Sinica BOW (Bilingual Ontological Wordnet): Integration of Bilingual WordNet and SUMO

Chu-Ren Huang, Ru-Yng Chang, Hsiang-Pin Lee

Academia Sinica.

130 SEC.2 Academia Road, Nankang, Taipei, TAIWAN 11529, R.O.C

churen@sinica.edu.tw, ruyng@gate.sinica.edu.tw, dreamer@hp.iis.sinica.edu.tw

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 & 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. Third, multiple linguistic indexing is built in to allow additional versatility. Fourth, domain information allows another dimension of knowledge manipulation.

1 Background and Motivation

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.

2 Resources and Structure

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 withvlexical 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.

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.

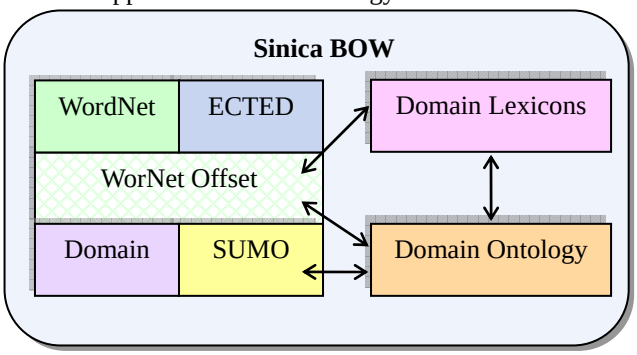

Figure 1: The resource and structure of Sinica BOW

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, & 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 lexicon, 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 domain 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 'unknown' 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

3 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..

3.1 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 learner's Lexical KnowledgeNet), and to online monolingual and bilingual dictionaries. In addition to online access of multiple sources information, each lemma's distribution in these resources is also a good indicator of its usage level.

Figure 2: Initial Return of Lemma Search

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.

3.2 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 3. 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 lemma query result of Sinica BOW

4 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. Out Domain Lexico-Taxonomy approach attempts to assign a domain tag to a word whenever applicable. We also encourage users of SUMO to give feedback with their own domain use of lexical items because domain specifications can not be covered by any single knowledge source. Hence 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.

5 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 4. 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 a 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 reflected 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 civilization's fascination with flying by looking at the dominance of animal references in the texts.

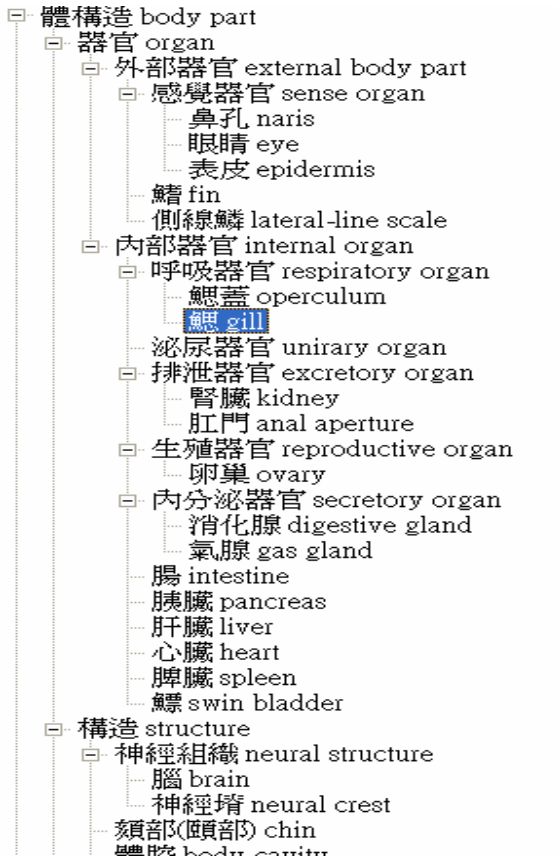

Figure 4: A sample domain ontology: Fish

6 Conclusion

Building a linguistic resource that bridges intuitive lexical use to structured knowledge is one current challenge facing computational linguists. The success of this resource will have a positive impact on both on the Semantic Web and the Computational Linguistics community. Sinica BOW is one of earliest attempts in this direction. An important future development will be to construct and integrate more domain ontologies to build up a more complete knowledge map.

Online Resources

Sinica BOW: http://BOW.sinica.edu.tw/
SUMO: http://ontology.teknowledge.com/
WordNet: http://www.cogsci.princeton.edu/~wn/

Referneces

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

Huang, Chu-Ren, Li, Xiang-Bing, & Hong, Jia-Fei. (2004). Domain Lexico-Taxonomy:An Approach Towards Multi-domain Language Processing. To be presented at 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, & 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.