

中央研究院中英雙語知識本體資料庫

The Academia Sinica Bilingual Ontological Database

1 簡介

「中央研究院中英雙語知識本體資料庫」(The Academia Sinica Bilingual Ontological Database), 網址: <http://bow.sinica.edu.tw/> 為一涵蓋約十一萬中文詞之中英雙語知識本體資料庫。本資料庫以IEEE批准執行的SUMO (Suggested Upper Merged Ontology, <http://www.ontologyportal.org>) 知識本體為基礎架構, 並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知識本體架構與概念內容, 中文語言資訊與概念架構(知識本體)的連結。在使用語言與詞彙資料的基礎上, 提供了知識運籌的基本架構(infrastructure)。讓不同來源的典藏知識內容, 可以轉換成互通的 (inter-operable) 訊息。SUMO資料庫之英文原始內容由Institute of Electrical and Electronics Engineers, Inc.所有並公開授權; 中文原始資料由語數位典藏國家型科技計畫技術分項計畫下之語言座標計畫, (NSC94-2422-H-001-009-)開發。經費由國科會補助, 中央研究院執行並持有中央研究院中英雙語知識本體資料庫。公開授權資料分別以純文字以及XML檔案格式儲存。

The SUMO is freely available, subject to the following IEEE license :

Copyright © 2004 by the Institute of Electrical and Electronics Engineers, Inc.

http://suo.ieee.org/SUO/SUMO/SUMO_173.kif

The mappings from SUMO to WordNet are copyrighted by Teknowledge and Articulate Software 2002-2006. They are released under the GNU Public License:

<http://www.gnu.org/copyleft/gpl.html> .

2 引用與致謝

授權本資料庫發表著作, 引用本資料庫內容, 請引用參考文獻 4,5,以及網頁 6. 致謝對象為中央研究院 (Academia Sinica)。

引用 SUMO 內容, 請引用參考文獻 1 以及網頁 7。致謝對象為 IEEE。

引用 SUMO 與 WordNet 對應內容, 請引用參考文獻 2 以及網頁 7。

3 資料內容說明

開放授權的中英雙語領域分類樹資料包括下列所示的三大類訊息：

A. SUMO 知識本體中文版：

依據 SUMO2002 版本的中文版資料, 涵蓋 11 大類的概念, 每大類又分為二至五個類別, 總共囊括 3,912 個概念, 每項概念下附有解釋與定理, 詳細資料請參閱參考文件一。

B. 中英雙語領域分類樹：

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 427 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)。

C. 中文與 SUMO 知識本體對應資料庫：

由中文詞彙出發，經由以英文 WordNet 定義的同義詞集 (synset) 為基準，對應到 SUMO 知識本體的概念節點。資料內容包括：

- (1) 中文詞彙：109,982 詞形；部份詞彙對應到一個以上英文同義詞集，以每個同義詞集對應計一次，共有 149,780 筆資料。
- (2) 每個詞形所對應的英語 WordNet 同義詞集記錄碼(offset)。共對應到 99,642 個同義詞集。
- (3) 同義詞集與 SUMO 知識本體的對應關係與知識本體分類 (SUMO 概念) 間的關係：SUMO 工作小組將 WordNet1.6 每個同義詞集對應 SUMO 概念，並針對兩者之間對應的關係分成三種，分別是一上位 (Hypernym)、同義 (Synonymy) 以及示例 (Instantiation)，詳細資料請參閱參考文件四。
- (4) 知識本體分類 (SUMO 概念英文)：SUMO 概念的英文版。
- (5) 知識本體分類 (對應 SUMO 概念中文)：SUMO 概念的中文版。
- (6) 詞彙領域分類：針對同義詞集給與相對應於領域分類樹之的領域訊息，詳細資料請參閱參考文件四。

4 範例

- A. SUMO 知識本體中文版：檔案名稱 SUMO_Chinese_2002.txt
(termFormat cn Abstract "抽象的")
(termFormat cn AbstractionFn "描述函數")

- B. 中英雙語領域分類樹：檔案名稱 20050311domain.xls
人文學科 Humanities
語言學 linguistics

- C. 中文與 SUMO 知識本體對應資料庫：由中文詞彙出發，經由以英文 WordNet 定義的同義詞集 (synset) 為基準，對應到 SUMO 知識本體的概念節點。

中文與SUMO知識詞網

每筆資料包括:

- (1) 序號 (無特殊意義，主鍵為「中文詞形」+「詞類」+「相對應的 WordNet1.6 同義詞集offset」)
- (2) 中文詞形
- (3) 詞類

(4) 與 WordNet 對應關係—ISA 對應關係 OF 相對應的 WordNet1.6 同義詞集 offset

(5) 知識本體分類-- @ SUMO 概念 @ SUMO 中文

I. 純文字檔：檔案名稱 Ontological_License 中英雙語知識本體.txt

- 1 安布利亞 Noun ISA Instantiation OF 06467131N @ LandArea @ 陸地
- 2 楓香樹 Noun ISA Hypernym OF 08619764N @ FloweringPlant @ 開花植物

II. XML 檔：檔案名稱 Ontological_License 中英雙語知識本體.xml

```
<Record Conut="1">
  <ChineseLemma>安布利亞</ChineseLemma>
  <POS>Noun</POS>
  <WordNetSynsetOffset Version="1.6">06467131N</WordNetSynsetOffset>
  <SUMO>
    <SUMORelation>Instantiation</SUMORelation>
    <SUMOConcept>LandArea</SUMOConcept>
    <SUMOChi>陸地</SUMOChi>
  </SUMO>
</Record Conut="1">
<Record Conut="2">
  <ChineseLemma>楓香樹</ChineseLemma>
  <POS>Noun</POS>
  <WordNetSynsetOffset Version="1.6">08619764N</WordNetSynsetOffset>
  <SUMO>
    <SUMORelation>Hypernym</SUMORelation>
    <SUMOConcept>FloweringPlant</SUMOConcept>
    <SUMOChi>開花植物</SUMOChi>
  </SUMO>
</Record>
```

5 參考文件

1. Niles, I., & Pease, A. (2001). "Toward a Standard Upper Ontology". In Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001). Ogunquit, Maine, October 17-19, 2001. pp.2-9.
2. WordNet Reference Manual, <http://www.cogsci.princeton.edu/~wn/doc.shtml>
3. Niles, I., and Pease, A. (2003). "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology". Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003). Las Vegas, Nevada, June 23-26, 2003. pp.412-416.
4. 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.

5. 張如瑩,黃居仁. 2004. 中央研究院中英雙語知識本體詞網 (Sinica BOW) : 結合詞網,知識本體,與領域標記的詞彙知識庫。發表於第十六屆自然語言與語音處理研討會 (ROCLING XVI) September 2-3.
6. Sinica BOW: <http://BOW.sinica.edu.tw/>
7. SUMO: <http://www.ontologyportal.org>