中央研究院中英雙語詞網

The Academia Sinica Bilingual WordNet

1 簡介

「中央研究院中英雙語詞網」(The Academia Sinica Bilingual Wordnet)，簡稱「研究院雙語詞網」(Sinica BW)，網址：http://bow.sinica.edu.tw/為一涵蓋約十萬英文同義詞集（SynSet）之中英雙語電子資料庫。本資料庫以英文WordNet架構為基礎，並以以台灣地區的語言使用為經驗基礎。提供的訊息包含中英雙語跨語言資訊轉換、詞義的區分與詞義關係的連結以及使用領域（含使用之頻率）。讓不同來源的典藏知識內容，可以轉換成互通的 (inter-operable) 訊息。所引用的資料主要為中央研究院文獻語料庫（語言所），詞庫小組（資訊所）開發的資料外。另外引用了普林斯頓大學（Princeton University）的WordNet（http://wordnet.princeton.edu/），以及遠見科技股份有限公司與中研院共同開發資料。本資料庫之原始內容（除英語WordNet單語資料庫由普林斯頓大學開發擁有並公開授權。）其智財權由中央研究院與遠見科技股份有限公司共同持有。公開授權資料分別以純文字以及XML檔案格式儲存。

2 資料內容說明

開放授權的資料包括下列所示的二大類訊息：

A. 領域分類樹：

參考「中國圖書分類法」為基準，並參考各知識分類與實際研究經驗，提出：包含九大類的知識分類（Knowledge Content），涵蓋 438 個領域，並因應語言資源特性加入下列語言使用（Language Usage）的各類訊息：專名（說明文字符號的指涉）（Proper Name）、語體（說明文字符號的使用）（Genre/Strata）、各種語言／詞源（Language/Etymology）、各國地名（Country Name）。知識分類（Knowledge Content）的九大類分別是：人文學科（Humanities）、社會科學（Social Science）、形式科學（Formal Science）、自然科學（Natural Science）、醫療科學（Medical Science）、工程科學（Engineering Science）、應用產業（Production Industry）、藝術（Fine Arts）以及休閒娛樂（Recreation）。

B. 英漢雙語詞網對應資料庫：英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類為基準，每一筆紀錄皆標示以下訊息：

（1） 英文詞形

（2） 英文 WordNet1.6 同義詞集 offset：該英文詞形對應的 WordNet 1.6 版本同義詞集 offset，詳細資料請請參閱參考文件一。

（3） 詞類：詞類，包含名詞（Noun）、動詞（Verb）、形容詞（Adjective）以及副詞（Adverb），詳細資料請請參閱參考文件一。

（4） 英文詞形搭配詞類所屬的頻率等級：據詞彙分佈情況可分為三層次依序為核心詞彙、通用詞彙以及參考詞彙。英文區分核心、通用與參考詞彙的原則為：(1)、核心詞彙是 BNC, Brown, CIDE 等幾個語料庫取累計頻率 66%時的所有詞，再把各語料庫的個別詞表交集，得到的結果便是核心詞彙。(2)、通用詞彙：同上，但累計頻率取到 80%。(3)、參考詞彙：其餘在各英文資源有收的詞。

（5） 中文對譯詞形：在 WordNet1.6 同義詞集中該英文詞形對譯的中文詞形，詳細資料請請參閱參考文件二。

（6） 中文對譯詞形搭配詞類所屬的頻率等級：根據中文詞彙分佈情況可分為四層次依序為核心詞彙、通用詞彙、參考詞彙以及一般詞彙。參考詞彙包含了通用詞彙與核心詞彙，通用詞彙則包含了核心詞彙。(1)、核心詞彙是指所參考的五本辭典都列的詞且出現在中研院平衡語料庫語料庫十次以上。(2)、通用詞彙則收錄在任意三本辭典以上的詞且出現在中研院語料庫四次以上。(3)、參考詞彙指的是收錄在三本以上辭典的詞，或收錄在五本辭典中任一本且出現在中研院語料庫一次以上，或者是同義詞詞林的標題詞。區分原則的詳細資料請請參閱參考文件三。

（7） 詞彙領域分類：針對該英文詞形在 WordNet 1.6 同義詞集給與相對應於領域分類樹之領域訊息，詳細資料請請參閱參考文件四、五。

WordNet1.6 同義詞集原本共 99,642 筆，英文詞形為 122,045 個，中文詞形為 109,970 個，以英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類為基準，總共有 173,941 筆資料，以英文詞義為基礎其中 24,222 筆有領域訊息。

3 範例

A. 領域分類樹：檔案名稱 20050311domain.xls

人文學科 Humanities

語言學 linguistics

B. 漢英雙語詞網對應資料庫：以英文詞形、對應的 WordNet1.6 之同義詞集（synset）以及詞類為基準，每一筆紀錄皆標示以下訊息：

I. 純文字檔：檔案名稱 SinicaBOW_License1.0 英漢雙語詞網.txt

65 exercise Noun 00469856 通用詞彙 例題 ◎ 通用詞彙
66 exercise Noun 00411620通用詞彙 體操 ◎ 核心詞彙 、 運動 ◎ 核心詞彙 體操 ◎ gymnastics

II. XML 檔：檔案名稱 SinicaBOW_License1.0 英漢雙語詞網.xml

```
<Record Conut="65">
    <EnglishLemma>exercise</EnglishLemma>
    <POS>Noun</POS>
    <WordNetSynsetOffset Version="1.6">00469856</WordNetSynsetOffset>
    <EnglishFrequancyRank>通用詞彙</EnglishFrequancyRank>
    <ChineseTransList>
        <ChineseTrans>
            <ChineseLemma>例題</ChineseLemma>
            <ChineseFrequancyRank>通用詞彙</ChineseFrequancyRank>
        </ChineseTrans>
    </ChineseTransList>
</Record>
<Record Conut="66">
    <EnglishLemma>exercise</EnglishLemma>
    <POS>Noun</POS>
    <WordNetSynsetOffset Version="1.6">00411620</WordNetSynsetOffset>
    <EnglishFrequancyRank>通用詞彙</EnglishFrequancyRank>
    <ChineseTransList>
        <ChineseTrans>
            <ChineseLemma>體操</ChineseLemma>
            <ChineseFrequancyRank>核心詞彙</ChineseFrequancyRank>
        </ChineseTrans>
        <ChineseTrans>
            <ChineseLemma>運動</ChineseLemma>
            <ChineseFrequancyRank>核心詞彙</ChineseFrequancyRank>
        </ChineseTrans>
    </ChineseTransList>
```

```
        <DomainList>
            <Domain>
                <ChineseDomain>體操</ChineseDomain>
                <EnglishDomain>gymnastics</ChineseDomain>
            </Domain>
        </DomainList>
    </Record>
```

4 參考文件

1. WordNet Reference Manual, http://wordnet.princeton.edu/doc/
2. Huang, Chu-Ren. Elanna I. J. Tseng, Dylan B. S. Tsai, Brian Murphy. (2003). “Cross-lingual Portability of Semantic relations: Bootstrapping Chinese WordNet with English WordNet Relations”. Language and Linguistics. 4.3.509-532.
3. 中文分詞標準計畫相關資料下載區一四、成果效益檢討內容 http://ckip.iis.sinica.edu.tw/ROCLING/4.htm, Accessed：2003/04/15
4. Echa Chang, Chu-Ren Huang, Sue-Jin Ker, Chang-Hua Yang. (2002). “Induction of Classification from Lexicon Expansion :Assigning Domain Tags to WordNet Entries“. 19th COLING 2002 Post-Conference Workshop --SEMANET: Building and Using Semantic Networks Processing. Center of Academia Activities, Academia Sinica. Taipei. Taiwan. September 1, 2002. pp.80-86.
5. Chu-Ren Huang, Xiang-Bing Li, Jia-Fei Hong. (2004). “Domain Lexico-Taxonomy: An Approach Towards Multi-domain Language Processing“. The 1st International Joint Conference on Language Language Processing (IJCNLP-04) Asian Symposium on Natural Language Processing to Overcome Language Barriers. Sanya City. Hainan Island, China. 25-26 March,2004 . pp.54-60.
6. Chu-Ren Huang, Ru-Yng Chang, Shiang-Bin Lee. (2004). “Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO”. 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon. Portugal. 26-28 May, 2004. pp.1553-1556.

Sinica BOW: http://BOW.sinica.edu.tw/
WordNet: http://wordnet.princeton.edu/