From General Ontology to Specialized Ontology: A study based on a single author historical corpus

Ru-Yng Chang
National Cheng Kung University
Tainan, Taiwan
p7894124@ccmail.ncku.edu.tw

Chu-Ren Huang
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
churen@gate.sinica.edu.tw

Feng-Ju Lo
Yuan-Ze University
Chungli, Taiwan
gefjulo@saturn.yzu.edu.tw

Sueming Chang
Institute of Linguistics
Academia Sinica
Taipei, Taiwan
kati@gate.sinica.edu.tw

Abstract

This paper proposes to deal with the construction of a specialized ontology as the discovery of a new knowledge structure, based on the premise that ontology is a structured knowledge. Our study integrates the following approaches: the mental lexicon approach, the Shakespearean-garden approach, and the ontology-merging as ontology-discovery approach. In particular, WordNet is used both a source of lexical knowledge and a (linguistic) ontology. SUMO (Suggested Upper Merged Ontology), on the other hand, is used as an upper ontology that provides fragments of well-structures knowledge. These two resources are compared and merged through Sinica BOW (Academia Sinica Bilingual Ontological Wordnet). The domain knowledge that we model our specialized ontology on is the collection of Su-Shi's poems from Song dynasty. 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. Based on this result, we further outlined an architecture for a workbench for semi-automatic construction of specialized ontologies.

1 Motivation: General and Specialized Ontologies

Gruber et al. (1994) described ontology as "an explicit specification of conceptualization." Ontologies can be used to explicitly represent structured information and support knowledge sharing and reuse. As a model for knowledge formation, the architecture of ontology critically depends on the type of knowledge to be represented. In terms of coverage, there are two contrasting types of ontology, general and specialized ontology. General Ontology is the upper ontology shared by all domains such as SUMO (Suggested Upper Merged Ontology). Specialized ontologies represent exhaustive information for certain domains, more specialized schemata must be created to make the data useful in making real world decisions. A specialized ontology may signify an ontology specific to a domain, historical period, an author etc. According to real conditions in a specific domain, re-construction and verification of conceptual structure should be done. Another problem that arises in different space, time, and domain is knowledge processing with mismatched knowledge structure.

One of the greatest challenges to the research on ontology is how to ensure both the felicity and compatibility of a new specialized ontology. Felicity here refers to the faithful and comprehensive representation of domain knowledge. Compatibility refers to the interchangeable and interpretability of the domain knowledge with regard to a shared upper ontology. In this paper, we propose that this challenge can be met when we take the creation of a specialized ontology as the merging of the segments of domain lexical knowledge to SUMO and WordNet. The mapping to SUMO ensures compatibility to an upper ontology, while the mapping to WordNet allows comprehensive representation of domain knowledge. Merging of the partially mapped ontology segments reduces the portion of knowledge representation which is missing from either prototypical ontology.

2 Research Methodology

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

Second, we call our text-based approach a Shakespearean garden approach (Huang et al. 2004). A Shakespearean garden collects all the plants referred to in Shakespearean texts by identifying plants mentioned in the plays and sonnets. A Shakespearean garden is used to illustrate the flora of the Shakespearean England and gives scholars a context in which to interpret his work. For instance, a Shakespearean garden helps to illustrate how plants played important roles in medicine, religious, and history (Keyser, 2004). In our text-based approach, we do not actually grow the plants in a garden. In stead, 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.

Third, we take the ontology-merging as ontology-discovery approach. This approach deals with the 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 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.

A previous study that is perhaps most relevant to this current work was Huang et al.'s (2004) 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. The first generalization is supported by the fact that the node of marine mammals as well as the dominance of hoofed animals. The second generalization is supported by the dominant frequency of birds among vertebrates, as well as the fact that insects are the only attested invertebrates (since they are the only winged invertebrates). In short, the knowledge structure of a specialized ontology helps to form new knowledge.

3 Basic Resource

Sinica BOW (Academia Sinica Bilingual Ontological Wordnet) (Huang et al., 2004), provides the basic infrastructure for our resoruces. 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 tool for merging two complimenting ontologies: a well-structured upper ontololgy 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. Note that Su Shi's language is almost 1,000 years old, hence there exists a substantial lexical difference. Another important fact is that each dictionary will have varying degree of coverage of domain terms. Hence there will be gaps in the Chinese version of the wordnet that can sometimes be filled by looking up the English wordnet.

4 Construction of a specialized ontology Based on Su-Shi Poems

There is no general ontology available from the Song dynasty. However, because of the continuous writing and textual tradition of Chinese, we can establish clear lexical correspondences between Song and contemporary lexical knowledge. Hence, we use Sinica BOW to try to map and locate conceptual atoms to both SUMO and WordNet (as a linguistic ontology).

4.1 Su Shi's Poems

Su Shi (A.D.1036-1101) is one of the most prominent scholars in Song dynasty who is very knowledgeable and well-traveled. In one memorable string of incidents, he was send into exile further and further away from the capital until he reached Chinese version of Land's End (tian1ya2hai3jiao3). This is Hainan, an island lying to the south of the southern-most point of mainland China. Hence Su Shi has unique firsthand knowledge of the fauna and flora of all China, including places where most scholar shun. In addition, the Northern Song dynasty was interesting in that they enjoyed a period of prosperity that allowed them to show strikingly modern characters. The Song people were in many respects similar to modern Western life of the same time (Huang, 1999). Poetry and prose were regarded as a part of everyday life and a normal medium for representing feelings and thoughts. The works of Su-Shi well illustrate these qualities, so we also selected Su Shi poems as materials to carry out experiments and to construct a domain ontology.

4.2 The Construction of Domain Ontology

First of all, word segmentation and classification are implemented. Forty-five (45) volumes (out of 50) of Su-Shi's works have already been digitized, segmented, and classified under the direction of Feng-ju Luo (http://cls.hs.yzu.edu.tw/cm). All lexical entries are extracted. The above textual database contains a total of 98,430 word types. Three sub-lexica of Su-Shi's poems are extracted for domain ontology construction: animals, plants, and artifacts.

In addition, this study matches lexicon, WordNet synset and SUMO concept automatically. Each Chinese lexical item is assigned a unique Chinese sense. Because Sinica BOW had integrated WordNet, SUMO, and ECTED, each sense is mapped through Sinica BOW to an English synset of WordNet and a SUMO concept. When no direct mapping is available, we consult other lexical resources (such as Tongyici Cilin, Mei et al. 1984) to improve the recall rate.

All mappings were manually double-checked by human. Ambiguous, obscure and changed meanings in the historical lexicon are solved during this stage. WordNet lexical semantic relations are adopted to extend conceptual hierarchy. After bootstrapping, inconsistencies and redundancies are pruned. Finally, all data are transformed into ontology browser format.

4.3 Su-Shi's Poetic Lexicon and Ontology

It this section, we examine some generalizations derived from Su-Shi's poetic lexicon and ontology.

The first possible source of information can be derived from the distribution of terms within a certain domain. For in stance, animal terms consist of 1.4294 %, plant terms consist of 1.7393%, and artifact terms consist of 1.4467% of Su-Shi's lexicon in terms of types. Furthermore, flowering plants consist of over 98% of the instances of plants in the texts. This fact reinforces the fact that flower presents strong poetic image. In **Figure 1** and **Figure 2**, the distribution of hoofed mammal in animal and transportation device in artifact supports the know fact that transportation plays an importnat role in Su Shi's life.

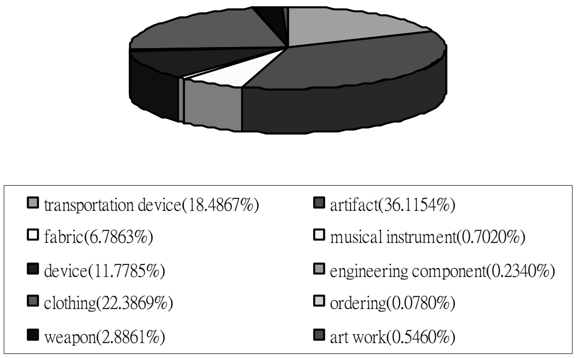


Figure 1. Distribution of artifact concept in Su-Shi's poems.

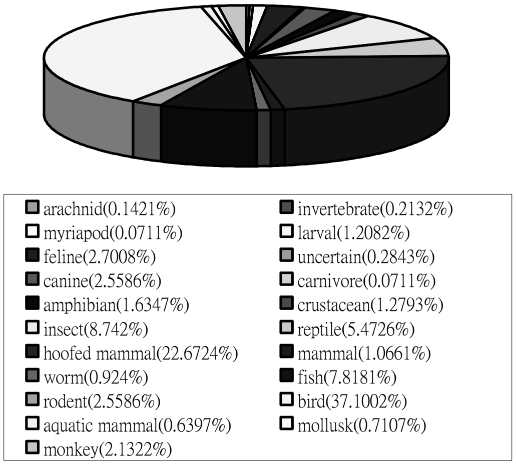


Figure 2. Distribution of animal concept in Su-Shi's poems.

Another possible source of information comes from unique lexical concepts that are not shown in other textual databases. **Table 1** lists lexical concepts found in Shu Shi's but not in Tang 300. Words marked with star sign mean there are other synonyms attested in these texts. The reference to aquatic mammals and crustaceans attested not only to Shu Shi's fame as a real gourmand but also to the fact that, unlike all major Tang poets, he has lived among the sea fishermen in Guangdong and Hainan Island.

Concept	Words
Aquatic mammal	whale 鯨*
Amphibian	frog 蛙*, toad 蟾蜍*, salamander 鯢
Mollusk	clam 蛤*, gastropod 螺*, oyster 蠔, snail 蝸牛*, earthworm 蚯蚓*
Crustacean	crab 蟹*, shrimp 蝦

Table 1. Concepts found in Su Shi's but not in Tang 300.

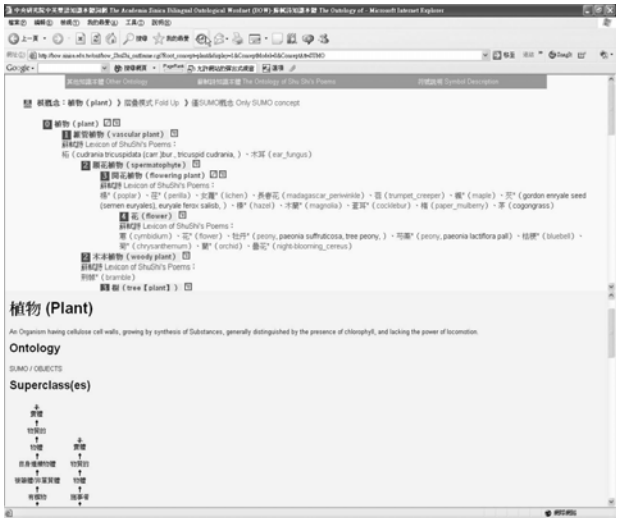
4.4 The Ontology Browser of Su-Shi's Poems

The ontology browser prototype displays the relation of concepts and the lexicon association with concepts in tree format. Different options can be selected then different type information can be displayed. Users can look up any concept as root node to view the relation between concepts, and the connections between concepts and lexicon. Concepts are taking turns or expending all to display. Original concepts model and plus additional concepts model can be changed to show. A hierarchical numeral code system is hired to reach inheritance of concept. For instance, *physical* is a sup-concept of *object*, then the hierarchical numeral code of *physical* is 1.1. , and *object* is 1.1.1. As a result, a concept name given in a different font color denotes whether it or its sub-concepts include the lexicon. Different background color of concept indicates original or additional concept in SUMO. In the bottom of the prototype browser, when users look up a concept, information from SUMO is shown. It represents Chinese translation, class, definition, superclass, subclass, and axiom of this concept. Either Chinese or English lexicon can be searched for the following kinds of information: sentence citation, synonym, concept in SUMO, additional concept, keyword in related resource, and information on WordNet. Synonyms are also allowed in queries.

4.5 Towards a Workbench for Specialized Ontology: Browser and Editor

Based on the largely manual construction of specialized ontologies reported above and in, we conclude that a uniform process for constructing domain ontologies can be semi-automated. A prototype ontology browser was set up. At this time, a semi-automatic word concept extraction and construction ontology workbench for browsing and editing is being constructed.

The workbench for specialized ontology is composed of three modules: add, edit, and browse ontologies. Both texts and lexica can be imported to add ontology, as well as for setting the source of domain ontology. The user can set up style of writing and category of word list for word segmentation and select ontologies to be consulted to match SUMO and lexicon. A word segmentation algorithm and lexicon are selected according to writing style. Automatically, after word segmentation, the system proceeds to match concept and synset by consulting WordNet, SUMO, Sinica BOW, and related existing ontologies. Subsequently, embryo ontology, the suggestion list, and missing item list for ontology construction are shown. The suggestion list provides, inter *alia*, candidate synset, candidate synset synonyms, explanation of candidate synset, and concept of candidate synset. The missing item list denotes no synset or concept automatically assigned. While a user edits an ontology, the suggestion and missing item lists are displayed for handy reference.



中央研究院中英雙語知識本體詞網 The Academia Sinica Bilingual Ontological Wordnet (BOW)-蘇軾詩知識本體 The Ontology of - Microsoft Internet Explorer

其他知識本體 Other Ontology　蘇軾詩知識本體 The Ontology of Shu Shi's Poems　符號說明 Symbol Description

根概念：植物 (plant) 》摺疊模式 Fold Up 》僅SUMO概念 Only SUMO concept

0 植物 (plant)
1 維管植物 (vascular plant)
蘇軾詩 Lexicon of ShuShi's Poems：
柘 (cudrania tricuspidata (carr.)bur., tricuspid cudrania,) ‧木耳 (ear_fungus)
2 顯花植物 (spermatophyte)
3 開花植物 (flowering plant)
蘇軾詩 Lexicon of ShuShi's Poems：
楊* (poplar) ‧荏* (perilla) ‧女蘿* (lichen) ‧長春花 (madagascar_periwinkle) ‧苕 (trumpet_creeper) ‧楓* (maple) ‧芡* (gordon euryale seed (semen euryales), euryale ferox salisb.,) ‧榛* (hazel) ‧木蘭* (magnolia) ‧蒼耳* (cocklebur) ‧楮 (paper_mulberry) ‧茅 (cogongrass)
4 花 (flower)
蘇軾詩 Lexicon of ShuShi's Poems：
蕙 (cymbidium) ‧花* (flower) ‧牡丹* (peony, paeonia suffruticosa, tree peony,) ‧芍藥* (peony, paeonia lactiflora pall) ‧桔梗* (bluebell) ‧菊* (chrysanthemum) ‧蘭* (orchid) ‧曇花* (night-blooming_cereus)
2 木本植物 (woody plant)
蘇軾詩 Lexicon of ShuShi's Poems：
荊棘* (bramble)
樹 (tree [plant])

植物 (Plant)

An Organism having cellulose cell walls, growing by synthesis of Substances, generally distinguished by the presence of chlorophyll, and lacking the power of locomotion.

Ontology

SUMO / OBJECTS

Superclass(es)

Figure 3. Ontology browser prototype.

The edit module provides a friendly interface to update word concept, to update the mapping between WordNet synset and lexicon, and to edit other information in the lexicon. Users can collect information on suggestion list, missing item list, embryo ontology, and other open custom-made specialized ontologies to add, update, and delete the concept node. They also can change the mapping between WordNet synsets and the lexicon, and edit other lexical information which is imported from related resources or added by users.

In browse mode, this proposal retains all functions of the prototype that search and view all information. Also available are WordNet linguistic ontology, Chinese WordNet linguistic ontology, SUMO, and custom-made specialized ontology automatically presented in graph format through Hypernym, Hyponym, and *isa* relations. Chinese WordNet from Sinica BOW is transformed. Various style formats are used to identify different kinds of information. This visual model assists users quickly to find the location of a lexical item in SUMO, in linguistic ontology, and in custom-made ontology and to view clearly the relation between other concepts and words.

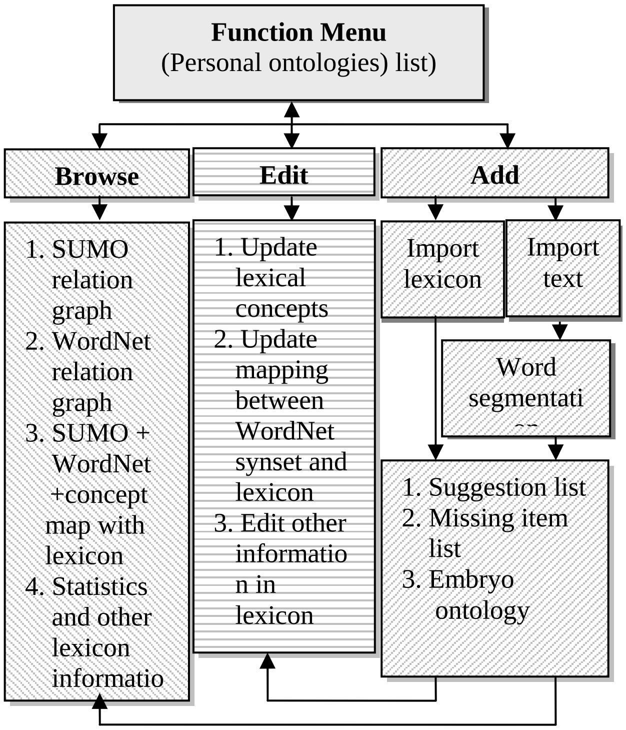

Function Menu
(Personal ontologies) list)
Browse
Edit
Add
1. SUMO relation graph
2. WordNet relation graph
3. SUMO + WordNet +concept map with lexicon
4. Statistics and other lexicon informatio
1. Update lexical concepts
2. Update mapping between WordNet synset and lexicon
3. Edit other informatio n in lexicon
Import lexicon
Import text
Word segmentati on
1. Suggestion list
2. Missing item list
3. Embryo ontology

Figure 4. System architecture of workbench for ontology.

This workbench greatly differs from tools described by Denny (2004) in three aspects: showing bilingual linguistic and general ontology in a graphic model, processing different text styles, and combining lexical and ontology information. Word segmentation, word concept extraction, and construction ontology are essentials to establish such a system. Word sense disambiguation can improve the precise rate of word concept extraction.

5 Conclusion

Ontologies represent the knowledge structure of a domain or historical period. In this paper, we base on WordNet, SUMO, and Sinica BOW to take Su-Shi's poems to construct a domain ontology which includes lexical and ontological information. The methodologies that we adopted are the mental lexicon approach, the Shakespearean-garden approach, and the ontology-merging as ontology-discovery approach. According to the construction of Su-Shi's and 300 Tang poems ontologies, we have provided an online interface to browse ontologies and lexica. In the future, we will complete the online ontology editor and browser, which will enable easier accesses to domain lexica and ontology, while being linked to WordNet, and SUMO through Sinica BOW. Last but not the least, our proposed Domain Ontology Workbench will allow semi-automatic construction of domain ontologies hence facilitates comparative studies of various domain ontologies.

References

Jean Aitchison. 2003. Words in the mind: an introduction to the mental lexicon. Malden, MA: Blackwell Pub.

Michael Denny. 2004. Ontology Tools Survey, Revisited. XML.com. http://www.xml.com/pub/a/2004/07/14/onto.html

Thomas R. Gruber, Gregory R. Olsen. 1994. "An ontology for engineering mathematics." The Fourth International Conference on Principles of Knowledge Representation and Reasoning, Bonn, Germany.

Bob Huang. 1999. "Su Shi's Greatest Works in His Own View. "http://www.geocities.com/WallStreet/Floor/2391/essays/essay29.htm

Chu-Ren Huang, Feng-ju Lo, Ru-Yng Chang, Sueming Chang. 2004. "Reconstructing the Ontology of the Tang Dynasty: A Pilot Study of the Shakespearean-Garden Approach". The 4th International Conference on Language Resources and Evaluation (LREC2004)-- Workshop on Ontologies and Lexical Resources in Distributed Environments (OntoLex 2004). pp. 43-49. Lisbon. Portugal.

Chu-Ren Huang, Ru-Yng Chang, Shiang-Bin Lee. 2004. "Sinica BOW (Bilingual Onto-logical Wordnet): Integration of Bilingual WordNet and SUMO." The 4th International Conference on Language Resources and Evaluation (LREC2004). pp. 1553-1556. Lisbon. Portugal.

Joseph M. Keyser. 2004. Bring your bards to your yard. http://www.montgomerycountymd.gov/deptmpl.asp?url=/content/dep/greenman/shakespeare.asp

Jia-Ju Mei, Yi-Ming Zhu, Yun-Qi Kao, Hong-Xiang Yan. 1984. TongYiCiCiLin. Shanghai: Shanghai Lexicographical Publishing House.

Ian Niles, Adam Pease. 2003. "Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology". Proceedings of the IEEE International Confer-ence on Information and Knowledge Engineering. (IKE 2003). pp.412-416. Las Vegas, Nevada.

Online Resources

Suggested Upper Merged Ontology, http://www.ontologyportal.org

The Academia Sinica Bilingual Ontological Wordnet, http://bow.sinica.edu.tw

The Research of Building a Content Markup System for Su-Shi's Poems in Extensible Markup Language, http://cls.hs.yzu.edu.tw/cm/