A DAML+OIL-Compliant Chinese Lexical Ontology

Yu-Sheng Lai, Ren-Jr Wang and Wei-Tek Hsu
Advanced Technology Center,
Computer & Communications Research Laboratories/Industrial Technology Research Institute,
E000, 195-11, Sec. 4, Chung Hsing Rd. Chutung, Hsinchu, Taiwan 310, R.O.C.
{laiys, rjwang, whsu}@itri.org.tw

Abstract

This paper presents an ongoing task that will construct a DAML+OIL-compliant Chinese Lexical Ontology. The ontology mainly comprises three components: a hierarchical taxonomy consisting of a set of concepts and a set of relations describing the relationships among the concepts, a set of lexical entries associated with the concepts and relations, and a set of axioms describing the constraints on the ontology. It currently contains 1,075 concepts, 65,961 lexical entries associated with the concepts, 299 relations among the concepts excluding the hypernym and hyponym relations, 27,004 relations between the lexical entries and the concepts, and 79,723 relations associating the lexical entries with the concepts.

Introduction

The Semantic Web relies heavily on formal ontologies to structure data for comprehensive and transportable machine understanding [Maedche and Staab 2001]. Therefore, constructing applicable ontologies influences the success of Semantic Web largely. An ontology mainly consists of a set of concepts and a set of relations that describe relationships among the concepts. An upper ontology is limited to concepts that are abstract, generic, domain-broad, and articulate. Cycorp constructed an upper ontology – Upper Cyc® Ontology. It consists of approximately 3,000 terms, i.e. concepts and relations. It has been used for organizing the upper structure of a knowledge base – the Cyc® KB. A working group of IEEE (P1600.1) is also trying to standardize the specification of the upper ontology. An upper ontology, called IEEE SUO (Standard Upper Ontology), is expected to enable computers to utilize it for applications, such as natural language understanding and generation, information retrieval and extraction, Semantic Web services [McIlraith et al. 2001], etc. It is estimated to contain between 1,000 and 2,500 terms plus roughly ten definitional statements for each term.

This paper presents an ongoing task that will construct an upper-ontology-like ontology for Chinese research and applications. We refer to it as CLO (Chinese Lexical Ontology). In addition to the structural portion, the CLO will contain Chinese lexicons associated with the concepts and relations. A pure ontology containing concepts only (without lexicons) is abstract. A lexicon-associated ontology makes the substantiation of abstract concepts easier. HowNet defines 65,961 Simplified Chinese lexical entries by a set of predefined features including 6 categories of primary features and 100 secondary features, and several symbols, in which the primary features are in a taxonomy with single inheritance. The taxonomy is essentially regarded as the taxonomy of the CLO. However, the Chinese lexical entries defined in HowNet are simplified version. They are not suitable for Traditional Chinese research and applications. A traditional version of Chinese dictionary released by Sinica of R.O.C. is frequently used for Traditional Chinese NLP. By combining the Traditional Chinese dictionary and the HowNet, we attempt to construct the CLO and represent it in the semantic markup language DAML+OIL since DAML+OIL is currently a basis of Web ontology language.

The task of constructing the CLO can be divided into three portions. Firstly, a hierarchical taxonomy of concepts including relations among the concepts is required. In our case, we utilize the hierarchical primary features of HowNet to form the structure. Secondly, a set of lexical entries should be associated with the concepts and relations. Thirdly, a set of axioms that describe additional constraints on the ontology are required. This paper addresses the ongoing construction task and a brief introduction of Web ontology languages.

1 An Overview of Ontology and Its Languages

This section will describe the definition of ontology from different viewpoints and several semantic markup languages frequently used for representing ontologies.

1.1 What is Ontology?

In WordNet, the word "ontology" has a sense of "*the metaphysical study of the nature of being and existence*." The term has a long history in philosophy. It refers to the subject of existence. For AI community, it seems to generate a lot of controversy. One definition by Gruber (1993) is that "*an ontology is an explicit specification of a conceptualisation*." He considered that every knowledge base, knowledge-based system, or knowledge-level agent is committed to some conceptualization explicitly or implicitly. The conceptualization includes the objects presumed or hypothesized to exist in the world and their interrelationships [Genesereth and Nilsson 1987].

According to Gruber's definition, an ontology basically consists of a set of *concepts*, i.e. the so-called objects, which represent classes of objects, and a set of *relations*, i.e. the so-called interrelationships, which are binary relations defined on concepts. A special transitive relation "**subClassOf**" represents a subsumption relation between concepts and structures a taxonomy. In addition to the taxonomy, an ontology typically contains a set of inference rules. The inference rules enhance the ontology's power in reasoning.

Maedche and Staab (2001) proposed an ontology-learning framework for the Semantic Web. In their case, they formally defined an ontology as an 8-tuple $<L, C, H_C, R, H_R, F, G, A>$, in which the first primitive L denotes a set of strings that describe lexical entries for concepts and relations, the middle 6 primitives structure the taxonomy of the ontology, and the last primitive A is a set of axioms that describe additional constraints on the ontology. Staab and Maedche (2001) considered that the axioms make implicit facts more explicit. The ontology is actually a lexical ontology since it comprises a set of lexical entries. We adopt the ontology's definition for constructing the CLO.

1.2 Ontology Languages for the Semantic Web

The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [Berners-Lee et al. 2001]. The goal of developing the Semantic Web is to facilitate the communications of people-to-machine, machine-to-people, and machine-to-machine. A way to achieve this goal is to give the information the Web provided a well-defined meaning. Several markup languages are developed for this purpose. Fig. 1 shows the layer language model for the Web [Fensel 2000].

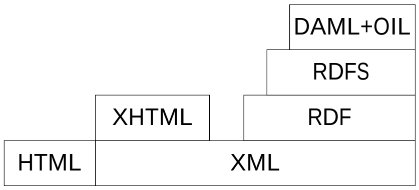

Fig. 1 The layer language model for the Web.

HTML is the most popular markup language for structuring documents. Its simplicity enabled fast growth of the Web, but also resulted in information overload and that knowledge could not be shared conveniently and efficiently. An extensible metalanguage – XML was born for this reason. One of its instances – XHTML redefined from HTML is more extensible than HTML, but is still not powerful enough for the Semantic Web.

RDF (Resource Description Framework) developed by the W3C (World Wide Web Consortium) is also an instance of XML. It is a foundation of Web metadata processing and used for describing relationships between resources. In general, any Web resource could be described in RDF. The formal model for RDF can be represented as triples: < *predicate*, *subject*, *object*>. The instances of the model can be also viewed as directed labeled graphs, which resemble semantic networks [Quillian 1968]. It also provides interoperability among applications, which enables applications to exchange and share machine-understandable information on the Web.

RDFS stands for RDF Schema. It is an extensible, object-oriented type system, which is introduced as a layer on top of the basic RDF model [Brickley and Guha 2000]. RDFS can be viewed as a set of ontological modeling primitives on top of RDF [Fensel 2000]. For example, two semantic modeling primitives "**rdfs:Class**" and "**rdfs:subClassOf**" can be used for defining the taxonomy of an ontology.

A semantic markup langauge – DAML+OIL, derived from RDF and RDFS, defines more primitives, such as **daml:complementOf** for complement relation, **daml:sameClassAs** for equivalence relation, etc., to represent more relationships among resources. DAML+OIL was built originally from the DAML ontology language – DAML-ONT. It combines many language components of the OIL with the formal seamtics and reasoning services provided by description logic. Summarily, compared to other languages, such as XML DTD, XML Schema, and RDF(S), DAML+OIL possesses richer language features, such as *class expressions*, *defined classes*, *formal semantics*, *inference*, *local restrictions*, and *qualified constraints* (see more at www.daml.org/language/features.html). It is currently the basis of W3C's Web ontology language. Therefore, we also follow this specification for the CLO.

2 Construction of the Chinese Lexical Ontology

As mentioned above, the CLO mainly consists of three components: a hierarchical taxonomy of concepts, a set of lexicons associated with the concepts and relations, and a set of axioms. We do not intend to explore the axioms in this paper. In the following, we will describe how to construct the hierarchical taxonomy and how to associate lexicons with the concepts.

2.1 Conversion of HowNet

The hierarchical taxonomy is actually a conversion of HowNet. One of the important portions of HowNet is the methodology of defining the lexical entries. In HowNet, each lexical entry is defined as a combination of one or more *primary features* and a sequence of *secondary features* with prefixed *symbols*. The primary features indicate the entry's category, which are in a hierarchical taxonomy as shown in Fig. 2. Based on the category, the secondary features make the entry's sense more explicit, but they are non-taxonomic. Some of the secondary features are prefixed with symbols. The symbols describe the relationships among the lexical entry, the primary feature, and the secondary features.

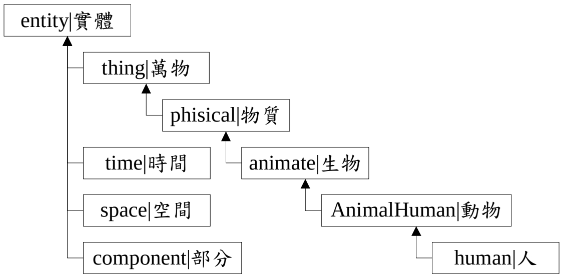

Fig. 2 Taxonomy of partial primary features in HowNet.

For example, HowNet defines the lexical entry "股票經紀人" (stockbroker) as follows:

(D.1) "*Human|人, #occupation|職位, commercial|商, *buy|買, *sell|賣, #fund|資金*"

The first term "*Human|人*" is the so-called primary feature. The remaining terms are the so-called secondary features, in which the secondary feature "*buy|買*" is prefixed with the symbol "*." It indicates that "股票經紀人" (stockbroker) can be an agent of "*buy|買*."

In the following, we will describe how to extract hierarchically structured concepts, relations among the concepts, and relations between the lexical entries and the concepts, and how to associate the lexical entries with the concepts.

Hierarchically Structured Concepts

Totally 1,521 primary features are divided into 6 upper categories: **Event**, **Entity**, **Attribute Value**, **Quantity**, and **Quantity Value**. After eliminating replica, we obtain 1,075 distinct primary features. These primary features are well organized into a hierarchical taxonomy. Each primary feature can be viewed as a concept for an ontology. By the taxonomy, we construct a fundamental ontology that consists of a set of concepts and a special relation – “**subClassOf**.” The “**subClassOf**” relation realizes the *hypernym* and *hyponym* relationships among the concepts.

Relations among the Concepts

In the two categories, **Event** and **Entity**, some primary features have auxiliaries for describing the relationships to other primary features. For example, the primary feature “*human|人*” has auxiliaries: [*!name|姓名, !wisdom|智慧, !ability|能力, !occupation|職位, *act|行動*]. It means that “*human|人*” has attributes: “*name|姓名*,” “*wisdom|智慧*,” “*ability|能力*,” “*occupation|職位*,” and can be an agent of “*act|行動*.” For the CLO, these auxiliaries are used for constructing the relations among the concepts.

Relations between the Lexical Entries and the Concepts

A noticeable thing is that, in HowNet, many primary features appear in the secondary features of many lexical entries to assist describing the senses of those lexical entries. That is, they play the roles of secondary features. For example, in (D.1), the fourth term “*buy|買*” is a secondary feature for the lexical entry “股票經紀人” (stockbroker). And it is also a primary feature of the taxonomy. In other words, these secondary features are concepts for the ontology. For each of them, its prefixed symbol provides a relation between the lexical entry and the concept to which it corresponds.

Associating the HowNet Lexical Entries with the Concepts

As mentioned before, a lexical entry in HowNet is defined as a combination of one or more primary features and a sequence of secondary features with prefixed symbols. Its primary features being taxonomic indicate its category. And we took the taxonomy as the taxonomy of the CLO. Therefore, each HowNet lexical entry can be well associated with one or more concepts according to its primary features.

2.2 Classification of the Lexicons

From HowNet, we constructed the ontology taxonomy and obtained Simplified Chinese lexical entries. It is still lack of Traditional Chinese lexical entries. For the completeness of CLO, we need a dictionary supporting Traditional Chinese.

A Traditional Chinese dictionary compiled by Academia Sinica of R.O.C. (www.sinica.edu.tw) was released for computational processing of Chinese natural language. It consists of 78,322 words; each is associated with one or more parts-of-speech (POS). Taking into account the POS, there are totally 80,491 lexical entries. The dictionary is an available and necessary complement to Traditional Chinese's research and applications. We will refer to it as Sinica Dictionary for short.

Associating the lexical entries in Sinica Dictionary with the concepts is equivalent to classifying them into the ontology taxonomy. In order to ensure the correctness of the classification task, we proceed in a semi-automatic approach. Fig. 3 illustrates the flow diagram of semi-automatically classifying the lexical entries into the ontology taxonomy.

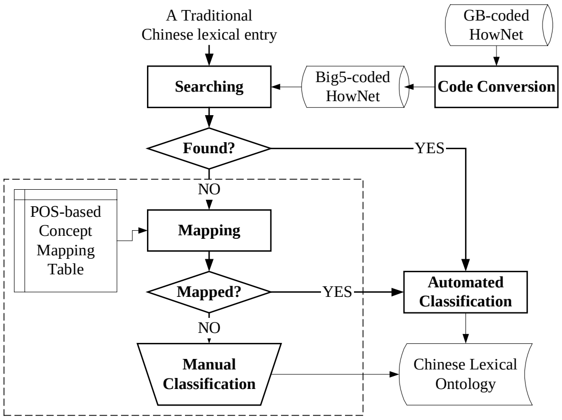

Fig. 3 The flow diagram of manually classifying the Sinica lexical entries into the CLO.

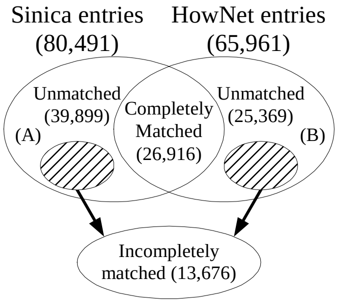

Fig. 4 A diagram matching Sinica's lexical entries with HowNet's.

For each lexical entry in Sinica Dictionary, we search it in a pool of Big5-coded HowNet lexical entries, which are transformed from the HowNet originally encoded in GB code. If it can be found and its POS is the same as its corresponding HowNet lexicl entry's, then we associate it with the concepts which the HowNet lexical entry belongs to. There are totally 26,916 completely matched entries. The remaining Sinica lexical entries contains 39,899 unmatched entries and 13,676 incompletely matched entries whose characters are matched but POSs are not matched. (see Fig. 4)

Observing the incompletely matched entries, in fact, all of them are multi-conceptual. To classify each of them into the ontology taxonomy according to its primary feature in HowNet is resonable. For example, the lexical entry "扣人心弦" (exciting) is classified into the adjective category in HowNet, but has a POS "VH11" in Sinica Dictionary. It is a verb in (S.1) and an adjective in (S.2).

(S.1) "這部電影扣人心弦。"
(The movie excited everybody.)

(S.2) "我看了一部扣人心弦電影。"
(I saw an exciting movie.)

Since the incompletely matched entries are multi-conceptual, they must be classified into other concepts. Totally 53,575 lexical entries (unmatched and incompletely matched) should be classified yet. It is very hard to manually classify such a large number of lexical entries into the large scale ontology consisting of 1,075 concepts. Therefore, an efficient approach is still required.

Traditional and Simplified Chinese languages are originated from the same people. The languages should not be much different. An assumption is that most of the lexical entries in the two unmatched groups, i.e. regions (A) and (B) in Fig. 4, should be almost identical in semantics and syntax. Under this assumption, we can produce a mapping table between the two groups of unmatched entries according to their POSs. Thus, we can shorten the time for manual classification. This task, i.e. the dash-blocked area in Fig. 3, is ongoing.

Conclusion and Future Work

The CLO currently contains 1,075 concepts, 65,961 lexical entries associated with the concepts, 299 relations among the concepts excluding the "**subClassOf**" relations, 27,004 relations between the lexical entries and the concepts, and 79,723 relations associating the lexical entries with the concepts.

We are working toward the classification of the unmatched Tradictional Chinese lexical entries into the CLO. Besides, the relations are not associated with lexical entries yet, therefore we will research into this problem in the future.

Acknowledgements

This paper is a partial result of Project A311XS1211 conducted by ITRI under sponsorship of the Ministry of Economic Affairs, R.O.C.

References

Berners-Lee T., Hendler J. and Lassila O. (2001) *The Semantic Web*, Scientific American.

Brickley D. and Guha R. V. (2000) *Resource Description Framework (RDF) Schema Specification 1.0*, W3C Candidate Recommendation 27 March 2000.

Chinese Knowledge Information Processing Group (1993) *Analysis of Chinese Part-of-Speech*, Technical Report no. 93-05, Institute of Information Science, Academia Sinica, R.O.C.

Fensel D. (2000) *The Semantic Web and Its Languages*, IEEE Intelligent Systems, vol. 15, no. 6, pp. 67-73.

Genesereth M. R. and Nilsson N. J. (1987) *Logical Foundations of Artificial Intelligence*, Los Altos, California: Morgan Kaufmann Publishers, Inc., Chap. 2, pp. 9-13.

Gruber T. R. (1993) *A Translation Approach to Portable Ontology Specifications*, Knowledge Acquisition, vol. 5, no. 2, pp. 199-220.

Maedche A. and Staab S. (2001) *Ontology Learning for the Semantic Web*, IEEE Intelligent Systems, vol. 16, no. 2, pp. 72-79.

McIlraith S. A., Son T. C. and Zeng H. (2001) *Semantic Web Services*, IEEE Intelligent Systems, pp. 46-53.

Quillian M. R. (1968) *Semantic Memory*, Semantic Information Processing, The MIT Press, pp. 227-270.

Staab S. and Maedche A. (2001) *Knowledge Portals – Ontologies at Work*, AI Magazine, vol. 21, no. 2.