Towards Learning Domain Ontology from Legacy Documents

Yijian Wu, Shaolei Zhang, Wenyun Zhao
School of Computer Science and Technology,
Fudan University,
Shanghai, China
{wuyijian, 072021121, wyzhao}@fudan.edu.cn

***Abstract*—Learning ontology from text is a challenge in knowledge engineering research and practice. Learning relations between concepts is even more difficult work. However, when considering only a particular domain in which the concept hierarchy and relations can be modeled manually within an acceptable period of time, the learning process may be simplified. We focus on learning composite concepts and building up a knowledge base from existing documents. Our approach tries to make the machine understand the documents sentence by sentence and finally fit the knowledge conveyed by the document in our pre-defined ontology. Basic semantic units are defined for reasoning with higher-level concepts, including classes and instances. An agricultural case study on learning instances from plant disease descriptions is presented with a web-based ontology learning tool.**

Keywords-ontology learning; agriculture; domain ontology

I. INTRODUCTION

Ontology provides a sound semantic ground of machine-understandable description of digital content [1]. Domain ontology enables automatic document processing within specific domains. Carefully developed domain ontology provides a consistent and exchangeable way to express domain knowledge. But developing ontological concepts and relations is still labor-intensive and time-consuming. Conventionally, domain experts should be participating in ontology development with knowledge engineers throughout the entire building process. Due to the scale and complexity of domain knowledge, the process of ontology building would last months, even years. But the fact is that working with domain experts, especially for a long period of time, is expensive and sometimes not quite possible.

Another problem is that human experts tend to make mistakes, even if they are available. Even the most experienced human expert may miss some important concepts or make ambiguous assumptions, which may lead to inconsistency in the ontology under development. To avoid potential errors and possible misunderstandings, interaction between knowledge engineers and domain experts should be very intensive. Long-lasting cooperation is helpful to fix errors, but as is mentioned, it is not always possible.

Additionally, human experts have to learn how do express concepts, relations and rules using ontology modeling language, which could be a tough task for certain elder experts. Knowledge engineers are responsible for explaining how knowledge is represented formally and help domain experts to express knowledge explicitly and unambiguously.

While human experts play an important role in ontology development, researchers and practitioners seek automatic and more efficient ways to construct domain ontology. Ontology learning provides a potentially efficient way for constructing domain ontology. As Zhou stated in [1], there are different levels of ontology learning. At a high level, ontological concepts, relations and axioms are to be recognized and learned, while instances of concepts and relations can also be extracted from text-based documents at a lower level. Whether modeling pure highly-abstracted information or enriching low-level knowledge base elements is an application-oriented decision to make.

In our research, the emphasis is put on learning low-level knowledge elements to build up a knowledge base for agricultural production. To our experiences, very detailed information (such as details of plant diseases) may not come completely from any human expert without referring to any existing documents. When building up a knowledge base for plant diseases, we tend to refer to an encyclopedia or some authoritative documents (because they do exists), rather than to any single human expert.

In order to make these documents machine-understandable, we employ a fine-grained approach which simulates human cognition process.

The rest of the paper is organized as follows: Section 2 lists some related work on building domain ontology. Section 3 states the assumptions that we make in our study. The assumptions specify under what circumstances the proposed method is applicable. Section 4 describes the methodology and detailed algorithms. Section 5 illustrates the methodology with a featured example within agricultural domain. The case study is mainly based on Chinese materials and environments. A web-based tool for ontology learning is also introduced. Section 6 concludes our approach and gives a brief description of future work.

II. Related Work

Although many researches still focus on building ontology manually [5], ontology learning still promises an efficient way to construct ontology automatically or (at least) semi-automatically. It leads to an automatic way to discover and create ontological knowledge from vast existing documents [1].

The learning objectives and granularities differ in various approaches. A classification can be found in [1]. Some research focuses on discovering concepts, their semantic definitions and the relations between concepts. The goal is to construct the basic concept hierarchy of the ontology. Guizzardi showed several ways to do it [6]. But from an engineering perspective, automatically developing a generic ontology is only too costly and risky. Thus some researchers turned to learning detailed contents or instances in the ontology and constraint the ontology in a specific domain [4][7][11], instead of learning the structure of general-purpose ontology.

Both approaches need a core ontology for incremental learning [12]. When learning instances, the core ontology acts as a basic terminology. Compared to learning concept hierarchy, learning instances has more precision and better coverage to the original text documents. The seed concepts proposed in [11] can be analogous to our basic semantic units, but our approach and supporting tool provide more support on continuous learning and evolution. Although some researchers argue that ontology should stay on a high abstract level, knowledge engineering practices show that building an ontology into a knowledgebase is feasible and reasonable [1].

Research shows that the basic structure of ontology can be carefully designed to be error-free [3]. Based on this basic structure, it is plenty of instances that make the ontology become useful in real applications. Instances in ontology are intensively researched. Ding reported that most RDF documents consist of only instances [3]. Tao showed that potential inconsistency problems exist among instance data in ontology and proposed an approach to detect them [2]. However, how all these instances can be efficiently created and how they evolve are not considered. There are also many other researches focusing on ontology instances, such as [9], showing that enriching instances in domain ontology is worth.

Work on ontology learning is language-specific. We found some work based on Chinese materials. In [8], Hu proposed a domain ontology-based collaborative annotation system, using Chinese NLP tools and techniques to deal with Chinese news texts. The work is similar to ours in that Chinese NLP tools are used for text understanding, but the difference is that our goal is to learn the content of text and build the underlying knowledge into ontology. Similar understanding-and-summery work can be found in [10], which was also Chinese-oriented work.

III. Problem Statement

We are trying to build up a domain-specific ontology-based knowledge base from legacy documents (e.g., encyclopedia texts). In our experiment, we deal with Chinese documents. However, the whole methodology should be applicable in dealing with any documents in other languages, despite of some linguistic differences. The content of the knowledge base is to be automatically extracted from text documents. To clarify the method's applicability, we state our assumption in domain knowledge in Subsection A and then discuss the main learning idea in Subsection B and C.

A. An Assumption on the Stability of Domain Knowledge

Philosophically, domain knowledge is unique, although there are various ways to express it. Ontology presents domain knowledge. The ontology description language is a way that human use to make ontology visible. Thus, when we express some domain knowledge, we actually choose a subjective way to express it. The expression could be good or bad, right or wrong, suitable or not suitable for a particular purpose. Nevertheless, the expression of the real domain knowledge is bounded. The complexity of establishing such an expression is not overwhelming, but can be afforded by human labor. From this point of view, if human are able to construct basic hierarchy of ontological concepts and relations, it would be less important to learn structure of ontology than to learn instances from given texts. In other words, if we intend to build up a knowledge base, we may assume that the basic structure of the ontology is ready and the learning efforts focus on finding relations between concepts or instances.

Although human experts may propose different ways to express the nature of domain knowledge, the structure of domain ontology is comparatively stable once it is established. Take agriculture domain for example. The concept hierarchy is not changing much when a document about Blight disease is processed and understood by the machine. The disease Blight can be modeled as either an instance or a class, but in either case, the upper-level concept hierarchy and relations remain untouched. Thus, when learning ontology from the existing text, an instance about the disease Blight is created and the corresponding relations are instantiated (i.e., the attributes are assigned values), but the class Disease and the related concept hierarchy do not have to change. The basic structures are stable and we just add contents to the basic structures and build up our knowledge base. The basic structures play a core-ontology role in our approach.

B. Basis for learning from natural language texts

The core ontology, similar to upper ontology, represents basic understandings of the machine to the domain under consideration. This is similar to how a human understands a piece of text. He must have a basic understanding of the concepts (i.e., the ontology), and then he must have knowledge about the words (i.e., the vocabulary). After that,

he may get to understand the piece of text by mapping the vocabulary to the ontology. The ontology is comparatively stable and the understanding process is mainly to create the connections between the vocabulary and the ontology. In order to achieve this, the prerequisites include the following obviously: 1) we have to create a domain vocabulary, and 2) we have to build up core ontology for understanding the words in the vocabulary. We spend some efforts on these two prerequisites, which are briefly introduced in the following sub-sections.

1) The vocabulary

There are several general-purpose libraries of words available on the Internet. But when we try to deal with agricultural texts, we found that it is almost impossible to get precise result, especially when we deal with Chinese texts. Thus we employ FAO agricultural vocabulary AGROVOC which is multilingual and structural. Another source of our vocabulary is sample web-pages. We selected a subdirectory of several agricultural websites (in Chinese) and analyze the correlation among neighbor words. We built up a model for Chinese to find combined words and selectively add these combined words into our vocabulary. An example of finding combined words in Chinese is illustrated in section 2.3.

2) The ontology

The initial version of the ontology includes only higher-level concepts and relationships between them (object properties of concepts). The structure describes concept hierarchy of the domain, but the meaning of each meaningful word is not included in the structure yet. It is not possible to add all words to the ontology and to create a full-fledged relation network among them at once. A feasible way is that we build up the ontology with core words incrementally.

In early versions of the ontology, a meaning is given to each of the essential words. The essential words are collected from our sample web-pages in particular web-sites. The sample web-pages are selected within the domain (i.e., the agricultural domain in our case) and the websites are representative. Thus, the number of essential words is quite limited in early development. But to some extent, they will work. It is similar to the situation that a child knowing only a few words will not perfectly understand all articles in newspapers, but he/she does understand some articles written in simple words or part of the articles. If he/she wants to understand more, he/she will have to learn more words (the meaning of the words). So does the machine. The original texts are integrated into the ontology.

Once new knowledge is added to the ontology, the original texts should be re-processed for further understandings. The process will be described in detail in the next section.

C. Finding Combined Words and Basic Semantic Units

In Latin-style languages, words are split naturally. But in character-based languages, such as Chinese language, words may be created freely by combining different characters. This is one of the reasons that the vocabulary may be incomplete in certain domains. Before we begin to find semantic units in Chinese text, we would like to build up a more domain-specific vocabulary (a thesaurus) by analyzing domain-specific documents.

A basic way is to adopt an existing thesaurus. We found an ontology and vocabulary from FAO. But there could be some other words that are not covered in FAO vocabulary. Part of the reason is that Chinese words are comparatively flexible. That is, new words may be easily created without affecting human understandability. For example, in agriculture domain, the Chinese word *Ruanfu*(软腐, soft and rotten) is not collected in FAO word list. But human may literally understand the word by splitting the word into *Ruan*(软, soft) and *Fu*(腐, rotten) and make a sense of the whole word. This *is* a word, for it appears many times in the given documents (when describing the appearances of some plant diseases). The characters *Ruan*(软) and *Fu*(腐) are words listed in generic vocabulary, but the combined one is not. Each of the newly-found words typically has a particular meaning, which should be recorded in the ontology. Our goal is to find the combined words (*Ruanfu* in this example) and add them into the vocabulary, and finally into the ontology.

There are several ways to find the combined words by computing the interrelations between the neighboring Chinese characters. We tried several models to find these "hidden" Chinese words, including using Lucene platform in language processing. In our case study, we first added over 100 newly discovered words for a preliminary experiment. The detailed approach is beyond the scope of this paper.

IV. Understanding Existing Text Documents

A. A Semantic Overview

If ontology is used for understanding articles, the semantic should be presented by the concepts mapped from the ontology to the text. But pure lexical-level mapping is not accurate and may suffer from ambiguity. For example, we have an ontology consisting concepts Apple and AppleComputer. The concept Apple is of type Plant and AppleComputer is of type Laptop. In an article where the word Apple appears a lot of times, how do we know whether the article is about agriculture or computer? The answer is obvious. We read the article, understand other words in it, and finally decide what the word "Apple" means within the whole article.

We believe that the machine should follow the same pattern that human understands texts. We first "teach" the machine which words it should be sensitive to by adding them in the ontology. In our example, the word Apple is a meaningful word, and other words such as Computer, Laptop, Agriculture, Plants, etc. are all meaningful words. They form up at least two concept hierarchies: one is Agriculture based, and the other is (possibly) Computer based. If mapped concepts are all within the agriculture concept hierarchy, the word Apple is more likely to have the agricultural meaning. Note that even if all other concepts fall in the agriculture concept hierarchy, the word Apple does not definitely

represent an agricultural concept. More precise mathematical models can be established for a quantitative analysis, which is not concerned in this paper.

Our goal of understanding texts is to add new concepts (typically, instances) to the ontology. The texts describe facts within the corresponding domain. In agricultural domain, texts that describe plant diseases contain knowledge about symptoms, causes and cures of plant diseases. The texts are written in natural language (Chinese in particular). They are pure MSWord documents without any explicit structures. The contents describe valuable information about many diseases, but the descriptions are difficult to be integrated directly in the ontology for automatic process. This triggered our attempt to make the machine understand the content of these materials, elicit the meanings automatically and build up a knowledge base from these text descriptions.

In order to make the text understandable for the machine, we first prepared a basic ontology including main classes and the corresponding relations. This is the base concept hierarchy. Then we prepared a small set of basic semantic units for the machine to "understand" the text. They act as the small, meaningful words that should be understood by the machine. Finally the words in the text are analyzed against basic semantic units, mapped to concepts in the ontology, and the mapped concepts are inferred to find implicitly associated classes, to create appropriate instances, and to assign values to properties.

B. The Architecture of the Learning Tool

The ontology learning tool consists of an NLP component, an ontology maintenance component and an inference component.

The NLP component is responsible for extracting word from the text. In a Chinese-language circumstance, extracting word needs a particular thesaurus (besides a standard vocabulary). The reason is that, in Chinese, creating a word is sometimes as simple as combining two or three Chinese characters. Such "newly-created" words may be common in a specific domain (e.g., agriculture), but not common at all in everyday life.

In the domain-specific thesaurus, some important words are installed in the ontology by an ontology maintenance component. This process is manual and may lasts for a long time. The more words are integrated in the ontology, the more complete understanding the machine will achieve.

After the NLP component extracts the words (usually the position-of-speech (POS) are identified as well) and the words are mapped to the ontology, an inference procedure takes place. In the inference procedure, associated classes and instances (or expected instances, if not present) are identified. The text is then associated with a corresponding class-instance pair in the ontology. The class-instance pair is the result of understanding of the text. The result may be refined or automatically modified by adding or modifying basic semantic units ("important words") via the ontology maintenance component.

The whole approach is detailed in the next section.

C. Detailed Learning Process

In order to obtain semantic information about symptoms of plant diseases from text descriptions, we take the following steps to deal with the given text.

Step 1. Create basic semantic units in the ontology.

This is a step that we specify which words are meaningful in the text under consideration. As is discussed in the previous sections, we have to "teach" the machine what are meaningful. Traditionally the relationship between words can be referred to in semantic web tools such as WordNet. This is another consideration about how machines associate other words that are not explicitly listed in the ontology. What we care about now is which words the machine should be sensitive to. Take plant disease for example. A plant disease is described by the following three aspects: a) basic description of the symptom, such as mildew (*Mei*, 霉), speckle (*Bandian*, 斑点), rotten (*Fulan*, 腐烂), wither (*Kuwei*, 枯萎), etc.; b) colorful appearance, such as gray (*Hui*, 灰), white (*Bai*, 白), taupe (*Huihe*, 灰褐), etc.; c) on which part of plant, such as leaf (*Ye*, 叶[1]), stem (*Jing*, 茎), root (*Gen*, 根), etc. These very simple words (Chinese characters and words) are the basic semantic units that the system should recognize. So in this step we just "teach" the machine to understand the meaning of these basic words by adding them into the ontology. There could be words that may literally match more than one basic semantic unit. In this case, we just keep the match and deal with this the latter steps.

Step 2. Create semantic mappings.

In this step we turn to natural language processing (NLP) tools for text processing. NLP tools are language-specific. We tried ICTCLAS proposed by Institute of Computing Technology, Chinese Academy of Science. ICTCLAS provides word segmentation, part-of-speech tagging and unknown words recognition. In order to process agricultural document more precisely, we added terms from AGROVOC and words learned from sample texts to ICTCLAS vocabulary. Once the Chinese characters and words are literally matched with the basic semantic units, a semantic mapping can be created.

Note that not all words in the vocabulary contribute to semantic mapping. In fact, the mapping rate (number of words mapped over number of words processed) relies on the size of the ontology and the mapping algorithm. Nevertheless, given an ontology in Step 1, we get a mapping from text (words) to ontology (concepts). If more words are

[1] The word "leaf" in English can be presented in Chinese in several words, such as *Ye*(叶), *Yezi*(叶子), *Yepian*(叶片), etc. All these words are semantically equivalent. How to express the equivalency is not considered in this paper. But the single character *Ye*(叶) can be used in all other cases, which is an acceptable approximation. Other words we take in our ontology follow the same pattern.

"taught" to the machine (i.e., added to the ontology), the mapping rate will rise correspondingly.

Step 3. Make semantic reasoning.

In Step 1, we add basic semantic units which correspond to simplest meaningful words (or Chinese characters). In Step 2, we have the mappings between words and basic semantic units. In this step, an integrated reasoning facility is introduced to create machine understandable results. The reasoning is based on the relationships between concepts (typically, classes). A sentence *snt* is mapped to *n* basic semantic units su_i, where i is positive integer and $i \le n$. Every su_i is an instance in the ontology. Csu_i is a base class of instance su_i. Now we have the basic semantic units for sentence *snt*, and we will create corresponding instance(s) to present the meaning of the whole sentence.

Let the reasoning depth be 1. Then the sentence *snt* corresponds to a semantic unit: instance *I* of class *C*, where *I* and *C* satisfy the following: For each Csu_i, there exists an object property *a*, such that the domain of *a* is class *C*, the range of *a* is Csu_i, and instance *I* has property *a* whose value is su_i. The C-I-tuple is defined as the following.

$\{(C, I) \mid I \text{ instance\_of } C,$
$\forall Csu_i \exists a \in \text{ObjectProperty}, a.domain = C \wedge$
$a.range = Csu_i \wedge I.a = su_i\}$

If a word *w* has more than one meaning, i.e., there exists two semantic units su_j, su_k ($su_j \neq su_k$) in the ontology and both are mapped by word *w*. In this case, we have to decide which meaning the word *w* really means in the given sentence *snt*. Therefore su_j and su_k should be treated separately. Let the set of basic semantic units be *SU*. We consider (*SU*-{su_k}) and (*SU*-{su_j}) separately. For each set of semantic units, we get a set of C-I-tuples (CI_1 and CI_2).

It is also possible that there is no such a class and/or instance that satisfy all conditions above. Thus, we have to relax the constraints by taking a subset of the set $\{Csu_i\}$. Given a set of classes of basic semantic units $\{Csu_i\}$, the C-I-tuple is defined as the following.

$\{(C, I) \mid I \text{ instance\_of } C, \exists SC \subseteq \{Csu_i\}\ \forall SC_i \in SC$
$\exists a \in \text{ObjectProperty}, a.domain = C \wedge a.range = SC_i \wedge$
$I.a = sc_i \wedge sc_i \text{ instance\_of } SC_i\}$

There could be another problem that the instance *I* is not present at the time. Actually this is a normal situation when we are constructing the ontology as a knowledge base. The process of building up the knowledge base is to create instances and to instantiate the relations.

Step 4. Create instance automatically.

In this step, instances are created automatically based on the results (expected results, actually). If the expected

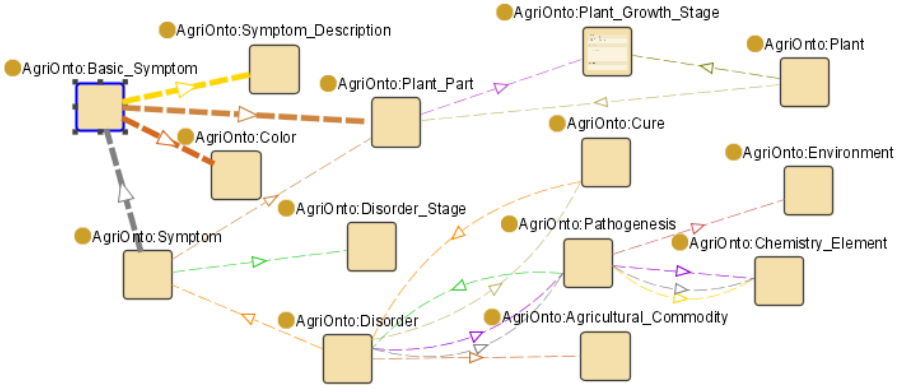

Figure 1 Classes and Object-properties in the Ontology (a fragment)

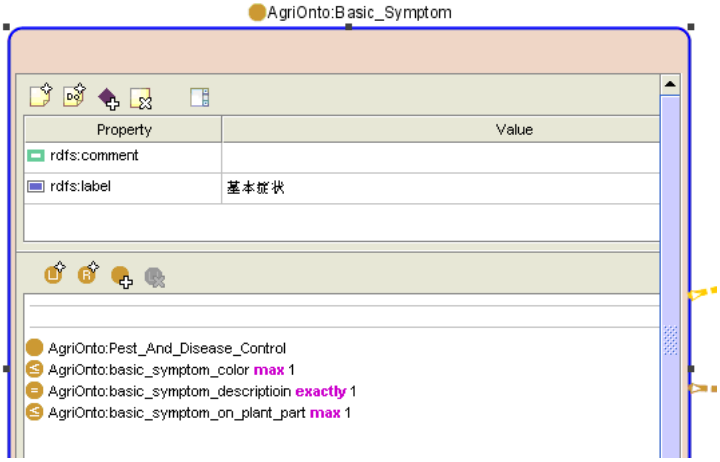

Figure 2 A depiction of definition of class Basic_Symptom

instance is not present, then an instance *I* of class *C* is created. The object properties are assigned values according to the expected results from Step 3.

V. A Case Study in Agriculture Domain

A. The Agriculture Ontology

The construction of agricultural ontology is part of a project in National High-tech Research and Development Program of China. One of the purposes of this project is to build up a knowledge base for agricultural decision-making systems. The construction effort is mainly based on the analysis of pieces of text descriptions. Take disease symptoms for example. A manual approach is to read text description for each disease and create corresponding instances in the ontology. There could be several defects when applying the manual approach: 1) the process of analysis is not repeatable and thus difficult for evaluation and evolution; 2) basic terminology is not explicit or bounded (there could be several forms expressing the same appearance, especially in Chinese); and 3) the original description (the most precise and complete addressing) is not preserved for further references. Therefore, we try to develop an automatic method for the machine to understand the text documents, create appropriate concepts in the ontology, and enable continuous updates when basic semantic units are added.

Firstly, basic semantic units are defined in the ontology. Theses units include 14 basic descriptions of disease symptoms, 12 colors and 7 parts of plant. Each of these units is described in a Chinese word (or character) . Each kind of basic semantic units is of a class type. Every unit is an instance of that class. The relations between classes are depicted in Figure 1. We focus on the class Basic_Symptom. The Basic_Symptom is defined as a sub-concept of Pest_and_Disease_Control, each of whose instances should be related with exactly one symptom description, at most one color and at most one plant part, as shown in Figure 2. Every instance of this class describes a basic symptom observed on the plant.

For example, the description “White mildew can be seen on leaves” is modeled as a basic symptom described by instances Color:white, Symptom_Description:mildew and Plant_Part:leaf. The primary aspects of the given sentence are then elicited. Usually, the symptoms of a plant disease are described in several sentences. For each sentence, we try to find the words according to the basic semantic units and find a proper instance of Basic_Symptom. If there does not exist such an instance, we create an instance associated with given basic semantic units (such as Color:white, Symptom_Description:mildew and Plant_Part:leaf in the above example). Due to the constraints casted on class Basic_Symptom, a sentence describing the symptom of the disease may be mapped to zero, one, or more basic symptoms. Therefore, the symptom (an instance of class Symptom) of a disease (an instance of class Disorder) is associated with several instances of Basic_Symptom. The instances of Basic_Symptom can be created automatically, given that the associated instances (basic semantic units) are previously added in the ontology.

B. The Web-based Ontology Learning Tool

A web-based ontology learning tool is developed based on the given approach. The graphic user interface is depicted in Figure 3. The text description is filled in the right panel. The text is treated sentence by sentence. For each sentence, the ranges of related properties are used for extracting basic semantic units (see Figure 3).

The right panel is where users may enter (usually copy and paste) texts. Based on the basic semantic units mentioned in the previous section, three instances of Basic_Symptom are identified, as is shown in Figure 4.

In our example, documents about plant diseases are analyzed automatically and ontology instances are identified. The ontology, especially the instances learned from legacy documents, acts as a knowledge base for agriculture users. Each instance represents a simple basic symptom describing plant diseases. Plant diseases may share some basic symptoms, which helps users to distinguish different plant diseases by identifying disease symptoms. Whether an instance is different from another is judged by its properties (values of object properties, actually). The usage of the ontology is of great importance, but beyond the scope of the paper.

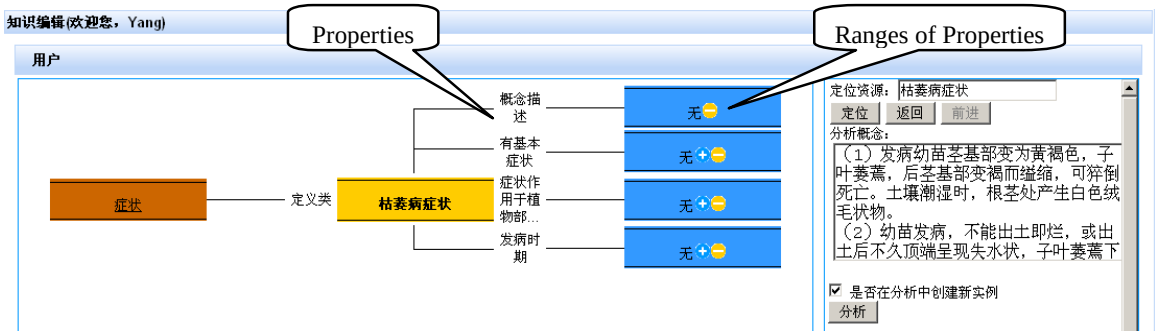

Figure 3 Learning the symptoms of a plant disease (The ranges of related properties are used for extracting basic semantic units from the text.)

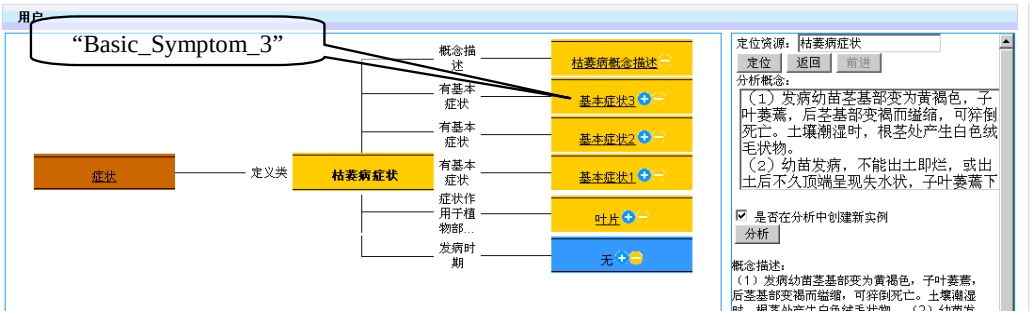

Figure 4 Instances of the class Basic_Symptom are created. (Quoted comments are English translation to the corresponding part in the screen-shot.)

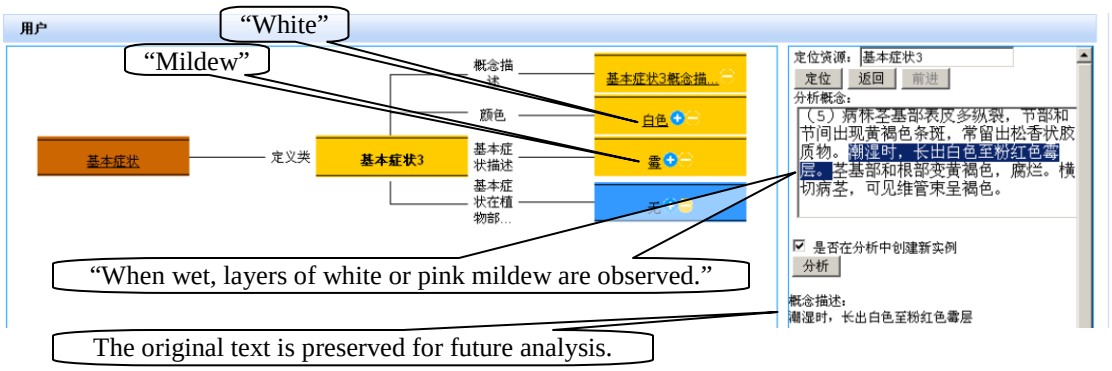

Figure 5 Some semantic information may be missing if basic semantic units are incomplete. (Quoted comments are English translation to the corresponding part in the screen-shot.)

There are cases that not all “important” words are pre-installed in the ontology, hence cannot be recognized by the tool. Take the instance Basic_Symptom_3 for example. The corresponding sentence is “when wet, layers of white or pink mildew are observed”. The identified instance is associated with basic semantic units (also instances in the ontology) Color:white and Symptom_Description:mildew. Because the color “pink” (粉红色 in Chinese) is not installed as a basic semantic unit, only “white mildew” is recognized and elicited from the sentence. See Figure 5.

There are also other sentences that are not understood by the tool. The reason is also the lack of basic semantic units. Once the semantic units are enriched, the whole process may be re-executed and better results will be achieved.

C. *Discussions on Ontology Evolution*

As is mentioned in the last section, the learning result is affected by how many basic semantic units are installed in the ontology. The more basic semantic units are pre-installed, the more basic symptoms can be recognized by the program.

Initially, the basic semantic units are the most interested words within the domain under consideration. In agriculture domain, the words are provided by domain experts or extracted from existing documents. But they are far from completeness. When there are only a few words in the ontology, the result of the learning process is also coarse-grained. In other words, some details in the document may be lost. In our case study (in the previous section), for example, the “pink” information is missing in the learning result. But the missing information will be complemented after the “pink” concept is added to the ontology (as an instance of class Color) and the learning procedure is re-executed. Therefore, as long as the team of knowledge engineering keeps maintaining the basic semantic units (basic instances in the ontology), the learning precision will increase gradually.

When we take the keep-maintaining approach, another potential problem emerges. It is much easier to maintain a small number of basic words and their classifications, relations or mutations, than to maintain an increasing number of words. Exact meanings of the words, including semantic similarity between words, should be inspected and expressed with great care. A simple and trivial approach to manage the complexity of basic semantic units (simple words) is by classification, i.e. the corresponding class in the ontology. Semantics of all instances of a particular class are maintained by a single person or team to achieve maximum conceptual integrity. Nevertheless, other approaches for managing the complexity of basic simple words can be developed independently.

VI. Conclusion and Future Work

When considering a specific domain, we may find that constructing the basic structure of the domain-specific ontology is not as difficult as we expected. The real engineering problem is how we can create the instances of

classes before we put the ontology into use. Rather than dealing with the ontology as a high-level abstraction of domain knowledge, we deal with ontology as a fusion of abstract knowledge and concrete facts. The challenging work here is no longer the creation of conceptual hierarchy, but integrating the content of natural-language documents into the ontology as knowledge facts. In this paper, we took agricultural legacy documents for example, constructed a basic ontology of concept hierarchy and concept relations, and proposed a method to learn plant disease symptoms expressed in Chinese by defining basic semantic units. The learning purpose can be extended to other regions of agriculture domain, such as production management, sales management and marketing, as long as the basic semantic units (basic words) are defined and installed in the ontology. With the help of NLP tools, other natural languages also work in our approach.

The work is still in its infancy. We are still trying to continue our work by adding more meaningful words in the agriculture domain and expecting much more fine-grained understanding of legacy documents for machines.

ACKNOWLEDGMENT

The work presented is supported by National High Technology Research and Development Program of China (863 Program) under grant No. 2007AA01Z179, National Natural Science Foundation of China (NSFC) under grant No. 60903013 and Shanghai Leading Academic Discipline Project No. B114.

REFERENCES

[1] Lina Zhou, Ontology Learning: State of Art and Open Issues, Information Technology Management, Springer, 2007, 8(3): 241–252

[2] Jiao Tao, Li Ding, Deborah L. McGuinness, Instance Data Evaluation for Semantic Web-Based Knowledge Management Systems, Proceedings of the 42nd Hawaii International Conference on System Sciences, 2009

[3] L. Ding, T. Finin, Characterizing the Semantic Web on the Web, International Semantic Web Conference (ISWC), 2006, 242-257

[4] Jon Atle Gulla, Vijay Sugumaran, An Ontology Creation Methodology: A Phased Approach, International Workshop on Ontology Dynamics (IWOD), 2008.

[5] Antonio De Nicola, Michele Missikoff, Roberto Navigli, A software engineering approach to ontology building, Information Systems, Elsevier, 2009, 34, 258-275

[6] Giancarlo Guizzardi, Ricardo de Almeida Falbo, José Gonçalves Pereira Filho, Using Objects and Patterns to Implement Domain Ontologies, Proceedings of the 15th Brazilian Symposium on Software Engineering, Rio de Janeiro , Brazil, 2001, http://www.loa-cnr.it/Guizzardi/SBES2001vf.pdf

[7] F. Xu, D. Kurz, J. Piskorski, S. Schmeier. Term extraction and mining of term relations from unrestricted texts in the financial domain, presented at Business Information Systems, Poznan, Poland, 2002.

[8] He Hu, Xiaoyong Du, ConAnnotator: Ontology-Aided Collaborative Annotation System, Proceedings of the 10th International Conference on Computer Supported Cooperative Work in Design, 2006, 850-855.

[9] Mark Craven, Dan DiPasquo, Dayne Freitag, Andrew McCallum, Tom Mitchell, Kamal Nigam, Sean Slattery. Learning to extract symbolic knowledge from the World Wide Web, Proceedings of the 15th National Conference on Artificial Intelligence (AAAI), 1998, 509-516.

[10] Chang-Shing Lee, Zhi-Wei Jian, and Lin-Kai Huang, A Fuzzy Ontology and Its Application to News Summarization, IEEE Transactions On Systems, Man, And Cybernetics (Part B: Cybernetics), v. 35, n. 5, October 2005, 859-880

[11] Hazman, Maryam; El-Beltagy, Samhaa R.; Rafea, Ahmed, Ontology learning from domain specific web documents, International Journal of Metadata, Semantics and Ontologies, 2009, v. 4, n 1-2, 24-33

[12] T. Dietterich, Machine learning research: Fore current directions, AI Magazine, 1997, 18, 97-136..