大規模詞彙語意關係自動標記之初步研究: 以中文詞網 (Chinese Wordnet) 為例

謝舒凱 Petr Šimon 黃居仁

中研院語言學研究所

{shukai, petr.simon, churenhuang}@gmail.com

Abstract

近年來, 以知識資源爲本的自然處理技術已成爲一種重要的研究取向。對於各種詞彙語意資源之建構, 包括電子辭典 (Lexicon)、同義詞詞林 (Thesaurus)、詞彙網路 (WordNet), 甚至知識本體 (ontologies), 已成爲一個不可抵擋的趨勢。其中, 詞彙網路是在計算語言學相關領域中, 目前最爲普遍利用之一項詞彙語意資源。

然而, 詞彙網路之建構是一項耗時費力之基礎工程。對於世界上許多使用頻度不高的語言而言, 更是一項艱鉅之任務。本文提出一個借力於普林斯頓英語詞網 (Princeton WordNet) 與歐語詞網 (EuroWordNet) 之 bootstrapping 方法, 應用在正在發展的中文詞網詞彙語意關係之自動標記工作上。實驗的結果與初步評估證明, 此法對於詞網建構是一個相當可行的方式。

1 前言

近年來, 以知識資源爲本的自然處理技術已成爲一種重要的研究取向。對於各種詞彙語意資源之建構, 包括電子辭典 (Lexicon)、同義詞詞林 (Thesaurus)、詞彙網

路 (WordNet), 甚至知識本體 (ontologies), 已成爲一個不可抵擋的趨勢。其中, 詞彙網路更已成爲計算語言學相關領域中, 最爲普遍利用之一項標準 (de facto) 詞彙語意資源。

詞彙網路是以同義詞集 (synset), 以及詞彙語意關係 (lexical semantic relation) 所架構出的詞彙知識系統。也就是說, 詞彙網路架構表達的不僅是詞彙本身的概念性知識, 它亦表達了詞彙之間的語意關係。然而, 從普林斯頓英語詞網以及歐語詞網 (EuroWordNet) 的建構經驗來看, 這是一項費時耗力的龐大語言工程。對於經費取得困難、使用頻度較低之語言而言, 建立此項語言資源更爲不易。從詞彙語意與知識表達的角度觀察, 我們認爲不同語言對於**概念原素** (conceptual atoms), 可能有著不同之表達方式, 但是在**詞彙語意關係**的表達上, 則應具有更大程度之「普同性」。因此「借力」於已發展成熟之英語、歐語詞網之語意關係, 以加速新的詞網雛形成形, 就成了一個自然而然的另類選擇。

基於以上動機, 本文之組織如下: 第二節描述目前有關不同之詞彙語意關係所採用之各種自動、半自動判定演算法。在第三節中, 我們提出了一種便捷而有意義的方法與實驗設計。文章之第四節討論實驗結果之評價工作, 最後一節則鋪陳我們的結論與未來的展望。

2 詞彙語意關係之自動判定法

近年來, 關於自動判定或學習詞彙語意關係的研究文獻越來越豐富。爲了重點強調之方便, 本文將之粗分爲兩大走向: 利用單語資源, 與利用雙語或多語資源之研究。

2.1 使用單語資源

此類方法大多以語料庫及網頁資料爲主, 利用詞彙語法模式 (lexical-syntactic patterns) 或是「叢集」(cluster) 來抽取詞彙語意關係。目前這樣的作法之所忽略的兩

個問題是: (1). 所抽取之語意關係不夠全面, 大半部分還是受限在少數之關係上, 例如 is-a, part-of 等等。[1] (2).所需要的學習實例 (seed instances) 太多。最近 (Pennacchiotti and Pantel 2006) 所提的 Espresso 演算法, 欲針對此兩個問題提出新解法, 但是在評價上尚未完整, 仍有待觀察進一步之發展。

2.2 使用雙語或多語資源

這個方向包括使用雙語語料庫 (Diab 2004); 利用同義詞集之雙語對譯 (bilingual correspondences)[2] 直接藉助於以存在之詞網 (通常係普林斯頓詞網)。後者已有相當多之探討 (Pianta, et al 2002; Huang et al 2002, 2003, 2005), 西班牙詞網與義大利之 MultiWordNet 計畫皆是此想法下之實作產物。

3 實驗設計與方法

基於前面之文獻討論, 本文認爲, 著眼於當前多語處理之需求, 在策略上, 應該採取**先求同再求異**。因此, 借力於已存在之多語資源應該是第一步的工作。

3.1 我們提出之 Model: Bootstrapping from Multilingual Wordnets

本文提出的模型, 是基於 (Huang et al. 2002, 2003, 2005) 的擴充版本。先前之文獻, 已就借力於其他成形之詞網的跨語詞義關係預測所涉及之邏輯條件加以闡述, 在 (Huang et al 2003) 中, 並曾針對 210 個中文詞形 (lemma) 做過小規模之試驗與評價。本文則接續之前的基礎, 進一步在**規模**上與**多語擴充**兩個面向上作延伸試驗 。亦即, 在規模上, 我們將目前在中研院中文詞網小組所定義完成之七千多筆

[1]晚近亦有處理所謂 Textual Entailment 之關係。

[2]但是, 在此我們同時必須先理解到, 一組雙語對譯詞不一定是「同義關係」。此外, 它們可能在各自語言系統中與不同的詞/synset 間有不同的詞彙語意關係。

中文同義詞集爲主; 在多語擴充上, 我們將歐語詞網 (Vossen 1998) 亦納入實驗對象。其中包括了德語、法語、捷克語、荷語、西語、義語與愛沙尼亞語等七種歐洲語言。

本文提出之 model 在方法論上有兩個意涵:

- 對於多語之詞網發展, 可提供一個**符應** (correspondence) 與**協作** (collaborative) 架構: 在不同語種之詞網之間建立符應關係, 是多語知識處理工作之一重要環節。我們認爲, 這種符應性應該是表現在詞彙語意關係上, 而非在上層知識本體 (top ontology) 或詞彙翻譯上 (word translations)。此外, 在資料的標誌與管理上, 我們亦採用了全球詞網協會 (Global WordNet Association) 所建議之 XML/Schema 格式, 以便於將來國際詞網網格 (global wordnet grid) 環境架構。

- 另一方面, 此 model 可作爲一個詞網之**快速原型** (rapid prototyping) 發展。加速詞網核心部分建構之過程。

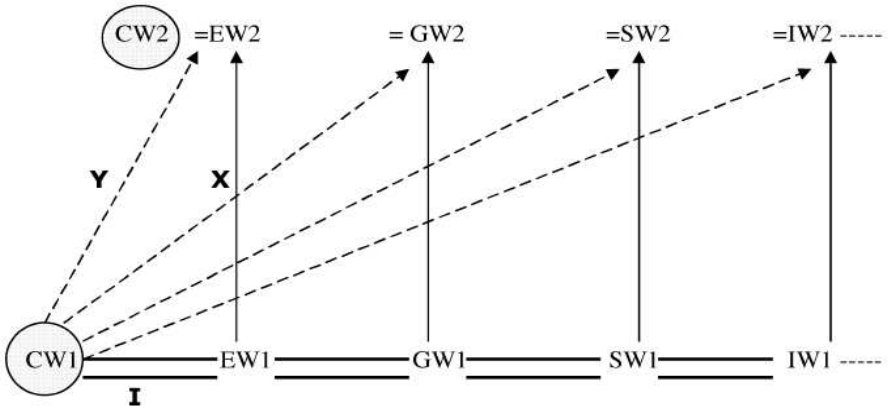

Figure 1: 本文提出之基於多語之大規模詞義關係抽取模型

Figure 1 中展示了這個模型。此模型是 (Huang et al 2002) 之擴充版本。圖中的變項 Y, I, X 間的關係, 可以用簡式來表達 $Y = I + X$ 。中文詞形 $CW1$ 與 $CW2$ 間的詞彙語意關係 Y 可被視爲是 I 與 X 的**機能結合** (functional combination)。最簡單的例子是, 如果 $CW1$ 與 $EW1$ 間的 I 是同義關係, 那麼英語詞形 $EW1$ 與 $EW2$ 間的語意關係 X , 就可被預測爲是 $CW1$ 與 $CW2$ 的關係。同理, 藉由 synset 之中介, 可進一步拓展到其他語種的詞網資源 (如德語、西語等等)。

不過, 就如之前所提到的, 一組雙語對譯詞不一定是「同義關係」。當不同關係 (例如上下位、部分與全體關係等) 涉入時, 我們就需要一組具邏輯性的自動推理規則。表格一則列出了我們所使用的詞彙語意關係邏輯推理規則。[3]

3.2 使用之詞彙資源

本實驗所需要使用的資源包括如下:

- **中文詞網之中頻詞形、詞義資料** 這是目前由中研院語言所中文詞網小組所完成之以中頻詞爲主的詞形(lemma) 及詞義區分資料。[4]
- **中英雙語** synset **對譯資料庫及評價標注** (Huang et al 2003)。此資料庫係由中央研究院與遠見科技股份有限公司共同開發。包括了以 WordNet 1.6 之 99,642 筆同義詞集 (synset) 爲基準之中文對應翻譯。
- **歐語詞網及普林斯頓** WordNet 1.5-1.6 **對應表**

3.3 步驟

- 生成中文詞網之 synset

  因爲正在發展中的中文詞網並無正式之 synset 格式設計, 我們採用 gloss

[3]詳細實例說明請參見 (Huang et al 2005)。

[4]網路展示版請見 http://cwn.ling.sinica.edu.tw

	I	X	Y	Bootstrapped Results
1	HYP	ANT	ANT	{CW1, ANTONYM, CW2}
2	HYP	HYP	HYP	{CW1, HYPONOMY, CW2}
3	HYP	NSYN	HYP	{CW1, HYPONYM, CW2}
4	HYP	HOL	HOL	{CW1, HOLONYM, CW2}
5	HYP	all other LSRs	undecided	?
6	HPO	ANT	ANT	{CW1, ANTONYM, CW2}
7	HPO	HPO	HPO	{CW1, HYPONYM, CW2}
8	HPO	NSYN	HPO	{CW1, HYPONYM, CW2}
9	HPO	MER	MER	{CW1, MERONYM, CW2}
10	HPO	all other LSRs	undecided	?
11	NSYN	ANT	ANT	{CW1, ANTONYM, CW2}
12	NSYN	HYP	HYP	{CW1, HYPERNYM, CW2}
13	NSYN	HPO	HPO	{CW1, HYPONYM, CW2}
14	NSYN	NSYN	NSYN	{CW1, NEAR-SYNONYM, CW2}
15	NSYN	MER	MER	{CW1, MERONYM, CW2}
16	NSYN	HOL	HOL	{CW1, HOLONYM, CW2}
17	HOL	ANT	ANT	{CW1, ANTONYM, CW2}
18	HOL	HYP	HYP	{CW1, HYPONYM, CW2}
19	HOL	NSYN	HOL	{CW1, HOLONYM, CW2}
20	HOL	HOL	HOL	{CW1, HOLONYM, CW2}
21	HOL	all other LSRs	undecided	?
22	MER	ANT	ANT	{CW1, ANTOMY, CW2}
23	MER	HPO	HPO	{CW1, HYPONYM, CW2}
24	MER	NSYN	MER	{CW1, MERONYM, CW2}
25	MER	MER	MER	{CW1, MERONYM, CW2}
26	MER	all other LSRs	undecided	?

Table 1: 詞彙語意關係邏輯推理規則

matching 的方式, 抽出、過濾並作流水編號。[5]

- 依照本文提出之模式從 WordNet 及 EuroWordNet 萃取詞彙語意關係

4 實驗結果與評價

4.1 基本數據與結果

以下則簡列出以 xml 格式標記的部分結果。

```
<synset id="00002517-x">
<gloss>預估費用並承諾以該費用履行合約。</gloss>
<variants>
<variant sense="01">估價</variant>
<variant sense="01">報價</variant>
</variants>
<ILIRelations>
<relation type="SYN" targetID="00692314-v"/>
<relation type="HYP" targetID="01529684-v"/>
<relation type="HPO" targetID="00692437-v"/>
</ILIRelations>
<internalRelations>
<relation type="HYP" targetID="01529684-v"/>
<relation type="HPO" targetID="00692437-v"/>
</internalRelations>
</synset>
<synset id="00002004-x">
<gloss>形容正常的, 只用於疑問句或否定句, 假設其不正常。</gloss>
<variants>
<variant sense="02">對</variant>
</variants>
```

[5]目前已發展了更有意義的編碼方式, 目前的設計純粹爲了實驗目的之故。在此過程中, 剛好附帶的作了詞網品管 Quality Control。包括處理了幾個問題: (1). 不同的中文 synsets 卻有相同的 synset offset; (2). 相同的 gloss 卻有不同的普林斯頓詞網 offset; (3). 類似錯誤的註解。如: 下列註解可能是相同的。

- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除... 以外”。
- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除... 之外”。
- 將前述對象排除或已知前述對象包含在所屬範圍。常用 “除..... 外”。

```
<ILIRelations>
<relation type="SYN" targetID="00138021-a"/>
<relation type="x_similar" targetID="00137150-a"/>
</ILIRelations>
<internalRelations>
<relation type="x_similar" targetID="00137150-a"/>
</internalRelations>
</synset>
<synset id="00001749-x">
<gloss>普通名詞。憤怒的情緒。</gloss>
<variants>
<variant sense="05">氣</variant>
</variants>
<ILIRelations>
<relation type="SYN" targetID="05587878-n"/>
<relation type="HYP" targetID="05560878-n"/>
<relation type="HPO" targetID="05588413-n"/>
<relation type="HPO" targetID="05588725-n"/>
<relation type="HPO" targetID="05588822-n"/>
<relation type="HPO" targetID="05588960-n"/>
<relation type="HPO" targetID="05589074-n"/>
<relation type="HPO" targetID="05589169-n"/>
<relation type="HPO" targetID="05589301-n"/>
<relation type="HPO" targetID="05589430-n"/>
</ILIRelations>
<internalRelations>
<relation type="HYP" targetID="05560878-n"/>
<relation type="HPO" targetID="05588413-n"/>
<relation type="HPO" targetID="05588725-n"/>
<relation type="HPO" targetID="05588822-n"/>
<relation type="HPO" targetID="05588960-n"/>
<relation type="HPO" targetID="05589074-n"/>
<relation type="HPO" targetID="05589169-n"/>
<relation type="HPO" targetID="05589301-n"/>
<relation type="HPO" targetID="05589430-n"/>
</internalRelations>
</synset>
```

4.2 評價

對結果之評價上, 我們採取蔡(2002) 所提出之中文詞義關係的判定原則, 並參酌 EuroWordNet 技術報告。爲了便利評估, 我們亦發展了一個簡便之評價系統。Figure 2 是此系統之操作介面。表格三則列出人工評價結果。[6]

[6]因資料龐大, 目前歐語詞網部分僅是目前跑出之部分結果。

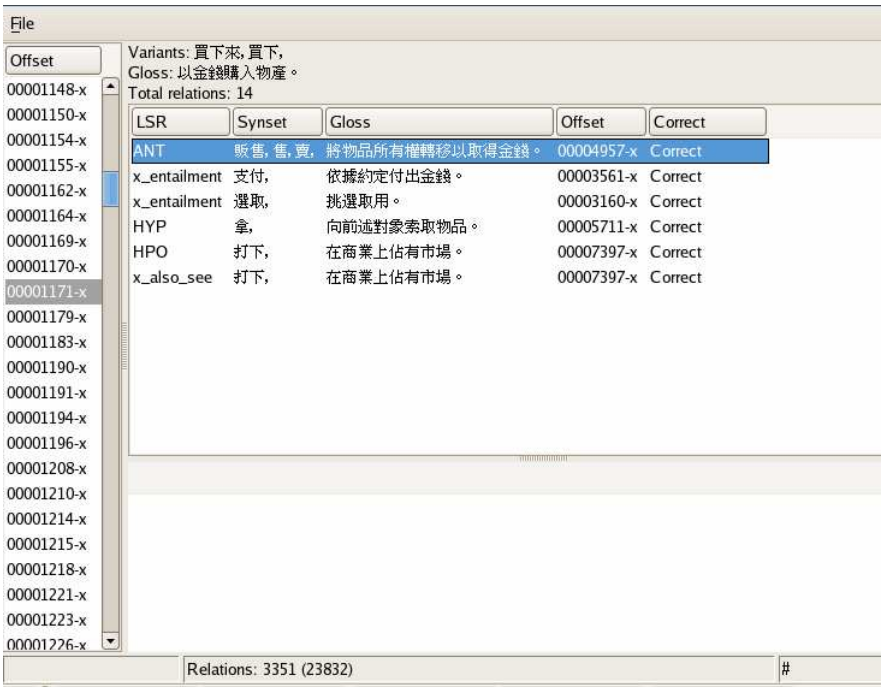

Figure 2: 檢測工具之圖形介面

我們由此得到的準確率如下:

$$Precision\ \ rate_{pwn} = \frac{3057}{3252} * 100\% = 94\%$$

$$Precision\ \ rate_{ewn} = \frac{2673}{2981} * 100\% = 89\%$$

	correct	incorrect	other	total retrieved with Chinese syn.	total retrieved
英語詞網	3057	124	71	3252	23832
歐語詞網	2673	210	98	2981	19711

Table 2: 評價結果

4.3 錯誤分析

從結果上看, 我們可說此法是一個相當可靠的方式。從各項語意關係類型之錯誤分佈來看, 我們的觀察結果大致與先前之小規模試驗相似。預測之正確率由高到低的排列大概都是 { MPT, x _ similair to, x _ pertainym, x _ also _ see ... } ≻ ANT ≻ HYP ≻ HPO ≻ { OTHER RELATIONS } 。詳細之數據留待完整實驗結果再行分析。此外, 此種理論模式要面臨到的主要難題, 包括詞彙的缺隔 (lexical gaps) (同一個概念在甲語言中與乙語言所使用的詞彙表達單位不同)、指稱差異 (denotation differences) (對等翻譯存在, 但是概念的抽象度卻不同) 等等, 如何在技術上克服這些非同義關係對應的語言現象, 值得往後進一步細究探討。

5 結論與未來展望

按我們的設想, 詞網之語意關係自動建構可以分由**全域的** (global) 與**區域的** (local) 特徵兩個面向來著手。本文展示了利用全域特徵之可行性, 並由此建立了中文詞網詞彙語意關係網路之雛形。在此基礎上, 我們下一步將細究在地詞網的特徵。包括利用詞彙模式 (lexical patterns) 從辭典釋意 (gloss)、語料庫等資源萃取各種詞彙語意關係, 亦包括利用漢語之構詞語意模式與漢字知識本體資源 (例如 HanziNet/Hantology), 使機器自動學習。我們相信這個成套的半自動詞網自動建構方法, 亦可供其他語種建構詞網參考。

References

[1] 蔡柏生、黃居仁等 (2002). 中文詞義關係的定義與判定原則。*Journal of Chinese Information Processing 16.4*.

[2] Diab, Mona. (2004). The Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet. *Proceedings of the Arabic Language Technologies and Resources*, NEMLAR, Cairo.

[3] Huang, Chu-Ren et al. (2002). Translating Lexical Semantic Relations: The First Step Toward Multilingual Wordnets. *COLING 2002*, Taipei.

[4] Huang, Chu-Ren et al. (2003). Cross-lingual Portability of Semantic Relations: Bootstrapping Chinese Wordnet with English WordNet Relations. *Language and Linguistics 4.3:509-532*.

[5] Huang, Chu-Ren et al (2005). Cross-lingual Conversion of Lexical Semantic Relations: Building Parallel Wordnets.

[6] Pennacchiotti, Marco and Patrick Pantel. (2006). A Bootstrapping Algorithm for Automatically Harvesting Semantic Relations. *LREC 2006.* Italy.

[7] Vossen, P. (ed). (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic.