OntoLTCn: A Chinese Text Oriented Semi-auto Ontology Knowledge Discovery Tool

Ying Jiang[1], Hui Dong[1], Haoyi Xiong[2]
[1]School of Information Management, Wuhan University, Wuhan, 420072, P.R. China
[2]College of Electrical and Electronic Engineering, Huazhong University of Science and Technology, Wuhan 430074, P.R China
jpz6311whu@w3china.cn, lhjdh@126.com, xhyccc@sina.com

Abstract

Ontology (RDF/OWL) plays a foundational role of Semantic Web for knowledge representation. But nowadays there are few Chinese ontology bases available, which hinders the research and development of Chinese Semantic Web applications. This paper introduces an ontology knowledge discovery tool, named OntoLTCn, which supports semi-auto domain ontology acquisition from Chinese corpus. In brief, OntoLTCn is a Protégé plug-in based on OntoLT platform, which integrates Chinese NLP and XML pattern mapping technologies for knowledge discovery. A case study in Chinese e-Government domain is discussed as well, which shows the usability of OntoLTCn for ontology construction from digital archives.

1. Introduction

Semantic Web [1] is an evolving extension of the World Wide Web in which the semantics of information and services on the web is defined, making it possible for the web to understand and satisfy the requests of people and machines to use the web content. In Semantic Web layered architecture proposed by W3C, the layer of ontology (RDF/OWL) is intended to provide a formal description of concepts, terms, and relationships within a given knowledge domain. In a word, ontology is the foundation and prerequisite of Semantic Web.

Many efforts are taken on ontology construction in different domains. For example, the Linking Open Data project [2] is a lead effort to create openly accessible, and interlinked, RDF Data on the Web. The data in question takes the form of RDF Data Sets drawn from a broad collection of data sources. As of April 2008, the ontology base surges to more than 2 billion RDF triples. But Falcons [3] is the only one dataset contributed from China. Even in Falcons, there're few data in Chinese language.

We now lack of Chinese ontology greatly. Chinese researchers have to use foreign ontology for Semantic Web studies, such as ontology aligning, ontology evaluation, and ontology evolution. What's more, there're few Semantic Web applications in Chinese industrial track for the same reason. China seems to fall behind again in the early stage of Semantic Web development.

On the other hand, China has made rapid progress in informationization in recent years. There are abundant legacy data in different domains such as databases, electronic documents, digital archives and etc. Manual extracting knowledge from them to construct Chinese ontology base is a tedious task. We also lack of such automatic or semi-auto tools for Chinese as Text2Onto [4] for English and OntoLT [5] for German.

In this background, this paper presents a Chinese text oriented semi-auto ontology knowledge discovery tool: OntoLTCn. The remainder parts of this paper are structured in this way: Section 2 gives an introduction of OntoLTCn and describes in details the mechanism of ontology knowledge discovery process of OntoLTCn; Section 3 deliveries a case study in Chinese e-Government domain; the conclusion is drawn in Section 4.

2. OntoLTCn

OntoLTCn is based on OntoLT platform, which is a Protégé Plug-in [6] for semi-auto ontology learning from text. OntoLT supports only German, while OntoLTCn extends it with Chinese NLP technology for Chinese text.

978-0-7695-3336-0/08 $25.00 © 2008 IEEE
DOI 10.1109/CSSE.2008.788

Figure 1 illustrates the ontology knowledge discovery process mechanism of OntoLTCn.

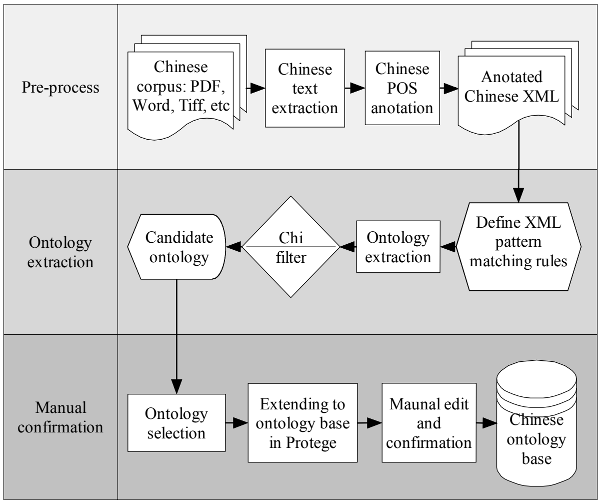


Figure 1. Ontology knowledge discovery process from Chinese corpus with OntoLTCn.

2.1. Pre-process

Towards Chinese text, OntoLTCn aims at ontology knowledge discovery from Chinese corpus. But text based Chinese legacy data sources are of different formats. For example, XML, PDF, Word, even Tiff/Tif pictures are widely used in different domains in China. Extracting Chinese plain text from them is the first step of pre-process.

Table 1. Supported formats of OntoLTCn.

Format	Tool	Type	Description
XML	-	.xml	Supported by OntoLT
PDF	PDFBox *(Java)*	.pdf	Extend the text extraction ability from two Bytes based Chinese PDF documents
Word	Apache POI *(Java)*	.doc	Integrate POI for Chinese text extraction from Word documents
Tiff/Tif	MODI *(C#)*	.tiff/.tif	Adopt JNI to wrap Chinese OCR C# library in Java through Managed C++ Library
Plain text	-	.txt	Directly use plain text for further process

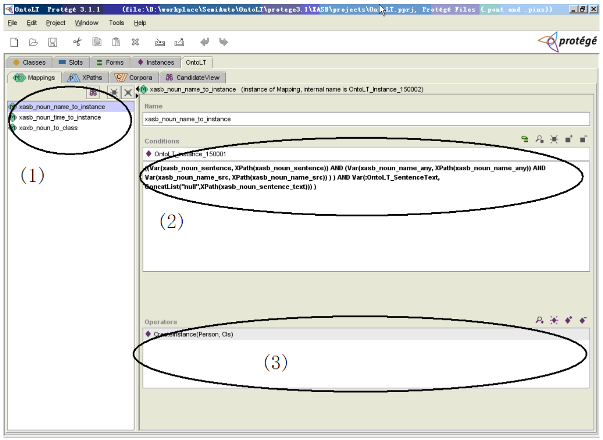
OntoLTCn extends and integrates several existing tools in a unified Java based environment for Chinese plain text extraction, as is showed in Table 1. Note that OntoLT can only accommodate annotated XML as the data source without supporting Chinese。

Not like English, there are no separators between words in Chinese text. OntoLTCn uses ICTCLAS [7] (Chinese Lexical Analysis System, C++ library to be wrapped with JNI for Java environment in OntoLTCn) as a word segmentation parser. Generally, most terminologies and concepts of ontology are nouns or partial nouns. In pre-process step, we also use ICTCLAS as a Part-of-Speech (POS) tagger. For example, "*根据罗干同志的讲话精神，省委组织部于1999年1月下发了《表彰十佳干警通知》*" should be annotated as "*<any type="p"><src>根据</src></any><any type="nh"><src>罗干</src></any><any type="n"><src>同志</src></any><any type="u"><src>的</src></any><any type="n"><src>讲话精神</src></any><any type="w"><src>，</src></any><any type="n"><src>省委</src></any><any type="n"><src>组织部</src></any><any type="p"><src>于</src></any><any type="nt"><src>1999年1月</src></any><any type="v"><src>下发</src></any><any type="u"><src>了</src></any><any type="w"><src>《</src></any><any type="v"><src>表彰</src></any><any type="j"><src>十佳</src></any><any type="n"><src>干警</src></any><any type="n"> <src>通知</src></any><any type="w"><src>》</src></any>*" (*"n"* for noun, *"v"* for verb and *"p"* for preposition). Such annotated XML can be used in the next step of ontology extraction.

2.2. Ontology extraction

The ontology extraction process is implemented as follows. OntoLTCn provides a precondition language with which the user can define XML pattern mapping rules. Preconditions are implemented as XPATH expressions over the Chinese linguistic annotation. If the precondition is satisfied, the mapping rule activates one or more operators that describe in which way the ontology should be extended if a candidate is found.

As is showed in area (2) in Figure 2, the precondition of the example showed as first rule in the list of area (1) is "*((Var(xasb_noun_ sentence, XPath (xasb_noun_sentence)) AND (Var (xasb_noun_name _any, XPath(xasb_noun_name_any)) AND Var(xasb _noun_name_src, XPath(xasb_noun_ name_src)))) AND Var(:OntoLT_SentenceText, ConcatList("null",* *XPath(xasb_noun_sentence_text)))) ". "XPath (xasb_noun_name_any)"* stands for the XML fragments of *"any[@type='nh']"* (*"nh"* marks nouns of person name), such as "*<any type="nh"> <src>* 罗干 *</src></any>*".


OntoLT Protégé 3.1.1 (file:\D:\workplace\SemiAuto\OntoLT\protege3.1\XASB\projects\OntoLT.pprj, Protégé Files (.pont and .pins))
File
Edit
Project
Window
Tools
Help
protégé
Classes
Slots
Forms
Instances
OntoLT
Mappings
XPaths
Corpora
CandidateView
xasb_noun_name_to_instance
xasb_noun_time_to_instance
xaxb_noun_to_class
xasb_noun_name_to_instance (instance of Mapping, internal name is OntoLT_Instance_150002)
Name
xasb_noun_name_to_instance
Conditions
OntoLT_Instance_150001
Operators
CreateInstance(Person, Cls)
(1)
(2)
(3)

```
((Var(xasb_noun_sentence, XPath(xasb_noun_sentence)) AND (Var(xasb_noun_name_any, XPath(xasb_noun_name_any)) AND
Var(xasb_noun_name_src, XPath(xasb_noun_name_src)) ) ) AND Var(:OntoLT_SentenceText,
ConcatList("null",XPath(xasb_noun_sentence_text))) )
```

Figure 2. Define XML pattern mapping rules.

In area (3) of Figure 2, it defines an operation of "CreateInstance(Person, Cls)", which stands for creating a instance of Person class. It includes a sub operation of "FillSlot(Person_Name, Value)", which means transferring the value of XPATH to the property "Person_Name" of this instance. For example, this rule will create a Person instance and adopt the string of "罗干" as the property value "Person_Name" for the XML fragment of "*<any type="nh"><src>* 罗干 *</src> </any>*"

OntoLTCn provides 4 operators to create classes, slots and instances, as is showed in Table 2:

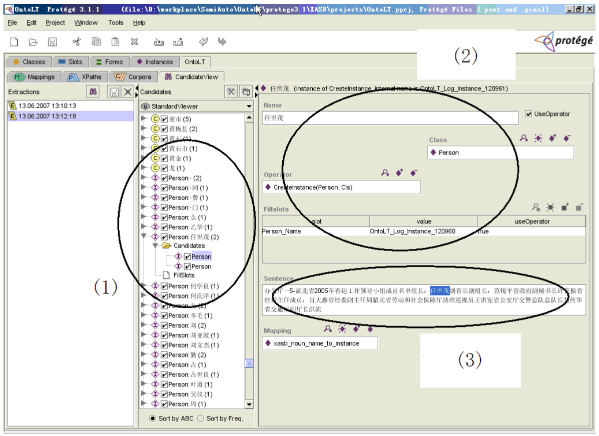
Table 2. Supported operators of OntoLTCn.

Operator	Description
CreateCls	create a new class
AddSlot	add a property to a class or create it if non-existing
CreateInstance	introduce a new instance for an existing or new class
FillSlot	set the property value of an instance

OntoLTCn executes all mapping rules collectively. Therefore, according to which preconditions are satisfied, all corresponding operators will be activated to create a set of candidate classes and slots that are to be validated by the user. According to this interactive process, classes and slots will be automatically generated into a new ontology or integrated into an existing ontology.

In order to use only extracted linguistic information that is relevant for the domain, the approach includes a statistical preprocessing step. Here we base our approach on the use of the "chi-square" function for determining domain relevance. This function computes a relevance score by comparison of frequencies in a domain corpus under consideration with that of frequencies in a reference corpus. In this way, word use in a particular domain is contrasted with that of more general word use. OntoLTCn adopts the People Daily news corpus of January 1998 as the reference corpus (including 28603 Chinese sentences, 58201 Chinese words and covering many different domains in China).

2.3. Manual confirmation



slot	value	useOperator
Person_Name	OntoLT_Log_Instance_120960	true

Figure 3. Manual selection from ontology candidates.

The extracted ontology needs manual confirmation by domain experts, because the automatic extraction may not be perfectly correct due to Chinese OCR, POS annotation and parameters in chi-square process.

As is illustration in Figure3, domain experts are responsible for selection from ontology candidates according the followings:

- Domain experts' experiences whether the candidates are strongly related to the domain.
- The term frequency of each candidate in this corpus in area (1) of Figure 3.
- The mapping rule that triggers the operation of creation in area (2) of Figure 3.
- The context information of each candidate in the data source in area (3) of Figure 3.

After confirmation, domain experts can import the selected ontology candidates into the ontology base. They can further edit the ontology base as usual in Protégé.

3. Case study

OntoLTCn has been successfully applied in a Chinese e-Government project. The goal of this project is that the knowledge hidden in the archives should be represented and exposed in such a proper way that civil servants feel convenient to obtain and use.

In this project, Hubei Provincial Archives Bureau provides a collection of 11887 digital archives (nearly 20 million Chinese characters) in Tiff/Tif format. OntoLTCn is used for ontology base construction from the archives. There are six root classes manually defined by domain experts (civil servants in the bureau): Location, Person, Organization, Document, Spirit, and Event. With the help of OntoLTCn, totally 76 ontology subclasses, 47 ontology properties and 5427 ontology instances are semi-automatically constructed.

Table 3. Effectiveness evaluation of OntoLTCn.

Root class	**Instances**	**Auto extraction proportion**
Location	252	83.33 %
Person	106	100.00 %
Organization	378	67.99 %
Document	1633	93.08 %
Spirit	655	18.31 %
Event	2403	3.45 %
Total/Average	*5427*	*42.30 %*

The effectiveness evaluation results of OntoLTCn is presented in Table 3. OntoLTCn shows different effectiveness for different ontology classes. Ontology classes like Location and Person have high auto extraction proportion, because ICTCLAS annotates them well in POS process. The reason of good effectiveness of Document is that Chinese archive names are always marked between Chinese book markers (" 《" and "》"), and the XPATH mapping rules are easy to be established and of high accuracy. Other classes contain no such particularity, so they require certain manual intervene of domain experts. Especially for Event, there're hardly mapping rules for knowledge discovery with low pattern recognition rate. Its instances are almost manually established.

The performance of OntoLTCn is also tested, and the processing speed is around 231KB/s. The bottle necks lie in the Chinese OCR process and Chinese POS annotation. ICTCLAS claims its annotation speed is 500KB/s, so OntoLTCn can not exceed that. In fact, we adopt a parallel approach: one computer server only perform OCR with another for annotation and a third for ontology extraction. In this way, we successfully constructed the ontology base from the corpus of 11887 digital archives within a week. More servers can be allocated to perform OCR or annotation for better performance of larger corpus.

4. Conclusion

This paper presents a Chinese text oriented semi-auto ontology knowledge discovery tool: OntoLTCn. It integrates Chinese NLP and XML pattern mapping technologies for knowledge discovery. A case study in Chinese e-Government shows that OntoLTCn is an effective and efficient tool for semi-auto Chinese ontology base construction.

5. Acknowledgement

This work is supported by NSFC (National Nature Science Foundation of China) project (70773087). It is supported by the project (2006-x-29) sponsored by the State Archives Administration of China as well.

6. References

[1] Grigoris Antoniou and Frank van Harmelen, *A Semantic Web Primer*, MIT Press. Cambridge, MA. 2004.

[2] Linking Open Data, http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[3] Honghan Wu, Gong Cheng, and Yuzhong Qu, "Falcon-S: An Ontology-Based Approach to Searching Objects and Images in the Soccer Domain", in proceedings of International Semantic Web Conference (ISWC 2006), 2006.

[4] P Cimiano, J Völker, "A framework for ontology learning and data-driven change discovery", in proceedings of 10th International Conference on Applications of Natural Language to Information Systems (NLDB 2005), 2005.

[5] Buitelaar P, Olejnik D, Sintek M. OntoLT, "a Protégé plug-in for ontology extraction from text based on linguistic analysis", in proceedings of the 1st European Semantic Web Symposium, 2004.

[6] Protégé plug-ins & applications, http:// protege.stanford .edu/doc/pdk/plugins/overview.html

[7] HP Zhang, HK Yu, DY Xiong, Q Liu, "HHMM-based Chinese Lexical Analyzer ICTCLAS", in proceedings of Second SIGHAN Workshop on Chinese Language, 2003