Sinica BOW and 300 Tang Poems:

An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry

研究院知識詞網與唐詩三百首

—雙語知識本體詞網簡介及唐詩知識本體之初步構建

Chu-Ren Huang (Academia Sinica), Feng-ju Lo (Yuan Ze University),
Ru-Yng Chang (Academia Sinica), Sueming Chang (Academia Sinica)
黃居仁（中央研究院），羅鳳珠（元智大學），張如瑩（中央研究院），張舒茗（中央研究院）

Abstract

The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology). 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). ECTED encodes both equivalent pairs and their semantic relations (Huang et al. 2003). With the integration of these three key resources, Sinica BOW functions both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO.

Sinica BOW allows versatile access and facilitates a combination of lexical, semantic, and ontological information. Versatility is built in with its bilinguality, and the lemma-based merging of multiple resources. First, either English or Chinese can be used for the query, as well as for presenting the content of the resources. Second, the user can easily access the logical structure of both the WordNet and SUMO ontology using either words or conceptual nodes. That is, users can use words to search for ontology or use ontological nodes to search for linguistic realizations in both languages. Third, multiple linguistic indexing is built in to allow additional versatility. Fourth, domain information allows another dimension of knowledge manipulation.

In addition to serving as the reference and infrastructure for the construction of specific knowledgebases, the Sinica BOW model can also be applied to encode and represent a particular knowledge system, such as Tang civilization. This application will allow comparative studies of a historical conceptual system with our modern conceptual system. Our pilot study on the 300 Tang Poems is

reported here. The segmented and classified lexicon of the 300 Tang Poems (Chang and Luo 1999) serve as the basis of this study. Three domain ontologies are constructed and studied: animals, plants, and artifacts. Each domain is mapped to the SUMO/BOW structure. The resultant ontological representation is taken as a slice of the knowledge structure of Tang civilization. For instance, from the ontology of animals of Tang 300 (see attached file), we reach some broad generalizations about the familiar fauna of Tang. With further examination, we also found a fascination with flying in Tang is confirmed by the poetsí choice of poetic animals.

In sum, we argue that the Sinica BOW model will not only be a useful resource but also a productive model for the construction of a knowledgebase that will greatly facilitate our understanding of Tang civilization.

1 Background

The construction of an ontology from a knowledge background which is substantially different from ours can be challenging yet rewarding. We will refer to this type of ontology as "Non-Standard Ontology" for lack of better terms. Work on non-standard ontology presents a dilemma. On one hand, the structure of knowledge is often neither explicated nor represented before the non-standard ontology is constructed. On the other hand, to construct such an ontology, one needs to start with at least some pre-defined terms and conceptual taxonomy, which is in practice a small (upper) ontology. For historical ontologies, it is very rare to find a synchronous ontology from the same period, such as Wilkins (1668). In this case, the structure of the synchronous ontology can be adopted and mapped to a modern system for study. However, for the knowledge domains with no existing ontologies available, the greatest challenge also underlines the greatest potential to gain new knowledge. For instance, seventh century Chinese does not have the same scientific knowledge or the philosophical tradition that the current academic world holds to be common. Hence, even though there is much knowledge to be gained, there is also very little to fall back to as the working hypothesis. We will show in this paper how such dilemma can be resolved with successful integration of lexical resources and upper ontology.

The target ontology of this study is the ontology of the Tang dynasty (618-907AD). In this pilot study, we work with the text of the collection of the Tang 300 Poems. We adopt SUMO as our upper ontology. The lexical resources used include the domain lexica extracted from the text and the English-Chinese bilingual wordnet system Sinica BOW.

2. Sinica BOW: lexicon based bilingual knowledgebase

Lexicons can perform the bridging function between documents and conceptual categorisation. This position is motivated by both language engineering concerns as well as psychological felicity. In addition, when the issues and needs of multi-linguality are taken into consideration, it becomes obvious that the lexicon is the only level where generalizations as well as variations across different languages can be captured efficiently and comprehensively. In this demo, we will show our work on integrating multiple lexical

resources with ontology such that the linguistic-to-conceptual representation and language-to-language gaps can be bridged simultaneously.

The ***Sinica BOW*** (Academia Sinica Bilingual Ontological Wordnet) is intended as a linguistic infrastructure for knowledge representation and knowledge engineering. It is built upon the relation-based structure of WordNet. On one hand, a bilingual wordnet is constructed with the crucial design feature of treating bilingual translation correspondences as lexical semantic relations (Huang et al. 2003). On the other hand, SUMO (Suggested Upper Merged Ontology) is adopted as the shared system of conceptual categorization (Niles and Pease 2001). SUMO is also one of the first conceptual categorization systems to be mapped to an English lexicon (Niles and Pease 2003). Since SUMO is mapped to WordNet 1.6 (and most recently to WordNet 2.0), the English WordNet has become the cornerstone for linking across languages and between a language and its conceptual system. In addition, domain tags are assigned to lemmas when necessary in order to ensure domain inter-operability.

By the combination of ontology and wordnet, we hope that Sinica BOW will 1) give each linguistic form a rigorous conceptual location, 2) clarify the relation between conceptual classification and linguistic instantiation, and 3) facilitate genuine cross-lingual access of knowledge.

The Sinica BOW allows lexical searches in either language to return ontological information (in either language). Searches on Sinica BOW can return the following information: Sense-based English-Chinese translation equivalency, English word-sense-based ontology and inference, Chinese word-based ontology and inference, Word-sense-based domain specification (under construction).

In addition to the integration of Wordnet and ontology, it is also an important goal of Sinica BOW to integrate lexical resources. Sinica BOWís design is lemma-driven. A lexical database of word forms is first compiled by integrating multiple lexical resources. This becomes the central database for lexical management for Sinica BOW. Making use of this lexical database, a lexical search may link to either the main BOW knowledgebase or any of the corresponding entries in an online lexicon.

2.1. The Multilingual and Cross-Domain Properties of (Semantic) Relations

In addition to relying on lemmas as retrieval keys, a crucial step in establishing synergy between language and knowledge resources is to identify the conceptual atoms that apply equally effectively to knowledge and language resources. Lexical semantic relations are exactly such a set of atoms. Sinica BOW implements this idea by encoding the lexical semantic relations between English-Chinese translation equivalent pairs. In addition to more precisely describing the relationship between two translation equivalents, this also allows better cross-lingual inferences. Explicitly allowing lexical semantic relations to be coded cross-lingually also will facilitate the transferring to a structured set of tree relations from one language to the other.

2.2. Taking Advantages of Lexical Structure*s*

In addition to the integration of bilingual WordNet and SUMO, Sinica BOW also integrates the rich structural information of the integrated lexical resources. Glyph,

phonological, and morphological structures can all be used to help access the ontological wordnet. This work has implications far beyond being convenient search tools. It is often claimed that the glyph composition (e.g. radicals) in Chinese has its semantic base. This can also be said about the morphological composition (and to a much lesser degree, phonological composition). In other words, the integration allows us to study the possible links between these lexical structures and conceptual classifications.

2.3. Conclusion

Integrating and interpreting information from multiple and varying sources will be the main challenge for information processing for the current generation. Taking lexicon as the bridging knowledgebase and ontology as the overall knowledge structure seems to be a logical choice. Integrating the two resources with multilingual capacity will add to the versatility and open new possibilities.

3. Resources and Structure of Sinica BOW

Conceptual structure and lexical access are two essential elements of human knowledge. Bilingual representation of both conceptual structure and lexical information will enable language independent knowledge processing. In this paper, we introduce a new type of integrated language resources: Bilingual Ontological Wordnet. The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) was constructed in 2003. We argue that such combination of ontology and wordnet will 1) give each linguistic form a rigorous conceptual location, 2) clarify the relation between the conceptual classification and its linguistic instantiation, and 3) facilitate genuine cross-lingual access of knowledge.

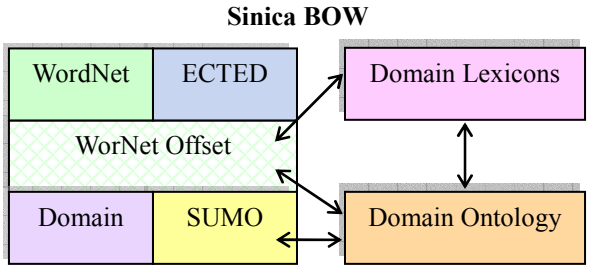
The Academia Sinica Bilingual Ontological Wordnet (Sinica BOW) integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED), and SUMO (Suggested Upper Merged Ontology).

WordNet is a lexical knowledgebase for English language that was created at Cognitive Science Laboratory of Princeton University in 1990 (Fellbaum 1998). Its content is divided into four categories based on psycholinguistic principles: nouns, verbs, adjectives and adverbs. WordNet organizes the lexical information according to word meaning and each synset groups together a set of lemmas sharing the same sense. In addition, WordNet is a semantic network linking synsets with lexical semantic relations. WordNet is widely used in Natural Language Processing applications and linguistic research. The most updated version of WordNet is WordNet 2.0. We adopted WordNet 1.6., the version which is used by most applications so far.

ECTED was constructed at Academia Sinica as a crucial step towards bootstrapping a Chinese wordnet with English WordNet (Huang et al. 2002, Huang et al. 2003). The translation equivalence database was hand-crafted by the WordNet team at CKIP, Academia Sinica. First, all possible Chinese translations of an English synset word (from WN 1.6.) are extracted from several available online bilingual (EC or CE) resources. These translation candidates were then checked by a team of translators with near-native bilingual ability. For each of the 99,642 English synsets, the translator selected the three most appropriate translation equivalents whenever possible. The translation equivalences were defaulted to lexicalized words, rather than

descriptive phrases, whenever possible. The translation equivalences were then manually verified. Note that after the first round of translation, there were about 5% of the lemmas whose Chinese translation can neither be found in our bilingual resources nor be filled by the translators. We spent another 2 person-year consulting various special dictionaries to fill in the gaps.

Figure 1: The resource and structure of Sinica BOW:



SUMO is a upper ontology constructed by the IEEE Standard Upper Ontology Working Group and maintained at Teknowledge Corporation. SUMO contains roughly 1,000 conceptual nodes for knowledge representation. It can be applied to automated reasoning, information retrieval and inter-operability in E-commerce, education and NLP tasks. Niles & Pease (2003) mapped synsets of WordNet and concept of SUMO in three relations: synonymy, hypernymy and instantiation. For instance, the synset "animal" (a living organism characterized by voluntary movement) in WordNet is synonymous with the SUMO concept of "Animal". In "bank" (a financial institution that accepts deposits and channels the money into lending activities) this case, bank is a corporation that is a hypernym of the associated synset. President of the United States (the office of the US head of state) is an instantiation of "position" concept. Through the linking and the interface available at the SUMO website (http://ontology.teknowledge.com), each English lemma can be mapped to a SUMO ontology node.

The three above resources were originally linked in two pairs: WordNet 1.6 was mapped to SUMO by Niles and Pease. ECTED maps English synsets in WordNet to Chinese lexical equivalents, which encodes both equivalent pairs and their semantic relations (Huang et al. 2003). WordNet synsets thus became the natural mediation for our integration work. Thus, with the integration of these three key resources, Sinica BOW can function both as an English-Chinese bilingual wordnet and a bilingual lexical access to SUMO. In other words, Sinica BOW allows a 2x2x2 query design, where a query could be in either Chinese or English, either in lexical lemmas of SUMO terms, and the query target can either be the wordnet content or the SUMO ontology.

The design of Sinica BOW has an additional domain information layer, as shown in figure 1. The domain information will be represented by a set of Domain Lexico-Taxonomy (DLT, Huang, Li, and Hong 2004). In this design, our main concern is domain inter-operability. It can be safely assumed that domain exclusive words (i.e. lemma-sense pairs) are recorded only in domain lexica, hence there will be no ambiguity and no inter-operability issues. We concentrate instead on the lexical items that intersect with the general lexicon. On one hand, since these are the lemmas that may occur in

more than one domains with one or more different meanings, domain specification would help resolving the ambiguity. On the other hand, these general lemmas with domain applicability can be effective signatures for the applicable domains. The real challenge to domain inter-operability involves the ëunknowní domains where no comprehensive domain lexica/corpora are available. We argue that this problem can be greatly ameliorated by tagging the general lexicon with possible domain tags. When domain tags are assigned to lemmas whenever possible, the general lexicon will contain substantial partial domain lexica. Although we cannot expect to construct full-scale domain lexica within the general lexicon, these domain-tagged lexical items serve as a scalable basis for future bootstrapping for domain lexica.

4. Presentational Versatility

Sinica BOW allows versatile access and facilitates a combination of lexical semantic and ontological information. The versatility is built in with bilinguality, and lemma-based merging of multiple language sources.

The versatility and combinatory presentation is crucial to the presentation of a knowledge system.

4.1. Lexicon-driven Access

Since the main goal of Sinica BOW concerns knowledge representation, the lemma based or conceptual node based query results are directed linked to the full knowledgebase and expandable. The Sinica BOW access is lexicon-driven. Each query returns a structured lexical entry, presented as a tree-structured menu. A keyword query returns with a menu arranged according to word senses, as shown in Figure 2. The top level information returned including POS, usage ranking, and cross-reference links. In addition to wordnet information, cross-references to up to five resources are pre-compiled for either language. For an English word, the main resource is of course the bilingual wordnet information that our team constructed. Major outside references are listed for quick hyperlink. These include corpora and both EC and CE dictionaries. For Chinese, the main resource is again our bilingual wordnet. In addition, links are established to Sinica Corpus, to Wen-Land (a learnerís Lexical KnowledgeNet), and to online monolingual and bilingual dictionaries. In addition to online access of multiple sources information, each lemmaís distribution in these resources is also a good indicator of its usage level.

⊞ 詞義(Sense)4: 魚兒

領域 Domain	一般(General) 建議-fish Sense 4-的領域值
POS 詞類	名詞(Noun)
解釋 Explanation	any of various mostly cold-blooded aquatic vertebrates usually having scales and breathing through gills
翻譯 Translation	魚兒, 魚
同義詞集 Synset	fish
（整體）部件詞 Part meronym	milt, tail_fin, fishbone, fish_scale, fin, roe, caudal_fin, lateral_line_organ, lateral_line
上位詞 Hypernym	aquatic_vertebrate
下位詞 Hyponym	food_fish, game_fish, rough_fish, cartilaginous_fish, chondrichthian, bony_fish, mouthbreeder
（成員）群體詞 Member holonym	shoal, Pisces, school
SUMO	同 fish:Fish(魚類)

Figure 2: A sample lemma query result of Sinica BOW

The access to the ontology and the domain taxonomy are also lexicon-driven. That is, in addition to using the pre-defined ontology or domain terms (in either English or Chinese), a query based on a lexical term is also possible.

For SUMO, it will return a node where the word appears in. It can also be achieved by looking up the ontological or domain node the word belongs to.

One last but critical feature of the lexicon-driven access is the possibility to re-start a query with any lexical node. When expansion reaches at the leave node and results in a new word, clicking on the word is equivalent to start a new keyword search.

4.2. Multiple Knowledge Source

Sinica BOW preserves the logical structure of both WordNet and SUMO ontology yet links them together to allow direct accesses to the merged resources. This is shown in Figure 2. In a wordnet search, the return includes an expandable list of the complete bilingual wordnet fields. The fields are listed under each sense and include: POS, synset, sense explanation, translation, and list of lexical semantic relations. In addition, we add the domain information, translation equivalents, and link to the corresponding SUMO node. Each field is expandable to present the database content. For instance, Figure 2 shows the query return for the lemma "fish", with the Part_Meronym and Holonym of sense 4 expanded. The field of domain and SUMO will lead directly to the corresponding node in the domain taxonomy of the ontology and allow further exploration. For instance, the menu item of the mapped SUMO node links to the SUMO representation, as well browsing of the SUMO ontology and axioms.

Two more aspects of versatility can be achieved through the use of higher level linguistic generalizations and the use of domain taxonomy to organize information. These will be discussed in more details in the next section.

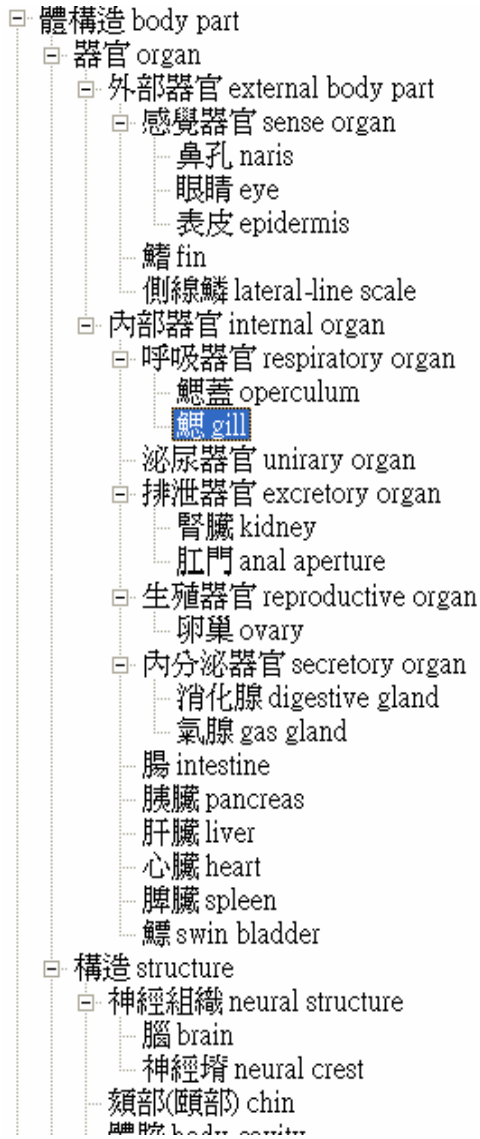


Figure 3: A sample domain ontology: Fish

5. Higher Level Generalizations

Linguistic as well as resources structures are utilized in Sinica BOW to facilitate formation of generalizations as well as to assist queries where the user is not sure of the precise lemma form. The non-lexical access includes

alphabetical (for English), prefix (for Chinese, including root compounds), suffix (for Chinese, including root compounds)), POS, frequency, domain, SUMO concepts, as well as a combination of the above conditions. With this additional level of resource integration, generalizations such as the semantic correlation of senses and morphological heads can be easily reached.

Domain taxonomy can also be utilized to organize and access information. Our Domain Lexico-Taxonomy approach attempts to assign a domain tag to a word whenever applicable. We also encourage users of SUMO to feedback with their own domain use of lexical items because domain specifications can not be covered by any single knowledge source. Hence we BOW contains rich domain information. Hence we also allow structured access to the Sinica BOW knowledge content by specifying a node on the domain taxonomy. This feature enables quick extraction and checking of a domain lexicon.

6. Domain Ontology

One of the most immediate and perhaps most powerful application of Sinica BOW is perhaps the construction of domain specific ontologies. This will be a crucial step towards providing a feasible infrastructure to implement web-wide specific ontologies, as required by the vision of Semantic Web. It is also a critical test to see if the upper ontology approach is really applicable to a wide range and diversity of knowledge domains. And lastly, for Sinica BOW, it provides a test ground for us to show that the combination of bilingual wordnet and ontology does provide a better environment for knowledge processing.

Two first attempts have been carried out. The first is a small fish domain ontology projected from the FishBase terms. This is mapped using Sinica BOW. Part of the ontology is shown in Figure 3. We would like to explore the possibility of using this domain ontology for non-expert to extract expert knowledge from the FishBase in the future.

The second attempt, reported in Huang et al. (2004), involves the Shakespearean-garden approach to domain ontology. In this approach, we collect domain lexicon from a target collection of texts (Tang poems in this case), and map them to the SUMO ontology. This approach allows us to examine the knowledge and/or experience of a specific domain as reflect in that collection of texts. This could be personal, historical, regional etc. This approach allows us to make generalizations based on the full knowledge structure, not just one lexical incident. For instance, we were able to confirm the Tang civilizationís fascination with flying by looking at the dominance of animal references in the texts.

7. The Shakespearean-garden Approach Toward Non-standard Ontology

We propose a Shakespearean-garden approach to the construction of non-standard ontology. This approach is both lexicon-based and domain-driven. A Shakespearean garden collects and grows all plants referred to in Shakespearean texts. The purpose of a Shakespearean garden is to replicate the botanic knowledge and flora experience of Shakespearean England. A Shakespearean garden works because we can reasonably assume that the plants we collect now are by and large identical to the Shakespearean plants and have the same functions. Similarly, when constructing a non-standard ontology, we propose to start with concrete sub-domains. A

chosen domain must have two properties: that it plays roughly equivalent roles in the knowledge backgrounds of the target ontology and the reference ontology (i.e. our contemporary ontology); and that it is empirically verifiable with lexical resources supporting the target ontology. Even though the Shakespearean-garden approach does not guarantee a complete ontology, it will lead to very reliable domain ontologies. When there is sufficient data and knowledge collected, these domain ontologies can be further linked to approach a complete ontology of the target knowledge domain.

Our approach requires a shared upper ontology as the anchor for bootstrapping and for comparative studies. We assume that when two knowledge systems are studied, there will be no meaningful comparison unless both of them can be put in the same representational framework. In the current work, we adopt SUMO (Suggested Upper Merged Ontology, Niles and Pease 2003) as the framework for ontological representations. SUMO was constructed with the explicit goal to serve as the upper ontology of varying knowledge domains by the IEEEís suggested upper ontology workgroup. In other words, SUMO is supposed to be versatile and has robust coverage of general concepts used by different ontologies. Since SUMO is attested with many contemporary knowledge domains, it offers a good foundation for our comparative study of non-standard ontology. In addition, our application to a temporally and culturally far removed knowledge source offers a genuine challenge to the robustness of SUMO. Lastly, as an upper ontology, SUMO avoids elaboration of lower level nodes. Hence there is only a very low probability that it will run into contradictions with the expanded nodes of a non-standard ontology.

While an upper ontology is adopted as the anchor for domain ontology construction, such an upper ontology may not contain all the finer-grained concepts necessary to fully represent the chosen domain. Hence, we propose to use Wordnet to supplement the knowledge. Wordnet as a lexical knowledgebase provides the natural interface between the domain lexica and SUMO (Niles and Pease 2003). In addition, for concepts not explicitly represented in the upper ontology, wordnet lexical semantic relations can be used to construct a conceptual taxonomy.

All the lexical and knowledge resources required for this approach are already integrated in Sinica BOW (Academia Sinica Bilingual Ontological WordNet, Huang et al. 2004). Hence we use Sinica BOW as the primary referential knowledgebase in this study. Sinica BOW integrates three resources: WordNet, English-Chinese Translation Equivalents Database (ECTED, Huang et al. 2003), and SUMO. Referring to Sinica BOW has three advantages. First, it allows access to both lexical semantic relation in WordNet and conceptual taxonomy in SUMO. Second, it allows lexical search in either Chinese or English. Third, it allows research information to be represented in either Chinese or English.

8. Mapping Lexical Data to Ontology

8.1. Preparing the Lexical Resources

Tang civilization (618-907AD) was one of the most vibrant periods of Chinese civilization. It welcomed and integrated elements from many of the neighboring non-Han civilizations. In turn, Tang civilization was also venerated and imitated by neighboring countries. The Japanese civilization, for instance, borrowed

generously from Tang, including the kanji writing system. It is not an exaggeration to claim that the classical roots of Japanese civilization are actually Tang civilization. Hence, the ontology of the Tang dynasty has far more implications than being an ontology of a long-gone historical period. It may shed light on how heterogeneous knowledge systems integrate, as well as how a borrowed knowledge system develops in the new cultural background.

As a pilot of the main study of constructing an ontology based on the more than 10 millions characters in textual archives from the Tang Dynasty, we construct an ontology based on the famous anthology of The 300 Tang Poems. The text of the 300 Tang Poems contains slightly more than 15,000 characters. This is one of the most important and popular collections of Chinese literature. Its importance far out-weights its relative small size. In addition, since it is poetry, the conceptual density, as represented by the lexical types contained, is high. In this pilot study, the words and classification of words in the text are hand-tagged. The choice of manual tagging is made because our tagger is not tested for domain classification, even though it performs the task of pos tagging very well. The relatively small size of the text also allows manual work to be done efficiently. The highly reliable result will serve as valuable training data for future automatic tagging classification. There is already a classical Chinese tokenizer combining segmentation and tagging available from Academia Sinica. This tokenization program, adopting the basic design of Chen and Liu (1992), is very robust and performed well in the first SigHAN Chinese segmentation bakeoff in 2003. It has also successfully segmented over 5 million words of classical Chinese texts for the language archives project at Academia Sinica.

Three sub-lexicons from the Tang 300 Poems were extracted for domain ontology construction: animals, plants, and artifacts. A total of 176 words were assigned to the three domain lexica: The animals lexicon contains 64 words; the plants lexicon contains 59 words; and the artifacts lexicon contains 53 words. The result from the animal and plant domains will be reported in this paper. These domains are chosen because their meanings are referential and rich. Since they are referential, it is more likely to uniquely determine the meaning of each term. On the other hand, these are familiar terms and important poetic devices used to invoke empathy or express feelings.

The second step in the preparation of the lexical resources for ontology-building is the identification of the appropriate sense of each word for the target knowledge domain. There are two issues involved here. First, as most words are assigned more than one senses in wordnet, we need to identify the correct sense. Second, as these words are used over 11 hundred years ago, some meanings may have become obscure or changed. We need to identify the intended meaning. A batch query on these 176 words was sent to Sinica BOW. Of the 176 words, only 100 words found complete matching entries in the Chinese part of the bilingual wordnet. We then expand the query to include words that share the initial or ending characters. The expanded query still left 24 words with no possible matches in the current version of BOW. These 24 words were later assigned correct translation and meaning with manual dictionary lookup. For words with direct sense assignment from WordNet, the link form BOW to SUMO

ontology is utilized. When a sense does not belong to the target knowledge domain, it is discarded. The senses that belong to the target domain by SUMO assignment is kept for next step. Even though there were in average 2.18 senses assigned for each word, the domain requirement quickly reduced the number of possible senses to close to one.

It is important to notice that expertise knowledge is crucial in the identification of word senses when dealing with a non-standard knowledge domain. A good example is the word *mei2*, with grass radical found in the Tang poems. Its dominant sense in contemporary Chinese equals to berry, as in strawberry "*cao3mei2*". However, further investigation showed that such sensed did not exist in Tang dynasty. The word refers to a kind of moss instead. In other words, although the Chinese character composition reinforces its position in the plants domain, its actual reference cannot be reliably determined by using standard lexical knowledge.

Expertise knowledge and manual editing is also crucial for the words that do not find direct match in Sinica BOW. For example, *hu2jia1* is a particular musical instrument that was first invented and played by the Tartar people and no longer commonly used. Hence its lack of an equivalent in the English language is not surprising. To solve this problem, we consult similar senses from Wordnet. Since *hu2jia2* is a kind of tubular wind instrument, we considered it to be a kind of pipe, which does occur in WordNet and is linked to SUMO.

8.2. Constructing Domain Ontology

Once each lexical item is assigned a unique correct Chinese sense and its corresponding English synset, it can be mapped through Sinica BOW to a SUMO conceptual node. When there is no exact match, lexical semantic relations from WordNet are consulted to establish relation between a lexical item and SUMO. For lexical items that are thus assigned to an appropriate SUMO node, the construction of the domain ontology is as simple as connecting two dots. This is largely the case for the animals ontology (Figure 4).

On the other hand, SUMO as an upper ontology does not necessarily offers sufficient knowledge structure for all domains. For instance, although plants can be considered to be equally salient as animals conceptually, SUMO only gives the very rough-grained classification of FloweringPlant and NonFloweringPlant. Hence we need to use the lexical semantic relations from WordNet to construct the hierarchical conceptual network, i.e. the proposed domain ontology. In this case, we cannot simply copy and connect the relations. Since WordNetís main goal is to record all cognitively relevant semantic relations, not all relations can fit in a rigorous conceptual classification and inference system. Hence, after bootstrapping with all WordNet synsets and relations marked, an important step is to prune the resultant tree for both inconsistency and redundancy. The plants ontology in Figure 5 is the wordnet-based ontology after extensive pruning.

In establishing the link between a sense and a ontology node, it is important to notice that the SUMO-WordNet link is established with the contemporary background knowledge of the English speaker world. Hence it is likely to find that a non-standard ontology based on a different system will require a totally different conceptual assignment. An instance of such mismatches involves *mou2hu2*, which is a kind of silk flag. A flag, according to both the literary context and the assigned lexical

sense, should be a piece of artifact, solid and substantial. However, the SUMO-WordNet link that Sinica Bow follows mapped it to the conceptual node of "Icon". This may be appropriate when a flag is used in signing, but not appropriate in the Chinese context. Hence we simply correct the link and assign it to artifact.

What is more interesting in terms of linguistic use involves words that seem to carry the same meaning, while involves fundamentally different conceptualization. The difference in conceptualization requires assignment to a different ontological location. One such example is *dai4mei4*, which is given the sense of "a beaded sea turtle", and seems to be a straightforward case of a kind of animal. However, when we refer to the context, the sentence actually refers to "a beam inlaid with *dai4mai4*". In other words, it refers to the materials used in decorating a building. It is the shell of the turtle that has been ground and polished like a piece of jade. It is also interesting to note the fact that these two characters used have a jade radical, rather than an animal or fish radical. Both the context and the written form suggest that the sense being used here is the material, and there is no evidence suggesting that Tang people know that the *dai4mei4* material comes from a turtle. Hence this word is not included in the animals ontology.

On the other hand, when metonymy is used, it is often possible to argue that the original sense is invoked. An example in our study is *shuang1li2*, double-carp, which refers to a letter since letters are traditionally sent in a wood box with two carps carved on top. In this case, even though the actual reference is not the animal, but the lexical metonymy necessarily involve the image of the fish. Hence we consider the concept of carp is used, and hence justifying our including carp as an attested case for the animals ontology for Tang.

9. Result and Discussions

The result of this pilot study will include three semi-automatically constructed sub-ontologies: animal, plant, and artifact. The first two are completed and will be discussed here. The top part of each ontology is mapped to SUMO. The lower part of each ontology is extended using WordNet relations. These ontologies as well as the attached lexical terms will have Chinese-English bilingual representation.

The first generalizations that can be obtained are from the distribution of these domain terms in the texts. The total frequency of these three domains ranges from 1.65% to 1.89%. These are relatively high compared to a balanced corpus. In a balanced corpus, the top 20 animal or plant domain terms comprise of less than 1%.

The second generalizations can be made from the distribution among the different terms within the domain. Among animal concepts, the total frequency of birds is over 38%, and hoofed mammals over 30%. These two kinds each far exceed all the other eight kinds of animals combined. This fact should have implications on either the fauna of Tang, or the poetic choice of images. Even more striking is the fact that of all plants, flowering plants consist of over 95% of the instances in the texts. This fact should not be surprising because of the strong poetic image that a flower presents.

After the sub-ontologies are constructed, comparative studies of the Tang ontological structure with our contemporary ontology (based on SUMO) will be conducted. For

instance, we found that among the order of mammals, the families of marsupials and marine mammals are missing. The absence of marsupials is expected since it is a fact of science history that they were discovered much later. The absence of marine mammals may point to the fact that the Tang civilization is mainly land-based. In addition, we also found two interesting facts in other branches. First, almost all invertebrates that are documented are (winged) insects. And among the non-mammal vertebrates, with only less than 5 exceptions, all documented lexical items refer to bird. A possible explanation of the idiosyncrasy is the Tang civilizationís fascination with flying. We know as a fact that flying is a recurring theme in paintings from this period, and occur in poetry too.

The plants ontology of Tang offers a good test case of how to bootstrap an ontology with lexical knowledgebases such as wordnets. We showed that when the lexical resource contains sense and lexical semantic relations information, it is possible to use the information to bootstrap a domain ontology. The crucial challenge here is how to turn the set of pair-wise and lexicon-driven relations to a taxonomical hierarchy. An issue that will recur is how to deal with same level nodes that are classified and assigned with diagonal criteria. One such example is the classification of plants in Figure 5. FloweringPlants and HerbaceousPlants and AcquaticPlants create partially overlapping classes. These are all linguistically and cognitively motivated and cannot subsume each other. Given the fact that even an upper ontology like SUMO acknowledes such human cognitive facts and allows multiple inheritance, there is still reservations that an ontology can quickly become non-trackable if no constraints are put on such cross-classification. This is an issue that merits in-depth formal and theoretical deliberation.

10. Conclusion

In this current study, we propose the Shakespearean-garden approach to the construction of non-standard ontology. We showed with a pilot study that such an approach is feasible, especially when supported by the right combination of lexical knowledge sources and upper ontology. In addition, we showed that the constructed sub-ontology allows us to have a comprehensive view of the knowledge system of a civilization that no longer exists. Such a representation will offer a unique opportunity to study how their world differs from ours and how they view the world differently from us.

A natural extension of the current work is to try to piece these sub-ontologies together to form a skeletal ontology for the Tang dynasty. In order to carry out this full-scale work, we have already started the design and construction of automatic tools to construct domain ontology based on domain lexicons and SUMO. This will integrate the knowledge we gain from the current work as well as modules from existing systems, such as Sigma system constructed by Adam Pease. Such a working environment will facilitate the ultimate goal of the Shakespearean-garden approach. In addition, we will also try to apply the simultaneous bilingual mapping approach to construct a modern domain. Ultimately, we would like to see if it is still plausible to construct ontology based on a shared upper ontology even if the background knowledge systems are drastically different.

The current work on the domain knowledge of Tang civilization willl also

provide solid foundation for future work on metaphor. Based on Lakoffís contemporary theory of metaphor, Ahrens et al. (2003) shows that the crucial step in predicting and explanation of the use of linguistic metaphors lies in capturing the rules governing the mapping between source domain and target domain knowledge. For the historical poetic work such as Tang poetry, an additional challenge to the study of metaphor would be the precise characterization of the source domain knowledge. Our non-standard ontology can be viewed as the foundational work defining source domain knowledge in Tang poetry. With the source domain knowledge described, we will be able to develop in-depth study of Tang poetic metaphors in the future.

Lastly, the issue regarding the relation between a wordnet and an ontology is also touched upon. In the Shakespearean-garden approach, it is crucial that the specific domain lexicon can be obtained and annotated with correct lexical semantic information. However, how can lexical semantic relations be best used in an ontological study remains a challenging and promising issue.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu/~wn/
Tender Lyrics-The 300 Tang Poems (in Chinese) http://cls.admin.yzu.edu.tw/300/HOME.HTM
CKIP Segmentation and Tagging Program
http://corpus.ling.sinica.edu.tw/project/LanguageArchive/lc_index.html

Reference

Ahrens, Kathleen, Chu-Ren Huang, and Siaw-Fong Chung. 2003. Conceptual Metaphors: Ontology-based representation and corpora driven Mapping Principles. Presented at the Workshop on Lexicon and Figurative Language. An ACL2003 Workshop. July 11, Sapporo, Japan.

Chang, Ru-Yng and Feng-ju Luo. 1999. Cross-platform Web-bases Learning Systemó the construction of Tender Lyrics-The 300 Tang Poems (in Chinese). Presented at 1999 Taiwan Symposium on Taiwan Academic network. Kaohsiung.

Chen, K.-J. and S.-H. Liu. 1992. Word Identificaiton for Chinese Sentences. Proceedings of COLING92. 501-505.

Fellbaum, Christine. Ed. (1998). WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

Huang, Chu-Ren, Ru-Yng Chang, and Shiang-bin Li. 2004. Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO. To be presented at the LREC2004 conference. May26-28. Lisbon.

Huang, Chu-Ren, Li, Xiang-Bing, Hong, Jia-Fei. (2004). Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing. Proceedings of the Asian Symposium on Natural Language Processing to Overcome Language Barriers. March 25-26, 2004. Hainan Island.

Huang, Chu-Ren, Feng-ju Lo, Ru-Yng Chang, and Sueming Chang. (2004). Sinica BOW and 300 Tang Poems: An overview of a bilingual ontological wordnet and its application to a small ontology of Tang poetry. Presented at the Workshop on Possibilities of a Knowledgebase of Tang Civilization. Institute for Research in Humanities, Kyoto University. February 20-21.

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.

Huang, Chu-Ren. Elanna I.J. Tseng & Dylan B.S. Tsai. (2002). Translating Lexical Semantic Relations: The first step towards multilingual Wordnets. In Proceedings of the COLING2002 workshop: SemaNet: Building and Using Semantic Networks. Taipei, Taiwan.

Niles, I. & Pease, A. (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In Proceedings of the IEEE International Conference on Information and Knowledge Engineering. (IKE 2003), Las Vegas, Nevada.

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.

Wilkins, J. (1668). An Essay Towards a Real Character, and a Philosophical Language. Reprinted in 2002. Thoemme Press.

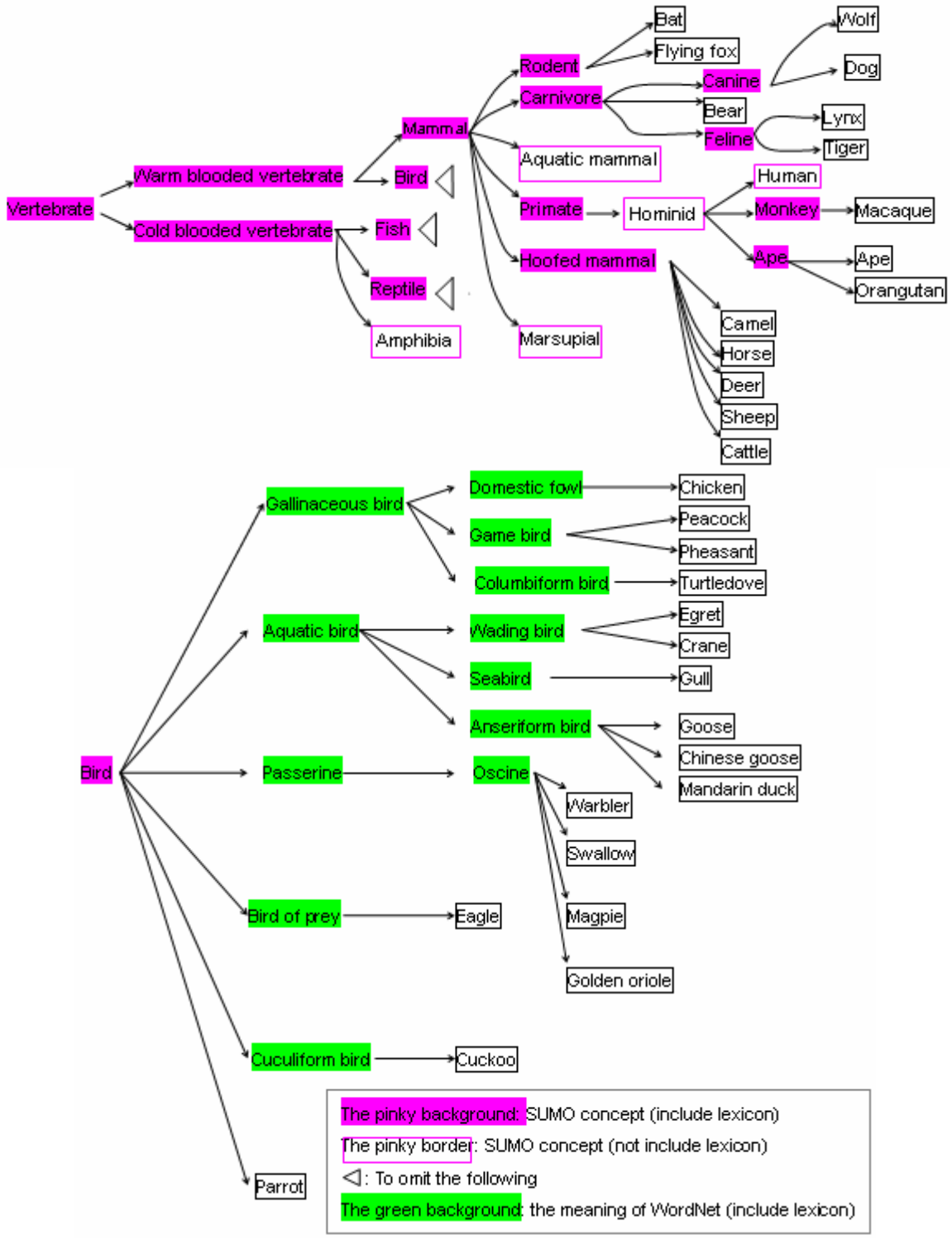


Figure 4: Tang Animals Ontology

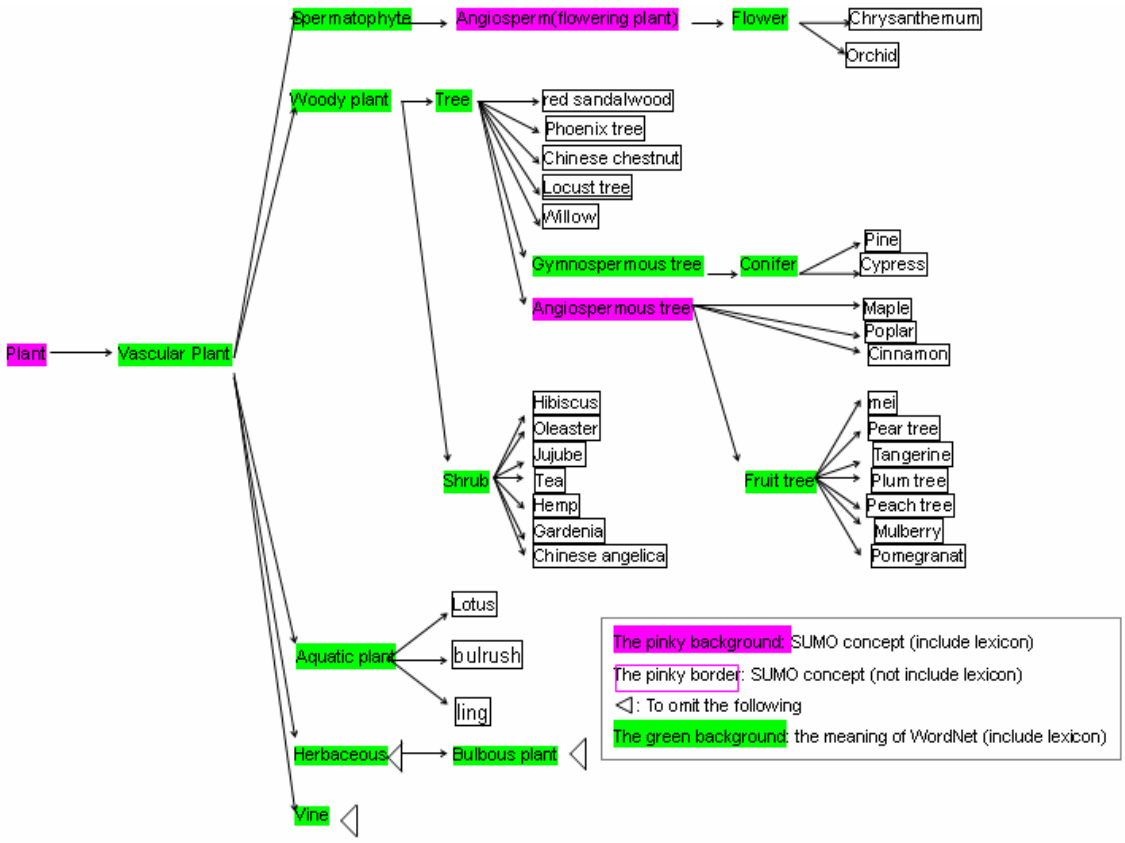


Figure 5: Tang Plants Ontology