ONTOLOGY LEARNING FOR CHINESE INFORMATION ORGANIZATION AND KNOWLEDGE DISCOVERY IN ETHNOLOGY AND ANTHROPOLOGY

Jing Kong

Institute of Ethnology & Anthropology, Chinese Academy of Social Sciences, Beijing, 100081, China
Email: kongjing@cass.org.cn

ABSTRACT

This paper presents an ontology learning architecture that reflects the interaction between ontology learning and other applications such as ontology-engineering tools and information systems. Based on this architecture, we have developed a prototype system CHOL: a Chinese ontology learning tool. CHOL learns domain ontology from Chinese domain specific texts. On the one hand, it supports a semi-automatic domain ontology acquisition and dynamic maintenance, and on the other hand, it supports an auto-indexing and auto-classification of Chinese scholarly literature. CHOL has been applied in ethnology and anthropology for Chinese information organization and knowledge discovery.

Keywords: Ontology learning, Domain ontology, Automatic ontology acquisition, Concepts extraction

1 INTRODUCTION

Since the 1990s, ontology has become a popular research topic studied by several artificial intelligence research communities. Ontologies, as shared conceptualizations for representing domain knowledge, are also becoming the key methods and tools in many fields, such as knowledge engineering, intelligent information integration, knowledge management, information retrieval, and the Semantic Web. Although ontology-engineering tools have matured over the last decade, manual ontology acquisition is a difficult, slow, time-consuming, tedious, and costly task that can easily result in a knowledge acquisition bottleneck. For this reason, it is necessary to develop methods and techniques that allow a reduction of the effort necessary for the ontology acquisition process, which is the goal of ontology learning.

The term *ontology learning* was coined in 2000, at the first workshop on ontology learning held in conjunction with the 14th European Conference on Artificial Intelligence (ECAI2000). Gómez-Pérez defined *ontology learning* as the set of methods and techniques used for building an ontology from scratch or enriching or adapting an existing ontology in a semi-automatic fashion using several sources (Gómez-Pérez & Manzano-Macho, 2003). Recently, there has been a surge of interest in studying ontology learning. In the past few years, many ontology learning tools such as TextToOnto (Maedche & Staab, 2004), OntoLearn (Velardi, Fabriani, & Missikoff, 2001), the ASIUM system (Faure & Nedellec, 1998; Nedellec, 2000), the Mo'k Workbench (Bisson, Nedellec, & Canamero, 2000), OntoLT (Buitelaar, Olejnik, & Sintek, 2004), Adaptiva (Brewster, Ciravegna, & Wilks, 2002), SOAT (WU & HSU, 2002) and DOGMA (Reinberger, et al., 2004) have been developed.

Despite the significant amount of work done on ontology learning in recent years, learning onology from Chinese text has not been widely applied in practice. This paper addresses a framework, an approach, and

techniques of ontology learning for Chinese information organization and knowledge discovery. We have developed a prototype system CHOL: a Chinese Ontology Learning tool, which semi-automatically extracts domain ontologies from Chinese domain specific texts. Further, we have tested CHOL in ethnology and anthropology and used it to find and extract unknown terms and the relationship between terms from Chinese texts about Chinese minority customs.

2 RELATED WORK

Because the term *ontology learning* was firstly proposed in Europe, ontology learning has been widely researched there in the past. However, in recent years, ontology learning has expanded to other areas of the world.

TextToOnto, OntoLT, Adaptiva and OntoLearn are typical ontology learning tools. TextToOnto (Maedche & Staab, 2001) developed at the AIFB Karlsruhe is a semi-automatic ontology learning tool, embedded in the Oi-Modeler platform and OntoEdit ontology engineering workbench. Its framework's modules serve different steps in the ontology engineering cycle. OntoLT is a plug-in for Protégé with which concepts (Protégé classes) and relations (Protégé slots) can be extracted automatically from a linguistically annotated text collection. Adaptiva (Brewster, Ciravegna, & Wilks, 2002) is an ontology building environment that implements a user-centered approach to the process of ontology learning. CHOL differs from TextToOnto, OntoLT, and Adaptiva as their functions do not support ontology learning modules that serve the information organization process in the information environment. Both TextToOnto and OntoLT are considered to support ontology only in the engineering environment, while Adaptiva is concerned with user-system interaction for ontology building, specifically in the context of knowledge management. CHOL emphasizes not only the interaction between ontology learning and ontology engineering environment, but also the interaction between ontology learning and information systems.

In China, ontology learning has been researched for the past two years. For example, WebOntLearn (Liu, 2005) developed in Zhejiang University is a prototype system for ontology learning from web pages. OntoSphere (Cheng, 2005) is a domain ontology engineering environment, which has been developed by the Institute of Computer Technology, Chinese Academy of Sciences. Its functions include corpus analysis, ontology learning, ontology editing, and ontology mapping. A hybrid approach to extracting domain-specific concepts proposed by Zhang makes use of rules, statistics, and semantic information about texts to identify concepts and introduces main verbs and semantic roles to extract concepts (Zhang, 2005). A collaborative mining approach to building ontology is proposed by Ming (2005). In this approach, domain experts, knowledge engineers, and domain end users work on the Internet cooperatively in order to build ontology. In addition, SOAT, a semi-automatic domain ontology acquisition tool from the Chinese corpus, has been developed in Taiwan.

3 AN ONTOLOGY LEARNING FRAMEWORK FOR INFORMATION ORGANIZATION AND KNOWLEDGE DISCOVERY

First, we propose an ontology learning framework for information organization and knowledge discovery. In this framework, ontology learning is applied to information management environments and knowledge management environments. The data processed or used by ontology learning are divided into two categories: information and knowledge. Information data are managed by information systems, called information layers. In these, knowledge is represented as ontologies that are built and edited in an ontology engineering environment such as

OntoEdit, Protégé, or WebODE, which are called knowledge layers. Ontology learning algorithms are used to semi-automatically extract knowledge and to organize information for ontology-based information systems, called ontology learning layers. Ontology learning is a cyclic process, which includes five stages: text processing, concepts extracting, relations extracting, formal representation, and application and evaluation. Figure 1 shows this process.

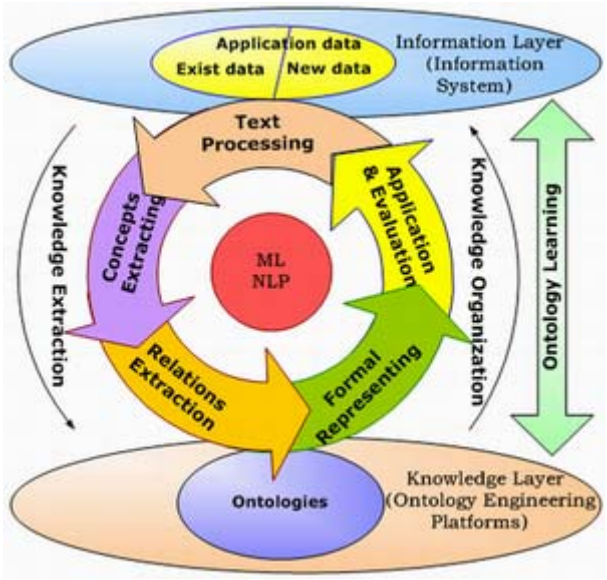

Figure 1. Ontology learning framework for information organization and knowledge discovery

- *Text processing* includes selecting and inputting data sources, such as domain text produced by information systems, discourse structure analyzing, text cleaning, POS tagging, morphological and lexical processing, and chunk parsing that use lexical knowledge bases to produce mixed syntactic/semantic information.

- *Concepts extracting* uses various term extraction and concept discovery methods on the annotated texts for concept acquisition.

- *Relations extracting* uses learning algorithms and background knowledge bases such as discovering association rules for discovering taxonomic and non-taxonomic relations among concepts in the text.

- *Formal representing* uses ontology formal language such as RDF, OWL, Ontolingua, and F-Logic to represent and store extracted ontology, which also can be stored in an RDBMS.

- *Application & evaluation.* The result of the ontology learning algorithm is provided as raw ontology for engineers to refine, update, and edit in ontology engineering environments. The automatic extracted ontology or manually updated ontology will be applied for knowledge organization in information systems. Meanwhile, the quality of the resulting ontology and the validity of the ontology learning algorithm will be evaluated in the application.

With the generation of new application data in information systems or the input of new data sources for learning selected by ontology engineers in ontology engineering environments, the above cyclic process will start again.

4 CHOL

CHOL is a system implementing the above ontology learning framework for Chinese text, which learns domain

ontology from the Chinese domain corpus.

4.1 Architecture

According to the above ontology learning framework, the function modules of CHOL will be applied in three kinds of environments: information environments, the ontology learning environment (CHOL), and knowledge environments. Figure 2 shows the CHOL architecture.

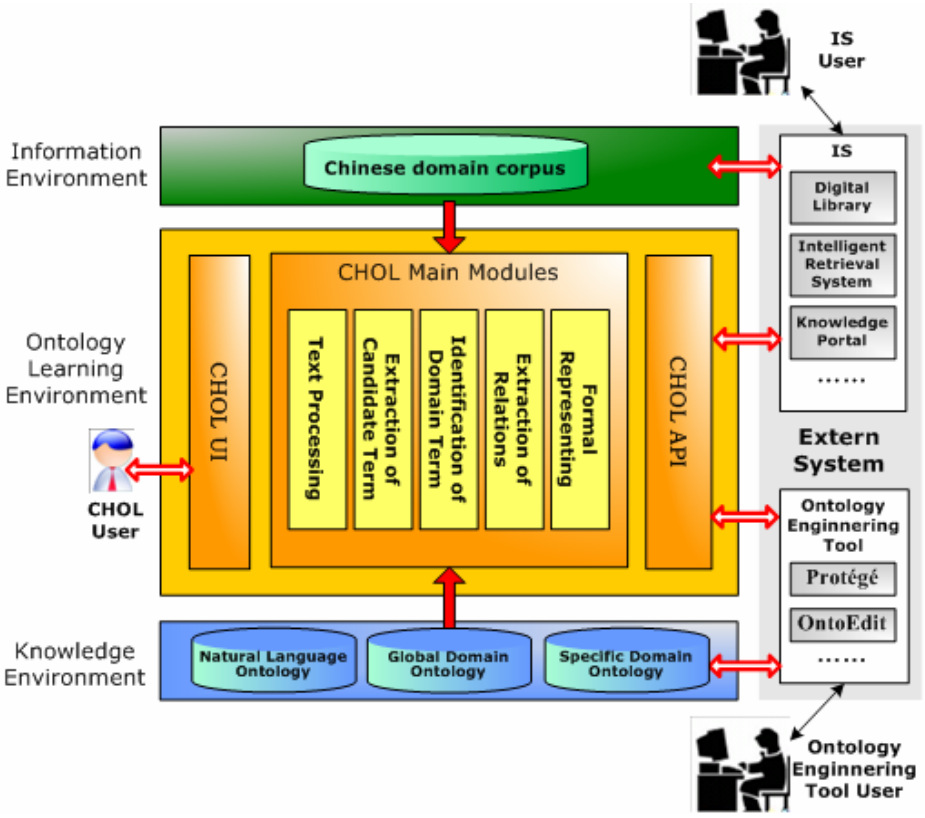

Figure 2. CHOL architecture

(1) The information environment includes different Chinese information databases, especially domain corpus, which are the source of ontology learning. These information databases are collected, organized, retrieved and managed by information systems such as knowledge portals, intelligent retrieval systems, and digital libraries. Given the task of constructing and maintaining domain ontologies for an information organization application in an ontology-based information system, CHOL provides some APIs (Application Programming Interface) to support use of the functions of ontology learning.

(2) The ontology learning environment (CHOL) includes all the processing and methods of ontology learning. It provides a stand-alone UI (User Interface) and some APIs to invoke ontology learning methods in its main function modules. With the stand-alone UI, users can manually start and customize the processing of ontology learning and also manually select the data source to input. Its main function consists of five modules: text processing, extraction of candidate terms, identification of domain terms, extraction of relations, and formal representation. For each function module, APIs are provided to information systems and ontology engineering tools for invoking the methods implemented by CHOL.

(3) The knowledge environment includes three kinds of knowledge: Chinese natural language knowledge, all

domain knowledge, and specific domain knowledge. The knowledge is represented as ontologies, respectively named natural language ontology, global domain ontology, and specific domain ontology. These ontologies accept the results of ontology learning as their initial ontology or proposition of ontology updating and also can be manually defined, updated, maintained, edited, and managed by ontology engineering tools that invoke CHOL APIs to implement semi-automatic ontology building.

In short, CHOL is designed to use as stand-alone system or embedded system of other applications. The data processed by the CHOL system can come from information systems, knowledge management platforms, and the stand-alone ontology learning tool (CHOL). Therefore, CHOL has three kinds of users: CHOL tool users, ontology engineering tool users, and information system users. All of these users can start an ontology learning process implemented by CHOL.

According to the above architecture, the CHOL system consists of two components: five main function modules and initial ontologies concerning Chinese natural language and domain knowledge. The definition and construction of initial ontologies in the CHOL system will be discussed in the next section. Figure 3 shows the data flow of the CHOL architecture and process.

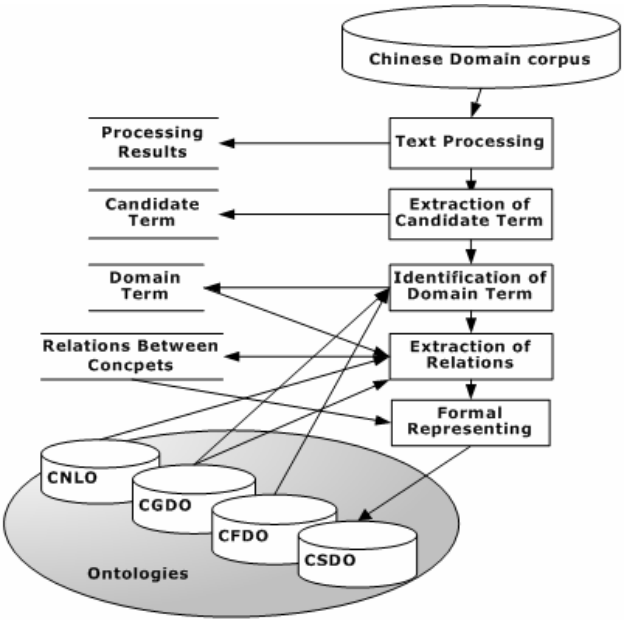

Figure 3. The data flow of the CHOL architecture and process

4.2 Initial ontologies construction

In the CHOL system, we have built some initial ontologies. First, we proposed a dynamic and hierarchical extended domain ontology model for knowledge organization and discovery. This facilitates the reuse of knowledge and represents the complete and concise extensibility of knowledge. This domain ontology has five levels: Natural Language Ontology, Global Domain Ontology, Foundation Domain Ontology, Specific Domain Ontology, and Domain Ontology Instances. Based on this model, we constructed our initial Chinese domain ontology using Hownet and the Chinese Classification Thesaurus for constructing upper domain sub-ontologies.

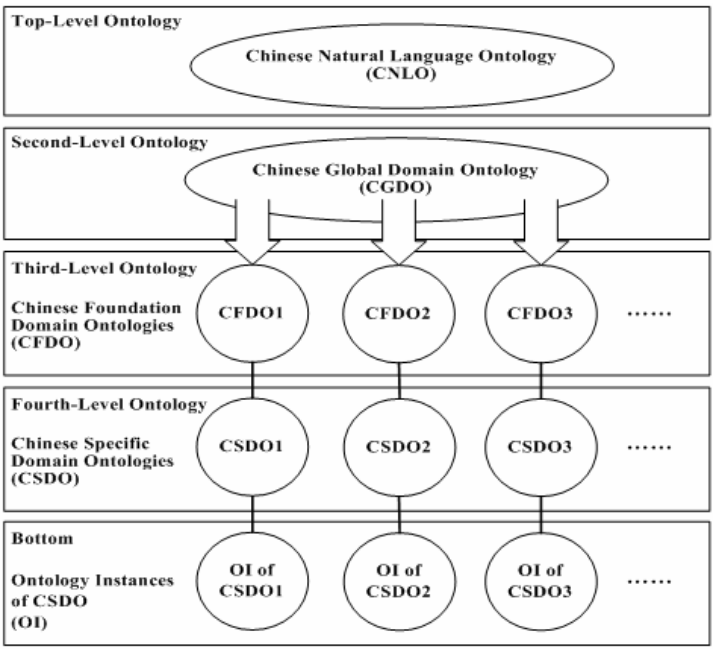

Figure 4. The components of domain ontology in CHOL

Initial ontologies consist of the following 4 levels of ontologies with domain ontology instances at bottom-level:

- The top-level ontology is the Chinese Natural Language Ontology (CNLO). It includes all the basic Chinese lexical words and the lexical relations between Chinese language concepts. It is used for text processing and extracting lower-level ontologies. It contains lexical knowledge of Chinese.

- The second-level ontology is the Chinese Global Domain Ontology (CGDO). It includes concepts of all specific domain and taxonomic relations among concepts. It is used for knowledge completeness and extracting lower-level ontologies.

- The third-level ontology is the Chinese Foundation Domain Ontologies(CFDO). For each specific domain, its foundation ontology is constructed. Each specific domain has some foundational domains. Its foundation ontology includes concepts of its foundational domains.

- The fourth-level ontology is the Chinese Specific Domain Ontologies(CSDO). It includes concepts of one specific domain. It provides a detailed description of the domain concepts from a restricted domain.

- At the bottom-level, there are many ontology instances of various specific domains.

The construction of the above initial ontologies used the following methods:

The Chinese Natural Language Ontology construction maps Hownet into Natural Language Ontology. HowNet is an on-line, common-sense knowledge base unveiling inter-conceptual relationships and inter-attribute relationships of concepts as meaningful in lexicons of the Chinese and their English equivalents (Dong & Dong, 2000). According to our mapping rules, we constructed our initial CNLO with 68,273 Chinese lexical concepts and relations such as synonym, act / result, and hierarchy.

The Chinese Global Domain Ontology maps the *Chinese Classification Thesaurus* into the Global Domain Ontology. According to our mapping rules, we constructed our initial CGDO with 115,142 Chinese terms, 128,747 concepts, and relations such as synonym, generality, and hierarchy

In CFGO construction, each CFGO of CSDO is dynamically constructed from CGDO by selecting the concepts of its foundational domains.

In CSDO construction, the initial CSDO is constructed from CGDO by selecting the concepts of each domain. Using ontology learning methods, the initial CSDO will be semi-automatic updated and enriched by CHOL.

5 OUR APPROACH

5.1 Concepts extraction

Terminology is the set of words or word strings that convey a single, possibly complex, meaning within a given community. In a sense, terminology is the surface appearance, in texts, of the domain concepts in a community. Because of their low ambiguity and high specificity, terms are also particularly useful to conceptualize a knowledge domain or to support the creation of a domain ontology. Therefore, a typical approach in ontology learning from a text first involves the extraction of (more or less complex) terms (concepts) from a domain-specific corpus. In general, some terms of specific domains are not found in initial ontologies, but they can be extracted from domain-related documents using natural language processing and statistical methods, as discussed below.

(1) Concepts extraction steps:

Step 1: Corpus Pre-processing. First of all, the automatic Chinese discourse structure analysis will extract the article title, keywords, paragraphs, and sentences and remove non-Chinese character words from the text.

Step 2: Chinese Words Segmentation with Part-Of-Speech (POS) Tagging. The Maximum Matching Method is used to segment the pre-processed text. Methods are used to recognize complex phrases such as time phrases and quantifier phrases and to remove conjunction phrases.

Step 3: Noun Keyword Extraction and Unknown Word Detection. Statistical methods, filtering rules, and word dictionaries, such as stop word lists, are used to extract existing terms of the initial Chinese Global Domain Ontology (CGDO) and candidate terms which are not in our initial CGDO.

Step 4: Identification of Domain Terms. A novel domain term identification formula is used to filter domain terminology and for each candidate term or existing term, to identify the domains to which the term belongs. If a term belongs to several domains, it will put this term in several domains. This formula will be discussed below.

Step 5: Domain Concepts Generation. A domain concepts extraction method based on term co-occurrence in the same sentence in this special domain corpus is used to generate a domain concepts represented as a weighted term vector.

(2) Domain term identification formula

We used a novel method for identifying a domain terminology, which improves the method proposed by Roberto Navigli for filtering “pure” terminology (Navigli & Velardi, 2004). Our method is based on three measures: $DR_{t,k}$, $DC_{t,k}$, and $DC_{t,k}$. The term weight for filtering non-terminological candidate terms is a

combination of these three measures. For each candidate term the following term weight $TW_{t,k}$ is computed as:

$$TW_{t,k} = \alpha DR_{t,k} + \beta DC_{t,k}^{norm} - \gamma GC_t \quad (1)$$

where $\alpha, \beta, \gamma \in (0,1)$, and if $DC_{t,k}$=0, then $TW_{t,k}$=0.

$DR_{t,k}$ measures the domain relevance of a term *t* in a domain D_k. It is computed as:

$$DR_{t,k} = \frac{P(t \mid D_k)}{\max_{1 \le j \le n} P(t \mid D_j)} \quad (2)$$

where the conditional probabilities $P(t \mid D_k)$ are estimated as:

$$E(P(t \mid D_k)) = \frac{f_{d,t,T_k}}{f_{d,t}} \quad (3)$$

where $f_{d,t}$ is the frequency of documents including the term *t* in the whole training corpus database that contains corpus of different domains. T_k is the known term set of D_k, which is stored in the initial ontologies of CHOL. f_{d,t,T_k} is the number of documents in which this term *t* co-occurs with any term in T_k. Because this method computes the conditional probabilities $P(t \mid D_k)$ based on the co-occurrence of term *t* and known terms in the domain D_k in the training corpus, it decreases the error by directly computing the frequency of term *t* in domain D_k.

$DC_{t,k}$ measures the distributed use of a term *t* in the domain D_k. It is computed as:

$$DC_{t,k} = \sum_{d \in D_k} \left(P_t(d) \log \frac{1}{P_t(d)} \right) \quad (4)$$

where the conditional probabilities $P_t(d_j)$ are estimated as:

$$E(P_t(d_j)) = \frac{f_{t,j}}{\sum_{d_j \in D_k} f_{t,j}} \quad (5)$$

where $f_{t,j}$ is the frequency of term *t* in the domain D_k.

GC_t measures the distributed use of a term *t* in all domains. It is computed as:

$$GC_t = freqD_{t,D} \quad (6)$$

where $freqD_{t,D}$ is the total number of domains that use this term *t*.

The above domain term identification formula has the following distinctive features:

- It can identify various domains of a term because one term can belong to several domains.
- It can partly solve the problem of sparse data when the distribution of different domain corpus in a training set is uneven.

5.2 Relations extraction

Extracted concepts (terms) are statistically processed to determine their relevance for the domain corpus at hand and clustered into groups with the purpose of identifying taxonomy relationships of potential classes. Usually, the acquisition approach of concept relationship has two primary categories: Pattern-based and Machine Learning-based. Neural networks methods are often skipped. Existing ontology learning systems do not use neural networks either. Therefore, we studied a neural networks method in ontology learning and propose a method of relationship extraction based on the SOM algorithm, combining the fuzzy clustering algorithm and domain term identification method in CHOL. This approach includes the following steps:

Input: a newly discovered term *t* & documents in which this term is used.

Output: Relations between term *t* and related terms

Step 1: Extract all existing terms in CGDO and new terms discovered by CHOL from documents. Each document is expressed as a weighted term vector for the SOM algorithm.

Step 2: Use SOM for term clustering and produce clusters of terms.

Step 3: Use the fuzzy clustering algorithm to generate a two-level hierarchy relationship of terms.

Step 4: Use our domain term identification method to identify the domains to which term *t* belongs. If term *t* belongs to different domains, for each domain, a term relations tree is generated.

Step 5: Trim and update these term relations trees using CGDO and CNLO.

6 AN EXPERIMENT IN ETHNOLOGY AND ANTHROPOLOGY

To test the qualitative and quantitative performance of CHOL, we designed an experiment to apply CHOL to ethnology and anthropology for finding and extracting unknown terms and the relations among terms from Chinese texts about minority customs in China. We traditionally used the *Chinese Classification Thesaurus* for information organization. The number of total Chinese minority custom terms in this Thesaurus is no more than 40. It is obvious that using this thesaurus as the knowledge organization system cannot meet the needs of an information organization of Chinese minority customs. Therefore, we used CHOL to semi-automatically extract unknown terms and the relationships for enriching and updating Chinese minority custom knowledge in the *Chinese Classification Thesaurus*.

We had built a Chinese minority database in the past, which has 70,000 data records. Now we built a new

database collecting the corpus of majority domains as the contrasting corpus, which has 80,000 data records. We used all data in both two databases as the training corpus. First, CHOL was applied to the Chinese minority festival field. We extracted unknown concepts such as “雪顿节” (Xuedunjie), “望果节” (Wangguojie), “法会” (Fahui), “三月街” (Sanyuejie), “采花山” (Caihuasan), and “姊妹节” (Zimeijie). A numerical evaluation of the terminology identification led to a precision ranging from 70 percent to about 80 percent and a recall from 27 percent to 83 percent. Figure 5 shows the precision and recall for the terminology identification. We extracted relationships between concepts such as “瑶族” (Yao) and “盘王节” (Panwangjie), “畲族” (She) and “乌饭” (Wufan), and “藏族” (Tibetan) and “转山会” (Zhuanshanhui). A numerical evaluation led to a precision relationship extraction ranging from 20 percent to about 55 percent.

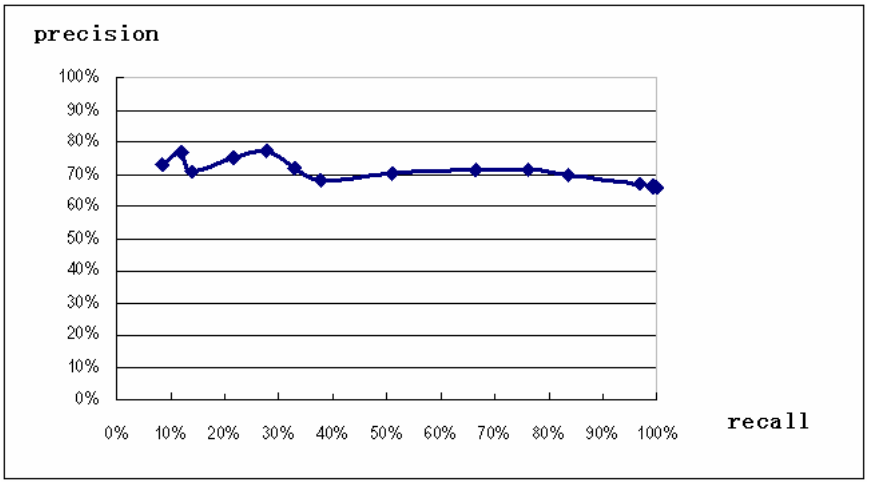

Figure 5. Precision and recall for the terminology identification

7 CONCLUSION

We have developed a prototype system for ontology learning from the Chinese corpus named CHOL. In CHOL, we propose methods to identify terms in a domain and extract taxonomic relationships between them. These methods have been proven to be feasible and effective in the application of information organization and knowledge discovery in ethnology and anthropology. At present, CHOL is a simple prototype system. In the future, we will use more methods, especially, deep semantic analysis, and CHOL will be applied in many different domains and larger datasets.

8 REFERENCES

Bisson, G., Nedellec, C., & Canamero, L. (2000) Designing clustering methods for ontology building - The Mo’K workbench. In Staab, S., Maedche, A., Nedellec, C., & Wiemer-Hastins, P. (eds.), *Proceedings of the First Workshop on Ontology Learning OL*. Berlin.

Brewster, C., Ciravegna, F., & Wilks, Y. (2002) User-Centred Ontology Learning for Knowledge Management. In Andersson, B., et al. (eds.) *NLDB 2002*, pp 203-207. Stockholm. Berlin, Heidelberg: Springer-Verlag.

Buitelaar, P., Olejnik, D., & Sintek, M. (2004) OntoLT: A Protégé plug-in for ontology extraction from text Based on Linguistic Analysis. In Davies, J.et al. (eds.), *ESWS 2004*, pp 31-44. Heraklion, Crete. Berlin,

Heidelberg: Springer-Verlag.

Cheng, Y. (2005) *Research on Ontology-Based Uncertain Knowledge Management*, PhD thesis, Institute of Computer Technology, Chinese Academy of Sciences, Beijing, China

Dong, Z., & Dong, Q. (2000) *HowNet*. Retrieved March 12, 2005 from the World Wide Web: Http://www.keenage.com/zhiwang/e_zhiwang.html.

Faure, D. & Nedellec, C. (1998) A corpus-based conceptual clustering method for verb frames and ontology. In Velardi, P. (eds.), Proceedings of the *LREC Workshop on Adapting lexical and corpus resources to sublanguages and applications*, pp 5-12. Granada, Spain.

Gómez-Pérez, A. & Manzano-Macho, D. (2003) *A Survey of Ontology Learning Methods and Techniques. OntoWeb Deliverable D1.5.* Retrieved July 24, 2005 from the World Wide Web: Http://www.deri.at

Liu, B. & Gao, J. (2005). A Study on Ontology Learning for the Knowledge Grid. *Computer Engineering and Applications 41(20)*, 1-5.

Maedche, A. & Staab, S. (2001) Ontology learning for the Semantic Web [J]. *IEEE Intelligent System 16(2)*, 72-79.

Maedche, A. & Staab, S. (2004) Ontology learning. In Staab, S. & Studer, R. (eds.), *Handbook on Ontologies*, pp 173-189. Berlin, Heidelberg: Springer.

Ming, Z., et al. (2005) A Collaborative-Mining Approach to Building Ontology. *Acta Scientiarum Naturalium Universitatis Sunyatseni 44 (3)*, 15-19.

Navigli, R. & Velardi, P. (2004) Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics 30 (2*), 151-179.

Nedellec, C. (2000) Corpus-Based Learning of Semantic Relations by the ILP System, Asium. In: Cussens, J. & Dzeroski, S. (eds.), *Proceedings of Learning Language in Logic*, pp 259-278, LLL'99, Bled. Berlin, Heidelberg: Springer-Verlag.

Reinberger M.-L., Spyns P., Pretorius A.J., & Daelemans W. (2004) Automatic Initiation of an Ontology. In Meersman, R. & Tari, Z. (eds.), *CoopIS/DOA/ODBASE 2004*, pp 600~617. Agia Napa, Cyprus. Berlin, Heidelberg: Springer-Verlag.

Velardi, P., Fabriani, P., & Missikoff M. (2001) Using Text Processing Techniques to Automatically Enrich a Domain Ontology. In Welty, C. & Smith, B. (eds.), *Proceedings of the International Conference on Formal Ontology in Information Systems* pp 270-284, Ogunquit. New York: ACM Press.

WU, S. & HSU, W. (2002) SOAT: A Semi-Automatic Domain Ontology Acquisition Tool from Chinese Corpus. In *Proceedings of the 19th International Conference on Computational Linguistics*, pp 1313 -1317. Taipei .

Zhan, C. (2005) *A Research on Methods of Knowledge Acquisition from Domain-Specific Texts and Their Application in Knowledge Acquisition from Archaeological Texts*, PhD thesis, Institute of Computer Technology, Chinese Academy of Sciences, Beijing, China