Computational Linguistics and Chinese Language Processing
Vol. 8 , No. 2 , August 2003, pp. 93-108 
© The Association for Computational Linguistics and Chinese Language Processing

# 從詞網出發的中文複合名詞的語意表達

**柯淑津***

## 摘要

WordNet 提供豐富的詞彙語意資訊，因此對於自然語言處理相關研究有很大的幫助。但是由於 Princeton WordNet 的語意資訊僅以英文的形式呈現，為了能讓 WordNet 所蘊含的豐富資源也能應用到中文相關處理，我們試圖利用雙語字典等多項已存在的資源做為橋樑，希望能將英文 WordNet 的豐富資源自動引介到中文。但是，在我們觀察這些連結英文 WordNet 與雙語字典所產生的初步結果後，發現由於語言之間的藩籬以及雙語字典的目標語詞彙大都偏向於解釋等多種原因，使得英文同義詞集(Synset)所對應到的中文翻譯，常是一些不具結構性的中文複合詞、片語、甚至是一長串的句子，而不是獨立的中文詞彙。這樣的現象與中文詞網應以詞彙為基本元件的要求相違背。因此，本研究將針對這種現象作進一步的處理。

本文的主要目標有下列兩項：首先，自中文複合詞找出最能代表其意義的中心詞彙，及若干個特徵詞彙。其次，將這些詞彙進一步以語意概念形式表達出來。第一個部分，我們透過語法結構分析來完成。至於，第二個部分，詞彙的語意我們透過知網的概念特徵來表示。當然，在中文詞彙轉為詞義概念的部分，是存在著歧義現象的。辨識語意歧義的方法，我們除了用到詞彙的詞性之外，還透過 WordNet 的上位關係來降低歧義度。我們以名詞部分進行實驗，實驗結果顯示在語意標示方面，可達到 93.5%的應用率以及 93.8%的正確率。

## Abstract

WordNet provides plenty of lexical meaning; therefore, it is very helpful in natural language processing research. Each lexical meaning in Princeton WordNet is presented in English. In this work, we attempt to use a bilingual dictionary as the backbone to automatically map English WordNet to a Chinese form. However, we encounter many barriers between the two different languages when we observe the preliminary result for the linkage between English WordNet and the bilingual dictionary. This mapping causes the Chinese translation of the English synonym

* 東吳大學資訊科學系 E-mail: ksj@cis.scu.edu.tw

collection (Synset) to correspond to unstructured Chinese compound words, phrases, and even long string sentence instead of independent Chinese lexical words. This phenomenon violates the aim of Chinese WordNet to take the lexical word as the basic component. Therefore, this research will perform further processing to study this phenomenon.

The objectives of this paper are as follows: First, we will discover core lexical words and characteristic words from Chinese compound words. Next, those lexical words will be expressed by means of conceptual representations. For the core lexical words, we use grammar structure analysis to locate such words. For characteristic words, we use sememes in HowNet to represent their lexical meanings. Certainly, there exists a problem of ambiguity when Chinese lexical words are translated into their lexical meanings. To resolve this problem, we use lexical parts-of-speech and hypernyms of WordNet to reduce the lexical ambiguity. We experimented on nouns, and the experimental results show that sense disambiguation could achieve a 93.8% applicability rate and a 93.5% correct rate.

## 1. 簡介

自然語言處理的研究與應用，隨著語言資源的快速增加，有愈來愈蓬勃的現象。這些語言資源包括：辭典、詞彙資料庫、語料庫等等，而其中相當引人注目的就是分類辭典，例如：GUM, CYC, ONTOS, MICROKOSMOS, EDR 和 WordNet [Gomez, 1998]。這些分類字典中，各自有不同的特徵，有些是專為某個特殊範疇設計，有些則是不限文體；它們的排列方式也各自有所不同，可能是根據詞彙關係(Lexical Relation)，也可能根據概念關係(Conceptual Relation)來排列。在這些分類詞典中，WordNet [Miller, 1990; Fellbaum, 1998]擁有最寬廣的應用空間，已然形成一種標準[Farreres, Rigau and Rodriguez, 1998]。因此，自 WordNet 推出之後，便被廣泛地應用在許多的相關研究中，像是文件檢索[Gonzalo *et al.*, 1998; Mandala, Tokunaga and Tanaka, 1998 ]，機器翻譯[Knight and Luk, 1994]，文件生成[Jing, 1998]，影像檢索[Aslandogam *et al.*, 1997]等等。WordNet 的成功，引發許多非英語系的國家，建置自己語言版本 WordNet 的構想，並且有不少計畫已開始實際進行。例如包含多種歐洲語言的 EuroWordNet 已經完成[Atserias *et al.*, 1997; Farreres, Rigau, and Rodriguez, 1998]。另外，韓語版本以及日語版本的 WordNet 建構計畫也都正積極進行中[Lee, Lee and Yun, 2000]。

WordNet 提供了豐富的語意相關資訊，因此對於自然語言處理相關研究有很大的幫助。但是由於 Princeton WordNet 的語意資訊僅以英文形式呈現，為了能讓 WordNet 所蘊含的豐富資源也能應用到中文相關處理，我們試圖利用雙語字典等多項已存在的資源做為橋樑，希望能將英文 WordNet 的豐富資源自動引介到中文。但是，在我們觀察這些連結英文 WordNet 與雙語字典所產生的結果後，發現由於語言間的藩籬以及雙語字典的目

標語詞大都偏向於解釋等多項原因，使得英文同義詞集(Synset)所對應到的中文翻譯，常是一些不具結構性的中文複合詞、片語、甚至是一長串的句子，而不是獨立的中文詞彙。因此，本研究對這種現象作進一步的處理，透過提供詞義的字典，將這些中文翻譯轉化成語意概念，使得連結資料可應用於語言相關處理。

本文的目標在將詞網的中文翻譯轉化成語意概念表示，主要的工作分為下列兩項：第一項工作為自中文複合詞中找出語意中心詞彙，而第二項工作則是將中文詞彙轉化成詞義概念表達。當詞網的中文翻譯本身就是詞彙時，我們只需進行第二項工作，透過知網[董振東、董強, 2002]的概念定義，將詞彙表達成適當的詞義概念。若中文翻譯不是詞彙，而是複合詞句時，我們需先作第一項處理，找出複合詞句的語意中心詞彙，再進行第二項工作，將中心詞彙進一步以語意概念形式表達出來。當然，在中文詞彙轉為詞義概念的部分，是存在著歧義現象的。辨識語意歧義的方法，我們除了用到詞彙的詞性之外，還透過 WordNet 的上位關係來降低歧義度。

本文第二節介紹相關研究，第三節對資料進行一些觀察，第四節與第五節分別提出標示中心詞彙與概念特徵的方法。實驗設計及結果在第六節，最後是結論以及未來研究方向。

## 2. 相關研究

自然語言處理研究，需要豐富的詞彙知識與語意關係作為基礎。這些重要的研究資源除了透過統計技巧，由語料庫中獲得以外[Gale, Church and Yarowsky, 1992; Yarowsky, 1992, 1995; Resnik, 1993; Dagan and Itai, 1994; Luk, 1995; Ng and Lee, 1996; Riloff and Jones, 1999]，還可粹取自機讀字典[Guthrie *et al.*, 1991; Slator, 1991; Li, Szpakowicz and Matwin, 1995; Chen and Chang, 1998, Yang and Ker, 2002]。

近來，有許多學者著力於建構含語意訊息的中英雙語資源，他們認為這些資源對於機器翻譯以及多語資訊檢索系統都有很大的幫助[Chang, Ker and Chen, 1998; Chen and Chang, 1998; Chen and Lin, 2000; Chen, Lin and Lin, 2000; Dorr *et al.*, 2000; Carpuat *et al.*, 2002; Wang, 2002]。其中，他們所使用的資源各自不一，有的是連結 WordNet 與同義詞詞林[Chen and Lin, 2000]，有的將 WordNet 與 HowNet 進行對應[Dorr *et al.*, 2000; Carpuat *et al.*, 2002 ]，也有的研究群將一般機讀字典與分類字典進行連結，設法由詞彙得到分類資訊[Chang, Ker and Chen, 1998; Chen and Chang, 1998]。

## 3. 觀察

透過觀察，我們發現不同的字典資源，雖收錄的詞彙、語意訊息各自有所差異，但當它們在表達同一詞彙所具有的相同語意時，常會存在著某些共通現象。這些共通現象包含有：中文複合詞之中心詞彙與上位詞之翻譯共用詞素，以及共用定義詞彙等等。以下，我們先介紹經連結後的資料，並觀察它們所包含的訊息，以及探討這些訊息如何使用於辨識中心詞彙或是進行歧義詞的詞義辨識工作。

### 3.1 連結上中文翻譯的WordNet資料

為了能讓 WordNet 所蘊含的豐富資源也能應用到中文的相關處理，在先前的研究我們利用雙語字典等多項已存在的電子資源做為橋樑，將 WordNet 的同義詞集連結上適當的中文翻譯。表 1 是部分的連結例子，其中第一個欄位是構成這個同義詞集的英文詞彙。第二個欄位是它們在 WordNet 的定義，而第三個欄位就是經過連結雙語字典後所得的中文翻譯。這些翻譯中像是「光線」、「大樓」以及「秋天」等都屬於中文詞彙，但是也有一些中文複合詞，例如：「一壘安打」、「扁桃腺切除術」以及「西洋棋騎士」等。這樣的現象與中文詞網應以詞彙為基本元件的要求相違背。因此，本研究將對這種現象作進一步的處理。

***表 1*** *連結 WordNet 同義詞集與中文翻譯之例子。*

| WordNet 同義詞集 | WordNet 定義 | 中文翻譯 |
|---|---|---|
| building, edifice | a structure that has a roof and walls and stands more or less permanently in one place | 大廈，大樓，建築物 |
| beam, beam of light, light beam, ray, ray of light, shaft, shaft of light | a column of light (as from a beacon) | 光束，光線 |
| clay, mud | water soaked soil; soft wet earth | 泥，泥巴，泥漿 |
| autumn, fall | the season when the leaves fall from the trees | 秋，秋天，秋季 |
| single | a base hit on which the batter stops safely at first base | 一壘安打 |
| tonsillectomy | surgical removal of the palatine tonsils; commonly performed along with adenoidectomy | 扁桃腺切除術，扁桃體切除術 |
| knight, horse | a chessman in the shape of a horses head; can move two squares horizontally and one vertically (or vice versa) | 西洋棋騎士 |

### 3.2 中心詞彙與上位翻譯詞彙共用詞素

在 WordNet 中上位詞代表一種泛稱。在一個同義詞集的定義與其上位詞集的定義常會有共用詞素的情形。這種原本出現於英文定義上的現象似乎也延續至它們的中文翻譯。這種情形對於中文複合詞翻譯尤其明顯。而且，如果我們將這些中文複合詞進行斷詞處理後，更可發現與其上位詞的中文翻譯擁有最多共用詞素的詞彙往往就是該中文複合詞的中心詞彙。例如，表 2 的中文複合詞「扁桃腺切除術」經斷詞處理後得到「扁桃腺」與「切除術」兩個詞彙。其中，「切除術」與該同義詞集的上位詞之中文翻譯「切除」擁

有較多的共用詞素。同時，「扁桃腺切除術」所談的主要是「切除術」，因此「扁桃腺切除術」的中心詞彙是「切除術」。同樣的情形也發生在表 2 的其他例子，如：「曲線球」、「一壘安打」、「西洋棋騎士」等中文複合詞，它們的中心詞彙分別是「球」、「安打」以及「西洋棋」。這些中心詞彙與其所屬的上位詞翻譯皆含有較多的共用詞素。

### 3.3 詞彙之概念特徵辨識

詞彙往往含有多個詞義，詞義的辨識是自然語言處理的核心工作。在此小節中，我們觀察連結上 WordNet 同義詞集的中文翻譯所含有的訊息，以便瞭解如何透過訊息處理，將這些中文翻譯轉化成知網的概念特徵。

***表 2*** *同義詞集之中文複合詞翻譯及其中心詞彙。*

| 同義詞集 | 同義詞集定義 | 中文翻譯 | 上位詞詞集 | 上位詞<br>中文翻譯 |
|---|---|---|---|---|
| testate, testator | a person who makes a will | 遺囑 **人** | mortal, individual, person, somebody, soul, someone, human | 人 |
| screwball | a pitch with reverse spin that curves toward the side of the plate from which it was thrown | 曲線 **球**, 內曲線 **球** | pitch, delivery | 投球 |
| single | a base hit on which the batter stops safely at first base | 一壘 **安打** | bingle, safety, base hit | 安打 |
| tonsillectomy | surgical removal of the palatine tonsils; commonly performed along with adenoidectomy | 扁桃腺 **切除術** | ablation, cutting out, excision, extirpation | 切除 |
| plumbing, plumbery | the occupation of a plumber (installing and repairing pipes and fixtures for water or gas or sewage in a building) | 鉛管 **業** | craft, trade | 職業 |
| knight, horse | a chessman in the shape of a horses head; can move two squares horizontally and one vertically (or vice versa) | **西洋棋** 騎士 | chessman, chess piece | 棋子 |

註：標示底線者為中文翻譯之語意中心詞彙

#### 3.3.1 單義詞部分

有些詞彙本身的語意很確定，不具有歧義性。這些詞彙在不同的資源中，雖可能會存在

不同的解釋用語。但是，它們大多是指同一事物，因此，我們直接讓它們對應。如：表 1 的同義詞集{autumn, fall}所對應的中文翻譯「秋，秋天，秋季」，經查詢知網的概念定義後，所得到的都是「time|時間, autumn|秋」（如表 3 所示）。另外，同義詞集{clay, mud}的中文翻譯「泥，泥巴，泥漿」對應知網後所得的也都是一致的概念定義「stone|土石」。像表 3 列出的這些單義詞彙，我們可以直接將它們標上唯一的概念特徵。

***表 3*** *部分中文單義詞及其在知網的概念定義。*

| 中文詞彙 | 知網的概念定義 |
|---|---|
| 秋 | time\|時間,autumn\|秋 |
| 秋天 | time\|時間,autumn\|秋 |
| 秋季 | time\|時間,autumn\|秋 |
| 秋海棠 | FlowerGrass\|花草 |
| 光束 | lights\|光 |
| 光線 | lights\|光 |
| 泥 | stone\|土石 |
| 泥巴 | stone\|土石 |
| 泥漿 | stone\|土石 |

### 3.3.2 岐義詞部分

對於歧義的中文詞彙，在知網中會存在一個以上的概念定義，例如「分號」與「目標」這兩個詞彙在知網中各自有兩個概念定義。如表 4 所示，「分號」可能是標點符號「；」，也可能代表商場上的分支機構。至於，表 5 是這兩個中文歧義詞當為 WordNet 同義詞集的中文翻譯之例子。其中，同義詞集{semicolon}因為它的英文詞彙與知網中具{symbol|符號}概念之詞條的英文詞彙相同。因此，我們可以將此中文翻譯「分號」標上{symbol|符號}概念。但是，這樣的作法對於歧義詞彙「目標」卻是行不通的。比對表 4 及表 5 的內容之後，我們可發現 WordNet 同義詞集{aim, object, objective, target}與「目標」在知網中的兩項定義皆有共同的英文詞彙。其中，與概念定義{purpose|目的}擁有相同的英文詞彙 ‘aim’ 以及 ‘objective’，而與概念定義{tool|用具,#weapon|武器，$AimAt|定向，$firing|射擊}也同時擁有相同的英文詞彙‘target’。另外，從表 6 所呈現的上位詞資訊中，我們也可以發現同義詞集{aim, object, objective, target}的上位詞中文翻譯「目的」與{purpose|目的}概念中的概念名稱完全相同。綜合上述資訊，我們可以判定同義詞集同義詞集{aim, object, objective, target}的正確知網概念為{purpose|目的}。

**表 4** 「分號」與「目標」在知網的概念定義與其英文詞彙。

| 中文詞彙 | 知網的概念定義 | 英文詞彙 |
|---|---|---|
| 分號 | InstitutePlace\|場所, branch\|支, commercial\|商 | branch |
| 分號 | symbol\|符號 | semicolon |
| 目標 | purpose\|目的 | aim, goal, objective |
| 目標 | tool\|用具,#weapon\|武器, $AimAt\|定向, $firing\|射擊 | target |

**表 5** 幾個以「分號」與「目標」為中文翻譯的同義詞集及其定義。

| WordNet 同義詞集 | WordNet 定義 | 中文翻譯 |
|---|---|---|
| semicolon | a punctuation mark (;) used to connect independent clauses; indicates a closer relation than does a period | 分號 |
| butt, target | an object set up for a marksman or archer to aim at | 目標 |
| aim, object, objective, target | the goal intended to be attained (and which is believed to be attainable) | 目標 |

**表 6** 表 5 的同義詞集所含的上位詞及其相關訊息。

| WordNet 同義詞集 | WordNet 上位詞同義詞集 | WordNet 上位詞定義 | 上位詞 中文翻譯 |
|---|---|---|---|
| butt, target | sports equipment | equipment needed to participate in a particular sport | 體育 裝備 |
| aim, object, objective, target | end, goal | the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it | 目的, 終極 |

## 4. 判定中文複合名詞的中心詞彙方法

對於中文複合詞，無法以直接連結字典的方式來進行語意標示。我們分兩個階段完成所需要的處理。首先，要作的就是區分它的中心詞彙與概念特徵，也就是要找出它的主要特徵詞彙與次要特徵詞彙。然後，再標示主要特徵詞彙之語意，以主要特徵詞彙之語意代表整個複合詞之主要語意。透過先前的觀察，我們發現同義詞集的中文翻譯複合詞與其上位詞之中文翻譯常擁有相同的詞素。在此，我們先將詞彙逐字拆解成詞素，以所得

的詞素集合來代表該詞彙。再透過 Dice 係數[Dice, 1945]估算兩個詞素集的相似度，當作兩個中文詞彙的相似度。針對給定的兩個中文詞彙 $W_1$，$W_2$，它們的相似度計算公式 Sim($W_1$, $W_2$)如公式－所示。

$$Sim\ (W_1, W_2) = \frac{2 \times |\ S(W_1) \cap S(W_2)\ |}{|\ S(W_1)\ | + |\ S(W_2)\ |} \qquad \text{（公式－）}$$

其中，$W_i$ ：中文詞彙，
$S(W_i)$ ：將中文詞彙 $W_i$拆解成詞素，所得的詞素集合，
$|S(W_i)|$ ：詞素集合 $S(W_i)$的長度。

## 5. 詞彙與詞義概念的對應方法

### 5.1 對應方法設計

在本節中，我們設法將 WordNet 同義詞集所連結的中文翻譯詞彙，對應到知網的詞義概念。透過觀察，我們知道單義詞的詞義概念直接標示即可。因此，以下僅就岐義詞部分討論它們與詞義概念的對應方法。

對於一詞多義的詞彙，我們需有一套辨識歧義的方法，來為詞彙找出最適當的詞義對應。透過第三節的觀察，我們發現不同的資源在定義同個詞彙的相同語意時，常常會使用共同的中英文詞彙。這種現象有時發生在中文翻譯，有時出現的卻是使用相同的概念語詞。

對於一個 WordNet 的同義詞集 $S$，以及它已標示的 n 個中文翻譯詞彙 $W_1$，$W_2$，…，$W_n$,，假設 $W_i$ 在知網中有 k 個不同的概念定義 $D_{i1,}$ $D_{i2,}$ …, $D_{ik}$，我們定義集合 $E_{ij}$ 表示詞彙 $W_i$ 在定義 $D_{ij}$ 中對應的英文詞彙所形成的集合。我們令同義詞集 $S$ 與 $E_{ij}$ 交集元素最多的定義 $D_{ij}$ 為 $S$ 的概念定義，即 CDEF($S$)，如公式二所示。

$$\text{CDEF}\ (S) = \arg\max_{i,j} |\ S \cap E_{i,j}\ |, \quad \forall\ i = 1, ..., n,\ \ j = 1, ..., k_i \qquad \text{(公式二)}$$

其中，$n$ ：同義詞集 $S$ 已標示的中文翻譯詞彙數，
$k_i$ ：同義詞集 $S$ 的第 i 個中文翻譯詞彙 $W_i$ 在知網中的概念定義數，
$E_{i,j}$ ：表示詞彙 $W_i$ 在定義 $D_{ij}$ 中對應的英文詞彙所形成的集合。

### 5.2 對應例子

下面我們以例子說明將 WordNet 同義詞集透過中文翻譯設定其詞義概念對應的方法。表 7 給出兩個同義詞集例子，其中，同義詞集{campaign, cause, crusade, drive, effort, movement}有兩個中文翻譯詞彙「運動」（稱 $W_1$）與「活動」（稱 $W_2$）。由表 8 可看到詞彙「運動」在知網中有三個詞義概念，分別為「fact|事情,function|活動,politics|政」（稱 $D_{11}$）、「fact|事情,exercise|鍛練,sport|體育」（稱 $D_{12}$）以及「fact|事情,AlterLocation|變

空間位置」(稱 $D_{13}$), 至於詞彙「活動」在知網中僅有一個詞義概念 $D_{21}$ 為「fact|事情,generic|統稱」。四個詞義概念所對應之英文詞彙分別為{campaign, drive, movement}, {athletics, exercise, sports}, {motion, movement}以及{activity, maneuve}, 它們與同義詞集之交集分別為{campaign, drive, movement}、空集合、{movement}以及空集合, 因此, 同義詞集{campaign, cause, crusade, drive, effort, movement}之詞義概念為 $D_{11}$「fact|事情,function|活動,politics|政」。同樣的方法, 我們可為同義詞集{flank, wing}標上詞義概念「place|地方,edge|邊,military|軍」。

***表 7*** *為同義詞集對應詞義概念方法的例子。*

| WordNet 同義詞集 | WordNet 定義 | 中文翻譯 |
|---|---|---|
| {campaign, cause, crusade, drive, effort, movement} | a series of actions advancing a principle or tending toward a particular end | 運動, 活動 |
| {flank, wing} | the side of military or naval formation | 翼 |

***表 8*** *表 7 之中文詞彙出現在知網的定義及其英文詞彙。*

| 中文詞彙 | 知網對應英文詞彙 | 知網詞義概念 |
|---|---|---|
| 運動 | {campaign, drive, movement} | fact\|事情,function\|活動,politics\|政 |
| | {athletics, exercise, sports} | fact\|事情,exercise\|鍛練,sport\|體育 |
| | {motion, movement} | fact\|事情,AlterLocation\|變空間位置 |
| 活動 | {activity, maneuve} | fact\|事情,generic\|統稱 |
| 翼 | {wing} | part\|部件,%artifact\|人工物,wing\|翅 |
| | {flank, wing} | place\|地方,edge\|邊,military\|軍 |
| | {wing} | part\|部件,%bird\|禽,wing\|翅,*fly\|飛 |

註： 標示底線者, 表示與同義詞集共用詞彙。

## 6. 實驗

### 6.1 實驗資料

我們將實驗分為由中文複合名詞標示出中心詞彙及中文詞彙語意概念標示兩個部分。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同義詞集個數比對於參與比對的總同義詞集數, 稱為應用率, 至於, 正確率定義為標示結果的正確比率。

本研究以標示中文翻譯的英文 1.6 版 WordNet 同義詞集進行實驗，實驗資料來自於標示中文翻譯的 53753 個名詞詞網同義詞集，經與知網所含詞彙比對後，其中 13628 個同義詞集的中文翻譯出現在知網中。至於含單義中文翻譯詞彙的同義詞集則有 12231 個，佔總同義詞集數的 22.8%，實驗資料分佈情形如表 9 所示。我們將中文翻譯含知網詞彙的同義詞集直接進行標示語意概念實驗，其餘的同義詞集資料則進行中心詞彙標示工作。

***表 9*** *實驗資料之分佈統計。*

| | |
|---|---|
| 標示中文翻譯的同義詞集數 | 53753 |
| 中文翻譯含知網詞彙的同義詞集數 | 13628 |
| 中文翻譯含單義知網詞彙的同義詞集數 | 12231 |
| 中文翻譯全為多義知網詞彙的同義詞集數 | 1397 |

## 6.2 標示語意概念實驗

### 6.2.1實驗設定與結果

在標示中文詞彙語意概念部分，我們先將中文詞彙依其在知網的定義，分為單義詞與歧義詞兩個部分，再進行不同之處理。若為單義詞彙，則詞彙語意直接連結知網的定義。若為岐義詞彙, 則依第 5 節之公式做出適當之語意連結。實驗結果在 1397 個不含單義中文翻譯詞彙的同義詞集中，有 546 個同義詞集可成功產生語意連結。因此，在語意連結之實驗共可產生 12777 筆連結，比對於參與語意標示實驗的 13628 個同義詞集，應用率為 93.8%(如表 10 所示)。為評估連結之正確率，作者自產生之 12777 筆連結資料中隨機選取 630 筆，經人工比對後計有 589 筆為正確之連結，正確率為 93.5%（詳見表 10）。

***表 10*** *詞彙語意標示實驗所得結果。*

| | |
|---|---|
| 進行語意連結的同義詞集數 | 13628 |
| 產生語意連結的同義詞集數 | 12777 |
| 應用率 | 93.8% |
| 人工比對資料筆數 | 630 |
| 正確連結資料筆數 | 589 |
| 正確率 | 93.5% |

### 6.2.2 實驗結果討論與錯誤分析

由實驗結果, 我們發現有些同義詞集無法成功的產生語意連結, 其原因主要有下列幾種：

1. 詞網的同義詞集與知網兩項資源未出現共用詞彙，2.多個概念定義與同義詞集擁有相同的交集個數，3. 不同詞義的中、英文詞彙皆擁有共用詞彙。

本文所提方法以共用詞彙作為相同詞義的徵象。但若兩份資源在同一詞義的詞彙表達上，作了不同的詞彙選擇，我們目前無法做出對應。這種情形共有 663 個同義詞集，比對於所有無法產生連結的 851 個同義詞集共佔 77.9%。

當多個概念定義與同義詞集擁有相同的交集個數時，我們現行的方法無法判斷出最合適的對應概念。例如表 11 一所呈現的同義詞集{pumpkin}及{Japanese plum, loquat}，即屬於這種情形。同義詞集{pumpkin}與其兩個語意概念「vegetable|蔬菜」及「part|部件,%vegetable|蔬菜,embryo|胚,$eat|吃」都同時擁有共同詞彙{pumpkin}，而且，這兩個概念應該都屬於可接受的連結。另外，詞彙「枇杷」應屬於單義詞彙，但是在知網上出現兩個定義，並且此兩份概念定義所對應的英文詞彙完全相同。若是將同義詞集{Japanese plum, loquat}對應至「fruit|水果」或「tree|樹」，應該都算是正確的對應。

通常不同詞義的原始語在翻譯至目標語時會有不同的詞彙選擇，例如，「river bank」與「money bank」在翻譯至中文時會分別使用不同的詞彙「河岸」及「銀行」。但是，表 11 的「鳶」在「tool|用具,*WhileAway|消閑」及「bird|禽」兩個很不同的語意概念中，它們的英文詞彙卻都是「kite」，這種情形，只以對應詞彙本身作為辨識歧義的依據，顯然是不夠的。

***表 11*** *無法成功產生語意連結的同義詞集例子。*

| 同義詞集 | WordNet 定義 | 中文翻譯 | 知網對應英文詞彙 | 知網詞義概念 |
|---|---|---|---|---|
| {pumpkin} | usually large pulpy deep-yellow round fruit of the squash family maturing in late summer or early autumn | 南瓜 | {pumpkin, cushaw} | vegetable\|蔬菜 |
| | | | {pumpkin, cushaw} | part\|部件, %vegetable\|蔬菜, embryo\|胚,$eat\|吃 |
| {Japanese plum, loquat} | yellow olive-sized semitropical fruit with a large free stone and relatively little flesh; used for jellies | 枇杷 | {loquat} | fruit\|水果 |
| | | | {loquat} | tree\|樹 |
| {kite} | any of several small graceful hawks of the family Accipitridae having long pointed wings and feeding on insects and small animals | 鳶 | {kite} | tool\|用具, *WhileAway\|消閑 |
| | | | {kite} | bird\|禽 |

註： 標示底線者，表示與同義詞集共用詞彙。

## 6.3 標示語意中心詞彙實驗

我們先將複合名詞進行斷詞拆解成詞彙，在 39228 個中文複合詞中，所拆解出的詞彙組合數分佈如表 12 所示，以含兩個詞彙之複合詞最多，佔 56.9%。對於由兩個詞彙組成的複合名詞，經第 4 節的詞彙相似度量公式，選出語意中心詞彙，實驗結果在詞彙組合數為 2 的部分, 複合詞之中心詞彙來自第一個詞彙的有 2120 筆, 來自第二個詞彙的有 10481 筆，而無法成功辨識出中心詞彙的同義詞集共有 9727 筆（見表 13）。

接著，我們將標上中心詞彙的同義詞集資料依第 6.2 節所述方法進行語意概念標示，結果有 5756 個同義詞集可標上語意概念。因此，比對於參與語意標示實驗的 11439 個同義詞集，應用率為 50.3%。為評估連結之正確率，我們自連結資料中隨機選取 288 筆，經人工比對後計有 261 筆為正確之連結，正確率為 90.6 %。實驗結果與 6.2 節比較正確率差異不大，但是應用率卻降低不少，主要原因為中文複合詞與其語意中心詞彙在對應之英文詞彙上，較難一致。若是考慮將對英文詞彙的一致要求，轉至概念上的一致性，應可提升應用率。

***表 12*** *參與中心詞彙標示實驗資料之複合名詞之詞彙組合數統計。*

| | 詞彙組合數 | | |
|---|---|---|---|
| | 2 | 3 | 4 |
| 複合詞數 | 22330 | 9742 | 4173 |
| 同義詞集總數 | 19323 | 8994 | 3991 |

***表 13*** *複合名詞中心詞彙標示之實驗結果。*

| | | 詞彙組合數 | | |
|---|---|---|---|---|
| | | 2 | 3 | 4 |
| 複合詞數 | | 22330 | 9742 | 4173 |
| 語意中心詞彙位置 | 無法判定 | 9729 | 3931 | 1482 |
| | 詞彙一 | 2120 | 599 | 240 |
| | 詞彙二 | 10481 | 790 | 207 |
| | 詞彙三 | ------ | 4422 | 299 |
| | 詞彙四 | ------ | ------ | 1945 |
| 應用率 | | 56.4% | 59.7% | 64.5% |
| 正確率 | | 84.8% | 75.2% | 70.0% |

## 7. 結論

本研究提出一套方法, 結合比對概念語詞、詞彙本身、以及上位關係的詞彙內容等技巧, 將已標示中文翻譯的 WordNet 同義詞集對應上知網概念定義。並且, 對於中文複合詞找出主要特徵所在。實驗顯示有不錯的結果。在未來的研究擬將詞彙相似度比對由詞彙擴充至概念, 克服同義詞問題, 以提升應用率。

## 致謝

本研究獲得國科會編號 NSC 90-2213-E-031-005 計畫之補助, 特此致謝。

## 參考資料

Aslandogam, Y. A., C. Their, C. T. Yu, J. Zou and N. Rishe, "Using Semantic Contents and WordNet in Image Retrieval," In *Proceedings of the 20th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, Philadelphia, 1997, pp. 286-295.

Atserias, J., S., Climent, X., Farreres, G. Rigau and H. Rodriguez, "Combining Multiple Methods for the Automatic Construction of Multilingual WordNets," In *Proceedings of International Conference of Recent Advances in Natural Language Processing* (*RANLP'97*), Tzigov Chark, Blgaria, 1997.

Carpuat, M., G. Ngai, P. Fung and K. W. Church, "Creating a Bilingual Ontology: A Corpus-Based Approach for Aligning WordNet and HowNet," In *Proceedings of the 1st International Conference on Global WordNet*, Mysore, India, 2002.

Chang, J. S., S. J. Ker and M. H. Chen, "Taxonomy and Lexical Semantics - from the Perspective of Machine Readable Dictionary," In *Proceedings of 3rd Conference of the Association for Machine Translation in the Americas* (*AMTA'98*), 1998, pp. 199-212.

Chen, J. N. and J. S. Chang, "TopSense: A Topical Sense Clustering Method based on Information Retrieval Techniques on Machine Readable Resources," *Special Issue on Word Sense Disambiguation, Computational Linguistics*, 24(1), 1998, pp. 61-95.

Chen, Hsin-Hsi, Chi-Ching Lin and Wen-Cheng Lin, "Construction of a Chinese-English WordNet and Its Application to CLIR," In *Proceedings of 5th International Workshop on Information Retrieval with Asian Languages*, Hong Kong, 2000, pp. 189-196.

Chen, Hsin-Hsi and Chi-Ching Lin, "Sense-Tagging Chinese Corpus," In *Proceedings of 2nd Chinese Language Processing Workshop*, Hong Kong, 2000, pp. 7-14.

Dagan, I. and A. Itai, "Word Sense Disambiguation Using a Second Language Monolingual Corpus," *Computational Linguistics*, 20(4), 1994, pp. 563-596.

Dice, L. R., "Measure of the Amount of Ecologic Association between Species," *Journal of Ecolog*, 26, 1945, pp. 297-302.

Dorr, B. J., G-A Levow, D. Lin and S. Thomas, "Chinese-English Semantic Resource Construction," In *Proceedings of 2nd International Conference on Language Resources and Evaluation*, (*LREC 2000*), Athens, Greece, 2000, pp. 757-760.

Farreres, X., G. Rigau and H., Rodriguez, "Using WordNet for Building WordNets," In *Proceedings of the Workshop of Usage of WordNet in NLPS*, COLING-ACL'98, 1998, pp. 65-72.

Fellbaum, C. ed., *WordNet: An Electronic Lexical Database*, MIT Press, May 1998.

Gale, W. A., K. W. Church and D. Yarowsky, "Using Bilingual Materials to Develop Word Sense Disambiguation Methods," In *Proceedings of the 4th International Conference on Theoretical and Methodological Issues in Machine Translation*, 1992, pp. 101-112.

Gomez, F., "Linking WordNet Verb Classes to Semantic Interpretation," In *Proceedings of the Workshop of Usage of WordNet in NLPS, COLING-ACL'98*, 1998, pp. 58-64.

Gonzalo, J., F. Verdejo, I. Chugur and J. Cigarran, "Indexing with WordNet Synsets can Improve Text Retrieval," In *Proceedings of the Workshop of Usage of WordNet in NLPS, COLING-ACL'98*, 1998, pp. 38-44.

Guthrie, J., L. Guthrie, Y. Wilks and H. Aidinejad, "Subject-Dependent Co-Occurrence and Word Sense Disambiguation," In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 1991, pp. 146-152.

Jing, H., "Usage of WordNet in Natural Language Generation," In *Proceedings of the Workshop of Usage of WordNet in NLPS*, COLING-ACL'98, 1998, pp. 128-134.

Knight, K. and S. K. Luk, "Building a Large-scale Knowledge Base for Machine Translation," In *Proceedings of The Twelfth National Conference on Artificial Intelligence*, 1994, pp. 773-778.

Lee, C., G. Lee and S. J. Yun, "Automatic WordNet Mapping using Word Sense Disambiguation," In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 142-147.

Li, X., S. Szpakowicz and S. Matwin, "A WordNet-Based Algorithm for Word Semantic Sense Disambiguation," In *Proceedings of the 14th International Joint Conference on Artificial Intelligence IJCAL-95*, Montreal, Canada, 1995.

Luk, A. K., "Statistical Sense Disambiguation with Relatively Small Corpora using Dictionary Definitions," In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 181-188.

Mandala, R., T. Tokunaga and H. Tanaka, "The use of WordNet in Information Retrieval," In *Proceedings of the Workshop of Usage of WordNet in NLPS, COLING-ACL'98*, 1998, pp. 31-37.

Miller, G. A., "Five papers on WordNet," *International Journal of Lexicography*, 3(4) 1990.

Ng, H. T. and H. B. Lee, "Integrating Multiple Knowledge Sources to Disambiguate Word Sense: an Exemplar-Based Approach," In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, Santa Cruz, 1996, pp. 40-47.

Resnik, P., "Selection and Information: A Class-Based Approach to Lexical Relationships," *Doctoral Dissertation, Department of Computer and Information Science, University of Pennsylvania*, 1993.

Riloff, E. and R. Jones, "Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping," In *Proceedings of 16th National Conference on Artificial Intelligence*, 1999, pp. 474-479.

Slator, B., "Using Context for Sense Preference," In Zernik (ed.) *Lexical Acquisition: Exploiting on-line Resources to Build a Lexicon*, Lawrence Erlbaum, Hillsdale, NJ, 1991.

Wang, Chi-Yung, "Knowledge-based Sense Pruning using the HowNet: An Alternative to Word Sense Disambiguation," *Thesis of Hong Kong University of Science and Technology*, *Computer Science*, 2002.

Yang, C. and S. J. Ker, "Considerations of Linking WordNet with MRD," In *Proceedings of the 19th International Conference on Computational Linguistics*, 2002, pp. 1121-1127.

Yarowsky, D., "Unsupervised Word Sense Disambiguation Rivalling Supervised Methods," In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995, pp. 189-196.

Yarowsky, D., "Word-Sense Disambiguation using Statistical Models of Roget's Categories Trained on Large Corpora," In *Proceedings of the 14th International Conference on Computational Linguistics*, Nantes, France, 1992, pp. 454-460.

董振東、董強：知網, 2000, http://www.keenage.com/.