 3.5。

實驗過程中經由演算法一所得之搭配資料包括有 45 個前置以及 18 個後置搭配，部分搭

配資料列於表二，其中，表二(a)為前置搭配，表二(b)為後置搭配。在測試資料中符合搭配條件之詞例數，請見表三，以詞彙「報紙」為例，在我們的測試資料中共有 243 個詞例，而其中詞義為「定期出版，報導新聞、提供各式訊息的出版品。」之資料共有 227 筆，目前我們的實驗共有 42 個符合搭配條件，因此應用率為 18.5%(42/227=18.5%)。經由人工檢查後，得知 41 個詞例為正確標註，因此其正確率為 97.6%。整體而言，利用搭配資訊進行詞義標註，在 2118 個測試資料中共有 449 例符合搭配條件，而其中有 444 筆為正確標註資料，因此應用率為 21.2%，正確率為 98.9%。

在這些新產生的高精準度標註資料，併入先前之人工標註資料後，再經公式二計算周邊詞彙共現機率後，再將機率模式運用於測試語料之標註實驗。綜合搭配資訊與機率模式後所產生之應用率與精確度，如表四所示。以含兩個詞義的詞彙「信」為例，在 258 個測試句子中，共可標註出 228 個測試句子，因此應用率為 88.0% (228/258)。系統標示結果與人工標註結果比對後，其中，標註正確的有 225 個句子，因此正確率為 98.7% (225/228) 。整體而言，則可達到 92.6%正確率與 74.9%應用率。

表一 標準答案檔所含的資料量

詞彙	詞性	詞義數	測試句數	訓練句數	詞彙	詞性	詞義數	測試句數	訓練句數
收	VC	18	238	12	好朋友	Na	2	44	102
兄弟	Na	6	287	72	明星	Na	2	123	104
按	VC	5	117	26	信	Na	2	258	71
股	Na	4	17	75	報紙	Na	2	243	65
面	Na	4	239	37	傳真	Na	2	10	39
群	Na	4	12	18	感	Na	2	5	44
電	Na	4	53	96	電影	Na	2	903	72
同志	Na	3	26	35	臉色	Na	2	16	46
命	Na	3	108	50	靈魂	Na	2	53	80
吵	VC	3	17	49	伸	VC	2	12	45
抱	VC	3	108	65	佔領	VC	2	10	42
解放	VC	3	48	84	服	VC	2	10	30

表二(a) 第一階段實驗所得的部分前置搭配

詞彙	詞性	前置搭配詞彙	詞義
信	Na	封 Nf	傳達消息的書札。
電	Na	帶 VC	指電子和質子運動所產生的能量。
電	Na	用 VC、供 VF、配 VD、缺 VJ、度 Nf	指發電機所產生的電能量。
明星	Na	影視 Na、偶像 Na	出色有名的人物。
股	Na	持 VC、類 Nf、換 VC	接尾詞。特指某一種或某一類的股票。股票的簡省。
傳真	Na	資訊 Na、電話 Na	指利用光電效應的通訊方式，通過有線電或無線電裝置的通訊方式所傳送到遠方的照片、圖表、書信和文件等。
服	VC	口 Na	以口食用藥物，吃藥。
報紙	Na	份 Nf、看 VC	指定期出版，報導新聞、提供各式訊息的出版品。

表二(b) 第一階段實驗所得的部分後置搭配

詞彙	詞性	後置搭配詞彙	詞義
靈魂	Na	深處 Nc、出竅 VA	指相對於肉體軀殼的抽象精神心靈。
靈魂	Na	人物 Na	指群體中的重要核心人物。
明星	Na	球員 Na	出色有名的人物。
明星	Na	學校 Nc	萬眾矚目。
伸	VC	手 Na	展開。
吵	VC	架 Na	各持己見而爭論。
抱	VC	在一起 Vh	用手(手掌及手臂)圈住某人或某物。
傳真	Na	伺服器 Na、號碼 Na	指利用光電效應的通訊方式，通過有線電或無線電裝置的通訊方式所傳送到遠方的照片、圖表、書信和文件等。
電	Na	輕子 Na	指電子和質子運動所產生的能量。
電影	Na	劇本 Na	利用強光映射在螢幕上，以供人觀賞。

表三 利用搭配詞彙標註詞義實驗結果

詞彙	詞性	測試句數	標示句數	正確句數	應用率(%)	正確率(%)
兄弟	Na	218	13	12	6.0	92.3
同志	Na	13	3	3	23.1	100
好朋友	Na	43	13	13	30.2	100
佔領	VC	10	1	1	10.0	100
吵	VC	6	1	1	16.7	100
抱	VC	70	2	2	2.9	100
明星	Na	123	5	5	4.1	100
服	VC	7	2	2	28.6	100
股	Na	17	2	2	11.8	100
信	Na	250	95	95	38.0	100
按	VC	101	2	2	2.0	100
面	Na	145	29	29	20.0	100
報紙	Na	227	42	41	18.5	97.6
傳真	Na	10	5	5	50.0	100
放	VC	33	14	13	42.4	92.9
電	Na	50	17	16	34.0	94.1
電影	Na	731	193	192	26.4	99.5
臉色	Na	11	2	2	18.2	100
靈魂	Na	53	8	8	15.1	100
Total		2118	449	444	21.2	98.9

表四 利用共現機率模式標註詞義實驗結果

詞彙	詞性	測試句數	標示句數	正確句數	正確率(%)	應用率 (%)
兄弟	Na	287	198	186	93.9	68.9
報紙	Na	244	232	216	93.1	95
好朋友	Na	44	41	40	97.5	93.1
傳真	Na	10	7	7	100	70
臉色	Na	16	8	7	87.5	50
靈魂	Na	53	30	30	100	56.6
信	Na	258	228	221	96.9	88.3
感	Na	5	2	2	100	40
伸	VC	12	6	4	66.6	50
佔領	VC	10	6	5	83.3	60
服	VC	10	5	4	80	50
明星	Na	123	110	99	90	89.4
電	Na	53	40	38	95	75.4
電影	Na	903	812	800	98.5	89.9
同志	Na	26	10	6	60	38.4
命	Na	108	55	51	96.4	50.9
收	VC	238	98	72	73.4	41.1
按	VC	117	92	60	65.2	78.6
股	Na	17	11	15	93.8	64.7
群	Na	12	10	8	80	83.3
面	Na	239	117	99	84.6	48.9
吵	VC	17	15	14	93.3	88.2
抱	VC	108	52	41	78.8	48.1
解放	VC	48	32	29	90.6	66.6
Total		2958	2217	2054	92.6	74.9

5. 結論

本研究顯示，利用引導方式，在有精確詞義區分資料庫的前提下，對大規模語料庫的全面詞義標記是可行的。我們的初步工作已標記了百萬語料庫中，600 個以上的詞義。因此，全語料庫詞義標記的瓶頸，在於精確詞義區分語料庫的建構。引導式的詞義標記可帶來高精準度的標記效果。嚴格條件的搭配標記法更可有效擴充訓練語料，使得周邊詞彙共現機率方法能突破資料稀疏的問題。

參考文獻

Atkins, S., Tools for Computer-Aided Lexicography: the Hector Project, *Acta Linguistical*

Hungarica, 41, 5-72, 1993.
Fellbaum, C., J. Grabowski and S. Landes, Matching Words to Senses in semantic annotation task, In: Fellbaum, C. ed., *WordNet: An Electronic Lexical Database*, MIT Press, 1998.
Gale, W., K. Church, D. Yarowsky, 1992, One sense per discourse, In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, 233-237.
Ker S. J. and J. S. Chang, A Class-base Approach to Word Alignment, *Computational Linguistics*, 23(2), 313-343, 1997.
Kilgarriff, A. and J. Rosenzweig, English SENSEVAL: Report and Results, In *Proceedings of 2nd International Conference on Language Resource and Evaluation*, 2000.
Ng, H.T. and H.B. Lee, Integrating Multiple Knowledge Sources to Disambiguate Word Sense: An Example-based Approach, In *Proceedings of the 34th ACL*, 1996.
Preiss, J., 2004, Probabilistic Word Sense Disambiguation, *Computer Speech and Language*, 319-337.
Stevenson, M. and Y. Wilks, 2001, The interaction of knowledge sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3), 321-349.
Yarowsky, D., 1993, One Sense Per Collocation, In *Proceedings ARPA Human Language Technology Workshop*, Princeton, NJ, 266-271.
中央研究院現代漢語平衡語料庫，http://www.sinica,edu,tw/SinicaCorpus/, 1995.
黃居仁，主編。中文的意義與詞義─I。中央研究院文獻語料庫與詞庫小組技術報告 03-01~03-04 合訂本。南港，中研院, 2004。
黃居仁，蔡柏生，朱梅欣，何婉如，黃麗婉，蔡宜妮, 詞義與義面：中文詞彙意義的區辨與操作原則 [Sense and Meaning Facet: Criteria and Operational Guidelines for Chinese Sense Distinction]. Presented at the Fourth Chinese Lexical Semantics Workshops. June 23-25 Hong Kong, Hong Kong City University, 2003.