AUTOMATIC THAI ONTOLOGY CONSTRUCTION FROM CORPUS, THESAURUS, AND DICTIONARY

INTRODUCTION

Ontology is a well-known term in the field of AI and knowledge engineering. Among numerous definitions, the most widely quoted term is "an explicit specification of a conceptualization." (Gruber, 1993). Ontology captures the structure, relationships, semantics and other essential meta information about the application. It can be used for many different purposes and applications. It allows interaction between software agents that use ontologies for knowledge representation, enhances the performance of information extraction and information retrieval that provided the word meaning by ontologies. It enables communications and interoperability on the next generation of web transformation in the form of the semantic web. Moreover, in order to accomplish any kind of linguistic task that involves understanding, a computational linguistic system must have a knowledge base of lexical semantic like ontology. Although there are existing resources such as WordNet (Miller, 1995), they are insufficient for many problems e.g. insufficient or non-coverage of terms in specific domain and language barrier when applied in the Thai language. In addition, intensive time and labor consumption create ontology through using expert, resulting in insufficient term coverage.

In order to reduce the costs and to support the open ended task, researches on ontology construction have been addressed in several activities. The major problems in building ontologies are the bottleneck of knowledge acquisition and time-consuming construction and integration of various ontologies for various domains/ applications. One of most interesting study is the automatic ontology building with a variety of resources, such as *raw text* (Hearst, 1992; Maedche and Staab, 2001; Kietz, 2000; Yamakuchi, 2001; Navigli *et al.*, 2003); Li *et al.*, 2007; Pustejovsky *et al.*, 2007), *thesauri* (Soergel *et al.* 2004; Clark, 2000; Wielinga, 2001; Pustejovsky *et al.*, 2007) and *dictionaries* (Janniak, 1999; Keitz, 2000; Aramaki *et al.*, 2007). Each of these resources has different characteristics, hence, each is based on various approaches, e.g. rules, natural language processing, statistics for term and relationship extraction, etc. Raw text consists of unstructured text containing huge amounts of frequently updated information both terms and relations. Dictionaries are semi-structured resources that are only occasionally updated; domain dictionaries, which have a certain structure, are suitable for extracting terms and their relationships (e.g., hyponyms, meronyms, and synonyms) as well as their definitions. Among the terminological resources considered, thesauri lend themselves best to ontology construction, because their explicit semantic structure eases natural language processing to extract terms and relationships such as converting BT/NT (*boarder term/narrow term*) to superclass/subclass relationship, and refining RT *(related term)* to more specific relationships.

Although there are already a lot of researches done in this area, there is lacking of studies that have been addressed with Thai ontology. Constructing a Thai ontology is attractive since many terms in Thai do not exist in other languages especially the term in leave level such as Thai native plant species name. And the Thai ontology is necessary knowledge for applications that processing Thai documents. Thus, in this thesis the appropriate methodologies for learning Thai ontology, particularly from raw text, are studied.

Research Question

The principal question addressed in this thesis is:

What is the appropriate method for learning Thai ontology?

The appropriate methodology should be evaluated with 3 criteria: accuracy, coverage and portability.

Approach

Among various approaches for ontology learning, each of them has disadvantages and some aspects that do not suitable for extracting Thai documents. The main disadvantage of rules-based and statistical-based approach is the problem of data sparseness that is the main problem in case of lacking corpora in Thai. Though, there are several works search the ontological elements (such as pattern) from WWW documents which are very large corpus, this solution can not be used in Thai language. Because Thai does not have delimiters to show word boundaries then it can not be directly extracted from WWW. Moreover, language resources in Thai such as WordNet are lacked. WordNet is the lexicon resource using for identifying words' meaning which can apply to extract the semantic relationship of nouns in NPs. Accordingly, this work proposes the combined methods of rules-based and statistical-based for learning the ontological concepts and relationships from small corpus and also propose techniques for applying WordNet in the task of extracting semantic relationship between nouns in Thai NPs. Although these methodologies are proposed for constructing Thai ontology but these techniques can be adapted for other languages. Even, the system is tested with the resources in the domain of agriculture because documents concerning this domain are very rich resources in Thai; the proposed methodology also works with other domains.

In this work, we rely on the following three resources: text corpora, a thesaurus, and a domain specific dictionary. However, text corpora are the important part of this thesis. Extracting terms and their relationships from corpora is the challenge since corpora contain new and up-to-date terms.

Ontology Extraction from Text Corpora: In order to construct and maintain ontology, we use NLP technologies, i.e. morphological analysis (Sudprasert and Kawtrakul, 2003), shallow parsing (Pengphon, 2002) and named entity recognition (Chanlekha and Kawtrakul, 2004), to identify potentially interesting terms. Moreover, in order to identify the intended ontological terms and their relationships, we use explicit cues, i.e. lexico-syntactic patterns and an item list (bullet list and numbered list). The main advantage of the approach is that it simplifies the task of concept and relation labeling since cues help for identifying the ontological concept and hinting their relation. However, these techniques pose certain problems, i.e. cue word ambiguity, item list identification, and numerous candidate terms. We also propose the methodology to solve these problems by using lexicon and co-occurrence features and weighting them with information gain. Moreover, machine learning techniques are used to mine the semantic relationships embedded in the texts' noun phrases by using the common super concepts of their head and modifier. Unfortunately, Thai lacks a resource or knowledge base like WordNet (Miller, 1995) to identify the super concept of terms. Hence, existing lexical resources, a Thai-English Thesaurus, AGROVOC (Food and Agriculture Organization [FAO], 2006), and a general Thai-English Dictionary, LEXiTRON (National Electronics and Computer Technology Center [NECTEC], 2007), would be beneficial for translating terms from Thai to English, and identifying the WordNet sense and the super concept of the ontological terms in order to support the mining of implicit ontological relationships in noun phrases with the machine learning techniques.

Ontology Extraction from Thesaurus: Concerning with the thesaurus-based ontology construction, we take AGROVOC (FAO,2007), a multilingual thesaurus that includes Thai (Thunkijjanukij *et al.*, 2005), as a seed. AGROVOC deals with two domains: food and agriculture. At present AGROVOC contains more than 28,000 descriptors and more than 10900 non-descriptors (synonyms) in English. In 2000, Thai National AGRIS Centre of Kasetsart University has developed Thai AGROVOC by translating from FAO AGROVOC. There are 16,607 descriptors and 2,302 non-descriptors. These agricultural vocabularies are repository in Thai language. The rest were terms mainly for local used that need further development. However, like all thesauri AGROVOC contains some explicit semantics resulting in straightforward ontology transformation. Unfortunately, like most other thesauri, AGROVOC's relationships are too coarse grained and too broad, and would challenge automatic transformation, into an ontology. Within AGROVOC, semantic relationships are poorly defined and inconsistently applied. For example, AGROVOC incorrectly uses ***NT***, approximately equivalent to 'superclass of,' or 'hypernym of', in *Milk* ***NT*** *Milk*

Fat, while a more specific, and correct, relationship could be 'containsSubstance'. In AGROVOC, ***RT*** is underspecified, subsuming numerous relationships; for example, it uses ***RT*** in *Mutton* ***RT*** *Sheep,* which should be refined to a more specific one, such as 'madeFrom' (Soergel *et al.* 2004) to distinguish from other uses of RT. Hence, the extraction of ontological relationships from a thesaurus requires data cleaning and refinement of the identified semantic relations. To achieve this, our system consists of three main modules: Rule Acquisition, Detection and Suggestion, and Verification. The module in charge of acquiring refinement rules draws on experts' knowledge and machine learning techniques. The Detection and Suggestion module performs an analysis of the terms' noun phrases and WordNet alignments in order to detect incorrect relationships. Once this is done it will suggest better relationships by using the acquired rules. The Verification module is a tool for confirming the proposed relationships.

Ontology Extraction from Dictionary: Another good resource for extending the ontology that we use here is "Thai Plant Names" Dictionary (Smitinand, 2001), a domain specific dictionary. In order to extract ontological terms and relationships from a specific dictionary, a task oriented parser is used to analyze the relational terms and convert them to the ontological tree.

Finally, all the ontological atomic or sub-trees collected by various techniques are integrated into a master tree by using the technique of term matching. This system, we use ontology extracted from thesaurus as a master tree because AGROVOC thesaurus has a number of concept hierarchies more than the other sources and it has the terms covered the domain of agriculture. After that, the ontology is organized in order to prune the inconsistent relationships.

Contributions

This research makes four contributions to the fields of Ontology Engineering:

1. Proposing the practical methodologies for Thai ontology construction from various resources

We propose the methodologies for ontology construction from text corpus by using cues based on the lexical and co-occurrence features (Imsombut and Kawtrakul, 2007) and extracting the relations of nouns in NPs in the terms of machine learning based on their common super concepts (Imsombut and Kawtrakul, 2005). For thesaurus-based ontology construction, we combine several methods: machine learning technique, noun phrase analysis and WordNet alignment, for thesaurus relationship cleaning and refinement (Kawtrakul and others, 2005). Concerning the dictionary, only a task-oriented parser is needed to extract the relevant terms and relations. Finally, all the ontological sub-trees are integrated into a master tree by using the technique of term matching and the ontology is also reorganized for pruning inconsistency relationships.

2. Learning and verification tools for ontology construction

These tools provide the modules for extracting and building ontologies from three resources i.e. text corpus, thesaurus and dictionary. Moreover, it allows users to verify the correctness of the ontologies.

3. Thai ontology in the domain of agriculture

This resource provides knowledge about terms, their synonym and their relations to other terms in the agricultural domain. It contains about 59,971 terms and 41,677 relationships. It is very useful for any kind of linguistic task that involves text understanding. It is vital for solving the problem of word sense ambiguity, which is the crucial problem of the application in the field of computational linguistic.

4. Annotated corpus for ontology learning

This resource contains a set of annotated tags that are composed of terms, their relations and cues. The corpus size is of 302,640 words from 90 documents. It is useful for studying the phenomena of the occurrence of the ontological terms, ontological relation and types of cues.

Thesis Organization

The rest of the document is organized as follows:

1. OBJECTIVES talks about the objectives of this research.

2. LITERATURE REVIEW presents background information in the area of ontologies and related terms. Next, the state-of-the-art about ontology learning from text corpus, thesaurus and dictionary and ontology merging are briefed. In addition, the crucial problems of Thai ontology construction based on the previous mentioned resources are described.

3. MATERIALS AND METHODS talks about materials that used in this study and methodologies for ontology construction and merging. It shows how ontological terms and relationships are acquired from text corpora by using cues and NPs component. Moreover, AGROVOC thesaurus refinement and Thai Plant Name dictionary conversion are introduced here. Finally, we also explain the merging technique for all sub-ontology trees based on linguistic matching.

4. RESULTS AND DISCUSSION presents a set of experiments of ontology construction from text corpus, thesaurus and dictionary and ontology merging in the domain of agriculture. Next, the results are analyzed and discussed.

5. CONCLUSION AND RECOMMENDATION provides conclusions and future work about the concepts presented in this thesis.

OBJECTIVES

1. To study problems and approaches for constructing and maintaining ontology from text corpus, thesaurus and dictionary and ontology merging from various sources.

2. To study methodologies for automatically Thai ontology construction and maintenance from text corpus, thesaurus and dictionary and ontology merging from various sources.

3. To develop algorithm and tools for Thai ontology construction and integration from text corpus, thesaurus and dictionary.