Reconstructing the Ontology of the Tang Dynasty
A pilot study of the Shakespearean-garden approach

Chu-Ren Huang*, Feng-ju Lo, Ru-Yng Chang*, Sueming Chang***
*Academia Sinica, **Yuan Ze University
130 SEC.2 Academia Road, Nankang, Taipei, TAIWAN 11529, R.O.C
churen@sinica.edu.tw, echinese@saturn.yzu.edu.tw, ruyng@gate.sinica.edu.tw, kati@gate.sinica.edu.tw

Abstract

We propose the Shakespearean-garden approach towards domain ontology construction in this paper. In sum, we suggest that domain lexica can be extracted and obtained for non-standard knowledge backgrounds. Once the comprehensive lexica are collected, a lexical interface between wordnet and our Sinica BOW can be applied. It will allow each lexical item to a conceptual location on Sinica BOW. With the WordNet and SUMO interface, as well as our bilingual correspondence program, each domain lexica can be mapped to a set of SUMO conceptual nodes. These nodes will each be linked to the ontology. We show that the domain ontologies can be constructed directly from synset-ontology pairs, or from the lexical information taken from Wordnet.

1 Background

1.1 Non-Standard Ontology

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 ontological 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.

1.2Some Basic Facts

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 The Shakespearean-garden Approach

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 IEEE's 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.

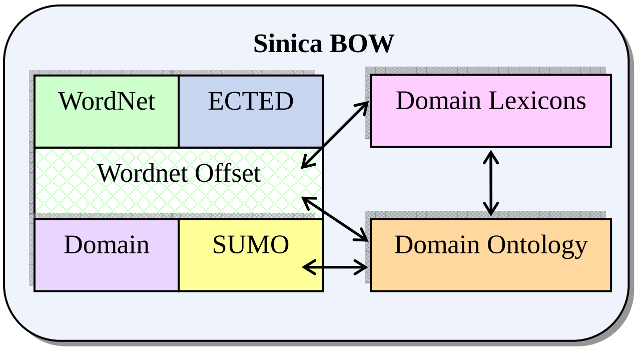

Figure 1: The resource and structure of Sinica BOW

3 Mapping Lexical Data to Ontology

3.1 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.

3.2Constructing 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 WordNet's 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 is 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 in 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 refer to a letter since letters are traditionally sent in a word 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.

4 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 civilization's 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.

5 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 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 Lakoff's 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.

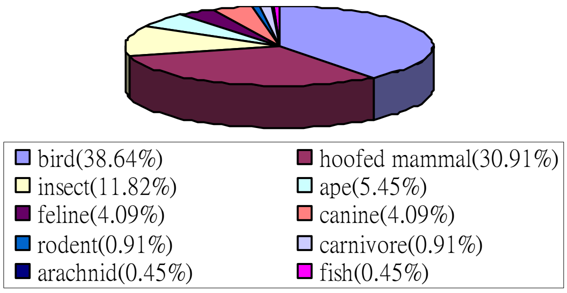

Figure 2: Distribution of animal concepts in Tang 300

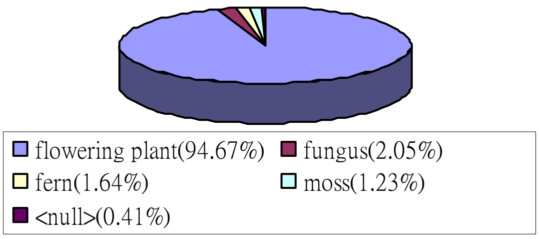

Figure 3: Distribution of plants concept s in Tang 300

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 System—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.

Pease, A., (2003). The Sigma Ontology Development Environment. In Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.

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

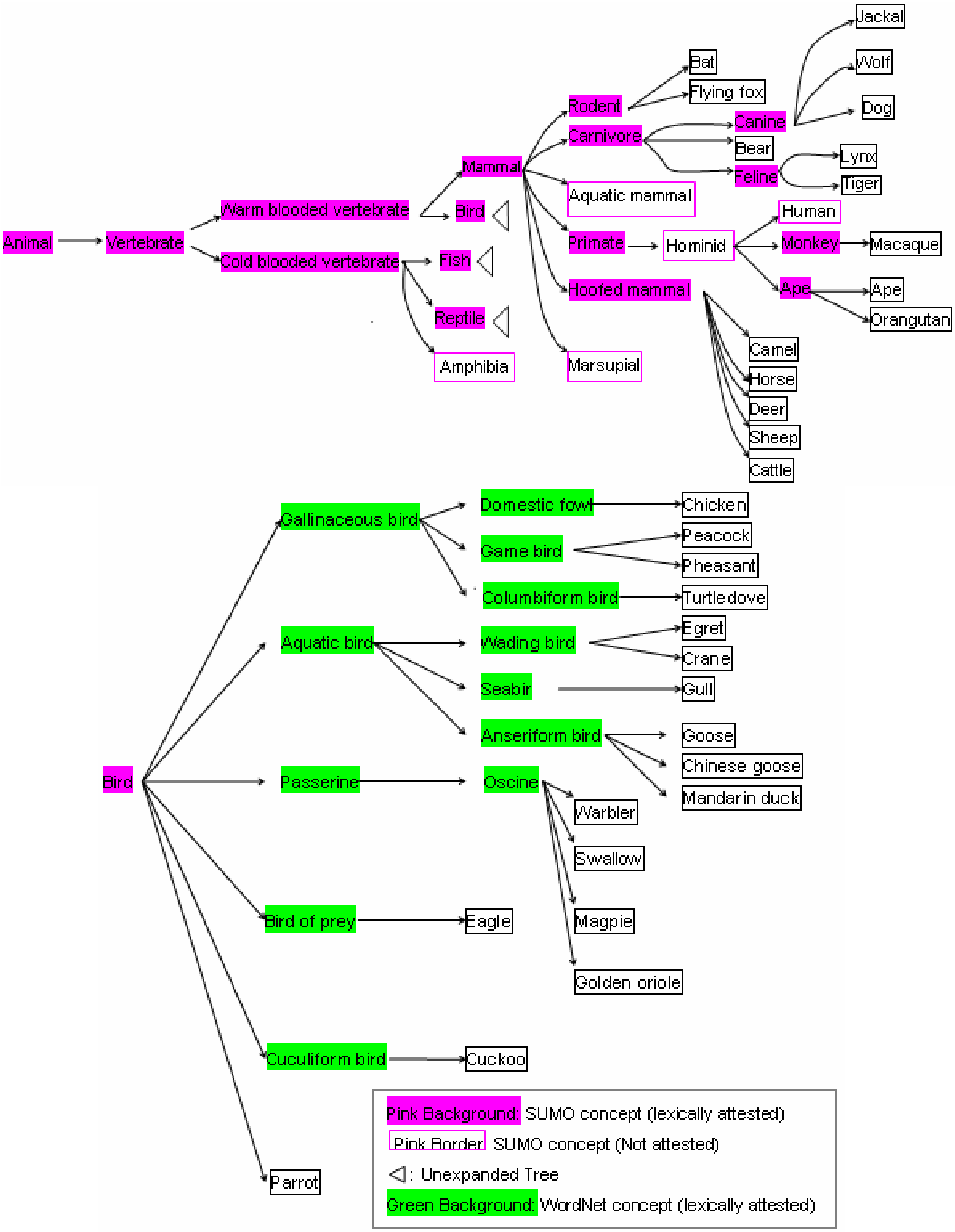

Figure 4: Tang Animals Ontology

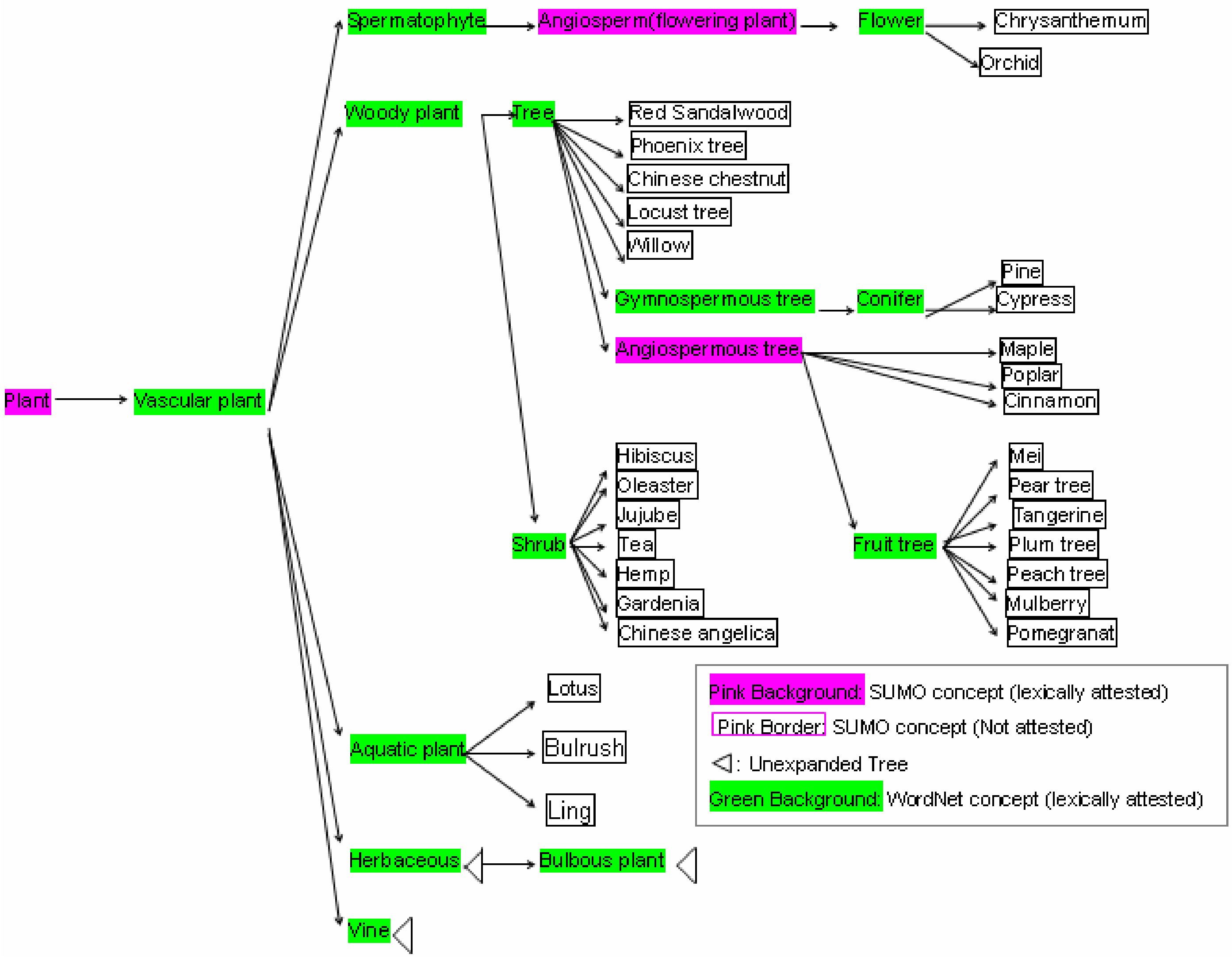

Figure 5: Tang Plants Ontology