Ontology Learning
for Chinese Information Organization and Knowledge Discovery in Ethnology and Anthropology

Kong Jing

Institute of Ethnology & Anthropology,
Chinese Academy of Social Sciences

Outline

- Introduction
  - Definition of Ontology learning
  - Development of Ontology learning
  - Our research objective
- Ontology learning frame for information organization and knowledge discovery
- CHOL(a Chinese Ontology Learning Tool)
  - Architecture
  - Components
  - Approaches
- Experiment in Ethnology and Anthropology
- Conclusion & Future Work

Definition

- Ontology learning is defined as the set of methods and techniques used for building an ontology from scratch, enriching, or adapting an existing ontology in a semi-automatic fashion using several sources.

(A. Gómez-Pérez, D. Manzano-Macho. A survey of ontology learning methods and Techniques. OntoWeb Deliverable D1.5, 2003,6)

Development

- Recently, there has been a surge of interest in studying on ontology learning. In 2000, the first workshop on ontology learning held in conjunction with the 14th European Conference on Artificial Intelligence (ECAI2000).
- In the past years, many ontology learning tools such as TextToOnto、OntoLearn、 OntoLT、Adaptiva、 the ASIUM system、the Mo'k Workbench、SOAT and DOGMA have been developed.

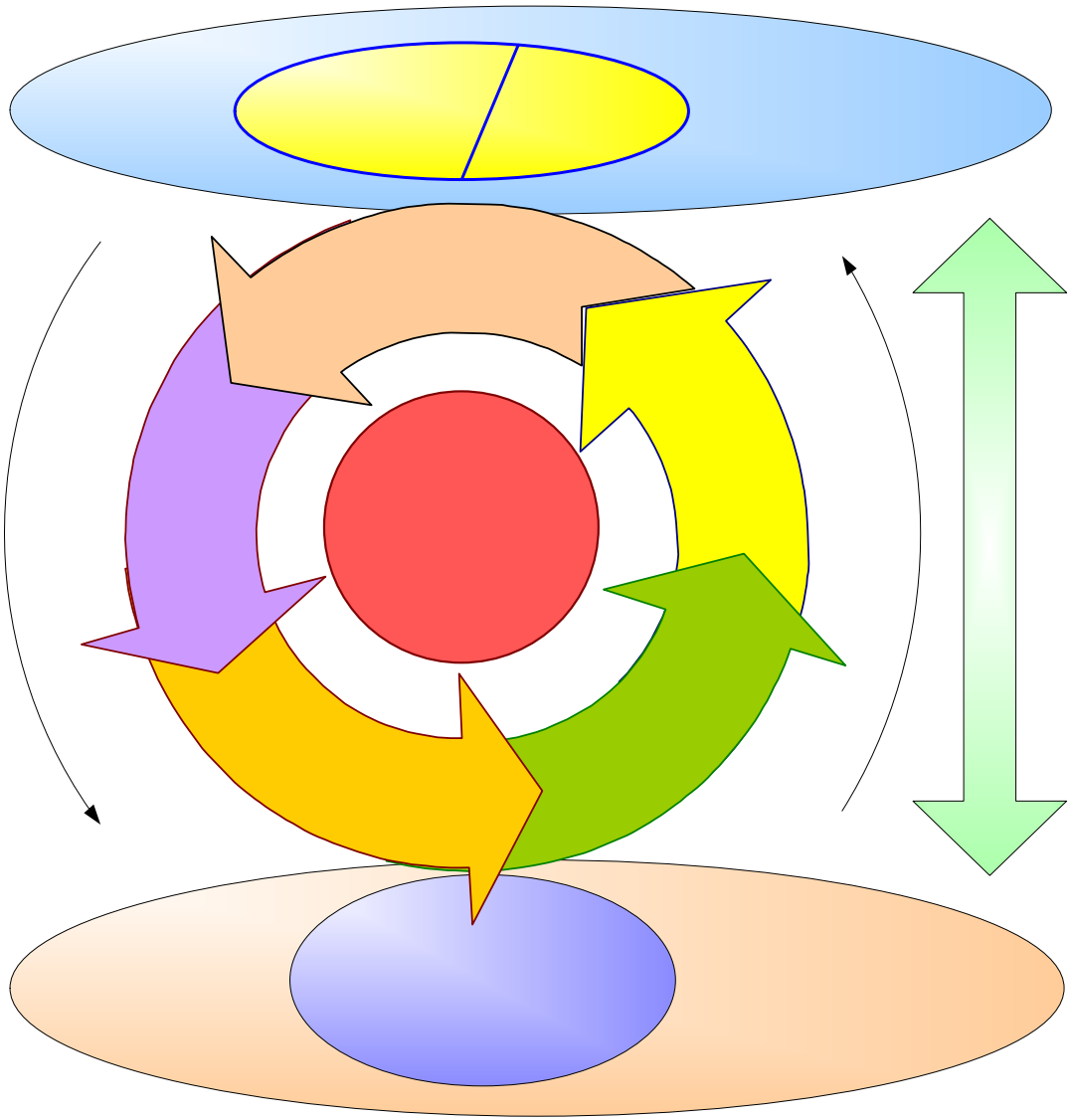
Our research objective

Despite the significant amount of work done on ontology learning in recent years, learning ontology from Chinese text hasn't been widely applied in practice.

So our research objective is to study the application of ontology learning in Chinese information organization and knowledge discovery.

Ontology learning frame

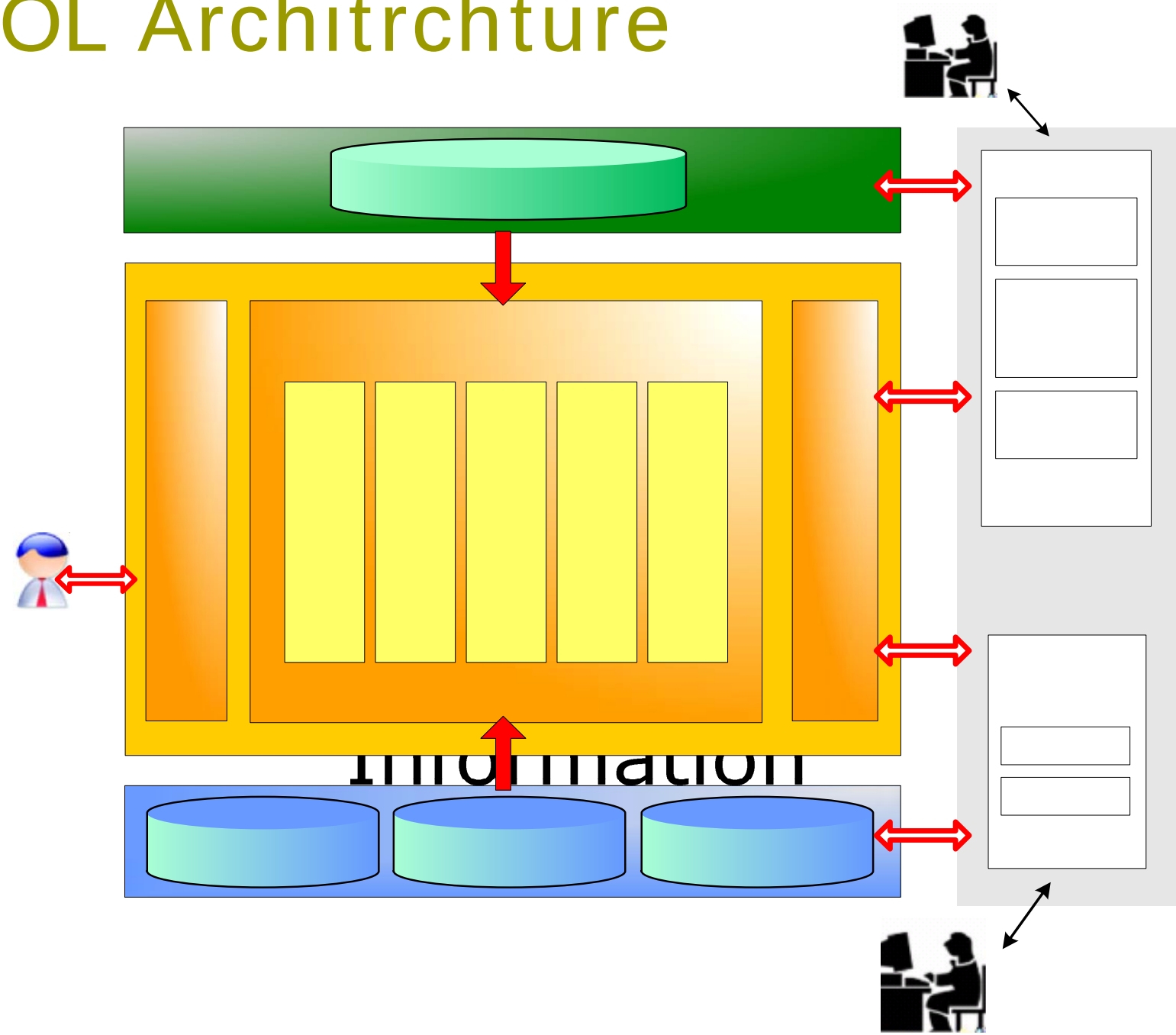
for information organization and knowledge discovery



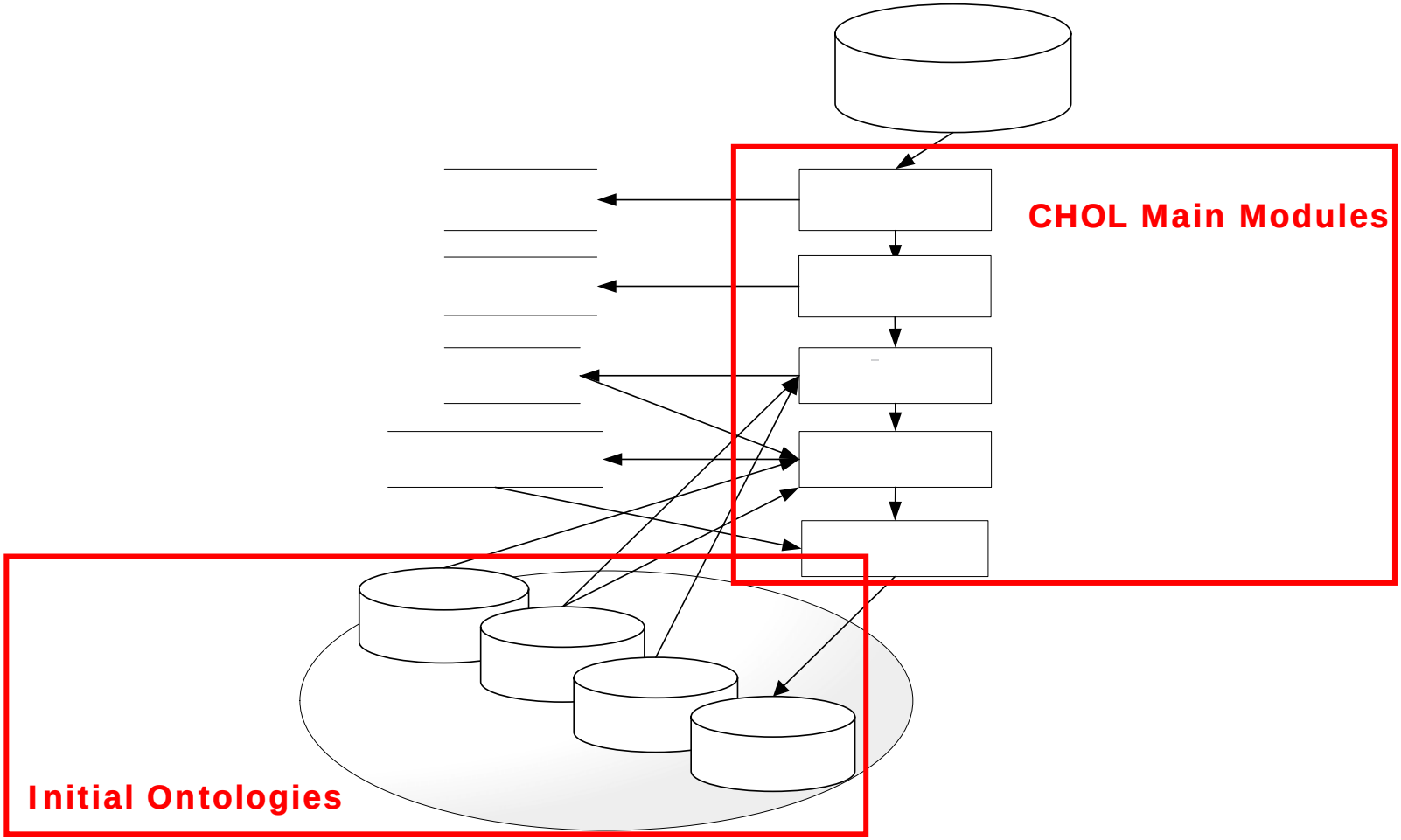
CHOL (a Chinese Ontology Learning Tool)

- Architecture
- Components
- Approaches

CHOL Architrchture



Components of CHOL



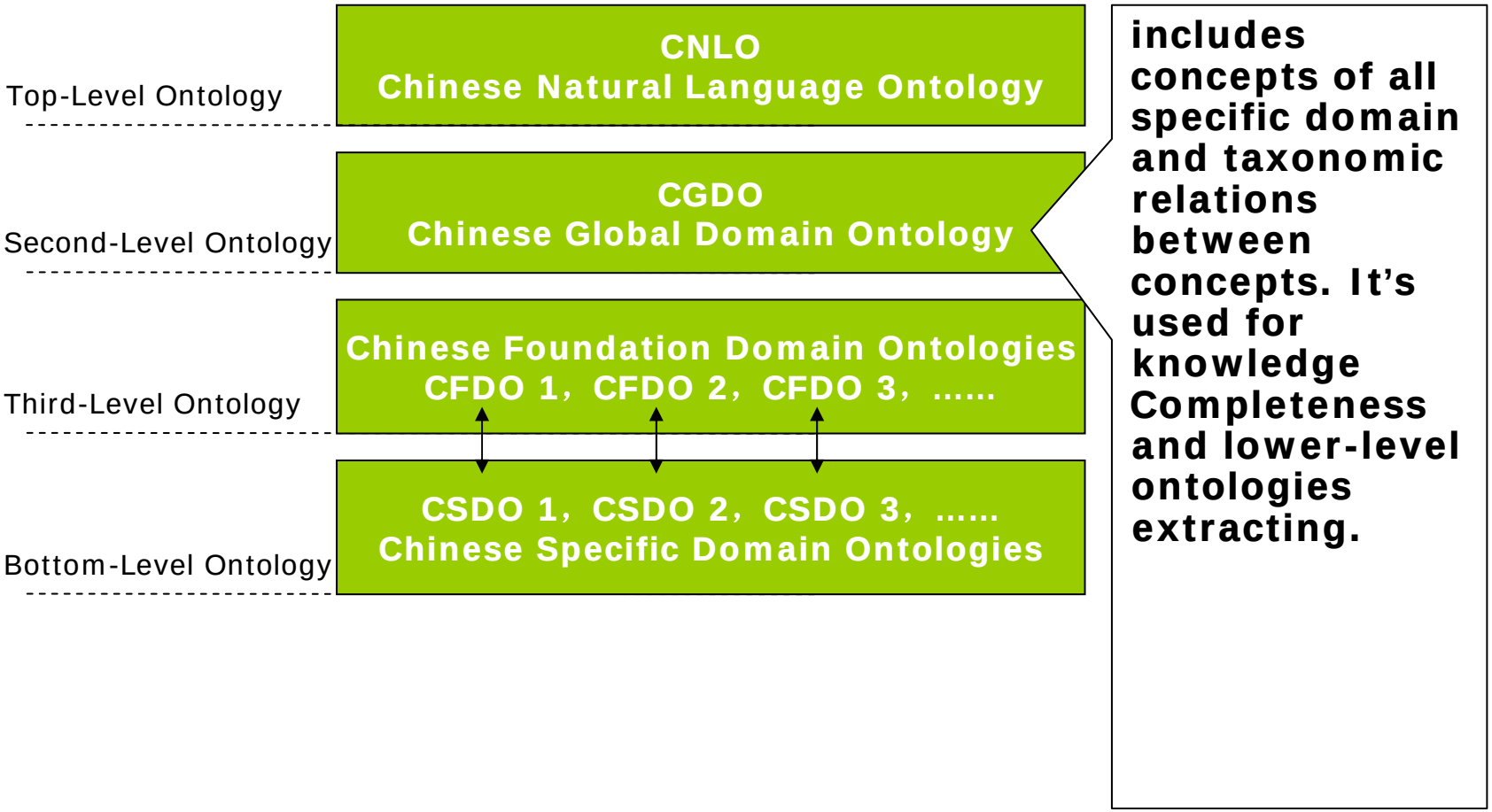
CHOL Main Modules

- Text Processing
- Extraction of Candidate Term
- Identification of Domain Term
- Extraction of Relations
- Formal Representing

Initial Ontologies

includes all the basic Chinese lexical words and the lexical relations between the Chinese-language concepts. It's used for text processing and lower-level ontologies extracting. It contains lexical knowledge of Chinese.

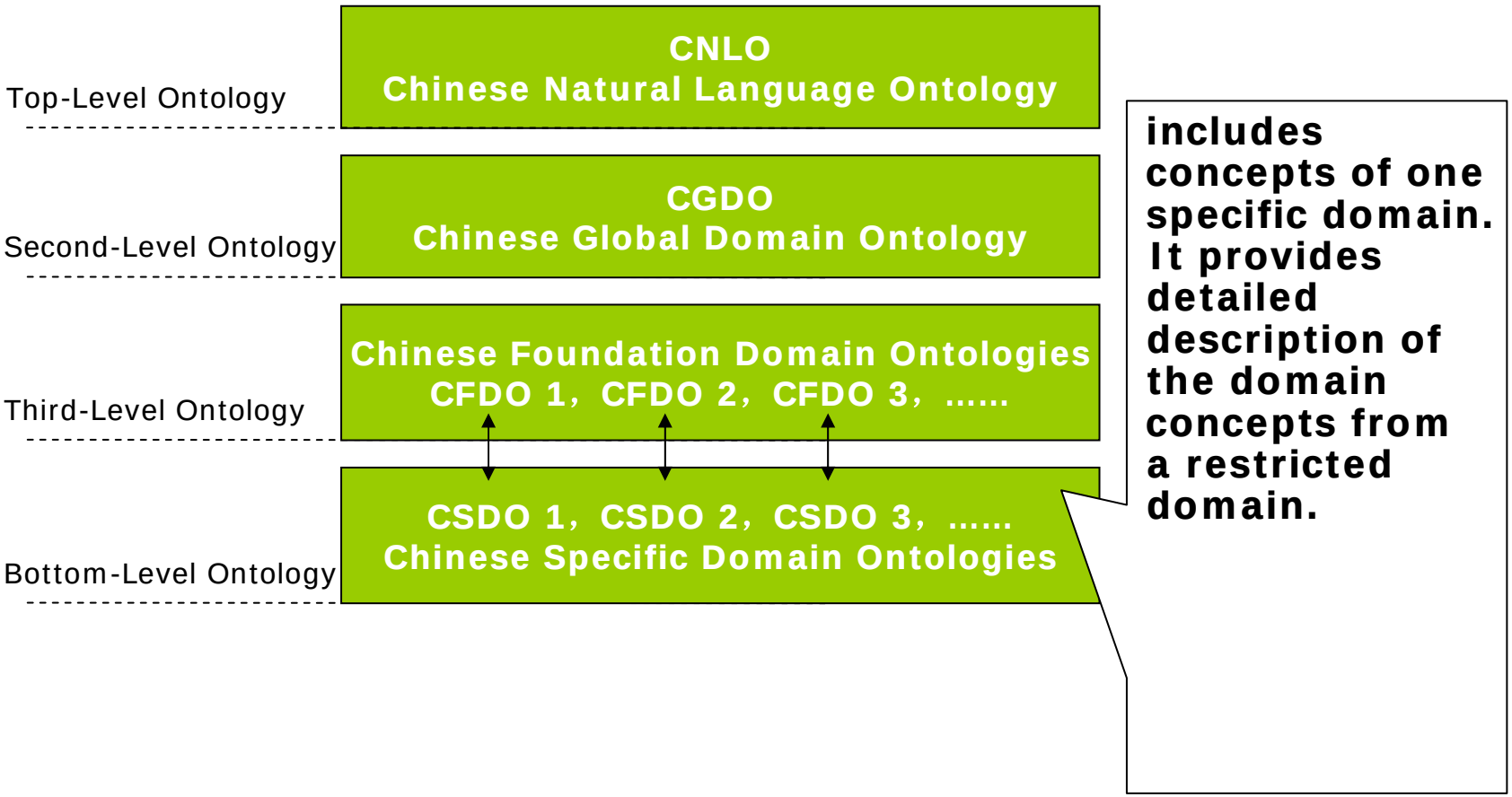
Initial Ontologies



Initial Ontologies

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.

Initial Ontologies



Our approaches

- Initial ontologies Constructing
  - CNLO
  - CGDO
  - CFGO
  - CSDO
- Concepts extraction Method
- Relations extraction Algorithm

CNLO Constructing

- Mapping Hownet into Natural Language Ontology.
- Results
  - Chinese lexical concepts: 68,273
  - Relations
    - Synonym: 60,310
    - Act / result : 7,121

CGDO Constructing

- Mapping *Chinese Classification Thesaurus* into Global Domain Ontology
- Results
  - Chinese Term: 115142
  - Concepts: 128747
  - Relations:
    - Synonym: 19158
    - Generality: 41714
    - Hierarchy: 67830

CFGO & CSGO Constructing

- CFGO Constructing

  Each CFGO of CSDO is dynamically constructed from CGDO by selecting the concepts of it's foundational domains.

- CSDO Constructing

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

Concepts extraction Method

- Domain term identification formula

For each candidate term the following term weight is computed:

$$TW_{t,k} = \alpha DR_{t,k} + \beta DC_{t,k}^{norm} - \gamma GC_t \qquad \alpha, \beta, \gamma \in (0,1)$$

$DR_{t,k}$ measures the domain relevance of a term t in a domain D_k.

$$DR_{t,k} = \frac{P(t \mid D_k)}{\max_{1 \le j \le n} P(t \mid D_j)} \qquad E(P(t \mid D_k)) = \frac{f_{d,t,T_k}}{f_{d,t}}$$

$DC_{t,k}$ measures the distributed use of a term t in a domain D_k.

$$DC_{t,k} = \sum_{d \in D_k} \left(P_t(d) \log \frac{1}{P_t(d)} \right) \qquad E(P_t(d_j)) = \frac{f_{t,j}}{\sum_{d_j \in D_k} f_{t,j}}$$

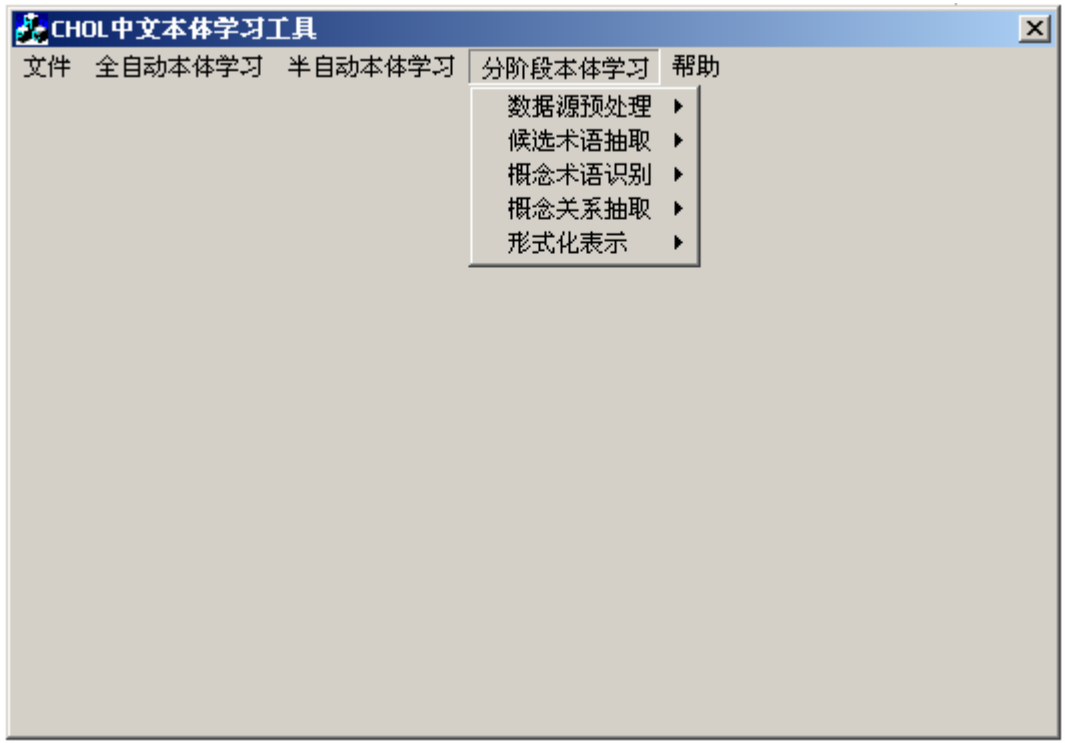
GC_t measures the distributed use of a term t in all domains.

$$GC_t = freqD_{t,D}$$

Relations extraction Algorithm

- Input: a new discovered term *t* & documents in which this term is used.
- Output: Relations between term *t* and related terms
- Step1: Extract all terms in CGDO and new terms discovered by CHOL from documents. Each document is expressed as a weighted keyword vector consisted of all terms for SOM algorithm.
- Step2: Use SOM for term clustering and produce clusters of term.
- Step3: Use the fuzzy clustering algorithm to generate the two level hierarchy relations of terms.
- Step4: Use our domain term identification method to identify the domains to which term *t* belong. If term *t* belong to different domain, for each domain generates a term relations tree.
- Step5: Trim and update these term relations trees using CGDO and CNLO.

Screenshot of CHOL



Experiment in Ethnology and Anthropology

- We have tested CHOL in ethnology and anthropology to find and extract unknown term and the relations between terms from Chinese text about minority custom in China.

Example:

- CHOL applied in Chinese minority festival database.
  - Extracted concepts:

"雪顿节(Xuedunjie)"、"望果节(Wangguojie)"、"法会(Fahui)"、"三月街(Sanyuejie)"、"采花山(Caihuasan)"、"姊妹节(Zimeijie)".....

  - Extracted relations:

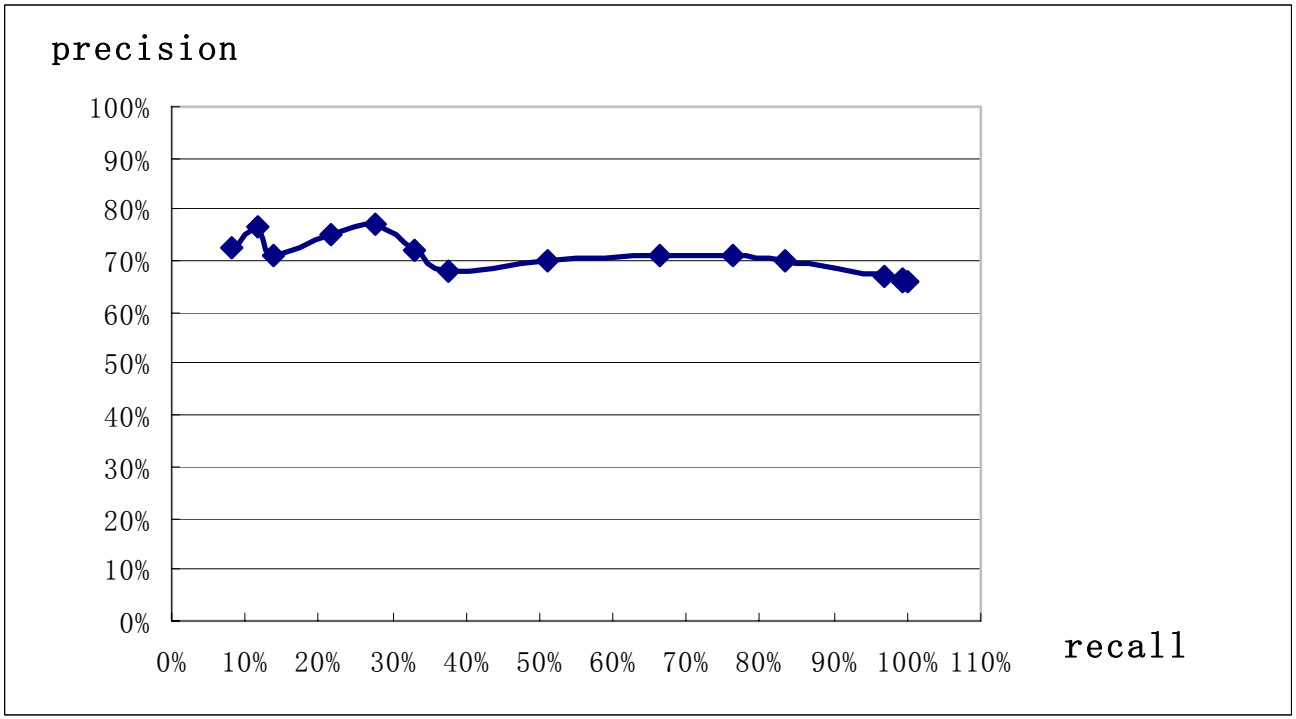
"瑶族(Yao)"-"盘王节(Panwangjie)"

"畲族(She)"-"乌饭(Wufan)"

"藏族(Tibetan)"-"转山会(Zhuanshanhui)"

.....

Precision and recall for the terminology identification



Conclusion & Future Work

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

Thanks

kongjing@cass.org.cn