GuangQunFangPu: e-Humanities Combining Textual and Botanic Information

Shu-Kai Hsieh Shu-Ming Chang Chun-Han Chang Yi-Shuan Zhou Chu-Ren Huang
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
churen@gate.sinica.edu.tw

Feng-Ju Lo
Yuan-Ze University
Taoyuan, Taiwan
getfjulo@saturn.yzu.edu.tw

Ru-Ying Chang
National Cheng-Kung University
Tainan, Taiwan

Abstract

In this paper, we propose a lexicon-driven and ontology-merging methodology of constructing diachronic domain knowledge via Sinica BOW, a bilingual ontological lexical resource based on WordNet and SUMO ontology. The main domain knowledge that we model our specialized ontology on is GuangQunFangPu, a Chinese classic literature of botany. Our studies yields promising result, we believe that the proposed research scenario will boost the on-going development of e-Humanities.

1 Introduction

In just a few short years, advances in digital technology have reshaped the way people read, write, and exchange knowledge information in traditional humanities disciplines. Along with these technical revolution, novel theoretical inquiry as well as practice engaged with this revolution has emerged. In this paper, we want to explore the issue of ontological knowledge acquisition in the context of *e-Humanities*.

The fact that people from different backgrounds may have knowledge structures unlike ours is a crucial issue to be addressed in knowledge engineering. In order to become sharable and reusable knowledge, all extracted information must first be correctly situated in a knowledge structure. In addition, the situated information must be allowed to transfer from knowledge structure to knowledge structure without losing its meaningful content. This is the vision behind the *Suggested Upper Merged Ontology* (SUMO) proposed by an IEEE working group. A shared upper ontology will both anchor the structured transfer of knowledge as well as set a standard for the construction of a middle and lower level ontology for each domain. This vision also has promising applications in the Semantic Web.

In the following sections we describe in more detail our approach to the construction of a specific domain ontology in the Chinese humanities: *GuangQunFangPu*. It is a classical Chinese archive which records diachronically an amount of textual and botanic information. The rest of this paper is organized as follows. In Section 2, a novel approach for the (re)-construction of diachronic domain knowledge is proposed. Section 3 gives a brief introduction to the classical work of *GuangQunFangPu* upon which our experiment performed. Some specific features of *GuangQunFangPu* are then discussed in Section 4. Section 5 shows the architecture of implemented system, and main results and other implications of our studies are presented in Section 6. Finally, Section 7 concludes with a description of the ongoing and future work.

2 Diachronic Domain Knowledge (Re)-Construction

The most salient factors dictating variations in knowledge structures are time, space, and domain. These factors are compounded with language, which is both the product and conduit of the conceptual structure of its speakers. In order to demonstrate the felicity of the shared upper ontology approach, we need to show that it can successfully applied to comparative studies of different knowledge structures regardless of their ontological variations.

Three important attributes characterize our methodology: **text-based**, **lexicon-driven**, and **ontology-merging**.

We call our text-based approach a Shakespearean-garden approach (Huang et al, 2004b). The Shakespearean-garden refers to the common practice in western museums of collecting in a garden all the plants referred to in the Shakespearean plays and sonnets. This garden then illustrates the flora of the Shakespearean England and will give us the context to interpret his work.

In our text-based approach, we do not actually grow the plants in a garden. Instead, we treat a collection of texts as an opus with an underlying knowledge structure. Since texts are composed of lexemes, we collects lexemes of a specific domain from the text just like plants are ‘collected’ from Shakespearean texts, and to ‘grow’ them to an ontology. That is, we will apply the shared ontology proposal to the interpretation historical texts by adopting the Shakespearean-garden approach towards construction of historical ontology.

2.2.1 WordNet as a Lexically Anchored Linguistic Ontology

The methodology proposed here is (mental) lexicon-driven. Mental lexicon is defined as a language user’s knowledge of words (Aitchison, 2003) The idea underlying our lexicon-driven approach is that concepts are stored in the mental lexicon and accessible through lexical access. In other words, we treat lexicon as a structured inventory of conceptual atoms. Therefore, the Princeton WordNet, a network-like lexical resource developed by the Cognitive Science Laboratory at Princeton University, can be a good candidate resource which mediates our lexical and conceptual knowledge.

2.2.2 SUMO as a Conceptually Shared Reference Ontology

Conceptually, ontology provides a structure for knowledge to be situated. However, there is a dilemma for the construction of a new ontology. On one hand, if no existing ontology was referred to, a new ontology could only be an reinvented wheel. On the other hand, when an existing ontology was referred to, errors could be introduced through pre-conceived conceptual structure and important generalizations could be missed.

To resolve the dilemma, we take the *ontology-merging as ontology-discovery* approach. We propose to map conceptual atoms to two (or more) reference ontologies. The merging of two ontologies leads to three possible scenarios: matched mapping, mismatched mapping, and complimentary mapping. Matched mapping simply confirms the knowledge structure. Mismatch mapping suggests that only one or neither is correct, and possibly lead to discovery of new knowledge structure. Lastly, when concepts are not attested in either ontology, we will have complimentary mappings. In this scenario, the coverage of either ontology can be increased coverage.

The choice of SUMO as the shared reference ontology is worth noting. SUMO represents the shared knowledge structure of our current time, which is in term the sum of human knowledge accumulated through history. It is true that a contemporary ontology necessarily differ from an historical ontology. However, in order to compare the knowledge systems of two historical periods or two domains, it is necessary to have one base reference. The contemporary time seems to be the natural reference not only because this is the knowledge system under which our scientific discourse takes place. The fact that it inherits knowledge from historical ontologies also makes is an effective reference. With this reference ontology, we will be able to observe and generalize systematically which part of the knowledge is different in the specific ontology.

In this approach, as first proposed in (Huang et al 2004a, 2004b), a lexicon of the targeted text, period, or domain is constructed first by segmentation and extraction of lexical items from the collected texts. Once the comprehensive lexicon of that period is collected, a lexical interface based on Sinica BOW [1] can be applied. It links each word to a conceptual class on the SUMO ontology, and a synset in WordNet. Since the lexicon from the text represents linguistically instantiated concepts, we use the linked conceptual nodes to construct an ontology for that text. The constructed ontology allows us to both interpret the conceptual structure of that text as well compare its knowledge with our contemporary knowledge.

2.2.3 Sinica BOW as a Integrated Base Resource

Sinica BOW (Academia Sinica Bilingual Ontological Wordnet) (Huang et al., 2004a), provides the basic infrastructure for our resources. It integrates three main resources: WordNet, SUMO, and the English-Chinese Translation Equivalents Database (ECTED). The three resources were originally linked in two pairs: WordNet 1.6 was manually mapped to SUMO (Niles and Pease, 2003), and also to ECTED (the English lemmas in WordNet were mapped to their Chinese lexical equivalents). With the integration of these three key resources, Sinica BOW functions both as an English-Chinese bilingual WordNet, and a bilingual lexical access point to SUMO. Sinica BOW plays a crucial role in our ontology-merging as ontology-discovery approach.

This is because we treat WordNet as a wide-coverage linguistic ontology. Hence Sinica BOW becomes a convenient

[1] http://bow.sinica.edu.tw

tool for merging two complimenting ontologies: a well-structured upper ontology with comprehensive levels of abstraction but with restricted lexico-conceptual coverage, as well as a linguistic ontology with almost complete lexico-conceptual coverage but incomplete levels of abstraction. In addition, the merging and comparison is lexically anchored. An additional aspect of Sinica BOW that we use only implicitly now is that fact that it is a bilingual wordnet. Taking a wordnet as a partially encoded ontology for a language again, we have the potential to explore the difference between these two linguistic ontologies in the process of ontology-merging. In our current work, we utilize the bilingual wordnet to fill in lexical gaps in either language.

Based on this model mentioned, a previous study that is perhaps most relevant to this current work was (Huang et al (2004))'s study of the ontology of *Tang* poetry. They segmented and classified a lexicon of 300 Tang Poems (618-907 A.D.), to build a small ontology of the Tang civilization. Three domain ontologies, animals, plants, and artifact, were manually constructed by mapping the extracted words to SUMO. The study was able to draw some tentative generalizations, including that the Tang civilization was primarily land-locked and that it was fascinated with flying. In short, the knowledge structure of a specialized ontology helps to form new knowledge.

3 Brief Introduction to GuangQunFangPu

The main domain knowledge that we model our specialized ontology on is *GuangQunFangPu*, a Chinese classic literature of botany. This work is chosen not only because the knowledge system is sufficiently different from the current one, but also because it is well-suited for the text-based and lexicon-driven strategy to discover knowledge structure. This study supports our text-based and lexicon-driven approach as an efficient way to build a specialized ontology as well as to infer domain knowledge.

GuangQunFangPu, published in 1708, was an officially composed document when Qing Kangxi was on the throne. Based on QunFangPu , GuangQunFangPu made a contribution towards botanical search and research as botany-related literature of all the past dynasties was collected and verified during the composition of GuangQunFangPu. GuangQunFangPu was later corrected, verbosity reduced and new material added. According to research, GuangQunFangPu was a literary agriculture and gardening masterpiece greatly influences agronomy and botany in China.

There are eighty volumes and ten 譜 (pu3,'pedigree') in GuangQunFangPu, including 穀 (gu3,'cereal'), 桑麻 (sang1 ma2, 'mulberry and hemp'), 蔬 (shu1,'vegetable'), 茶 (cha2,'tea'), 竹 (zhu2,'bamboo'), 花 (hua1,'flower'), 果 (guo3,'fruit'), 木 (mu4,' tree'), 卉 (hui4,'grass'), and 藥 (yao4,'medicine'). All materials were collected from literature and verified. Sources included histories, biographies, songs, poems, articles, commentaries on literary works etc. Each pedigree was attached with literary work or related allusions. Content length varied, so did knowledge involved. Myths, legends, art, literature, gardening, cuisine, medicine etc., these are fields one could find relative knowledge in GuangQunFangPu.

Except pedigrees tea and bamboo, there were subsets 譜名 (pu3 ming2) in each pedigree. 譜名 (pu3 ming2) referred to names of plants. Plant characteristics and appearance would be described in the beginning of each 譜名 (pu3 ming2), followed by subset categories. In some cases, plants were illustrated in more than one pedigree. 梅 (mei2, 'plum'), 杏 (xing4,'apricot') and 桃 (tao2, 'peach'), for example, were described in both flower and fruit pedigrees. In the flower pedigree, descriptions of the shapes, sizes and colors of flowers or branches were emphasized, while in the fruit pedigree, emphasis was put on fruits.

There were no general names like grass, flower or tree in GuangQunFangPu. Plants were sometimes given specific names, such as 春桂 (chun1 gui4, 'spring cassia') in the flower plant. At other times, plants were not specified.

Furthermore, alternative names were provided following each plant name. 芍藥 (shao2 yao4,'peony') had other names like 解倉 (jie3 cang1, 'peony') or 沒骨花 (mo4 gu3 hua1,'peony') while 羊桃 (yang2 tao2,'star fruit') could also be called 鬼桃 (gui3 tao2,'star fruit') or 羊腸 (yang2 chang2,'star fruit'). In some cases, one name could refer to different plants. 赤小豆 (chi4 xiao3 dou4) and 相思子 (xiang1 si1 zi3) both had an alternative name 紅豆 (hong2 dou4,'adzuki bean'). 蘭 (lan2, 'orchidaceous plant') could refer both to 蘭花 (lan2 hua1, 'orchid') or 蘭草 (lan2 cao3, 'fragrant thoroughwort'). These plant names and alternative names might reflect historical change, show geographical variance, or owe its name to allusions - these are three possible sources, revealing information a single name could contain, we would try to show in our GuangQunFangPu ontology.

4 Classification Scheme in GuangQunFangPu

In the following, we introduced some speciality in GuangQunFangPu as the background knowledge.

As mentioned, in GuangQunFangPu plants were divided into ten pedigrees, which were basically based on attributions and kinship. However, research showed that not only there were no definitions for these ten pedigrees but also some classifications were not categorical. This to some extent reflected GuangQunFangPu was not compiled purely out of botanical point of view. Some observations are presented as below:

- Some plants could be classified to more than one pedigree. 莕 (xing4,'amaranth'), classed with the grass pedigree, for example, could also be classified among the vegetable pedigree since 莕 (xing4,'maranth') is also a famous dish. Moreover, grass 蘭草 (lan2 cao3, 'fragrant thoroughwort') was mentioned in the flower pedigree, so did flower 菖蒲 (chang1 pu2, 'calamus') in grass, and vegetable 莕 (xing4,'amaranth') in grass. In cases when where could be alternative categorization, what criterion did it follow? The answer was left unclear.

- In the flower and fruit pedigrees or tree and fruit pedigrees, there were cases one 譜名 (pu3 ming2) was categorized as both. One could find 梅 (mei2, 'plum'), 杏 (xing4,'apricot') and 桃 (tao2, 'peach') in both flower and fruit pedigrees. There was still exceptions such as 栗 (li4, 'chestnut'), 榛 (zhen1,'hazel'), and 椰子 (ye2 zi5,'coconut') which were put only in the fruit pedigree but not in the tree pedigree.

- 筍 (sun3, 'bamboo shoot') was described in the bamboo pedigree but not in the vegetable pedigree while 箬 (ruo4, 'bamboo cuticle'), which is also bamboo, was classified as grass.

- Plants under certain 譜名 (pu3 ming2) would sometimes be given a new category. In GuangQunFangPu, 垣衣 (yuan2 yi1, 'wall moss'), which is a kind of moss, was given a 譜名 (pu3 ming2) other than 苔 (tai,'moss'). Under 桐 (tong2,'paulownia'), some were classified as new 譜名 (pu3 ming2) while some were not.

In GuangQunFangPu, under the ten pedigrees, there were detail categorizations based on shape or function instead of plant morphology, e.g. 水草 (shui3 cao3,'waterweeds'), 香草 (xiang1 cao3,'vanilla') in the grass pedigree; woody plants, such as 杜鵑 (du4 juan1,'azalea'), in the flower pedigree; bushes, such as 棘 (ji2,'thorn bushes') in the tree pedigree. Some of these concept differentiations were clear, some were not. Some appeared to be simple and clear, but actually very different from our common conception. For instance, 水果 (shui3 guo3,'fruit') in the fruit pedigree referred to juicy fruits like 荔枝 (li4 zhi1,'litchi'), 柑 (gan1,'tangerine'), 橙 (cheng2,'orange'). Juicy as 西瓜 (xi1 gua1,'watermelon') is, it was classified as 蓏果 (luo3 guo3), which referred to fruits of herbaceous plants or trailing plants. As for 蘋果 (ping2 guo3,'apple') and 葡萄 (pu2 tao2,'grapes') they were categorized as 膚果 (fu1 guo3), which meant fruits that had thin skin. Consequently, mistakes would be easily made if we try to understand GuangQunFangPu by current word senses.

All these characteristics of GuangQunFangPu leave space for the further exploration concerning with its way of domain knowledge structuring.

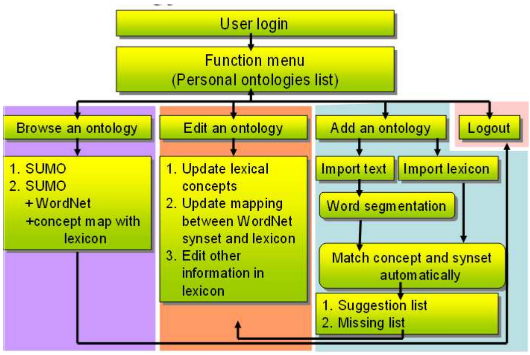
5 System Implementation

Based on previous discussion, Figure 1 outlined an architecture of our implemented workbench for semi-automatic construction of specialized ontologies. A graphical interface for online collaborative lexicon editing and ontology tuning is under construction.

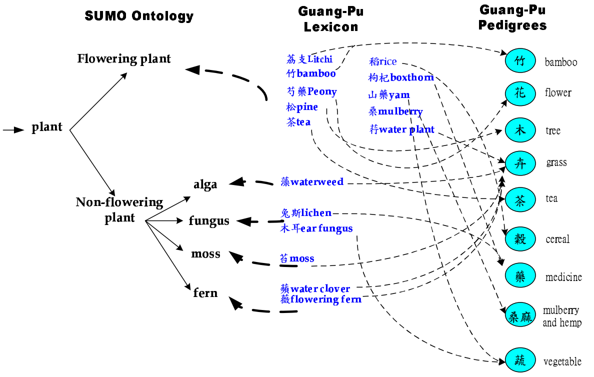
This workbench, which we call OntoLex Toolkits for e-Humanities, integrate available resources that include: WordNet, Sinica BOW, and segmentation tools for Chinese texts. This toolbox allows user to input a domain lexicon or specific lexical items. It will return all available information from our bilingual versions of WordNet and SUMO ontology. Lastly, it will allow automatically output of the tree representation of the specific ontology after it is constructed with verified lexical information.

6 Construction and Comparative Studies of Specific Ontologies

As proposed, our steps in exploring QuangQunFangPu are as follows: word segmentation → Match WordNet *synset* and SUMO *concept* automatically → Use WordNet

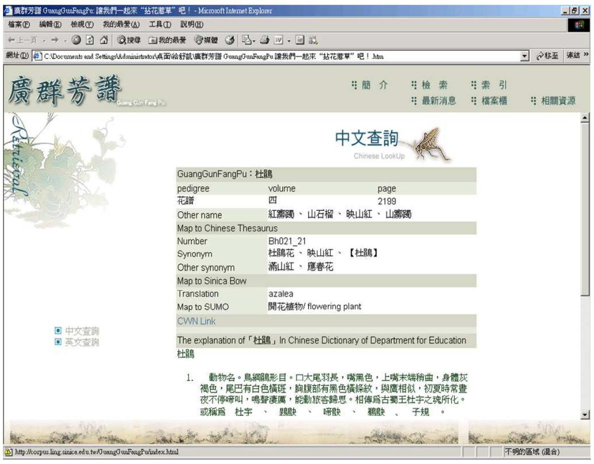

User login
Function menu
(Personal ontologies list)
Browse an ontology
Edit an ontology
Add an ontology
Logout
1. SUMO
2. SUMO
+WordNet
+concept map with lexicon
1. Update lexical concepts
2. Update mapping between WordNet synset and lexicon
3. Edit other information in lexicon
Import text
Import lexicon
Word segmentation
Match concept and synset automatically
1. Suggestion list
2. Missing list

information to check results and extend concepts (see Figure 2) → Transform into ontology browser format. Figure 3 shows the online browsing system of constructed QuangQunFangPu lexicon and ontology.[2]



In our previous first attempt at a text-based specific ontology (Huang 2004b), the 300 Tang poems (唐詩三百首) are analyzed based on the proposed model and system. A further study built on the foundation of the Tang 300 ontology is the ontology of poems by Su Shi (蘇軾詩) that is being completed. The choice of Su Shi offers more than historical comparison. Su Shi is from the Song dynasty, almost 500 years after the Tang. The time depth allow for comparative study. The collection is also a much larger text than the Tang 300, hence offers a good test case for our new approach.

Similarly, these lexical items were manually mapped to SUMO ontology. When there is no direct mapping to SUMO, Sinica BOW is consulted to give the lexical item a WordNet correspondence and relational structure.



GuangGunFangPu：杜鵑

pedigree	volume	page
花譜	四	2199

Other name 紅躑躅 、 山石榴 、 映山紅 、 山躑躅

Map to Chinese Thesaurus

Number Bh021_21

Synonym 杜鵑花 、 映山紅 、 【杜鵑】

Other synonym 滿山紅 、 應春花

Map to Sinica Bow

Translation azalea

Map to SUMO 開花植物/ flowering plant

CWN Link

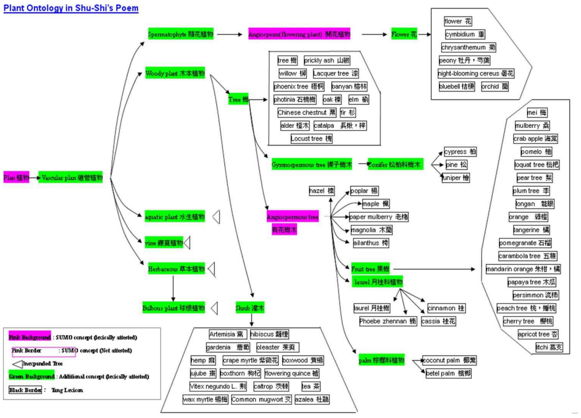
The explanation of「杜鵑」In Chinese Dictionary of Department for Education

杜鵑

1. 動物名。鳥綱鵑形目。口大尾羽長，嘴黑色，上嘴末端稍曲，身體灰褐色，尾巴有白色橫斑，胸腹部有黑色橫條紋，與鷹相似，初夏時常晝夜不停啼叫，鳴聲淒厲，能動旅客歸思。相傳爲古蜀王杜宇之魂所化。或稱爲 杜宇 、 鶗鴃 、 啼鴃 、 鵜鴃 、 子規 。

http://corpus.ling.sinica.edu.tw/GuangGunFangPu/index.html

杜鵑



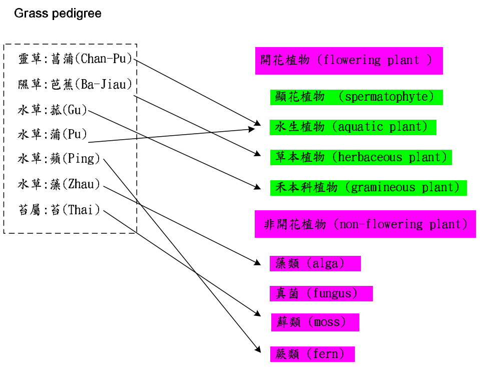
For example, Figure 4 and Figure 5 gives the domain ontology (plant) constructed from Su Shi poem and one of the ten pedigrees in QuangQunFangPu, respectively.(Concepts marked with pink are SUMO classes, while green one stand for the extension from WordNet synsets). These show how specific ontology construction facilitates systematic comparison of knowledge systems. The ontologies constructed from text collections allow us to compare and study the knowledge structure of different historical periods and gain perspective understanding of the different culture and time.

By this way, this model can harness domain archive knowledge *inter-operable*.

7 Conclusion

In this paper, we propose a potential advanced e-Humanities research scenario supported by existing lan-

[2] http://corpus.ling.sinica.edu.tw/GuangGunFangPu/



guage resources and technology. In order to facilitate the task of historical domain ontology acquisition, we use Sinica BOW, - a bilingual ontological lexical resource combining Princeton WordNet and IEEE SUMO ontology - as the backbone.

Our proposed model was tested on a specific Chinese classical botanic literature called QuangQunFangPu. Along with the promising result of our preliminary experiments, some interesting implications for the emerging e-Humanities discipline have been given. An easy-to-used OntoLex Toolbox for e-Humanities was developed to facilitate the construction and comparative studies of various domain ontologies.

References

[1] CKIP Segmentation and Tagging Program, `http://corpus.ling.sinica.edu.tw/project/LanguageArchive/lc_index.html`

[2] Fellbaum, C. (ed.). (1998). WordNet. An Electronic Lexical Database, Cambridge: The MIT Press.

[3] Wang, hao et al. (1708). (Pei wen zhai) GuangQunFangPu.

[4] Huang, Chu-Ren et al.(2004a). *Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO*. (Paper presented at the 4th International Conference on Language Resources and Evaluation (LREC2004). Lisbon: Portugal).

[5] Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. (2004b). *Reconstructing the Ontology of the Tang Dynasty: A pilot study of the Shakespearean-garden approach*. (Paper presented at the OntoLex 2004 Workshop. Lisbon.).

[6] Sinica BOW, `http://BOW.sinica.edu.tw/`

[7] SUMO: http://www.ontologyportal.org/

[8] The Ontolgoy of 300 Tang Poems `http://bow.sinica.edu.tw/ont/ts300_ont.html`