THESIS APPROVAL

GRADUATE SCHOOL, KASETSART UNIVERSITY

Doctor of Engineering (Computer Engineering)
DEGREE

Computer Engineering	Computer Engineering
FIELD	**DEPARTMENT**

TITLE: Automatic Thai Ontology Construction from Corpus, Thesaurus, and Dictionary

NAME: Ms. Aurawan Imsombut

THIS THESIS HAS BEEN ACCEPTED BY

Asanee Kawtrakul
THESIS ADVISOR
(Associate Professor Asanee Kawtrakul, D.Eng.)

Y. Poovarawan
COMMITTEE MEMBER
(Associate Professor Yuen Poovarawan, M.S.)

COMMITTEE MEMBER
(Associate Professor Kitsana Waiyamai, Ph.D.)

DEPARTMENT HEAD
(Assistant Professor Kemathat Vibhatavanij, Ph.D.)

APPROVED BY THE GRADUATE SCHOOL ON 29/06/07

Vinai Artkongharn
DEAN
(Associate Professor Vinai Artkongharn, M.A.)

THESIS

AUTOMATIC THAI ONTOLOGY CONSTRUCTION FROM CORPUS, THESAURUS, AND DICTIONARY

AURAWAN IMSOMBUT

A Thesis Submitted in Partial Fulfillment of
the Requirements for the Degree of
Doctor of Engineering (Computer Engineering)
Graduate School, Kasetsart University
2007

Aurawan Imsombut 2007: Automatic Thai Ontology Construction from Corpus, Thesaurus, and Dictionary. Doctor of Engineering (Computer Engineering), Major Field: Computer Engineering, Department of Computer Engineering. Thesis Advisor: Associate Professor Asanee Kawtrakul, D.Eng. 117 pages.

Ontology has a crucial role to play in information retrieval; however, its building by experts is an expensive task, and also a never-ending process which relies on evolution of knowledge especially in science. Hence, we suggest learning ontologies automatically in order to spare experts the bulk of the job.

We present here a hybrid approach for especially building and maintaining (semi-) automatically an ontology from corpus, and also thesaurus and dictionary. Concerning the corpus, we propose a methodology for extracting ontological concepts and taxonomic relations by using explicit cue expressions, i.e. lexico-syntactic patterns and an item list. However, this technique has several difficulties i.e. cue word ambiguity, item list identification, and numerous candidate terms. We solved these problems by using the lexicon and co-occurrence features and using information gain for weighting these features. Moreover, we fill the ontology with the semantic relations embedded in Thai NPs by translating Thai words into English, with a method of selecting the word sense from WordNet and by applying machine learning techniques to learn the semantic relations. In order to extract ontology from a specific dictionary, a task oriented parser is used to build the ontological tree. Moreover, we refine the thesaurus' relationships to ontological relations by using machine learning and some heuristic rules. Finally, we integrated all the ontological sub-trees collected by using the technique of term matching and then the ontology is reorganized for consistency. We tested the system by using Thai corpora in the domain of agriculture and the accuracy of the final result from 3 resources is 0.86.

Aurawan Imsombut
Student's signature

Asanee Kawtrakul
Thesis Advisor's signature

18 / Jun / 07

ACKNOWLEDGEMENTS

First of all, I would like to grateful thank and deeply indebted to my advisor, Assoc. Prof. Dr. Asanee Kawtrakul, for providing the basis and guidance that has enabled this work to emerge. She introduced me to the field of natural language processing and taught me a lot about research methodology. I am also grateful to the other members of my committees— Assoc. Prof. Yuen Poovarawan, Assoc. Prof. Dr. Kitsana Waiyamai and Assoc. Prof. Dr. Mathieu Lafourcade who provided useful comments and finalized my work.

I would like to present deeply thanks to Dr. Michael Zock for his patience to review my publication of this work. I would like to sincerely thank Dr. Patrick Saint-Dizier and Dr. Nigel Collier for their valuable comments and suggestion. My special thanks go to Mrs. Aree Thunkijjanukit for her suggestion and kindly providing the resources of AGROVOC for using in this research.

I am heartfelt thank to all the members of the NAiST Laboratory, which provided a home with a friendly and helpful research environment. Moreover, my friends: Chaveevan Pechsiri, Thana Sukvaree, Chaloempon Sirikayon and Trakul Permpool who help me prepare for and to overcome many obstacles along the way. I express my sincere gratitude to them.

I would also like to thank Dhurakij Pundit University for providing the financial support of my study. Moreover, this research was supported by the grant of NECTEC No. NT-B-22-14-12-46-06 and also funded in part by the KURDI; Kasetsart University Research and Development Institute.

Most of all, my appreciations devote to my family for their love and support, for continuing encouragements and for always believing in me.

Aurawan Imsombut

May 2007

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF TABLES (Continued)

LIST OF FIGURES

LIST OF FIGURES (Continued)

LIST OF ABBREVIATIONS

AI	=	Artificial Intelligence
adj	=	Adjective
AAT	=	Art and Architecture Thesaurus
BT	=	Boarder Term
χ^2	=	Chi-square testing
cl	=	Classifier
CBC	=	Clustering By Committee
nct	=	Collective noun
ncn	=	Common noun
det	=	Determiner
DODDLE	=	Domain Ontology rapiD DeveLopment Environment
FAO	=	Food and Agriculture Organization
vi	=	Intransitive verb
KA^2	=	Knowledge management ontology
LL	=	Log-Likelihood ratio
MLH	=	Most Likely Hypernym
MI	=	Mutual Information
NE	=	Name Entity
NER	=	Name Entity Recognition
NT	=	Narrow Term
NECTEC	=	National Electronics and Computer Technology Center
NLP	=	Natural Language Processing
NP	=	Noun Phrase
norm	=	Ordinal number marker
PP	=	Prepositional Phrase
npn	=	Proper noun
RBF	=	Radial Basis Function
RT	=	Related Term
RELCL	=	Relative clause

LIST OF ABBREVIATIONS (Continued)

prel	=	Relative pronoun
RDFS	=	Resource Description Framework Schema
S	=	Sentence
SUMO	=	Suggested Upper Merged Ontology
SVM	=	Support Vector Machine
vt	=	Transitive verb
UMLS	=	Unified Medical Language System
UF	=	Used For
VP	=	Verb Phrase

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.

LITERATURE REVIEW

Introduction to Ontology and Related Terms

In this section, introduction of ontology and related terms are presented. First, ontology and related terms definitions are briefed. Next section describes main components of an ontology and types of Ontology.

1. Ontology and Related Terms Definitions

There are many terms related to ontology. In this section, we will focus on thesaurus, dictionary and WordNet that are the resources used for constructing ontology.

1.1 Thesaurus

The thesaurus is the database representing the relationship among the terminology in a given domain. It can be used for indexing and retrieving information resources. Thesaurus consists of descriptor i.e. the priority word which is used to represent specific concept and non-descriptor which is non-priority word. They are linked together by the equivalence relation and reciprocal relation. For example, the Use (USE) and Used For (UF) relationship are used to describe the synonym of descriptor and non-descriptor. The reciprocal relationship is represented by hierarchical relation and association such as the Broader Term (BT) used to link to a more general term, the Narrower Term (NT) used to link to a more specific term and the Related Term (RT) used to represent the association of terms. Figure 1 shows the example of thesaurus represented by using these relationships of terms.

Roughly, there are two types of thesauri, i.e., general domain and specific domain. The example of general domain thesaurus is *Roget's Thesaurus* (Roget, 1962). It contains lists of words with similar meanings which are organized according to a system of thinking about the world and words. In Thai we have *Thai thesaurus*

(Kasetsart University, 1992) developed by working group of the department of Computer Engineering and department of Linguistics of Kasetsart University. Concerning specific domain thesaurus, there are many thesauri in various domain e.g. *AGROVOC Thesaurus (*Food and Agriculture Organization, 2007*), Art and Architecture Thesaurus* (Getty Institute, 2007), *Clinician's Thesaurus* (Zuckerman, 2005)

Cereals	
UF	Small grain cereals
BT	Plant products
NT	Oats
	Rice
	Rye
RT	Cereal crops

Figure 1 An example of AGROVOC Thesaurus

In ontology construction task, thesaurus is one of resources used for extracting concepts and relationships. The USE/UF relationships can be converted to synonym relationship and the BT/NT relationships can be converted to superclass/subclass relationships. However, the RT relationships can represent numerous relationships then it needs more techniques for refining them to a more specific relationship.

In this work, we applied AGROVOC thesaurus for constructing ontology and we propose the methodologies for cleaning and refining the AGROVOC's relationships by using machine learning, noun phrases analysis and WordNet alignments techniques.

1.2 Dictionary

The dictionary is a list of words with their definitions. It also provides pronunciation information, grammatical information, word derivations, histories, usage guidance and examples in phrases or sentences. Dictionary types include

general language dictionaries, subject dictionaries that cover the terms of a particular field, special purpose dictionaries that focus on a type of word such as slang (Texas State Library and Archives Commission, 2003).

Many research studied for extracting ontology from dictionaries (Janniak, 1999; Keitz, 2000; Aramaki *et al.*, 2007). Most of them analyzed the word and the definition of word from these dictionaries in order to extract the concepts and the relationships. The methods used in these works are similar to techniques of corpus-based ontology extraction since the definitions of word are unstructured texts as the corpus. These techniques will be discussed in this chapter. However, there are some subject dictionaries that have specific structure useable for the ontology extraction. Similarly, Thai Plant Name dictionary (Smitinand, 2001) used in this work can be analyzed the relationship of plant's family/sub-family/genus and converted to hypernym/hyponym relation of ontology.

1.3 WordNet

WordNet is an on-line electronic lexical database developed by a Princeton University group led by George Miller (Miller, 1995). In WordNet, words are organized into taxonomies where each node is a set of synonyms (a "synset") representing a single sense. There are four different taxonomies based on different parts of speech (noun, verb, adjective and adverb) and also there are many relationships defined among them. The basic relationships are hyponymy (is a kind of), hypernymy (this is a kind of), meronymy (part of this), holonymy (this is a part of), entailment for verbs (like meronymy for the nouns), antonymy, and synonymy. WordNet gives definitions (explanatory glosses) and sample sentences for the most of its synsets. It contains 152,059 unique strings, 115424 synsets and 203145 total word-sense pairs.

In this thesis, we aligned the relationships of WordNet to the AGROVOC's relationships and we also used WordNet for identifying sense of words

in order to extract the semantic relationships between head and modifier nouns in NPs.

1.4 Ontology

Ontology is the term that has been originally used in Philosophy where it is a systematic account of existence. More recently, the term has been used in various areas in Computer Science and Artificial Intelligence (AI) such as knowledge engineering, language engineering. Numerous definitions have been offered, and one of the most widely quoted definitions of "ontology" proposed by Gruber (1993) is that:

An ontology is an explicit specification of a conceptualization.

Borst (1997) modified Gruber's definition and proposed that:

Ontology is defined as a formal specification of a shared conceptualization.

Studer and colleagues (Studer and colleagues, 1998) merged Gruber's and Borst's definition as follows:

An ontology is a formal, explicit specification of a shared conceptualization. 'Conceptualization' refers to an abstract model of phenomena in the world by having identified the relevant concepts of those phenomena. 'Explicit' means that the type of concepts used, and the constraints on their use are explicitly defined. 'Formal' refers to the fact that the ontology should be machine readable. 'Shared' reflects that ontology should capture consensual knowledge accepted by the communities.

Another definition of ontology, often used in literature, has been given by Guarino in 1998:

An ontology is a set of logical axioms designed to account for the intended meaning of a vocabulary.

Swartout and colleagues (1996) have defined an ontology as the design of a knowledge base:

An ontology is a hierarchically structured set of terms describing a domain that can be used as a skeletal foundation for a knowledge base.

However, there are many other definitions that each ontological research group has tried to clarify their view on ontologies. These definitions depend on their purposes of ontological development and the applications of ontology. We define ontology here as "a general principle of any system to represent knowledge for a given domain, with information from heterogeneous sources. Information can be represented by concepts and semantic relationships between them." Although there are some minor differences, they refer to the ontology as a common understanding of a domain, and imply it as a repository of vocabulary for the knowledge of a domain.

2. Main Components of an Ontology

In (Gruber, 1993), ontology composes of five kinds of components:

$$(C, I, R, F, A)$$

where:

C is the set of the *concepts* that is the set of the abstractions used to describe the objects of the world. It also called *classes*. Each concept can have properties for describing them;

I is the set of *individuals* of an ontology, that is, the actual objects of the world. The individuals are also called *instances* of the concept;

$\mathcal{R}$ is the set of *relationships* that represent a type of association between concepts of the domain. It defined on the set C, that is, each $R \in \mathcal{R}$ is a product of n sets, R: $(C_1 \times C_2 \times \ldots \times C_n)$. Ontologies usually contain binary relations. For example `subclass-of` is the pair (C_p, Cc), where C_p is the parent concept and Cc is the child concept. For instance, `subclass-of` (*Animal, Cow)*, `Part-of` (*Cow, Horn*);

$\mathcal{F}$ is the set of *functions*. It is a special case of relations in which the n-th element of the relation is unique for the n-1 preceding elements. That is, each element $F \in \mathcal{F}$ is a usually expressed as F: $(C_1 \times C_2 \times \ldots \times C_{n-1} \mapsto C_n)$. For example, the function `Pay` is function of the concepts `Price` and `Discount`, and returns a concept `FinalPrice`, that is `Pay: Price×Discount`$\mapsto$ `FinalPrice`;

$\mathcal{A}$ is set of axioms that serve to model sentences that are always true. It used to verify the consistency of the ontology itself.

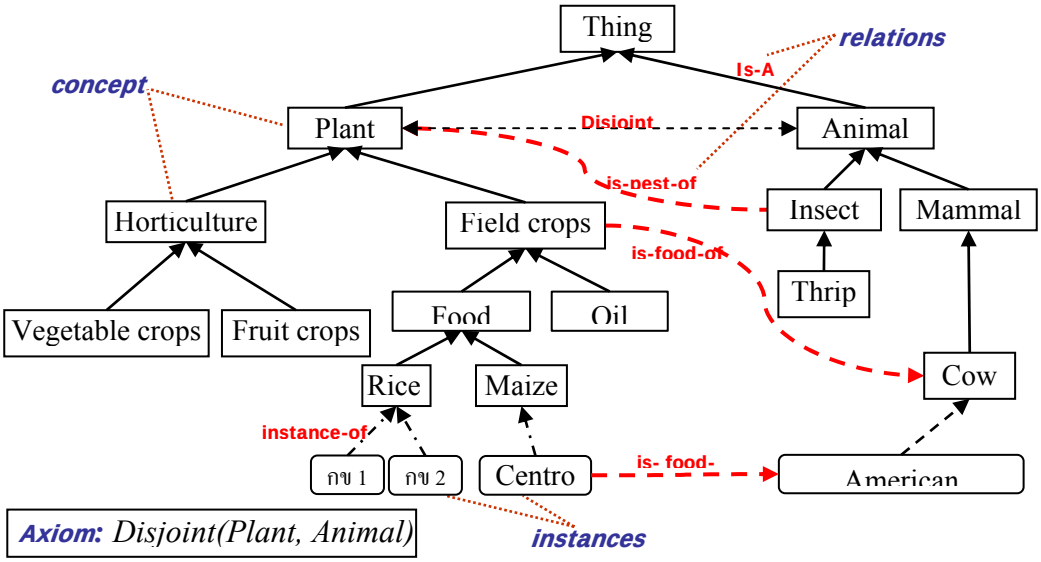


Figure 2 An example of ontology in the domain of agriculture

Some of these components are vital for an ontology, such as concept and relationship that are the component the simplest type of ontology (lightweight ontologies). For instance, Figure 2 is the example ontology in the domain of agriculture that has only four components: *Concepts, Instances, Relations* and *Axiom.* Although this limits the knowledge, it can be expressed about the domain. In this research, we focus on the extraction of concepts and some relationships, including is-a and part-of relationships which are the main structure of the ontology.

3. Types of Ontology

From the literature, the ontologies can be classified according to different dimensions. Here we present the most common type of ontologies. Guarino (1998) classified the ontologies based on the level of dependence of a particular task. The types of the ontologies are distinguished as shown in Figure 3.

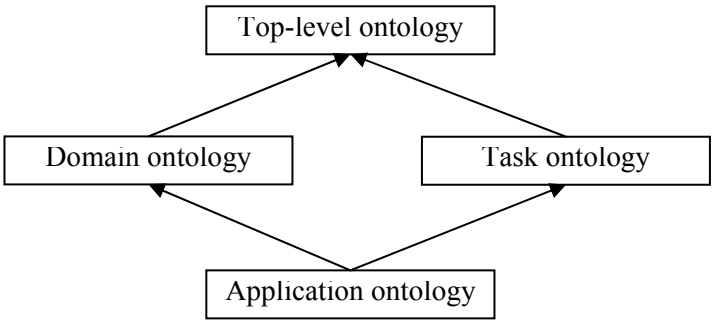


Figure 3 Categorization of ontologies classified by Guarino
Source: Guarino (1998)

- *Top-level ontologies*: This kind of ontology describes very general concepts such as space, time, matter, object, event, action, etc., which are independent of a particular problem or domain. The examples of top-level ontologies are: Top-level ontologies of universals and particulars built by Guarino and colleagues (Guarino and Welty, 2000) and SUMO: Suggested Upper Merged Ontology promoted by the IEEE Standard Upper Ontology working group (Schoening, 2003).

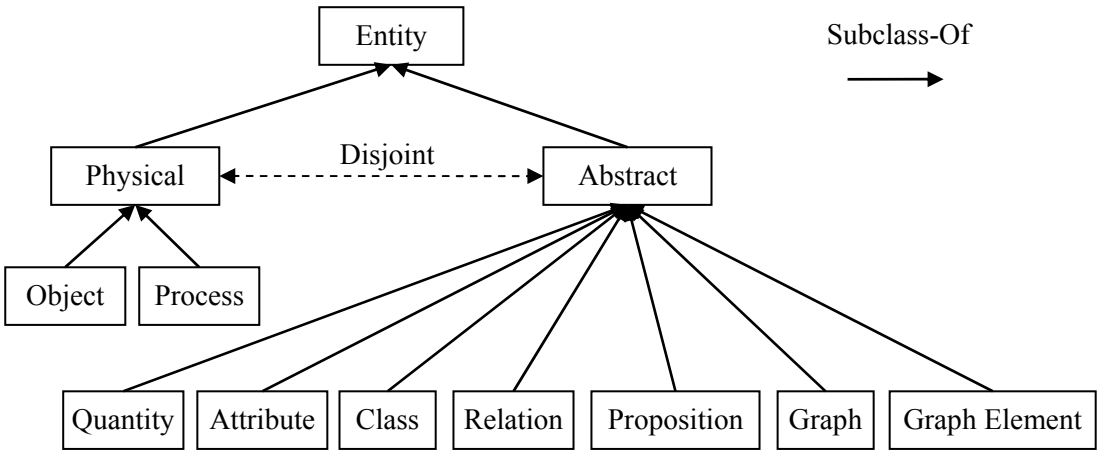


Figure 4 Structure of the first levels of SUMO
Source: Schoening (2003)

- *Domain ontologies*: This kind of ontology describes the vocabulary related to a generic domain such as medicine or physics by specializing the concepts introduced in the top-level ontology. For instance, UMLS (Unified Medical Language System) (Bodenreider, 2004) contains a lot of biomedical terms and KA^2 is Knowledge management ontology (Decker *et al.*, 1999).

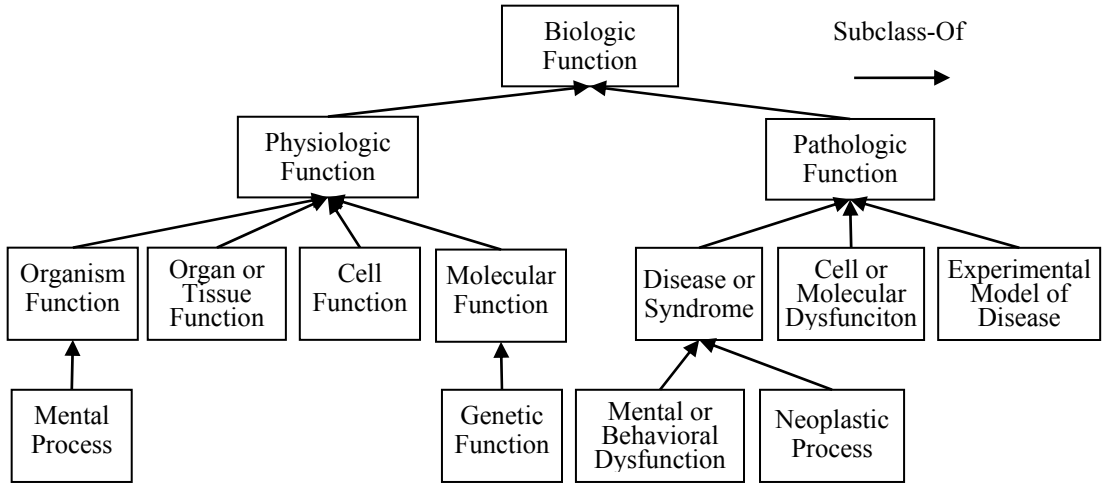


Figure 5 Partial view of the UML ontology
Source: Bodenreider (2004)

- *Task ontologies*: This kind of ontology describes the vocabulary related to a generic task or activity such as disease dispersion, diagnosis or selling by specializing the top-level ontology. Figure 6 shows task ontology about dispersion of disease (Kawtrakul *et al.*, 2007).

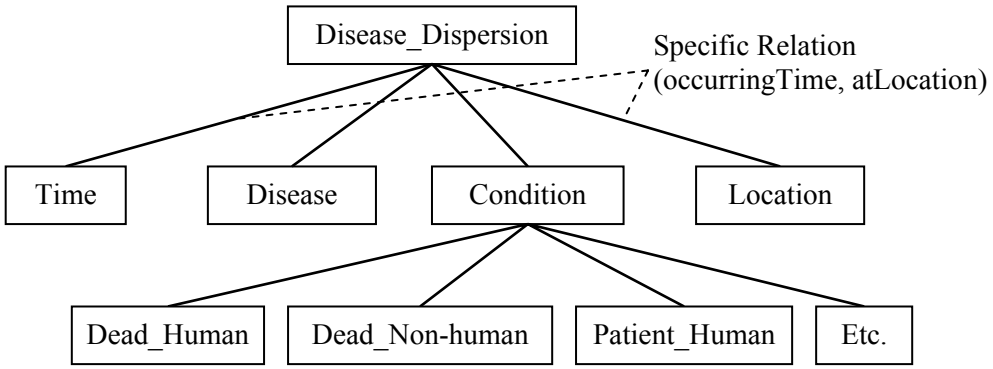


Figure 6 Task ontology about dispersion of disease
Source: Kawtrakul *et al.* (2007)

- Application ontologies: These are the most specific ontologies. It describes concepts depending both on a particular domain and on a particular task. Concepts in application ontologies often correspond to roles played by domain entities while performing a certain activity.

Figure 7 shows some parts of a simple ontology of Amazon web service for books (Scicluna *et al.*, 2005). This example service allows Searching by title, keyword(s) and price range. The terminology for semantic description of this service uses several ontologies. First ontology is the top-level ontology. The second is the domain ontology about books and further one is task ontology for requests. The last is application ontology defined involve amazonBooks and amazonRequests.

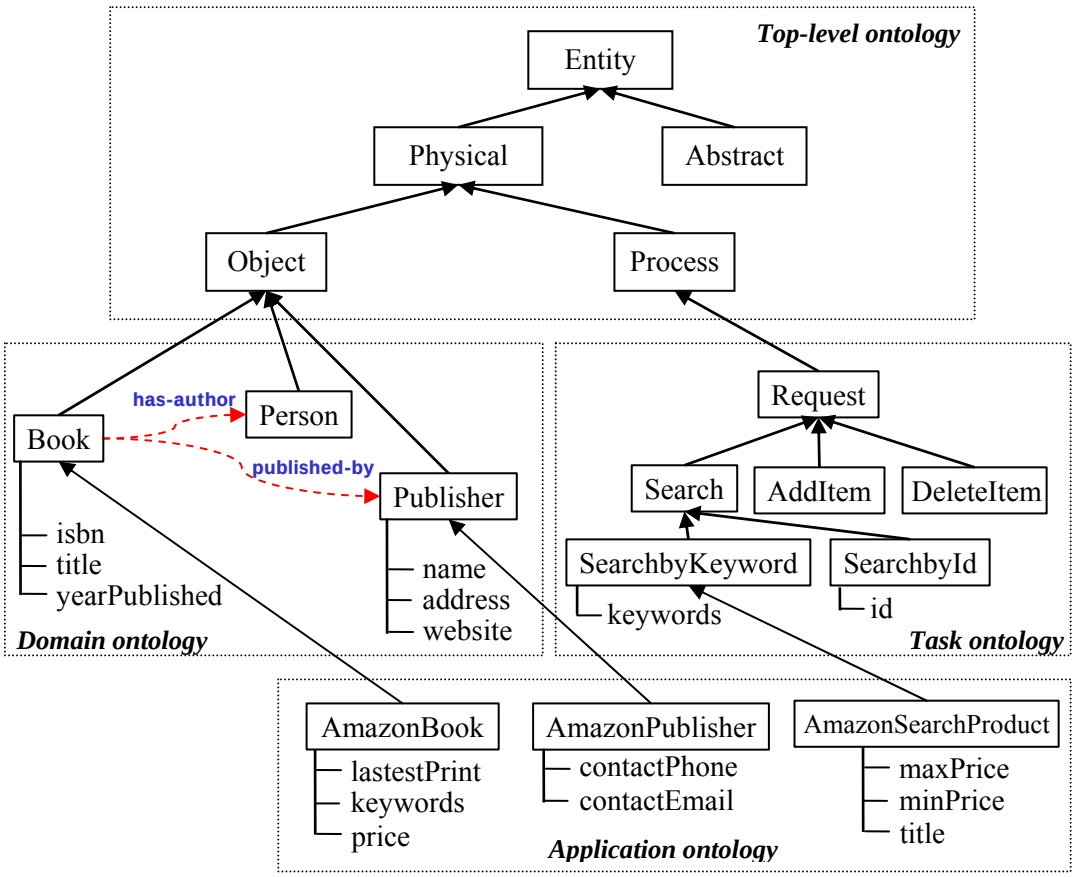


Figure 7 A simple ontology of Amazon web service for books

Source: Scicluna *et al.* (2005)

In addition, the ontology community classified ontology into *lightweight* and *heavyweight* ontologies, depending on the degree of formality used to express them. (Gomez-Perez, 2004)

- *Lightweight ontologies* are those ontologies that define a vocabulary of terms with some specification of their meaning. These ontologies include concepts, concept taxonomies, relationship between concepts and properties that describe concepts. Figure 8 shows the example of lightweight ontology that is in the domain of agriculture about plant concept.

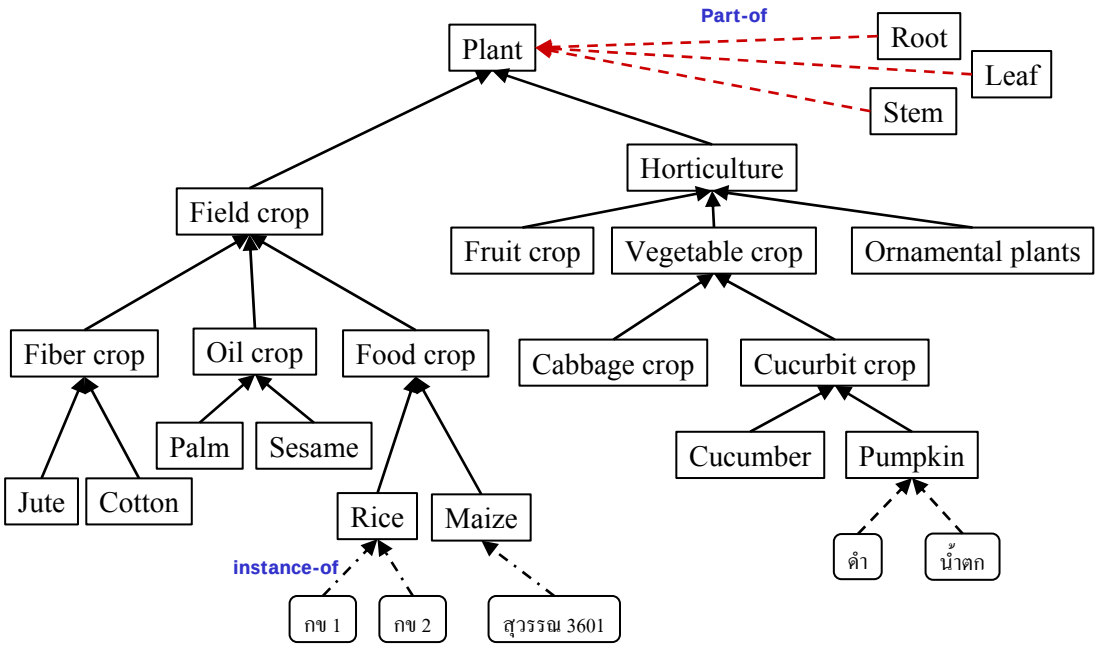


Figure 8 An example of lightweight ontology

- *Heavyweight ontologies* are those which are provided with restrictions on domain semantics, inference mechanisms aimed to equip ontologies with deductive power (e.g., inheritance), and that are characterized by a high degree of formality (e.g., underlying formal semantics). On the other hand, heavyweight ontologies add axioms and constraint to lightweight ontologies. The example of heavy weight ontology is shown in Figure 9. It is the ontology about researcher, topic and document that contains rules used for inferring the new knowledge.

Although there are many types of ontology and no matter which type an ontology is, it can be used as a tool to structure the knowledge of a given domain, let's say medicine (Aramaki *et al.*, 2007) or agriculture, our concern (Kawtrakul *et al.*, 2004a), (Kawtrakul *et al.*, 2005), (Imsombut and Kawtrakul, 2005). As such it plays an important role for enhancing the performance of systems addressing issues like information processing by and large, question-answering (Plas and Bouma, 2007), (Vargas-Vera and Motta, 2004), (Mann, 2002), knowledge sharing and knowledge management (Fensel *et al.*, 2000), (Davies *et al.*, 2002), (Aldea *et al.*, 2003), (Maedche *et al.*, 2002).

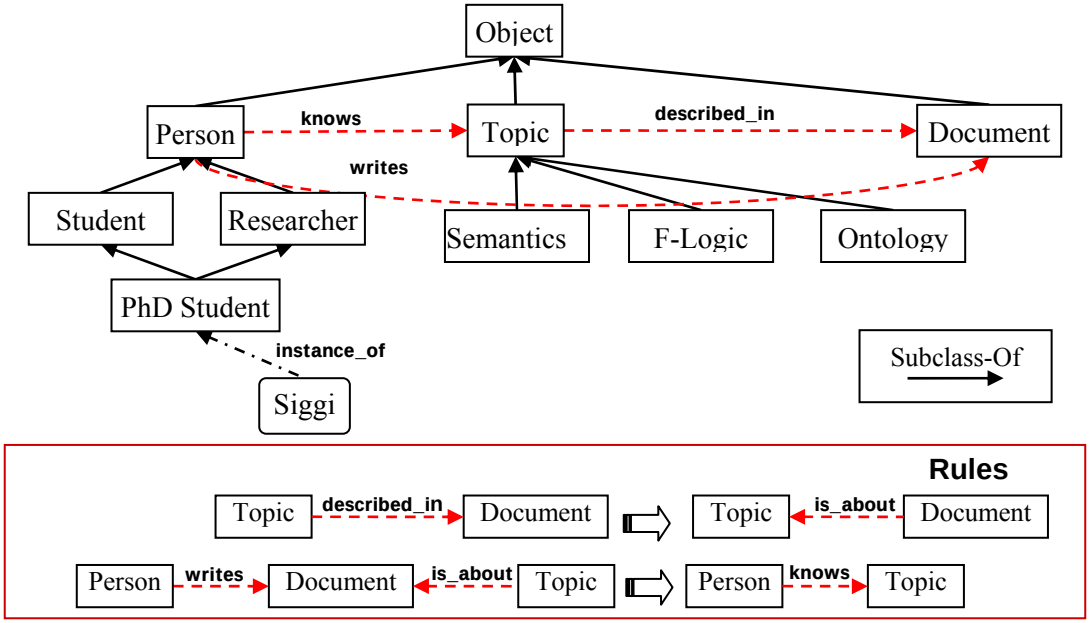


Figure 9 An example of heavyweight ontology

However, building ontologies requires much time and many resources. Hence, the systems for (semi-) automatically ontology constructing are needed. These influence to interesting of research in the area of ontology learning. Beside this, researches in ontology are ontology integration and ontology evaluation etc. In this thesis, we will focus on ontology learning and ontology integration. Concerning ontology learning, it can be 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 (Gómez-Pérez and Manzano-Macho, 2003). Other terms are also used to refer to the (semi-)automatic construction of ontologies like ontology generation, ontology mining, ontology extraction, etc. In addition, ontology can be constructed from existing ontologies since single ontology may contain insufficient information of the domain and this task is called as ontology integration or ontology merging. From the literature reviewing, we found that most of the research in ontology learning studied for automatically extracting domain specific ontology from text in order to build a new ontology or enrich the top-level ontology. Moreover, the researchers have been addressed primarily with the lightweight ontologies and the extraction of rules is probably the least addressed researched area

in ontology learning. Because this task is very difficult and if it can be done automatically, the generated rules are very worth knowledge for applying ontology in other applications. The existing approaches for (semi-)automatic ontology building are natural language analysis, statistical technique, heuristic rules and machine learning techniques, which will be discussed in the Related Works section.

In this research, we propose the methodologies for ontology learning from heterogeneous resources and integrating all extracted ontology to be the complete one. We tested the system by using Thai corpora in the domain of agriculture. The reason why we tested with Thai is the lacking of studies that have been addressed with Thai ontology. Moreover, Thai ontology is necessary for application of Thai documents processing. The details of the methodologies will be described in Materials and Methods Chapter.

Related Theories

This section describes background knowledge concerning ontology construction. First, natural language processing techniques, which are the important technique for ontology extraction, are discussed. Next section, we present the theories of Information Gain and Support Vector Machine that used for weighting features in the process of ontological relationship extraction.

1. Natural Language Processing

Natural Language Processing (NLP) is an important technique for extracting both concepts and relationships of ontology. Using NLP in Noun phrase (NP) extraction is a crucial technique that widely used in concept extraction task. Concerning relationship extraction, we can classify the level of ontological relationships in texts into three levels: phrase, sentence and paragraph.

1.1 Noun Phrase Extraction

The concepts are linguistically represented by terms (labels) and terms are predominantly represented by noun or NP (Jacquemin, 2001). Hence, NP analysis is one of the important tasks for ontology construction.

It is the same as other languages that Thai NPs can be formalized as many grammatical rules. Warawudhi (2006) proposed that there are 18 Thai NPs rules. However, we can summarize to 9 rules as illustrated in Table 1. They can be categorized into two main types:

Nominal NP is the NP constructed from verb phrase by using prefixes i.e. /kan/ (-ing), /kwam/.

Non-nominal NP is the NP constructed from the head noun (e.g. common noun (ncn), collective noun (nct)) and its modifiers could be common noun, proper noun (npn), pronoun (pron), adjective (adj), determiner (det), classifier (cl) followed by determiner, verb phrase (VP), prepositional phrase (PP) or relative clause (RELCL).

Among these patterns, only some patterns are used to construct ontology concept. In this research, we extract ontology concepts that are represented in patterns NP2, NP3, NP4 and NP5 because some NP rules could not be used to extract an ontological term. For example, [[/phak/(vegetable): ncn] /lae/(and): conj [/phonlamai/ (fruit): ncn]] (vegetable and fruit) can be formalized as NP2 and NP7. For NP2, we can extract 2 noun phrases i.e. [/phak/(vegetable): ncn] and [/phonlamai/(fruit): ncn]. For NP7, we can extract only one noun phrase: [/phak/(vegetable): ncn /lae/(and): conj /phonlamai/(fruit): ncn]. However, the whole NP from NP7 should not be an ontological term because it composed of two concepts. Then, the selected ontological terms should be separated into two terms, i.e. /phak/(vegetable) and /phonlamai/(fruit).

Table 1 Grammatical rules of Thai NPs

Pattern	Example
1. Nominal NP :	
NP1 = pref + VP	[/kan/(-ing):pref /song-ok/(export):vi] (exporting)
2. Non-nominal NP :	
NP2 = (ncn|nct+ncn|npn) + NP$^{?}$	[/chuea/(pathogen):ncn /wairat/(virus):ncn] (virus pathogen)
NP3 = NP2 + adj	[/kulap/(rose):ncn /daeng/(red):adj] (red rose)
NP4 = NP + VP VP = vi|(vt+NP)	[/a-ngun/(grape):ncn /tham/(produce):vi /wai/(vine):ncn] (vine grape)
NP5 = NP + PP PP = prep + NP	[/sinkha/(product):ncn /caak/(from):prep /tangprathet/ (foreign country):ncn] (product from foreign country)
NP6 = NP + RELCL RELCL = prel + (VP|S)	[/het/(mushroom):ncn /thi/(that):prel /than/ (eat):vi /dai/(be_able):vpost] (eatable mushroom)
NP7 = NP + conj + NP	[/ma/(dog):ncn /lae/(and):conj /maeo/(cat): ncn] (dog and cat)
NP8 = NP + (det|cl+det|norm| cl+norm|num+cl)	[[/phuet/(plant):ncn /samunphrai/(herb):ncn] /klum/(group):cl /ni/(this):det] (these herbs)

Remark: adj = Adjective ncn = Common noun nct = Collective noun
NP = NP1|NP2|NP3|NP4|NP5|NP6|NP7|NP8 norm = Ordinal number marker
npn = Proper noun pref = Prefix prel = Relative pronoun
prep = Preposition PP = Prepositional Phrase S = Sentence
vi = Intransitive verb vt = Transitive verb VP = Verb Phrase
x|y = either x or y x + y = x precede y x$^{?}$ = x can occur 0 or 1 time

1.2 Level of Ontological Relationship in Texts

As mentioned previously, we can extract the ontological relationships in texts with three levels i.e. phrase, sentence and paragraph.

1) Phrase Level

Phrase does not have any concrete word to hint the ontological relationship (here after called implicit cues). However, we can extract the embedded semantic relationship between terms in NP patterns that contain nouns more than one constituent such as NP2 and NP5. For example:

(1) /pik/(wing):ncn /kai/(chicken):ncn
(Chicken wing): 'part-of relationship'

(2) /nuy/ (cheese) /jak/ (from) /nom/ (milk) /kea/ (sheep)
(Sheep cheese): 'made-of relationship'

Moreover, the pattern NP4: NP + VP can embed some semantic relationship but it is a sentence-like structure then we will process it as the sentence.

2) Sentence Level

In sentence level, the ontological elements can be detected by using lexico-syntactic patterns and verbs. In this work, we focus to extract the ontological elements only in a sentence containing the cue word of the lexico-syntactic patterns. For instance, the cue word */chen/ (such as)* can hint the hyponym relationship. However, the cue phrase can modify NP in any position then it poses the problem of many candidate terms as shown in Figure 10 and in the examples (3) and (4).

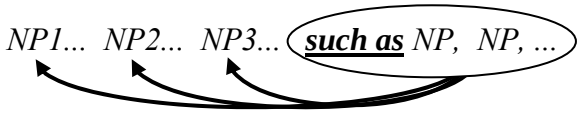


Figure 10 The cue phrase can modify any NP

(3) */pi thilaeo mi kan namkhao* ***kulap*** *chak tangprathet pen chamnuan mak daikae phan* ***sacha, mercedes*** *lea* ***gabrielle****/*

(Last year a lot of ***roses*** *have been imported from abroad such as variety of* ***Sacha, Mercedes*** *and* ***Gabrielle****.*

(4) */pi thilaeo mi kan namkhao kulap chak* ***tangprathet*** *pen chamnuan mak chen* ***itali nethoelaen sapen****/*

(Last year a lot of roses have been imported from ***abroad*** *such as* ***Italy, The Netherlands, Spain****.*

In sentences (3), the cue phrase modifies the NP that is not closed to the cue i.e. */kulap/ (rose)* but in sentence (4) the cue phrase modifies the nearest NP i.e. */tangprathet/ (aboard)*. The problem here is the attachment of the noun clause conjunction. Theoretically, this could be solved by a good parser, which can be challenging to obtain or to create one. However, creating a good parser is very expensive task then we propose the lexicon and co-occurrence features to select the correct hypernym term.

Concerning the detection of ontological elements by using verbs, it needs syntactic parser for analyzing the function of each constituent and semantic information of each term is also required for analyzing their meanings. For examples, some verbs represent the relationship of agent-patient like verb in sentence (5) and some verbs represent the semantic relationship of ontology as in phrase (6).

(5) /non/(worm):ncn /cho/(pierce): vt /kalampli/(cabbage): ncn
(Cabbage webworm): ‘agent-patient relationship’

(6) /a-ngun/(grape):ncn /tham/(produce):vt /wai/(vine):ncn
(Vine grape): ‘made-of relationship’

3) Paragraph Level

Ontology relationship that occurs in the paragraph level is very difficult to extract since it needs more process in the level of discourse processing to extract the relationship such as anaphora resolution. However, the more simple method, which has low cost, is the method that we use item list as the cue for extracting hypernym/hyponym relationship. The item list, we used here, can be classified to numbering list and item list as same as using the lexico-syntactic pattern, there are many candidate terms occurring in the preceding paragraph and these terms can be the hypernym term of the item terms. As shown in Figure 11, the preceding paragraph of the list contains many NPs, i.e. [following]$_1$, [common varieties]$_2$, …, [hundreds]$_{16}$, that can be hypernym term of the terms in the list.

There are [[hundreds]$_{16}$ of [[varieties]$_{15}$ of [**pineapple**]$_{14}$]$_{13}$]$_{12}$, ranging from [very large to miniature [size]$_{11}$]$_{10}$. There are also some [excellent [dwarf [varieties]$_9$]$_8$]$_7$ whose [core]$_6$ is edible. These mainly come from [Thailand]$_5$ and [South Africa]$_4$. Some of the [common [varieties]$_3$]$_2$ include the [following]$_1$:

1. **Sugarloaf** is a rather misleading term. Although large,…
2. **Cayenne** is relative large and cone-shaped. Its yellow flesh has …
3. **Queen** is an old variety miniature grown in South Africa. ..
4. **Red Spanish** is square-shaped, with a tough shell, and comes from …

Figure 11 An example of item list that has many ontological candidate terms

Moreover, this technique poses the problem of list identification since one document can contain many lists. There are some difficulties to identify the boundary of each list and classify the list when it contains other embedded lists. These problems will be discussed in the section of problems of automatic construction of Thai ontology in this chapter.

2. Information Gain and Information Gain Ratio

In this section, we brief the introduction of Information Gain and Information Gain Ratio based on the tutorial written by Nashvili (2004). In this work, we used the Information Gain and Information Gain Ratio for feature weighting in the process of selecting the hypernym term.

The *Information Gain* originally is the measure of goodness for attributes used in the decision tree learning algorithm C4.5 (Quinlan, 1993). It represents how precisely the attributes classify the classes (the target attribute) of data. Some attributes split the data up more purely than others, meaning that their values correspond more consistently with instances that have particular values of target attribute than those of another attribute. In another way, we can say that such attributes have some underlying relationship with the target attribute. By regarding this task as feature selection task, we can use the Information Gain as a feature weighting to decide which of the features are the most relevant in our ontology learning task.

Information Gain is defined in terms of *Entropy* that is a measurement used in Information Theory. Informally, the entropy of a dataset represents how disordered it is. It has been shown that entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, then the more information is required for describe the data. Information gain *Gain(S,A)* of attribute *A* can be calculated as follow:

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \quad (1)$$

$$Entropy(S) = \sum_{i=1}^{c} -p_i \log_2 (p_i) \quad (2)$$

where v is a value of A, S_v is subset of instances or examples of S when A takes the value v, and $|S|$ is the number of examples. The p_i is the proportion of examples in the dataset that take the i^{th} value of the target attribute.

When apply this technique to our work, S is the number of examples and A is represented by feature k. Where *Values(k)* is the set of all possible values for feature k and S_v is the subset of S for which feature k has value v. p_i is proportion of examples in class i i.e. positive (a hypernym term) and negative class (not a hypernym term).

An obvious way to negate the bias or "greediness" of Information Gain is to take into account the number of values of an attribute. This is exactly the approach that can be used. A new, improved calculation for attribute A over data S is *Information Gain Ratio*, defined as follow:

$$Gain - Ratio(S, A) = \frac{Gain(S, A)}{Entropy(A)} \tag{3}$$

A value of 0 for the gain ratio indicates that S and A have no association; the value 1 indicates that knowledge of A completely predicts S.

Information Gain and Information Gain Ratio has been widely used for feature weighting and feature selection. For instance, some studies (Ayan, 1999; Duch and Grudzinski, 1999; Mladenic, 1998) used information gain as feature weights to produce better classification accuracy. Mori (2002) utilized information gain ratio as term weighting for text summarization. Hall and Smith (1998) applied these techniques in feature selection algorithm in order to enhance the performance of machine learning.

In this work, we use Information Gain and Information Gain Ratio for weighting the features that relate to the important of each feature for selecting the ontological term. The examples of the feature (A) in this work are Name Entity (NE) term, properties term and co-occurrence feature. Concerning the NE term feature, it

has three possible values (v) i.e. 1 (when candidate term has same NE class as related term), -1 (when candidate term has different NE class as related term) and 0 (otherwise). The possible i values of the example are 1 (when it is a positive example or this candidate term is the ontological term) and 0 (otherwise). The details of the features and their weighting by using Information Gain and Information Gain Ratio are discussed in the Material and Method chapter.

3. Support Vector Machine

In this research, a support vector machine (SVM) is used to classify the semantic relation between nouns in NPs. Moreover, we applied SVM for weighting the features in the process of hypernym term selection.

SVM is a supervised machine learning technique applicable to both classification and regression proposed by Vapnik (1995, 1998). The main goal of SVM is to construct an optimal hyperplane to separate data in to two classes with a maximal margin which is the distance from the separating hyperplane to the closest data points. These points are called support vectors. Given a training set of instance-label pairs (x_i, y_i), $i = 1, \ldots, l$ where x_i is a n-dimensional feature vector $(x_i \in R^n)$ and y_i is a class label $(y \in \{1,-1\})$, SVM finds a hyperplane:

$$(w \cdot x) + b = 0 \tag{4}$$

where w is a weight vector and b is a threshold. By using this hyperplane, the examples are classified to positive class (+) or negative class (-) corresponding to decision functions:

$$f(x) = sign((w \cdot x) + b). \tag{5}$$

Figure 12 shows the hyperplane that is found by SVM for separating the data into two classes i.e. circle and rectangle.

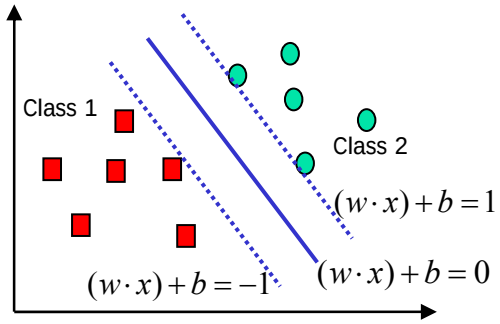


Figure 12 The optimal hyperplane separates circles from rectangles

For problems that can not be linearly separated in the input space, this machine offers a possibility to find a solution by non-linear mapping their n-dimensional input space into a high dimensional feature space, where an optimal separating hyperplane can be found. This non-linear mapping function is called the kernel function,

$$K(x_i, x_j) \equiv \phi(x_i)^T \phi(x_j). \tag{6}$$

There are the following four basic kernels:

- linear: $K(x_i, x_j) = x_i^T x_j$ (7)
- polynomial: $K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \gamma > 0$ (8)
- radial basis function (RBF): $K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|^2), \gamma > 0$ (9)
- sigmoid: $K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r)$ (10)

Here, γ , r, and d are kernel parameters.

In this work, we experiment by using only linear kernel since there is difficulty for setting the proper parameters of other kernels. Concerning the semantic relation classification of nouns in NPs, we have three sets of features: semantic classes of all head nouns, semantic classes of all modifier nouns and all semantic classes of prepositions. From the experiment with 1055 pairs of NPs' head and modifier, there are totally 936 features. In this work, we test with 10 semantic relations and the outputs of the system are 10 models or 10 hyperplanes for classifying the examples as positive or negative example of each relation. The system will select the relation that gives the maximum result (maximum distance between the example and the hyperplane) as the relation of nouns in NPs. The details of this experiment are discussed in the Material and Method chapter.

Related Works of Ontology Learning and Integration

There are a number of researches related to ontology learning. These researches focus on different types of source, methodologies and applied domain and this section presents a survey of the most relevant methods and techniques for building ontologies from multi-sources: text, dictionaries and thesaurus. Next section briefs the methodologies for integrating ontology from multi-sources to be the rich one.

1. Ontology Learning based on Unstructured-Text

There are many ontology extraction methods based on the usage of Pattern-based Technique, statistical techniques, natural language analysis techniques and the combination of these methods. The most well known techniques are 1) Pattern-based approach, 2) Statistical-based approach and 3) the Combination of methods i.e. linguistic approach, pattern-based approach, statistical-based approach and machine-learning approach.

1.1 Pattern-based Approach

Pattern-based approach is a heuristic method, where text is scanned for lexico-syntactic patterns that described hyponym/hypernym or meronym relation between concepts. Based on this technique, expert needs to define lexico-syntactic patterns which require a lot of time. Hearst (1992) describes a procedure, called hyponymy pattern approach, for automatic extracting relationships between concepts and adding them into an existing ontology (e.g. WordNet). It consists of 2 steps: 1) looking for concepts from texts that are related to an existing ontology and 2) determining whether they are associated to other concepts with a lexico-syntactic pattern. For example, the lexico-syntactic pattern of hyponym relation:

```
...NP {,NP} * {,} or other NP …
```

We can infer that NPs on the left of 'or other' are sub concepts of NP on the right of 'or other'. For example, the following sentence can be extracted three semantic relations: HYPONYM(*Bruise, Injury*), HYPONYM(*Wound, Injury*), HYPONYM(*Broken-bone, Injury*).

Bruises, wounds, broken bones or other injuries are common.

Later, Finkelstein-Landau and Morin (1999) add implementation to the Hearst's work by automatically generalizing of lexico-syntactic patterns. The generalizing relies on a syntactic distance between patterns. The system has two functionalities; the first functionality is the acquisition of lexico-syntactic patterns from corpus with respect to a specific conceptual relation. The experts define a list of terms' pairs linked by the conceptual relation. This list of terms is used to find sentences that contain the terms and the system will find a common environment that generalizes the lexico-syntactic expressions from collection of sentences extracted at the previous step. For instance, the relation HYPERNYM(*vulnerable area*, *neocortex*) is used to extract the sentence from the corpus:

Neuronal damage was found in the selectively vulnerable areas such as neocortex, striatum, hippocampus and thalamus.

The sentence is then transformed into the following lexico-syntactic expression:

```
NP find in NP such as LIST
```

Similarly, from the relation HYPERNYM*(complication, infection)*, the following sentence is extracted from the corpus.

Therapeutic complications such as infection, recurrence, and loss of support of the articular surface have continued to plague the treatment of giant cell tumor.

And it is produced to the lexico-syntactic pattern as following.

```
NP such as LIST continue to plague NP
```

The common pattern of these two patterns is:

```
NP such as LIST
```

After that, the expert will validate the lexico-syntactic patterns. The second functionality is the extraction of pairs of conceptual related terms through a database of lexico-syntactic patterns. They present in the paper that this method can find only small portion of related terms due to the variety of sentence styles and the inability to find a common environment to all those sentences.

The purpose of this ontology learning is to extend existing ontologies with new concepts and new relationships among the existing concepts in the original ontology. The pattern-based technique is applied in the ontology learning studies by Maedche and Staab (2000), Kietz *et al.* (2000), Shamsfard (2003) and others. The

crucial problem of this method is data sparseness. Some works overcame this problem by searching the patterns in the WWW by using the search engine e.g. Google. However, this solution can not be used with Thai language since Thai needs the pre-process to identify the word boundary but all search engines do not process this task. Then, we can not search the Thai patterns in the WWW. In addition, we overcome this problem by applying the combination methods of rule-based and statistical approaches. (See more detail in Materials and Methods chapter)

1.2 Statistical-based Approach

Many statistical-based approaches especially clustering methods are proposed for ontology construction. The methods are performed by transferring the information of term's occurrences in context into a feature vector of term. This feature vector of term is represented the meaning of terms. Clustering of feature vectors can be used to investigate the relations between groups of similar term.

Many studies are proposed by using clustering method based on different feature vectors as follows:

Agirre *et al.* (2000) present the method that exploits the text from the Web to enrich the concepts in the WordNet (Miller, 1995) ontology. The proposed method constructs lists of topically related words for each concept in the WordNet, where each word sense has one associated list of related words. For example, the word "*waiter*" has two senses: '*waiter in the restaurant*' and '*person who wait*'. The associated list of the first sense will contain *waiter-restaurant, menu, dinner, etc*. while the words in the associated list of the second sense are *waiter-station, airport, hospital, etc.* The system queries the web for the documents related to each concept from the WordNet and then extracts the words and their frequencies using a statistical approach. The words that have distinctive frequency are grouped in a list that is called topic signatures. Then the concepts are hierarchically clustered based on their topic signatures.

Another method, which has been introduced by Faure and colleague (Faure and Nedellec, 1998; Nedellec, 2000), implements the system ASIUM to learn sub-categorization frames of verbs and ontologies from syntactic parsing of technical texts in natural language (French). The inputs of the ASIUM are the results from syntactic parsing of texts. The outputs are sub-categorization examples and basic clusters formed by head words that occur with the same verb after the same preposition (or with the same syntactical role). On each level, it allows the expert to validate and label the concepts. The system generalizes the concepts that occur in the same role in the texts and uses generalized concepts to represent the verbs. However, ASIUM is suitable for technical corpora and considers only head nouns of terms.

Lin and Pantel (2002) present a clustering algorithm called CBC (Clustering By Committee) that automatically discovers concepts from text. It can handle a large number of elements, a large number of output clusters, and a large sparse feature space. It initially discovers a set of tight clusters called committees that are well scattered in the similarity space. The centroid of the members of a committee is used as the feature vector of the cluster. They proceed by assigning elements to their most similar cluster.

Furthermore, statistical analysis of co-occurrence data is used to learn conceptual relations from texts proposed by Yamaguchi (2001). He proposes DODDLE II, a Domain Ontology rapiD DeveLopment Environment. The system constructs domain ontologies by exploiting WordNet and domain-specific texts. The taxonomic relationships come from WordNet that match with the domain terms given by the users. Then, the system reconstructs the ontology tree by analyzing trimmed result. The non-taxonomic relationships come from domain specific texts with the analysis of lexical co-occurrence statistics, based on word space. The relation of each concept pair is considered by the similarity between their vectors in word space.

This statistical approach is automatic at the starting step but it needs user validation at final step for concept's cluster labeling and relation labeling and it needs a lot of corpus to measure statistical value for ontology construction. However, this

approach can process with a various types of data's features and it needs less preparation data than machine learning techniques that need a lot of training data.

1.3 Combination of Methods

Most systems use combination approaches to learn the ontology. They apply multi learning algorithms, e.g. decision tree and neural networks, to learn different components and to enrich the ontology. For example, Maedche and Staab (2001) using association rules and clustering techniques, Moldovan and Girju (2001) combining pattern-based technique and machine learning, Ontolearn (Roberto Navigli, *et al.*, 2003) applying linguistic approach and machine learning technique and HASTI (Shamsfard and Barforoush, 2003) applying a combination of logical, linguistic based, template driven and heuristic methods.

Maedche and Staab present an algorithm for semiautomatic ontology learning from texts. They apply data mining algorithm in the term of association rules to analyze statistical co-occurrences of term appearing in text. The input data is a set of transactions, each of which consists of a set of items that appear together in the transaction. For example, the following sentences can be generated the concept pairs as shown in Table 2.

```
All rooms have TV, telephone, modem and minibar.
Mecklenburg hotel is located in Rostock.
```

Table 2 Examples of concept pairs extracted from the sentences

$Term_1$	**$concept_1$**	**$Term_2$**	**$concept_2$**
room	room	TV	television
Mecklenburgs	area	hotel	hotel

After that, the super classes of each concept are added to each transaction e.g. $transaction_1$:= {room, television, furnishing}; furnishing is the super class of

television in WordNet. The algorithm extracts association rules represented by sets of items that co-occur sufficiently often and present the rules to the knowledge engineer. Table 3 shows the discovered relation and their confidence and support values that extracted from the previous sentences. The relations that have the lower confidence and support values than threshold will be pruning. The ontology learning system applies this method straightforwardly for ontology learning from texts to support the knowledge engineer in the ontology acquisition environment.

Table 3 Examples of discovered relation and their confidence and support values

Discovered relation	Confidence	Support
(room, furnishing)	0.39	0.03
(room, television)	0.29	0.02
(area, accommodation)	0.38	0.04
(area, hotel)	0.1	0.03

The next method is represented by Kietz *et al.* (2000). This method aims to prune an existing general ontology, e.g. WordNet, and to enrich it with new domain concepts and relations among them. It is a semi-automatic process. The method is based on the assumption that most concepts and concepts' relations of the domain to be are included in an ontology as well as the terminology of a given domain are described in documents. The authors propose to learn the ontology using as a core ontology (e.g. SENSUS, WordNet, etc.) that is enriched with new specific domain concepts. New concepts are identified using noun phrase analysis techniques over the sources previously identified by the users. The output ontology is pruned and focused to a specific domain by the use of several approaches based on statistics. For example, the terms that are more frequently occur in a domain-specific corpus than in a generic corpus should be proposed to be incorporated to the ontology. Finally, non-taxonomic relations between concepts are learnt applying learning methods based on the association rule's algorithm.

Moldovan and Girju (2001) present a method for discovering domain-specific concepts and taxonomic relationships in an attempt to extend an existing ontology, like WordNet, with new knowledge acquired from parsed text. The sources for discovering new knowledge are general domain corpora, and are augmented by using other lexical resources like domain specific and general dictionaries. The user provides a number of domain-specific concepts that are used as seed concepts to discover new concepts and relations from the sources. The users perform the validation of the process and confirm the correctness of the new concepts and relations learnt. In addition, they apply this approach for semi-automatically detecting part-whole relations (Girju and Moldovan, 2003). The system discovers the part-whole lexico-syntactic patterns (for example, the horn is part of the car.) and learns the semantic constraints needed for the disambiguation of these generally applicable patterns. Through this research, the system combines the learning results with the IS-A relation in WordNet for more accurate learning.

The other system, Ontolearn, is introduced by Navigli *et al.* (2003). The system has been developed and tested in tourism domain. It builds trees of domain concepts and combines them with existing core domain ontology. Terminologies from a corpus of domain text are extracted and filtered by using natural language processing and statistical techniques that perform comparative analysis across different domains corpora to extract terminologies. The authors use WordNet and SemCor (Miller *et al.*, 1993) as a source of prior knowledge for semantically interpreting the terms. WordNet and rule-based inductive-learning method are used for extracting domain specific relations of tourism e.g. ‘TIME’, ‘THEME’, etc. The system creates the domain ontology by integrating the taxonomy with core domain ontology. If the existing domain ontology is not available, the method proposes for creating a new one from WordNet, pruning concepts that are not related to the domain, and extending it with the new domain concept trees under the appropriate nodes.

Ketsuwan *et al.* (2000) propose the methodology for constructing Thai Thesaurus from dictionary and unstructured texts. They analyze the structure of

terms' definition in dictionary by using some heuristic rules and generate the taxonomic relation. The non-taxonomic relations are constructed by analyzing the frequencies of the co-occurrence of terms in the texts. Cimiano et al. (2004) learn taxonomic relations by considering various and heterogeneous forms of evidence. For example, they match Hearst patterns in a large text corpus and the WWW. Besides, they use linguistic technique for analyzing the head word of NPs. Schutz and Buitelaar (2005) proposes the *RelExt* system for extracting relevant verbs and their grammatical arguments (i.e. terms) from a domain-specific text collection and computing corresponding relations through a combination of linguistic and statistical processing. Pantel and Pennacchiotti (2006) proposes the *Espresso* system for harvesting binary semantic relations from raw texts by exploiting generic patterns for filtering incorrect instances from the Web and measuring the reliability of pattern and instance.

All of the approaches described above propose the methodologies for extracting concepts and relations between the concepts, except for HASTI (Shamsfard and Barforoush, 2003) that learns axioms beside concepts, taxonomic and non-taxonomic relations. It is an automatic ontology building system, which learns the ontology from Persian texts. The system starts from a small-scale ontology made by hand and the learning approach of the system is a hybrid approach i.e. a combination of linguistic, template (lexico-syntactic pattern) driven, logical and semantic analysis methods. The linguistic-based approach is applied for extracting case roles and template driven technique is used to extract concepts and relations between them. Logical approach is applied by inference engine to deduce new knowledge (new relations between concepts and new axioms). Furthermore, the system performs online and offline clustering to organize its ontology based on semantic analysis methods with several heuristics such as a pruning heuristic rule: *An unnecessary node, which is not referred to by any lexical unit and its own feature (property) set is empty, should be deleted and its children should be transferred to under its immediate father*.

This combined methods approach is among the most promising ones in this area because the combination technique can learn different ontological elements

(concepts, taxonomic relations and non-taxonomic relations) and increase accuracy and coverage of the knowledge in the ontology.

Figure 13 shows research map of ontology learning classified by approach and ordered by time. The Figure presents that most of researches usually studied both taxonomic and non-taxonomic relationships by using the combination of methods. In addition, the strengths and weaknesses of each approach are summarized in Table 4.

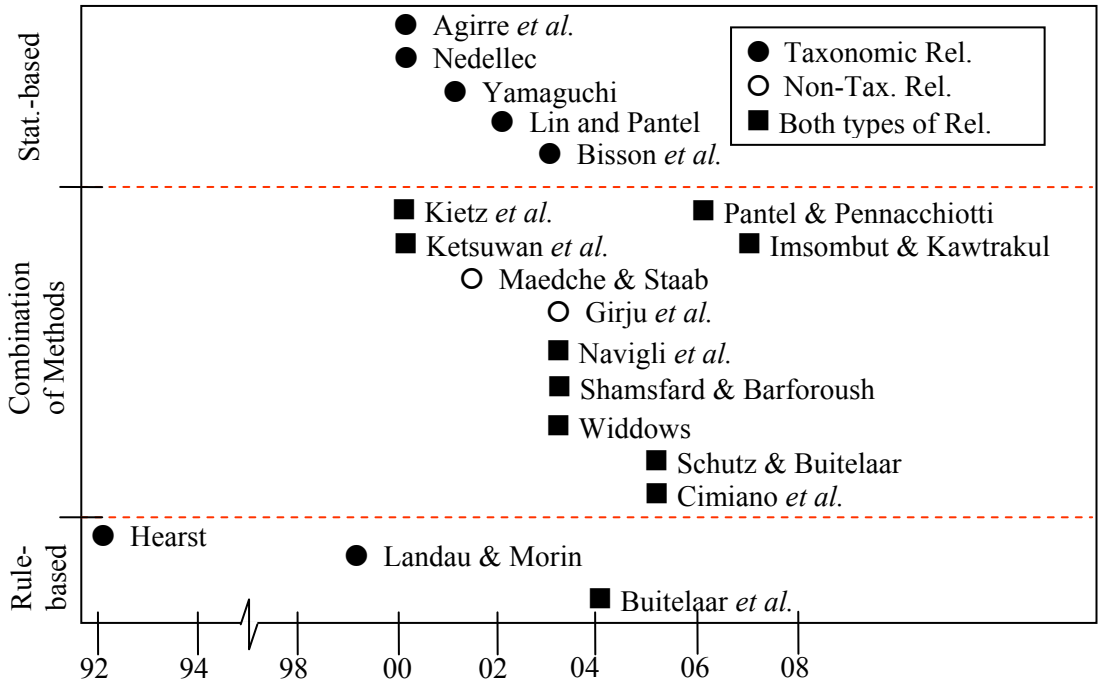


Figure 13 A research map of ontology learning

Table 4 Summary of ontology learning approaches from unstructured-text

Approaches	Researches	Strengths	Weaknesses
Pattern-based Technique	• Hearst (1992) [PC=N/A] • Morin (1999) [PC=79%] • Buitelaar *et al.*, 2004 [PC=N/A]	General patterns can be used in other domains.	- Must pre-defined extraction pattern - Cue word ambiguity - Data sparseness
Statistical-based Approach	• Bisson *et al.* (2003) [PC=N/A] • Nedellec (2000) [PC=48%] • Agirre *et al.* (2000) [PC=N/A] • Lin and Pentel (2002) [PC=N/A] • Yamaguchi (2001) [PC=N/A]	Can process on a huge data and a lot of features	- Need expert to label the cluster node - usually used to extract only taxonomic relation - Data sparseness
Combination of Methods	• Kietz *et al.*, 2000 [PC=20%] • Kedsuwan, 2000 [PC=N/A] • Maedche and Staab, 2001 [PC=11%, RC=13%] • Girju *et al.*, 2003 [PC=84%, RC = 98%] • Navigli *et al.*, 2003 [PC=73%, RC=53%] • Shamsfard and Barforoush, 2003 [PC=N/A] • Widdows, 2003 [PC=N/A] • Schutz & Buitelaar, 2005 [PC=24%,RC=36%] • Cimiano *et al.*, 2005 [PC=16%, RC=30%] Pantel & Pennacchiotti, 2006 [PC=N/A]	Can extract both taxonomic and non-taxonomic relation	- Must pre-defined extraction pattern for pattern-based technique - Need a lot of learning examples for machine learning technique

2. Ontology Learning based on Thesaurus

There are a few works that utilize thesaurus for ontology construction. They convert UF (Used For), BT (Broader Term) and NT (Narrow Term) relationship of thesaurus to ontology's Synonym, Hypernym and Hyponym relationship, respectively. For example, Clark *et al.* (2000) construct ontology from Boeing Thesaurus and this ontology is applied in document retrieving task. They convert the BT/NT relation to Hypernym/Hyponym relation and add more semantic relation between words in phrase by using the rules that expert defined. For example,

```
IF modifier is a Material AND head is a Physical-Object
THEN head is-made-of modifier
```

For instance, the phrase 'metal tube' can be applied with this rule by using the data in the thesaurus to identify the class of each term that 'metal' is Material and 'tube' is a Physical-Object.

Wielinga *et al.* (2001) convert AAT Thesaurus (Art and Architecture Thesaurus) into an ontology, where each concept has a labeled slot corresponding with the main term in AAT. It is represented in RDFS (Resource Description Framework Schema) for indexing and retrieving image information about art objects, in particular antique furniture. In the first step, the expert will construct a description template of antique furniture such as production-related descriptors (e.g., creator, style/period, etc.), physical descriptors (e.g., measurements, color, material, etc.). Next step, these descriptors are linked to specific subsets of AAT that can be used as values of furniture properties. Finally, the expert will describe additional domain knowledge, in particular about constraints between furniture-property values. For example, knowledge about the relationship between style periods and furniture characteristics (e.g. Late Georgian chests-of-drawers were typically made of mahogany).

Soergel *et al.* (2004) propose the rules-as-you-go approach where rules for semantic refinement are identified as experts work on the thesaurus and notice patterns in the occurrence of semantic relationships between terms. For example, milk NT milk Fat. This relation can be converted to 'containsSubstance' relation and the rule for converting this relation is:

```
IF Substance X NT Substance Y
THEN Substance X <containsSubstance> Substance Y
```

Since the patterns and rules are identified through human expertise, the refinements occur gradually and can deal with only a limited number of patterns.

For Thai language, the approach of rules-based can be applied in Thai because the constraint of the rules contains the concepts of terms in the thesaurus that can be applied in any language. In my research, the rules are automatically generated by applying machine learning technique instead of expert defining. The system generates the rules from the trained examples in order to identify the ontological relationship of terms (more details see in Materials and Methods chapter).

Table 5 Summary of ontology learning approaches from thesaurus.

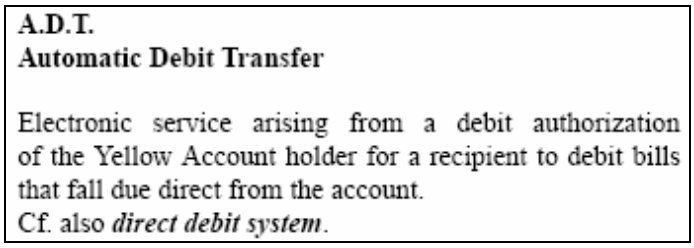
Approaches	Researches	Strengths	Weaknesses
Rule-based Approach	• Soergel (2004) • Wielinga (2001) • Clark *et al.* (2000)	Not complex	Need expert to define rules

3. Ontology Learning based on Dictionary

There are many studies on utilizing the dictionary in the task of ontology construction. Ontological terms and relations can be generated by analyzing terms and their definitions from dictionary with statistical technique and heuristic rules.

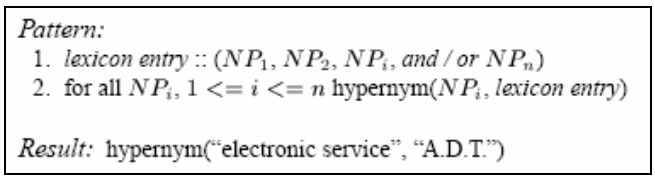
Janniak (1999) applies PageRank algorithms for automatically extracting hierarchical relationships from an on-line Webster's dictionary. Each head word and its definition group are nodes and each word in a definition of node is used to make an arc to the node of this head word. After that, the system computes a relative measure of arc strength and ranks them. The arc having the rank value over a threshold is accepted. The outputs are the nodes and their relations to other nodes.

Kietz (2000) uses several heuristic rules to build taxonomy. The first example is the heuristic rule that matching pattern to texts as shown in Figure 14 and 15. From these Figures, the lexicon entry is *A.D.T.*, the NP is *Electronic service* and the system can extract the *hypernym(electronic service, A.D.T.)* relationship. Another heuristic rule deals with compound noun, for example "unemployment benefits", which the head noun of the compound noun, i.e. "benefits" in this example, is a hypernym of the whole compound.


A.D.T.
Automatic Debit Transfer

Electronic service arising from a debit authorization of the Yellow Account holder for a recipient to debit bills that fall due direct from the account.
Cf. also ***direct debit system***.

Figure 14 An example entry in dictionary
Source: Kietz (2000)


Pattern:
1. *lexicon entry* :: (NP_1, NP_2, NP_i, *and / or* NP_n)
2. for all NP_i, $1 <= i <= n$ hypernym(NP_i, *lexicon entry*)

Result: hypernym("electronic service", "A.D.T.")

Figure 15 Example of patterns and the result by using this pattern
Source: Kietz (2000)

Kang (2001) derives case relations (e.g. agent, theme, recipient, etc.) between concepts from semantic information in the Sejong electronic dictionary by using specific rules, thesaurus and human intuition. The specific rules are inferred from training samples and the class in thesaurus is used to tag sense of word in dictionary. An example of rule:

```
IF Subject= life THEN relation=agent
```

As mentioned above, ontology construction based on dictionary can use both information from the terms and their definitions to analyze the relationship between terms. Moreover, dictionary can be extracted ontology by using the same methodologies as unstructured text based ontology construction. For example, the analysis of head word of NPs and the analysis of definitions of the terms that have difficult problems similar to unstructured text. However, some dictionaries have a specific structure that is useful for analyzing and generating ontological relation. Similarly, we can analyze the relationships of plant's family/sub-family/genus from structure of Thai Plant Name Dictionary (Smitinand, 2001) and convert them to hypernym/hyponym relations of ontology.

Table 6 Summary of ontology learning approaches from dictionary.

Approaches	Researches	Strengths	Weaknesses
Rule-based Approach	• Kang *et al.* (2001) • Kietz *et al.* (2000)	Not complex	Expert must defines the heuristic rules
Statistical-based Approach	• Jannink (1999)	Can process on a huge data	Can not define relation types

4. Ontology Integration

Since the ontologies extracted from various sources, i.e. unstructured text, thesaurus and dictionary, have many similarities and differences of concepts and relationships, the integration system is needed for integrating them together. There are

many existing researches worked with this term mismatch problem and they will be discussed throughout this section.

Ontology integration or ontology merging is an interesting issue of ontology research since there are many ontologies that are constructed and available on the web and single ontology is not enough to support distributed environment tasks. Multiple ontologies need to be accessed from several applications. The main approaches of the studies of ontology integration are rule-based, statistical-based and machine learning approach as follows.

There are two main researches appling the rules-based approach for ontology integration: PROMPT and Chimaera. PROMPT (Noy and Musen, 2000) is an algorithm that provides a semi-automatic approach for ontology merging and alignment. When an automatic decision is not possible, the system will provide the guidance for the user to performing the tasks. The algorithm starts with the process of automatically executes additional changes based on a set of knowledge-base operations. Next, the system will generate a list of suggestions based on the structure of the ontology to the user for selecting, and determines conflict of relations in the output ontology and finds possible solutions for those conflicts. McGuinness and colleagues (2000) propose similar tools as PROMPT, Chimaera, for merging ontologies. Chimaera supports users for reorganizing taxonomies, resolving name conflicts, browsing ontologies, editing terms, etc. The system suggested for merging names of classes or slots based on the similarity of the names. When comparing it with PROMPT, they are quite similar in that they are embedded in ontology editing environments, but they differ in the suggestions they made to their users with regard to the merging steps.

FCA-Merge (Stumme and Maedche, 2001) is a method for merging ontology by using statistical-based approach. It is based on Formal Concept Analysis (Ganter and Wille, 1999) and lattice of concept exploration. The assumption is that the concepts are identical or similar if they occur in the same set of documents. The inputs of the system are the two ontologies that will be merged and the set of

documents related to these ontologies. The instances of the ontologies are extracted from the documents. The concepts that have the instances occurring in the same document are merged together. The last process is pruning for deleting the incorrect relation based on the defined constraint.

Doan and colleagues (2004) develop a system, GLUE, which employs machine learning techniques to semi-automatically create semantic mappings between ontologies. GLUE uses a multi-learning strategies approach because there are many different types of information that can be represented as the membership of an instance e.g. its name, value format of instance's properties, the word frequencies of their values, and each of these information types is best utilized by a different learner with specific learning algorithm. The system combines all predictions from a set of learners by using a meta-learner. Finally, the system uses a relaxation labeling technique that assigns labels to nodes of a graph based on domain constraints and general heuristics.

Table 7 Summary of ontology integration approaches.

Approaches	**Researches**	**Strengths**	**Weaknesses**
Rule-based Approach	• Noy *et al.* (2000) • McGuinness *et al.*(2000)	High precision in specific domain	Does not work in general terminologies
Statistical-based Approach	• Stumme *et al.* (2001)	- Can process on a huge data - Do not prepare the training data	Can not identify relation types
Machine-learning Approach	• Doan *et al.* (2004)	- Can process on a huge data	Does not work well if training data is insufficient.

Problems of Automatic Construction of Thai Ontology

The problems of automatic construction of Thai ontology are described according to the resources (text and thesaurus) and the integration problems. Since specific dictionary-based ontology construction in this research uses only task-oriented parser to analyze the structure then it does not have any crucial problem. The general dictionary-based ontology construction usually analyzes from the definition of term then it can pose the problems as text-based ontology construction.

1. Problems with the Acquisition from Text

The main processes of ontology building are the identification of the related ontological terms and relations. In this research, we use cues: lexico-syntactic patterns and item lists to identify that the terms occurring with these cues are indeed related. Lexico-syntactic patterns are a frequently used technique, but they are not sufficient to allow the extraction of all ontological terms. We also use item lists as additional cues for the identification of hypernym-hyponym terms. We found that item lists are very frequent in a document; hence, they are a promising technique to be used for this task. Moreover, ontological relations can occur in corpora without any explicit cues such as complex NPs (the x of the y). However, these techniques pose certain problems and we can classify them into different groups as follows.

1.1 Concept and Concept Boundary Identification

Expected concept may be either phrase or some part of phrase. e.g.

(7) */phuet phak samunphrai/*
(herb vegetable plant)
(8) */phuet phak samunphrai thi niyom pluk tam ban/*
(herb vegetable plant that was usually cultivated at home)

In the example, phrase (7) is composed of many nouns and noun phrase and the system can generate many concepts from this phrase i.e. */phuet/(plant), /phak/(vegetable), /samunphrai/(herb), /phuet phak/(vegetable plant)* and */phuet phak samunphrai/(herb vegetable plant).* Concerning the phrase (8), only head word of noun phrase, i.e. */phuet phak samunphrai/(herb vegetable plant*), will be selected to be considered as the ontological term. The system needs to decide which term is an appropriate ontological term. We solve this problem by selecting the term that usually co-occurs in the corpus with the related terms.

1.2 Ambiguity concerning the sense of the 'cue words'

Using cue words, such as "/dai-kae/(i.e.)", "/chen/(for example)" and "/pen/(is)", for hinting relationships of terms is a technique for ontology learning, but a word might have several functions and several meanings. For example, a cue word like "/pen/(is)" might signal a "hypernym", a "disease" or a semantic "property":

(9) /*kalam-pli pen phuet phak chanit nueng* /
*(**Cabbage** is a kind of **vegetable**.)*.
*(10) /*kalam-pli pen rok-nao-le*/
(Cabbage has disease as Soft-rot.).
*(11) /*kap-bai pen si-namtan*/
(*Leaf is brown color*.)

In example (9), the cue word "/pen/(is)" signals a hypernym relation, while in the others it does not. We solve this problem by utilizing Name Entity and property list as features for pruning inappropriate relations.

1.3 Ambiguities concerning the ontological relation embedded in NPs

Ontological terms and their relations can be embedded not only in the sentence-, but also in the Noun-Phrase-level. The problem in the latter case is how to identify the semantic relation between the nouns, since it is implicit. Moreover, in the case of compound nouns, we need to identify the correct ontological terms before being able to mine their relationships. For example,

(12) */pui/(fertilizer):ncn /in see/(organic):ncn*
(organic fertilizer)
(13) */pui/(fertilizer):ncn /nai tro chen/(Nitrogen):ncn*
(Nitrogen fertilizer)
(14) */kuad/(bottle) /nam/(water) /plad-sa-tik/(plastic)*
(plastic water bottle)
(15) */kuad/(bottle) /nam/(water) /phon-la-may/(fruit)*
(fruit juice bottle)

The two nouns of noun phrases (12) and (13), have the same patterns; however, they could express different semantic relationships, namely '*made-of*' and '*composed-of*'. In (14) and (15), they have different segmentations, the one in (14) is /*plastic/-/water bottle*/, while the other is */fruit juice/- /bottle/* and they express different semantic relationships i.e. *'made-of*' and *'container'* relationships.

1.4 Problems of Item List Identification

Since the input of our system is plain text, we do not have any markup symbols to show the position and the boundaries of the list. Then, we used bullet symbols and numbers to indicate the list, but this technique has several problems (see Figure 16 and 17).

- *Long description in each list item.* Since some item lists may have long descriptions, it is difficult to decide whether the focused item is meant to continue

from the previous list or to start a new list. For example in Figure 16, these items can be classified into one list or two lists. If we consider only the bullet symbol, the 'Brown Spot' item can be continuous list of previous list or start a new list. Accordingly, we need to identify the meaning of each item of each list for identifying the boundary of the list. In this work, we applied NE class e.g. plant name, animal name and disease name for identifying the class of item list.

• *Embedded lists.* It frequently happens that a list contains another list, causing some identification problems. We solve this issue by detecting each list following the same bullet symbol or numbering order. Still, there is case which an embedded list may has a following number as the third item that is shown in Figure 17. In this case, we assume that different lists mention different topics; hence the meaning of each item of each list, e.g. plant, animal, etc., can solve this problem.

• *Ambiguity between non-ontological/ontological list items.* Authors frequently express procedures and descriptions in list form. However, the procedure list items are not the domain's ontological terms, and some description list items may not be ontological terms at all. For instance, as shown in Figure 17, the lists about treatment and protection of cabbage's pest are not the ontological list. Hence, the system needs to detect either the ontological list or the non-ontological list.

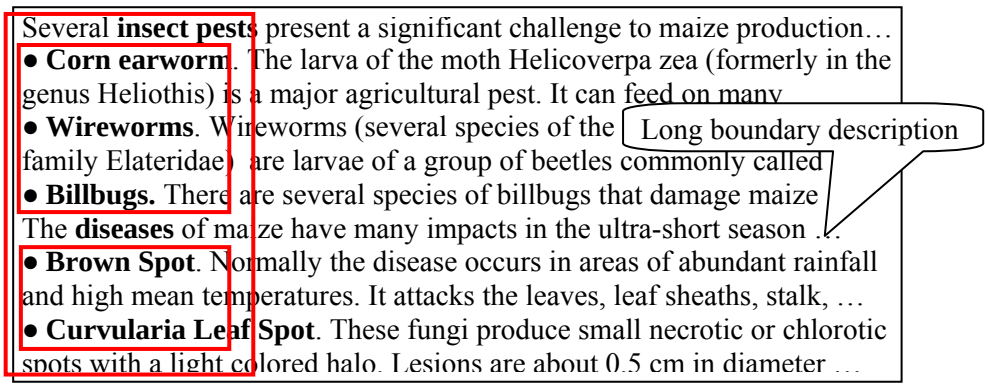


Figure 16 An example of the long description in each list item problem

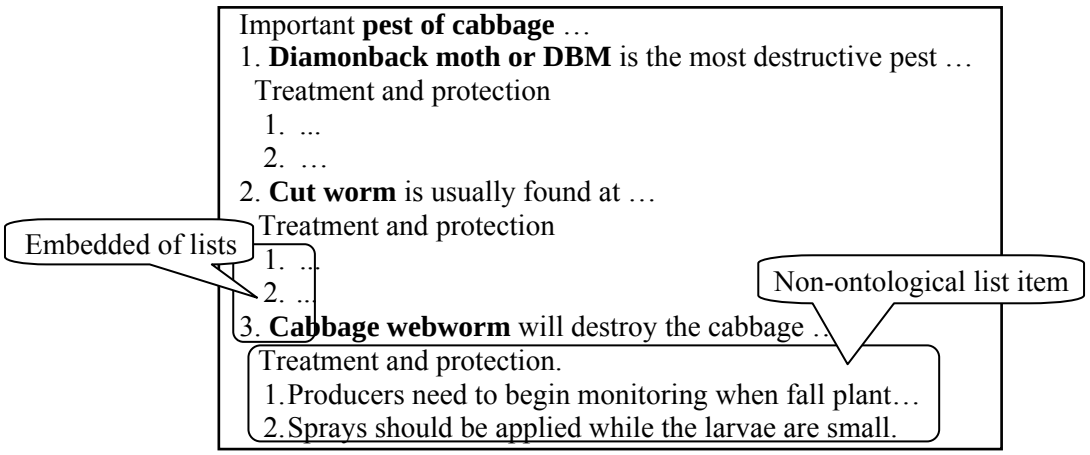


Figure 17 Examples of embedded lists and ambiguity between non-ontological/ontological list item problems

1.5 Candidate Term Selection

When both cues (lexico-syntactic patterns and item lists) are used to identify the related terms, they also pose a problem that there are many candidate terms for being an ontological term. In our texts, we often found that the term, which we are interested in, can be very far from the related terms. In addition, the ontological term can be in any position of the sentence. For example,

(16) */pi thilaeo mi kan namkhao* ***kulap*** *chak tangprathet pen chamnuan mak daikae phan* ***sacha, mercedes*** *lea* ***gabrielle****/*

(Last year a lot of ***roses*** *have been imported from abroad such as variety of* ***Sacha, Mercedes*** *and* ***Gabrielle****.*

(17) */pi thilaeo mi kan namkhao kulap chak* ***tangprathet*** *pen chamnuan mak chen* ***itali nethoelaen sapen****/*

(Last year a lot of roses have been imported from ***abroad*** *such as* ***Italy, The Netherlands, Spain****.*

Both sentences (16) and (17) have two candidate terms: *rose* and *abroad*, while the correct ontological term of (16) is *rose*, the correct ontological term of (17) is *abroad*. In addition, from the corpus observation, we found that there is 53% that

the related terms are far distance from each another, especially for the cue words /daikae/(such as) and /chen/(such as), as shown in Table 8. The problem here is the attachment of the noun clause conjunction. Theoretically, a good parser could solve this. However, it is very challenging to obtain or to create one and Thai parser does not exist right now. Hence, we propose to solve this problem by using lexical and contextual features. This solution is less expensive than generating a good parser. The next chapter will describe the details of this method.

Moreover, there is a problem concerning the selection of list item's hypernym. Since all the terms in the previous paragraph of the first item are candidates as a hypernym term. As shown in Figure 18, there are 16 candidate terms. The system also uses the lexical and contextual features for selecting the appropriate hypernym term.

Table 8 The statistics of the ontological relation occurrence classified by the distance of the related terms

Distance between the related terms	**Cue words (times of occurrence frequency)**			**Total**
	/pen/(is)	**/chen/(for example)**	**/dai-kae/(i.e.)**	
Far distance	10	81	79	**170 (53.12%)**
Adjoined	87	33	30	**150 (46.88%)**
Total	97	114	109	**320**

There are $[[\text{hundreds}]_{16}$ of $[[\text{varieties}]_{15}$ of $[\textbf{pineapple}]_{14}]_{13}]_{12}$, ranging from $[\text{very large to miniature } [\text{size}]_{11}]_{10}$. There are also some $[\text{excellent } [\text{dwarf } [\text{varieties}]_9]_8]_7$ whose $[\text{core}]_6$ is edible. These mainly come from $[\text{Thailand}]_5$ and $[\text{South Africa}]_4$. Some of the $[\text{common } [\text{varieties}]_3]_2$ include the $[\text{following}]_1$:

1. **Sugarloaf** is a rather misleading term. Although large,…
2. **Cayenne** is relative large and cone-shaped. Its yellow flesh has …
3. **Queen** is an old variety miniature grown in South Africa. ..
4. **Red Spanish** is square-shaped, with a tough shell, and comes from …

Figure 18 An example of item list that has many ontological candidate terms to be a hypernym term

Table 9 The statistics of the ontological relation occurrence classified by the characteristic of the occurrences

Characteristic of the occurrence	Numbers of the occurrence
1. Cue Word Expression	215 (38.19%)
2. Hypernym Relation in NP	131 (23.27%)
3. Semantic Relation in NP	133 (23.62%)
4. Bullet & Numbering	84 (14.92%)
Total	**564 (100%)**

Table 9 shows the statistics of the ontological relation occurrences, classified by the characteristics of the occurrence. Hence, in this study, we propose the methodology for the ontology construction by classifying to two main tasks. First, ontologies are extracted by using cues that are lexico-syntactic patterns and item list (i.e. bullet list and numbering list). The main advantage of the approach is that it simplifies the task of the concept and the relation labeling since using cues could help in identifying the ontological concept and hinting their relations. Second, Relations embedded in Thai NPs that included both hypernym and semantic relations are analyzed by applying machine learning.

2. Problems with the Acquisition from AGROVOC

AGROVOC is a good resource but it is imperfect, as some of its relations are assigned incorrectly and too broadly defined.

2.1 Incorrectly assigned relationships

A review of the data in AGROVOC reveals that some USE/UF (Use/Use For) and BT/NT (Broader Term/Narrow Term) relationships are incorrect or reflect inconsistent uses of the relationships. The USE/UF relationship links, not only synonyms but also quasi-synonyms, such as closely related and hierarchically related

terms (Soergel *et al.*, 2004). Likewise, the BT/NT relationship is highly ambiguous (see examples in Table 10).

As shown in Table 10, AGROVOC incorrectly uses ***NT*** (*narrow term*), approximately equivalent to '*superclassOf*' or '*hypernymOf*', in *Milk* ***NT*** *Milk Fat*, while a more specific, and probably more correct relationship would be '*containsSubstance*'.

Table 10 Examples of inappropriately defined relationships between terms

Relationship	Examples	Remark
UF	1. *Locomotion* UF *Walking*	Incorrect Relationship: *Walking* is not a synonym of *Locomotion*. WordNet shows that *Walking* is the hyponym of *Locomotion.*
	2. *Digestive juices* UF *Chyme*	Incorrect Relationship: *Digestive juices* is not a synonym of Chyme, and the two terms have different hypernyms in WordNet.
BT/NT	1. *Milk* NT *Milk fat*	Incorrect Relationship: *Milk* <containsSubstance> *Milk fat*.
	2. *Portugal* BT *Western Europe*	Incorrect Relationship: *Portugal* <spatiallyIncludedin> *Western Europe*

2.2 Vaguely defined (or underspecified) relationships

Because terms are very generally defined, they have been applied inconsistently. RT (Related Term) has been used to link any two, usually non-hierarchically, related terms that seem to be associated with each other. This relationship needs to be defined in order to reflect the more meaningful and specific associative semantics between the terms in the thesaurus. For example (see Table 11), ***RT*** *(related term)* is underspecified, subsuming numerous relationships like ***RT*** in *Mutton* ***RT*** *Sheep.* This relationship should be refined to a more specific one, such as '*madeFrom*' (Soergel *et al.*, 2004) to distinguish it from other uses of RT.

Table 11 Examples of the use of RT to represent different semantic relationships

Relationship	**Examples**	**Remark** (More Appropriate Relationship)
RT	1. *Mutton* RT *Sheep*	*Mutton* <madeFrom> *Sheep*
	2. *Rice* RT *Rice flour*	*Rice* <usedToMake> *Rice flour*
	3. *FAO* RT *UN*	*FAO* <memberOf> *UN*

3. Problems of ontology integration

The problem that underlies the difficulties in ontology merging is the different concept names that may exist between the new extracted ontologies and the existing one. Moreover, the relations of concept contained in the merged ontology may contain redundancy that requires the process to organize them.

3.1 Term Mismatch

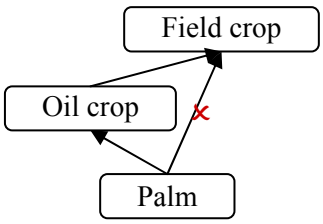
Term mismatch is the problem that concepts are represented by different names or they are synonym terms. For example, the term "*/mu/(pig)*" is contained in one ontology and the term "*/sukon/(pig)*" is in another ontology. We can infer this by using the cue word for identifying synonym relation such as /chao-ban-reak-wa/ (people call as), /ruchak-kan-nai-nam-khong/ (known as the name of). Moreover, this problem may occur when the noun phrase are composed of words that have similar meanings e.g. "/*phuet phak/(vegetable crop)*" and "*/phak/(vegetable)*". We can therefore infer that if labels are the same, the entities are probably the same by comparing labels with the edit distance (Levenshtein, 1966).

Another related problem involves with homonym terms that the problem is that sometimes two or more terms have the same label with different meanings. Visser *et al.* (1997) calls this a 'concept mismatch'. For example, the term "/*kaew*/" can mean a flower or a variety of mango. This inconsistency is much harder to handle; knowledge like rules or constraints are required to solve this ambiguity e.g. flower is a

disjoint concept of fruit. Hence, '/*kaew*/' can not be subclass of both flower and mango that is a kind of fruit. However, it is out of scope of this work for extracting the rules.

3.2 Redundancy in the class hierarchy

The problem of redundancy in the class hierarchy is caused by a node has a direct hierarchical relation to one of its ancestors (non-immediate parent). For example,

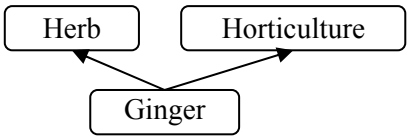


HYPONYM (Palm, Oil Crop),
HYPONYM (Oil Crop, Field Crop)
and *HYPONYM (Palm, Field Crop)*

Figure 19 An example of redundancy relation

In this case (Figure 19), the system can check that there is a redundancy relation of *HYPONYM (Palm, Field Crop)* which can be inferred from the other relations then the system will delete this redundancy relation.

Moreover, in some cases the concept can have multi-parents and these relations are not redundant. For instance, Figure 20 shows that *Ginger* has two parents: *Herb* and *Horticulture*. The system will keep both relations.



HYPONYM (Ginger, herb)
and *HYPONYM (Ginger, Horticulture)*

Figure 20 An example of multi-parent concept

3.3 Conflict relationships

The problem of conflict relationship or inconsistent relationship is occurred when the sub-ontology tree contains an incorrect relationship. The incorrect relationship can be caused by error in the process of ontology extraction from text especially when using the cue word /pen/ that has ambiguity meaning. The examples of conflict relationship are shown in Figure 21. In Figure 21(a), the *HYPONYM (Loam soil, Soil)* relationship can be extracted from the sentence (18) that the word /din/ (soil) is omitted the modifier, for instance, *'that should be used*'. The example sentence (19) can be extracted the relationship *HYPONYM (Plant, Rice)* as shown in Figure 21(b).

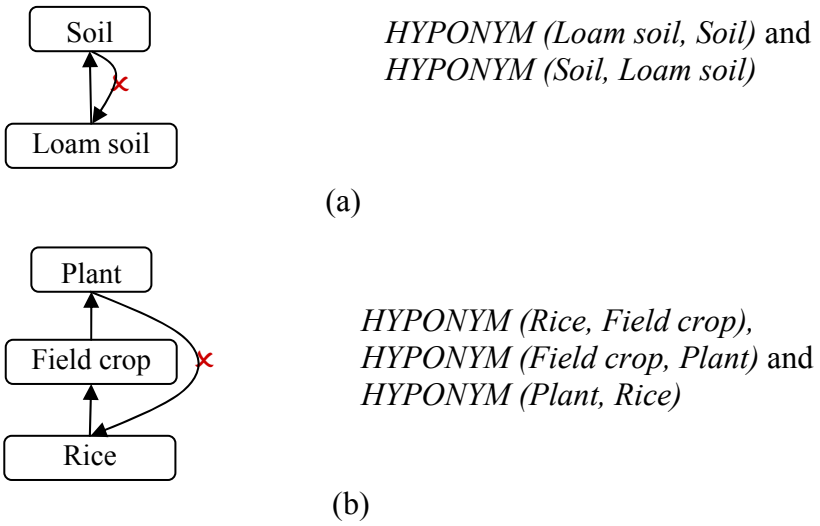


Figure 21 Examples of conflict relationship

(18) */din pen din-ruan/*
(Soil (that should be used) is loam soil.)

(19) */phuet thi pluk pen khao 50 poesen mai phon 50 poesen/*
(Plants that are cultivated are rice 50% and fruit 50%)

This problem is solved by comparing frequency of occurrence of each relation since we have the assumption that the correct relationship has more frequency than the incorrect relationship. The relationship that has less frequency is deleted. This process is also beneficial for pruning the incorrect relationships.

MATERIALS AND METHODS

Materials

1. Computer

The algorithms are implemented by using Visual Basic and PHP programming language. The experiments are run under the following computer specification.

- Pentium processor 1.73 GHz.
- RAM 1 GB
- Hard Disk 50 GB

2. Data

This research proposes the methodologies for automatic ontology building with a variety of resources i.e. text, thesaurus and dictionary.

2.1 Text

The corpus used to test the methodologies in this work deals with the domain of agriculture. It is the plain text in Thai and its size is of 302,640 words from 90 documents. The documents coming mainly from two resources as follows:

- Technical documents about plants from the Department of Agriculture and the Department of Agricultural Extension: It contains about 277,164 words from 85 documents.

- Thai encyclopedia on topic of plants: Its size is of 25,476 words from 5 documents.

2.2 Thesaurus

In this research, we emphasize to utilize AGROVOC Thesaurus (FAO, 2007), which is a multilingual agricultural thesaurus in English, French, Spanish, Portuguese, etc. that developed and maintained by the Food and Agriculture Organization (FAO) of the United Nations. It contains 16,607 descriptors and more than 10,706 non-descriptors (synonyms). It is used for indexing and searching information resources within the agricultural domains such as plant, fisheries, food, etc.

2.3 Dictionary

The system was based on the "Thai Plant Names" Dictionary, which developed by Smitinand and edited by the Forest Herbarium Royal Forest Department in 2001. It contains 37,110 words.

Methods

Figure 22 shows the overview of the Automatic Thai Ontology construction and Maintenance System consisting of Ontology Extraction, Ontology Integration and Reorganization and Verification. There are three sources for ontology extraction: unstructured (raw) texts, a semi-structured dictionary and a structured thesaurus. Unstructured texts should be dealt with by a hybrid approach: natural language processing, rule based and statistical based techniques being used in concert for identifying the related ontological terms and their relationships. For the semi-structured dictionary, only a task-oriented parser is needed to extract the terms and relations. However, the parser will work well if, and only if, the dictionary has a given structure. Since ontological terms can be transformed straightforwardly in the case of a structured thesaurus, we must make sure to have clean relationships between terms and possibly make certain refinements. However, even if we have gotten the appropriate ontological relationships, we still need to perform natural language processing at the phrase level and rely on machine learning techniques. Finally, all

sub-ontology trees are integrated to the core-ontology by using term matching technique. The ontology will be reorganized by pruning the redundancy relationships and merging the similar concepts. In addition, we develop tool for expert to verify and extend the Ontology.

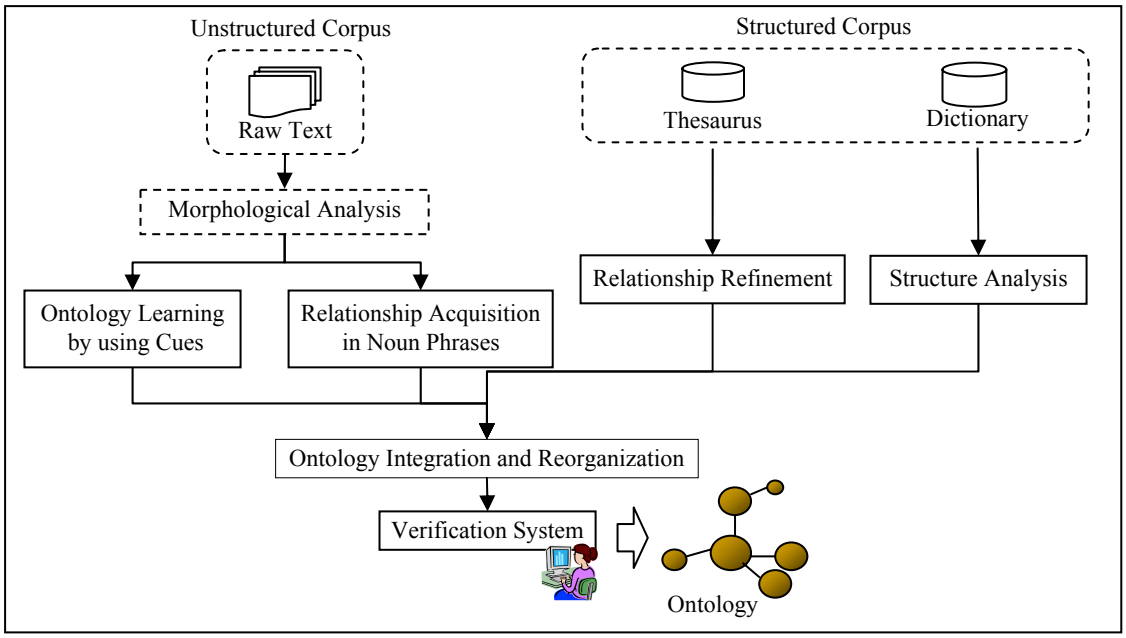


Figure 22 A conceptual framework of ontology construction and maintenance system

1. Ontology Learning based on Unstructured-Text

*Morphological Analysis and NPs Chunking***.** Similarly to many other Asian languages, in Thai there are no delimiters (blank space) to show word boundaries. Texts are a single stream of characters. Hence, word segmentation and part-of-speech (POS) tagging (Sudprasert, 2003) are necessary for identifying a term unit and its syntactic category. Once this is done, sentences are chunked into phrases (Pengphon, 2002) to identify noun phrase boundaries. In this paper, the parser relies on Noun Phrase (NP) rules, word formation rules, and lexical data. The accuracy of compound noun grouping is 92% and the accuracy of NP analysis with word formation is 90%. Before experimenting in the next process, the experts verified and corrected all the

NPs in the documents in order to test the actual performance of the ontological learning system.

1.1 Ontology Learning by using Cue (Imsombut and Kawtrakul, 2007)

Figure 23 gives an overview of the ontology learning by using cues which are lexico-syntactic pattern and item list. As far as the ontology learning is concerned, there are three main processes involved: ontological-element (concept and relation) identification by using cues: lexico-syntactic patterns and item list, candidate term generation, and candidate term selection.

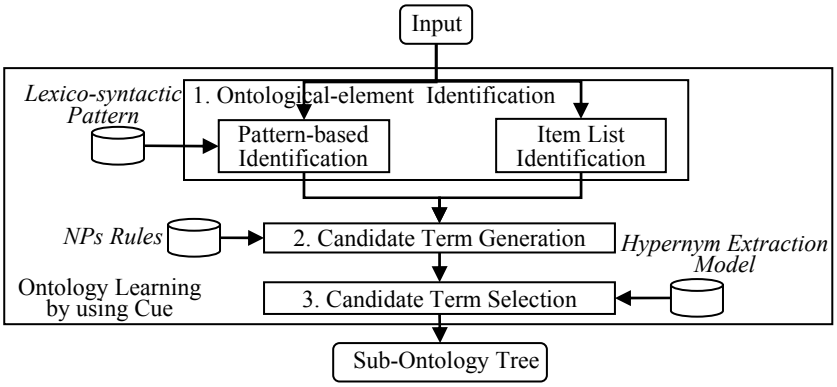


Figure 23 Architecture for building and maintaining an ontology of Thai

1.1.1 Ontological-Element Identification

We identify the ontological-element (concepts and relations) hinted by cues, which are lexico-syntactic patterns and item lists.

a. Pattern-based Identification

In order to collect hypernym relation patterns, we use IS-A relations with words (word pairs) from the AGROVOC thesaurus (FAO, 2007). From agricultural documents, we extract all sentences that contain the selected word pairs.

Finally, we manually select lexico-syntactic patterns from sentences by considering only the ones that the structure occurs often.

Table 12 Lexico-syntactic patterns with frequency of occurrence

Patterns	Cue-word meaning	POS	Occurrence Frequency	%
NP_0 …/chen/… NP_1, …,NP_n	(for example)	conjncl	392	41
NP_1… /pen/ NP_0	(is/am/are)	vcs	306	32
NP_0 .../daikae/… NP_1, …, NP_n	(i.e.)	conjncl	186	19
NP_0 .../champhuak/… NP_1	(kind of)	ncn	40	4
NP_1, …,NP_n /lea/ NP_0 /uen uen/	(and other)	conj	22	2
Other Patterns	-	-	16	2

Remark: vcs = complementary state verb, conjncl = noun clause conjunction, conj = conjunction, ncn = common noun.

The most occurring ones are focused (the top-5 patterns) in this article. By using the patterns above (see Table 12), the sentence anchoring process could identify plausible sentences whose content bare its ontological relation.

b. Ontological-list Identification

In this process, we propose a methodology for identifying an ontology element from item lists that we focus on bullet list and numbering list. Since item lists could be used to describe objects, procedures, and the like, this might lead to non-taxonomic lists. In order to identify object lists which contains ontological terms, the items of the list should be the Named Entity (NE), recognized by NER system (Chanlekha and Kawtrakul, 2004). Applying NER for identifying ontological lists works well in technical domain such as agriculture and bio-informatics since the growth of ontological terms almost come from the new entities. As shown in Figure 16 and 17, it still has two problems: long boundary description and embedded list that cause the ambiguity of item list members. In order to solve these problems, the items

that use the same bullet symbol and have same NE class will be considered as the same list. Like the bullet list, the items in the ordering number and having the same NE class will be considered as the same list.

1.1.2 Candidate Term Generation

In this work, we use linguistic information in the form of a grammar that mainly allows NPs to be extracted as candidate terms. Thus, this process checks whether some NPs occurred before the cue as candidate terms in order to generate the corresponding ontological terms. Thus, all NPs on the left hand side of the cue word in the pattern are generated as the candidate terms and the terms that occur in the preceding paragraphs of the item list are candidate hypernym terms. To do so it will consider only NPs corresponding to the NP's grammatical rules as shown in Table 13.

Table 13 Grammatical rules of noun phrases for ontological terms extraction.

Pattern	Example
NP2 = (ncn|nct+ncn|npn) + $NP^{?}$	[/chuea/(pathogen):ncn /wairat/(virus):ncn] (virus pathogen)
NP3 = NP2 + adj	[/kulap/(rose):ncn /daeng/(red):adj] (red rose)
NP4 = NP + VP VP = vi|(vt+NP)	[/a-ngun/(grape):ncn /tham/(produce):vi /wai/(vine):ncn] (vine grape)
NP5 = NP + PP PP = prep + NP	[/sinkha/(product):ncn /caak/(from):prep /tangprathet/ (foreign country):ncn] (product from foreign country)

Remark: adj = Adjective, ncn = Common noun, nct = Collective noun, npn = Proper noun, NP = NP2|NP3|NP4|NP5, prep = Preposition, PP = Prepositional Phrase, vi = Intransitive verb, vt = Transitive verb, VP = Verb Phrase, x|y = either x or y, x + y = x precede y, $x^{?}$ = x can occur 0 or 1 time

Even there are many NP rules, some NPs could not be an ontological term such as [ncn+conj+ncn], [ncn+det], where conj is conjunction and det is determiner. For example, [*/phak/(vegetable):ncn /lae/(and):conj /phonlamai/(fruit) :ncn*] (*vegetable and fruit*). The selected ontological terms should be separated into two terms, i.e. */phak/(vegetable)* and */phonlamai/(fruit)*.

1.1.3 Ontological Term Selection

Having generated the ontological candidate terms, the system will select the ontological term from a set of candidates. The most likely hypernym value (*MLH*) of term in the candidates list will be computed on the basis of an estimated function taking lexical and co-occurrence features into account. Let $h_i \in H$, H is the set of candidates of possible hypernym terms, while t_j is the related term j which is the term on the right hand side of lexico-syntactic pattern or the term in the item list. The features of the learning system are lexical and co-occurrence features. The estimate function for computing the most likely hypernymy term is defined as follows:

$$MLH(h_i, t_j) = \alpha_1 \cdot f_1(h_i, t_j) + \alpha_2 \cdot f_2(h_i, t_j) + ... + \alpha_n \cdot f_n(h_i, t_j) \quad (11)$$

Where α_k is the weight of feature k, f_k is the feature k, t_j is the related term j and n is total number of features (here, we use 5 features). f_1-f_4 are lexical features and f_5 is co-occurrence feature. The system will select the candidate term that has the positive and maximum *MLH* value in each candidate set to be the ontological term of the related terms.

For weighting each feature, we test with several techniques: Information Gain, Information Gain Ratio and SVM (with linear kernel) and we found that the feature weights from Information Gain Ratio gives the best result in the candidate term selection process. Information Gain Ratio is introduced to compensate the bias of the Information Gain. They are used to decide which features are the most relevant.

However, calculating information gain needs discrete value but the co-occurrence feature (f_5), is continuous value then it needs the method to convert continuous value to discrete value that is described in the detail of feature 5.

Features and their details are following.

f_1: Head word compatibility. This feature evaluates whether head word of candidate term is compatible with head word of related term or not.

$$f_1(h,t) = \begin{cases} 1 \text{ if } h \text{ is compatible with the head word term of } t. \\ 0 \text{ if otherwise.} \end{cases} \quad (12)$$

If the head word of a constituent is identical to the head word of another constituent, then these terms are related to each other. For more details, consider the following example.

(15) /**po-kra-chao** thi niyom pluk kan nai **prathed-thai** mi 2 chanit dai-kae **po-krachao-fak-yao** lae **po-krachao-fak-krom**/

(*There are 2 kinds of* ***Jute*** *generally planted in* ***Thailand*** *i.e.* ***Tossa Jute*** *and* ***White Jute***).

In this example, the candidate terms are *Jute* and *Thailand*. The head word of *Tossa Jute* and *White Jute* is *Jute,* then *Jute* has more possibilities to be an ontological term than *Thailand*.

f_2: NE class. This feature evaluates whether a candidate term of a hypernym belongs to the same NE class as related term or not.

$$f_2(h,t) = \begin{cases} 1 \text{ if } h \text{ belongs to the same NE class as } t. \\ -1 \text{ if } h \text{ belongs to the different NE class as } t. \\ 0 \text{ if otherwise.} \end{cases} \quad (13)$$

We consider the NE class as the feature because the cue word /pen/ might occasionally have the meaning “has symptom as”. For example,

*(16) /*kalampli pen rok-naole*/
(Cabbage has symptom as Soft-Rot.).

Here, *Cabbage* and *Soft-Rot* are NEs which have different classes, i.e. plant and disease, respectively. Accordingly, *Soft-Rot* is not a hypernym of *Cabbage*. In addition, we classify this feature’s value to three values, i.e., 1,-1 and 0, where 0 is assigned for the terms being at a high level of the taxonomy, e.g. /phuet trakun thua/(pulse crops) which are not NE.

f_3: Property list term. Since the cue word /pen/ (be) can be used to express the properties of an object. For example,

*(17) /*kap-bai pen si-namtan*/
(The *leaf is brown-color*.)

Brown-color is not the hypernym of *leaf*, but a property of the object leaf. This being so, we defined a set of properties to be able to determine which terms are concepts and which are properties. In the domain of agriculture, there are 3 types of property lists: colors, shapes, and appearances; e.g. *powder*.

$$f_3(h,t) = \begin{cases} -1 \text{ if } h \text{ is a property term.} \\ 0 \text{ if otherwise.} \end{cases} \quad (14)$$

f_4: Topic term. This feature evaluates whether candidate term is the topic term of the document (short document) or a topic term of the paragraph

(long document) or not. Here, topic term will be computed by using tf*idf where tf is the term frequency and idf is inverse document frequency (Salton, 1989).

$$f_4(h,t) = \begin{cases} 1 \text{ if } h \text{ is a topic term of the document (short document)} \\ \quad \text{or a topic term of the paragraph (long document).} \\ 0 \text{ if otherwise.} \end{cases} \quad (15)$$

f_5: Co-occurrence feature. Some statistical methods are used to analyze the co-occurrence of the candidate and the related terms. We explore three alternatives, Mutual Information (MI) (Church and Hanks, 1989), log-likelihood ratio (LL) (Dunning, 1994), and Chi-square testing (χ^2). By experimenting with Thai agriculture document, we found that Chi-square has the highest precision. Chi-square is based on hypothesis testing. It measures the divergence of the observed and expected data. Chi-square can be defined as follows:

$$f_5(h,t) = \chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad (16)$$

Where a represents the frequency of the term h occurring in the same sentence with the term t. The value b (respectively c) is the number of occurrences of term h (respectively t) in the corpus for sentences not containing term t (respectively h). The value d indicates the number of sentences that do not contain neither h nor t. The total number of sentences in the corpus is represented by N.

Since value for calculating Information Gain must be discrete value and the result of chi-square is continuous value, then the method for partitioning the continuous value to a discrete value is needed. We partition this feature value into two intervals at value x. However, chi-square value is very sparse. From observation, the minimal value of chi-square of our corpus is 0.001 and the maximum value is 206.667. Partitioning the data into two intervals by using this minimal and maximum value is not appropriate because in some candidate sets the chi-square values of all data are very low and this cut point can not separate data between positive and

negative class. Then we define x depending on each candidate term set as $x = (Max+Min)/2$ where *Max* and *Min* are the maximum and minimum chi-square value in the candidate term set. This partition value will separate data into two groups ($\leq x$ and $>x$; f_5=0 and f_5=1) for calculating Information Gain. Figure 24 shows an example sentence, the calculation of the partition value of this sentence and the co-occurrence feature of each candidate term. Table 14 shows examples of sentence, the candidate terms and their feature vectors and *MLH* values for selecting the ontological term.

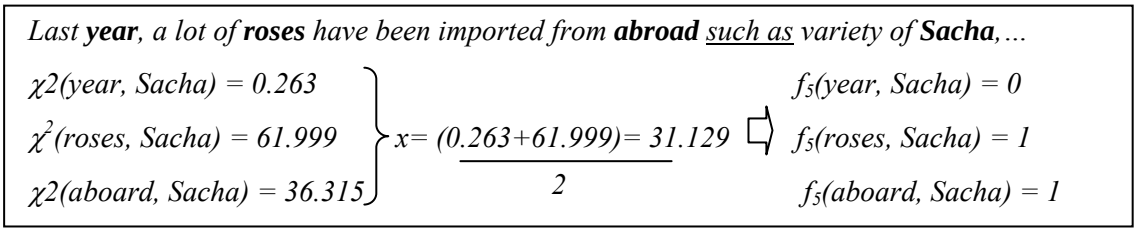


Figure 24 An example of calculation for the co-occurrence feature

Table 14 Examples of sentence, the candidate terms and their feature vectors and MLH values

	Head word	NE	Property Term	Topic Term	Co-occur.	MLH Value
weight	0.001	0.069	0.139	0.014	0.029	
*Last **year**, a lot of **roses** have been imported from **abroad** such as variety of **Sacha, Mercedes** and **Gabrielle**.*						
Year,Sacha	0	0	0	0	0	0
Rose,Sacha	0	1	0	1	1	**0.112**
Aboard,Sacha	0	0	0	0	1	0.029
*There are 2 kinds of **Jute** generally planted in **Thailand** i.e. **Tossa Jute** and **White Jute**.*						
Jute, Tossa Jute	1	1	0	1	1	**0.113**
Thailand, Tossa Jute	0	0	0	0	0	0
***Cabbage** has symptom as **Soft-Rot**.*						
~~*Cabbage, Soft-Rot*~~	0	-1	0	1	0	-0.55

The weighting of each feature in Figure 14 will be discussed in Results and Discussion chapter. The ontological term of each sentence is the candidate term that has the positive and maximum MLH value. The sentence that has the negative *MLH* value of candidate term will be pruning.

1.2 Relationship Acquisition in NPs (Imsombut and Kawtrakul, 2005)

The information concerning semantic relations can be extracted not only at the sentence level, but also at the noun phrase level. In Table 15, we list the semantic relations of NPs our system is able to analyze by taking as input a Thai corpus in the domain of agriculture. The semantic relations in the list are the most frequently ones found in the data, and they are based on relations given in (Vanderwende, 1994; Barker and Szpakowicz, 1998; Soergel, 2004; Kawtrakul *et al.*, 2005). Even though our analysis is on texts dealing with agriculture, the semantic relations are domain independent. An overview of learning system for discovery semantic relations in NPs is shown in Figure 25.

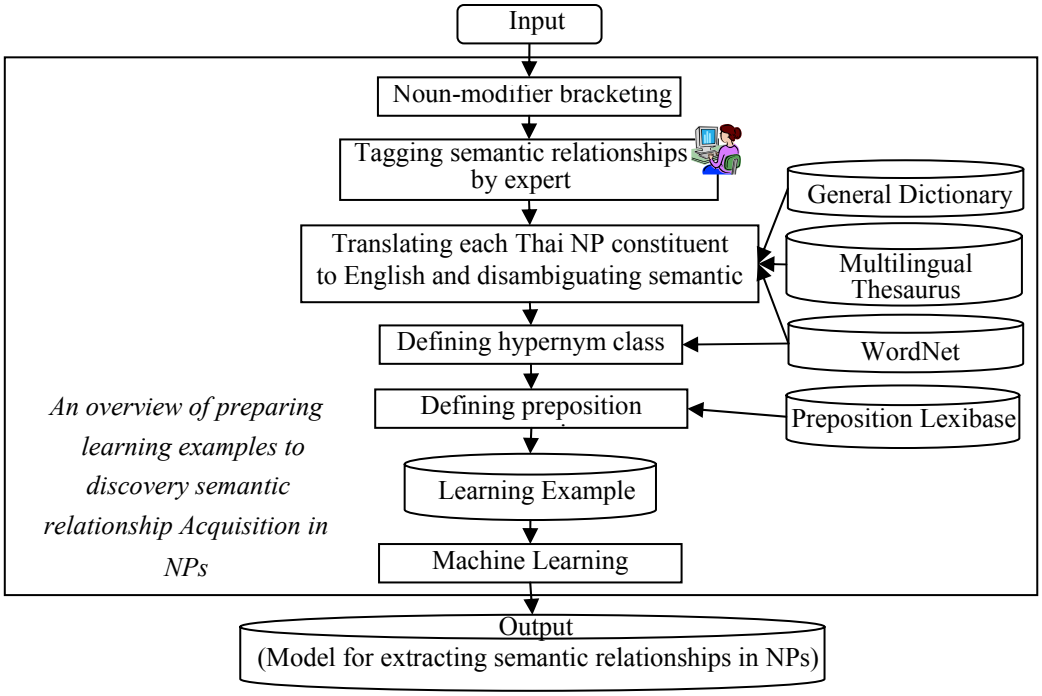


Figure 25 An overview of learning system for discovery semantic relations in NPs

Table 15 A list of semantic relations.

Relationships	Examples
1. **IS-A**	/ka-fae/(coffea) /ro-bus-ta/(Robusta) : (Robusta coffea)
2. Location	/phuet/(plant) /khet-ron/(tropical area) : (tropical plant)
3. **Purpose** :	/cream/(cream) /sa-lad/(salad) : (salad cream), /khai-mun/(fat) /sam-rub/(for) /prung/ (cook) /a-han/ (food) : (cooking fat)
4. **Possessor**	/rai/(field) /ka-set-ta-kon/(farmer) : (farmer field), /ran-kha/(shop) /khong/(belong_to) /chum-chon/(community) :(community shop)
5. **Made-of**	/sous/(sauce) /tua-luang/ (soybean) : (soybean sauce), /nuy/(cheese) /jak/(from) /nom/(milk) /kea/(sheep) : (sheep cheese)
6.**Source (from)**	/mun/(dung) /sat/ (animal) : (dung), /dok-mai/(flower) /jak/(from) /tang-pra-ted/ (foreign_country) : (flower from foreign country)
7. **Topic**	/kor-mul/(data) /pra-mong/(fishery) : (fishery data), /kor-mul/(data) /tang/(-) /pum-mi-ar-kart/(weather) : (weather data)
8. **Property**	/ku-lab/(rose) /see-deang/(red_color): (red rose)
9. **Part-Whole**	/perk/(husk) /kao/(rice) : (rice husk), /kao/(horn) /kong/ (belong_to) /sat/(animal) : (horn)
10. **Container**	/klong/(carton) /nom/(milk) : (milk carton)

During this step, the system will extract semantic relations from Noun Phrases (NPs) by learning the common ancestral concept of their head and modifier. The following features are taken into account by our learning component:

1. the semantic class of the head noun: The system will extract the head noun's sense and its hypernyms by using WordNet.

2. the semantic class of the modifier noun or head noun of a modifier phrase: Like the head noun, the system will extract the sense of the modifier noun or head noun of a modifier phrase and its hypernyms with the help of WordNet.

3. the semantic class of the preposition: This is applied only to NPs composed of a prepositional phrase. It is meant to provide information about the semantic role of the prepositions used in the NPs. The value of this feature is determined by Lexibase (Kawtrakul, 2004), a resource developed in our laboratory.

Since our learning features are based on the semantic information provided by WordNet, the learning examples have to be translated from Thai to English. To this end our system brackets first the head and the modifier of a given NP, to translate it into English then using a Thai-English Thesaurus, AGROVOC (FAO, 2007), and a Thai-English Dictionary, LEXiTRON (NECTEC, 2007). Next, the WordNet sense of nouns and the semantics of prepositions are identified. Here are some details concerning the algorithm.

1.2.1 Head noun and modifier segmentation

This step is similar to the approach taken by (Lauer, 1995; Barker, 1998). For a given sequence of X-Y-Z, segmentation is determined by comparing occurrences of X-Y with occurrences of X-Z in a corpus. If X-Z occurs in the corpus then the segmented phrase is [X-Y]-Z, otherwise it is X-[Y-Z]. If a phrase like '*/kuad/(bottle) /nam/(water) /plad-sa-tik/(plastic): (plastic water bottle)*', '*/nam/ /plad-sa-tik/ (water plastic)*' never occurs in a corpus then it will be segmented as [[*/kuad/(bottle) /nam/(water)*] */plad-sa-tik/(plastic)*].

If the phrase '*/kuad/(bottle) /nam/(water) /phon-la-may/(fruit): (fruit juice bottle)*', '*/nam/ /phon-la-may/ (fruit juice)*' occurs in a corpus then it will be segmented as follows [*/kuad/(bottle)* [*/nam/(water) /phon-la-may/(fruit)*]].

1.2.2 Translating each Thai NP constituent into English and disambiguation of the semantics.

At this step, the system will translate all constituents of Thai NPs. If the constituent was a group of words, then system will, translate it first into an English word. For example, [*/nam_phon-la-may/(water_fruit)*] is translated into '*fruit_juice*'.

If there is no such word in English, the system will translate only the head of the word group. For example, [*/kuad_nam/(bottle_water)*] will yield '*bottle*'. There are two techniques to accomplish this task.

a. Technique 1

The system uses a Thai-English thesaurus to translate a Thai word (*tw*) into its English correspondence (*ew*), every words having a one-to-one mapping, i.e. translation equivalent. However *ew* might have more than one word sense in WordNet. In such a case the system needs to disambiguate the senses by computing the most likely similarity between each sense *i* and the hyponyms of *ew* in a thesaurus.

$$ew^i = \arg\max_{ew^i} \sum_{j=1}^{n} similarity\ (ew^i, h_j) \qquad (17)$$

$$h_j \in Hyponym(ew)$$

similarity(x,y) = the amount of edges in common paths between x to root and y to root

Figure 26 shows an example of thesaurus-based semantic disambiguation of '*/phon-la-may/ (fruit)*'. Fruit has three senses in WordNet. Based on using thesaurus information, it can be decided that sense number one is the sense of fruit in the domain of agriculture. If the word does not exist in the thesaurus, it will be processed by using technique 2.

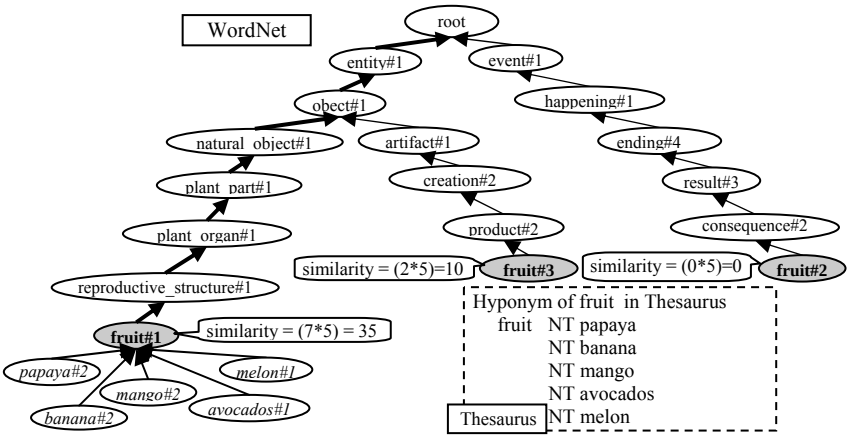


Figure 26 An example of thesaurus-based semantic disambiguation of '*/phon-la-may/(fruit)*'.

Remark. word#x stand for word in sense x.

b. Technique 2

A Thai-English dictionary is used to translate a Thai word (*tw*) into English (*ew*). By doing, we found that there are a lot of Thai words with a given meaning that could be translated into several words in English (about 70%). In this case we take the first English word or one of its synonyms in the dictionary as we assume that this is the most frequently one used. This word is then compared to the words in WordNet. If there are several senses in WordNet, the system will select the sense with the highest number of 'synset' terms similar to the set of translated words. Let S^{ew^j} be a synset ith of *ew* in WordNet and T be the set of translated terms of *tw* in a dictionary. In this case the system will select the *ew* sense by using the following equation.

$$ew^j = \arg\max_{ew^j} sizeof(S^{ew^j} \cap T) \tag{18}$$

Figure 27 shows an example of a dictionary-based semantic disambiguation of '/*krong*/ (cage)'

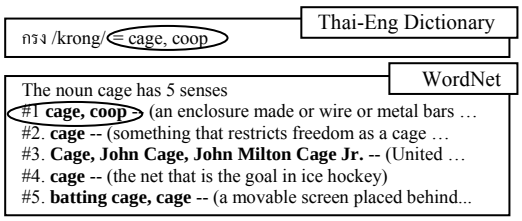


Figure 27 An example of dictionary-based semantic disambiguation of '/krong/(cage)'

If the Thai word had several meanings, or if it had only one translation, then the system will alert the expert to select the sense of the word in WordNet, manually.

In the case where the NP is composed of various words absent from WordNet, the system will select only the head noun for semantic disambiguation. For example, for the term '/*tang-pra-ted*/*(foreign country)*' the system will choose only '*country*' in '*foreign_country*' to disambiguate it semantically.

1.2.3 Defining hypernym class and preposition semantic

After translating and disambiguating the semantics of the head and the modifier of the NP, we extract from WordNet the complete hypernym path of each constituent in NP for the learning system. For example, the hypernym path of NP *"/kuad/(bottle) /nam phon-la-may /(fruit-juice): (fruit-juice bottle)*"

{bottle#1, vessel#3, container#1, instrumentality#3, artifact#1, object#1, entity#1}

{fruit_juice#1,beverage#1, food#1, substance#1, entity#1}

Moreover, for NPs containing a preposition, for example "*/dok-mai/(flower)* ***/jak/(from)*** */tang-pra-ted/ (foreign_country):(flower from foreign country)*", the system will define the semantic group to which it belongs by using Lexibase (Kawtrakul, 2004). In Thai, there are 10 semantic groups of prepositions such as location, purpose, time, etc. These semantic groups are mapped into 10

features. The feature values are set to 1 if the preposition has the same semantic group as the feature. Some prepositions belong to several semantic groups such as '/*nai*/ *(in)*" which can express a location or time meaning. Hence, the system will set the value of these features to 1. For NPs without prepositions, all values of these features are 0.

1.2.4 Learning of relationships

To obtain vectors of equal length, all hypernym list class of all examples are union to be list of hypernym class and the features will be converted into binary representations. Then, the features vector are the list of hypernym class of head, list of hypernym class of modifier and the list of the semantics, i.e.

features_vector{{list of hypernym class of head}, {list of hypernym class of modifier}, {list of preposition semantic}}

The features will be converted into binary representations to obtain vectors of equal length. The learning system will be applied to learn the common ancestral concept of the head and the modifier, to generate then a model to extract the semantic relationship of the NPs. Two machine learning techniques are applied by our system.

1) C4.5 of decision tree learning system by using the software package of Weka (The University of Waikato, 2007).

2) Support vector machine: We use several kernels but linear kernel is shown the best result. The software we used is the LIBSVM package (Laird, 2005).

For the experimental results, the decision tree learning system generated around 90 classification rules for discovery 10 semantic relations as mentioned in Table 3. For examples:

Rule: If head in container#1 then rel. Container.

Ex. */kuad/(bottle) /nam phon-la-may /(fruit-juice): (fruit-juice bottle)*

By applying this rule, the semantic relationship of noun "*/krong/(carton)*" and "*/nom/(milk)*" in "*/krong/(carton) /nom/(milk):(milk carton)*" will be *'container*' relation since carton is in the class of container#1.

In addition, the SVM learning system generated 10 learning models accordingly to the relations. The experimental results are shown in the next chapter.

2. Ontology learning based on a thesaurus

As mentioned in Section 2, the data of AGROVOC (FAO, 2007) should be cleaned before being used for ontology construction. We have divided this process of data cleaning and refinement of semantic relations into three main steps: Acquisition of Refinement Rules, Detection and Suggestion, and, finally, Verification. A system overview is given in Figure 28. (Kawtrakul *et al.*, 2005)

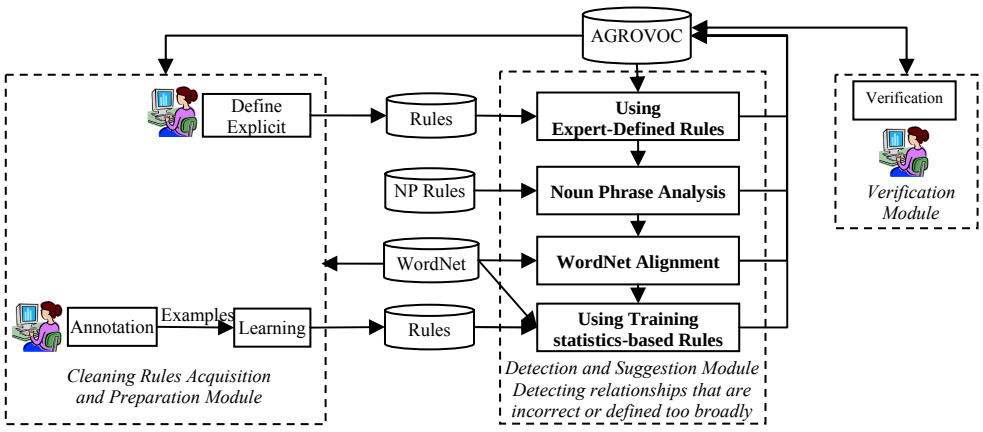


Figure 28 The process of cleaning and refining term relationships

2.1 The Rule Acquisition Module: Expert-defined Rules and Learning by Example

To mine the implicit relationships of some NPs, this module acquire a set of rules to suggest the most likely relationships in case that the relationship given by AGROVOC is underspecified (defined too broadly), especially RT. The rules will be provided by experts and by machine learning.

2.1.1 Expert-defined Rules

The experts can simply define a set of rules for allowing the correction of inappropriate relationships. They observe AGROVOC's data and define rules using data concerning concept types given in AGROVOC's database. For example, the rules constraint consists of the data in 'concept type data', the category of terms such as GC (Geographic term: Country level), and TP (Taxonomic term: Plant).

Given these rules, a relationship satisfying them will be revised automatically. For example, consider the following rule:

If X and Y are marked as ***"TP"*** *in the concept type field, and if X* ***BT*** *Y, then X<subclassOf> Y*

According to AGROVOC, the concept types of *Rosaceae* and *Malus* are TP, related by **BT**. Hence, the original relationship BT of "*Malus* BT *Rosaceae*" will be replaced by the <*subclassOf*>.

2.1.2 Learning rules-from-Examples

In this case, the rules are prepared to learn from examples in order to refine a relationship called ***RT.***

To prepare the learning set, we provide an annotation tool allowing the domain expert to manually tag term senses (labeled by a sense id number in WordNet). It allows also to specify the appropriate semantic relationship between some terms, for example, (*Sheep#1 <usedToMake> Mutton#1*).

In the case of compound nouns, only the noun heads are used. For example: *Rice* and *Rice Flour* will be annotated as follows: (*Rice#1 <usedToMake> Flour#1*)

Having prepared the examples, their complete hypernym path will be extracted from WordNet.

{*sheep#1*, *bovid#1, ruminant#1, mammal#1,vertebreate#1, animal#1, organism#1, living_thing#1, object#1,entity#1*}

{*mutton#1*, *meat#1, food#2, solid#1, substance#1, entity#1*}

The hypernym list given above will be used as the basis of the features' vectors, i.e.

Features'_vector{{list of hypernym class of all term1},{ list of hypernym class of all term2}}

The features will be converted into binary representations, in order to obtain vectors of equal length. The learning system, C4.5, will be applied to learn the common ancestral concept for term1, e.g., *animal#1,* and term2, e.g*., meat#1*, to generate then the rules. Figure 29 shows the example of the data set for training the <*usedToMake*> relationship. Table 16 displays the revision rules learnt from the training set.

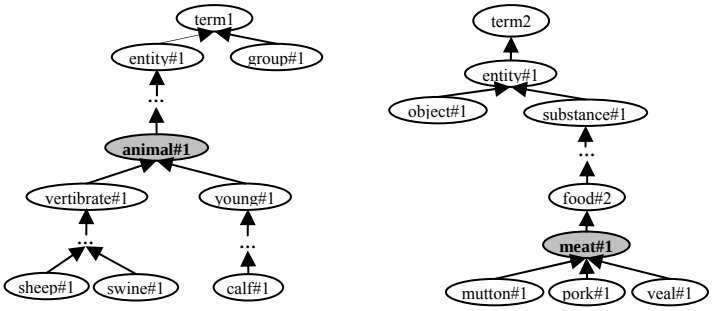


Figure 29 Examples of hierarchical data used for training the '*usedToMake'* relation

Table 16 Examples of training statistical-based rule

	Rule	**Example**
1	If class X is *animal#1* and class Y is *meat#1*, and X RT Y Then X <*usedToMake*> Y	*Sheep* RT *Mutton, Swine* RT *Pork*, *Calf* RT *Veal*
2	If class X is *plant*#2 and class Y is *food#1*, and X RT Y Then X <*usedToMake*> Y	*Rice* RT *Rice flour*, *Oat* RT *Oatmeal* , *Sugar Cane* RT *Cane Sugar*
3	If class X is *fruit#1* and class Y is *oil*#3, and X RT Y Then X <*usedToMake*> Y	*Castor beans* RT *Castor oil, Cottonseed* RT *Cottonseed oil*

By applying Rule 1, the original relationship RT of "*Chicken RT Chicken meat*" will be replaced by <*usedToMake*>.

2.2 The Detection and Suggestion Module

In this module, the system detects incorrect and inconsistently applied relationships and suggests appropriate relationships, waiting for the expert's confirmation. We propose three techniques to achieve this goal: rules for semantic relationships, noun phrase analysis, and WordNet alignment.

The outline of this algorithm is illustrated in Figure 30, where T_1, T_2 and Rel denote, respectively, Term1, Term2, and the AGROVOC relationship between them.

The relationship revision rules have been discussed in the previous section. Next, we briefly describe the procedures used for the analysis of noun phrases and the WordNet alignment**.**

2.2.1 Using Noun phrase analysis

This technique is used to analyze the surface form of a compound term's head word. If the head word of a term has the same surface form as its broader term, the system will apply the '*subclassOf*'/ '*superclassOf*' relationship. For example,

```
AGROVOC Cleaning_& Refinement (T1, T2, Rel)                    ;Return new__relationship
Input: Term1, Term2, Relationship
Output: New Relationship
1. If (Rel = BT or Rel = NT)
   Then If Agree_Expert_defined_Rules (T1, T2, Rel)
        Then return new_refined_relationship.                  ; following the rules
        Else If Headword-Is-Compatible (T1, T2)
             Then return subclass/superclass relationship.
             Else If Is_Wordnet_HypernymPath (T1,T2)
                  Then return subclass/superclass relationship.
                  Else If Agree_Revision_Rules (T1, T2, Rel)
                       Then return new_relationship             ; following the rules
                       Else return U.                           ; Un-refined
2. Else If (Rel=UF or Rel = USE)
        Then If Is_Wordnet_Synset (T1, T2)
             Then return synonym relationship.
             Else If Agree_Revision_Rules (T1, T2, Rel)
                  Then return new_relationship.                 ; following the rules
                  Else return U.                                ; Un-refined
3.      Else If (Rel=RT)
             Then If Agree_Revision_Rules (T1, T2, Rel)
                  Then return new_relationship.                 ; following the rules
                  Else return U.                                ; Un-refined
```

Figure 30 An algorithm for data cleaning and relationship refinement

Milk BT *Cow milk*

From the compound noun's analysis we see that the head word of *Cow milk* is *milk,* which obviously has the same surface form as *Milk*, the broader term of *Cow milk*. Hence, the system will apply the <*subclassOf*> relationship to *Cow milk* and *Milk*.

Milk BT *Milk fat*

The result of the analysis shows that the head word of *Milk fat* is *fat*, which is not compatible with the broader term, *Milk*. This will be detected, and the system will be trained by examples as mentioned before, in order to extract the rule for refining the relationship.

2.2.2 Using WordNet's Relationships

During this step, we use WordNet's hyper-hyponymy relationships to align the BT/NT relationship in AGROVOC. The synset of a term in WordNet is used to align the UF/USE relationship in AGROVOC. Since the relationships in WordNet are checked by experts and since it contains a great number of general, domain-specific terms, including agricultural terms, WordNet is a good resource for aligning certain relationships of AGROVOC, for example, taxonomic and synonym relationships.

At this stage the system starts retrieving the synset offset number of the AGROVOC UF/USE term in WordNet. If it can find these terms, and if they have the same synset id number, the system will consider that they are 'synonyms'. It will also check AGROVOC's broader term and the narrower term in WordNet. If it finds that the broader term is the ancestor of the narrower term in the WordNet hierarchy, it will conclude that we are dealing here with a '*subclassOf*' / '*superclassOf*' relationship. For example,

Cabbage BT *Vegetable*

Query results for *Cabbage* and *Vegetable* in WordNet show that *Cabbage* is a hyponym of *Cruciferous vegetable,* and *Cruciferous vegetable* is a hyponym of *Vegetable*. Figure 31 shows the relationship of *Vegetable* and *Cabbage* respectively in WordNet and AGROVOC.

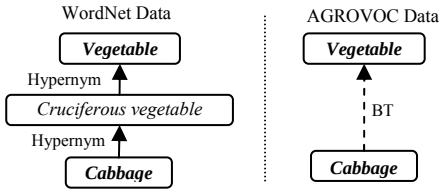


Figure 31 The relationship between '*Vegetable*' and '*Cabbage*' in WordNet and AGROVOC

Since *Vegetable* is an ancestor of *Cabbage*, the system will define *Vegetable* as a <superclassOf> *Cabbage*. In the case of *Milk* NT *Milk fat*, the relationship is not refined by this technique, because *Milk* and *fat* follow different hypernym paths in WordNet.

3. Ontology learning based on a dictionary

A Domain Specific Dictionary is the best way to extract relational information, as such kind of dictionary has a specific structure, as well as clear and accurate information. In this paper we use as case study the "Thai Plant Names" Dictionary, developed by Prof. Dr. Tem Smitinand and edited by the Forest Herbarium Royal Forest Department in 2001. The most frequent relationships of this specific dictionary are Hyponymy and Synonymy. Two steps are needed for extracting the ontological terms: Structure Analysis and Relation Analysis.

3.1 Structure analysis

In this system, the printed dictionary is digitalized using the optical scanner. The scanned-in image is analyzed to identify each alphabetic letter and converted into text document by OCR process (Kawtrakul and Waewsawangwong, 2000). After manually correcting OCR error, the text document is analyzed by the Task Oriented Parser to produce entity concept. The analysis of the structure of the dictionary is an important feature in order to be able to distinguish elements of word entries as sub-parts. The characteristic of term positions are analyzed and irrelevant

parts, such as author name, are filtered out. The needed parts are then transferred to a relational database by using a Task Oriented Parser.

Figure 32 illustrates the analysis of the dictionary's structure. Terms are clustered by using Alphabets' characteristics (see Table 17). The position of the terms in the text, such as, top and rightmost corner, top and leftmost corner, is also considered. A relational database's fields are predefined as Hierarchical relations such as Family, Sub-Family, Genus, Specific epithet and Formal Name, respectively.

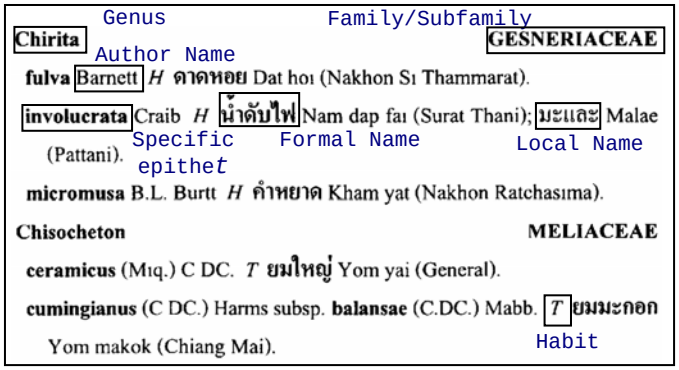


Figure 32 Dictionary structure

Table 17 Characteristics of the alphabet for dictionary conversion.

Feature	Database field	Example
All upper case at the top-rightmost corner	Family/Sub-Family	GESNERIACEAE
Starts with upper case at the top-left most corner	Genus	Chirita
All lower case	Specific epithet	involucrata
Thai alphabet in bold font	Formal Name	น้ำดับไฟ /Nam–dap-fai/
Thai alphabet	Local Name	มะและ /Malae/

3.2 Relation Analysis

After the parsing, Relation Analysis process will map each entity concept relation to the ontological relation. Figure 33 shows the process of ontology extraction based on specific dictionary. Figure 33a illustrates the output of the OCR system, Figure 33b shows the structure analysis output of the data in Figure 33a and the output of relation analysis is shown in Figure 33c. The system is able to extract 37,110 terms and 21,620 relationships. The experiment of dictionary-based-ontology extraction achieves 100%.

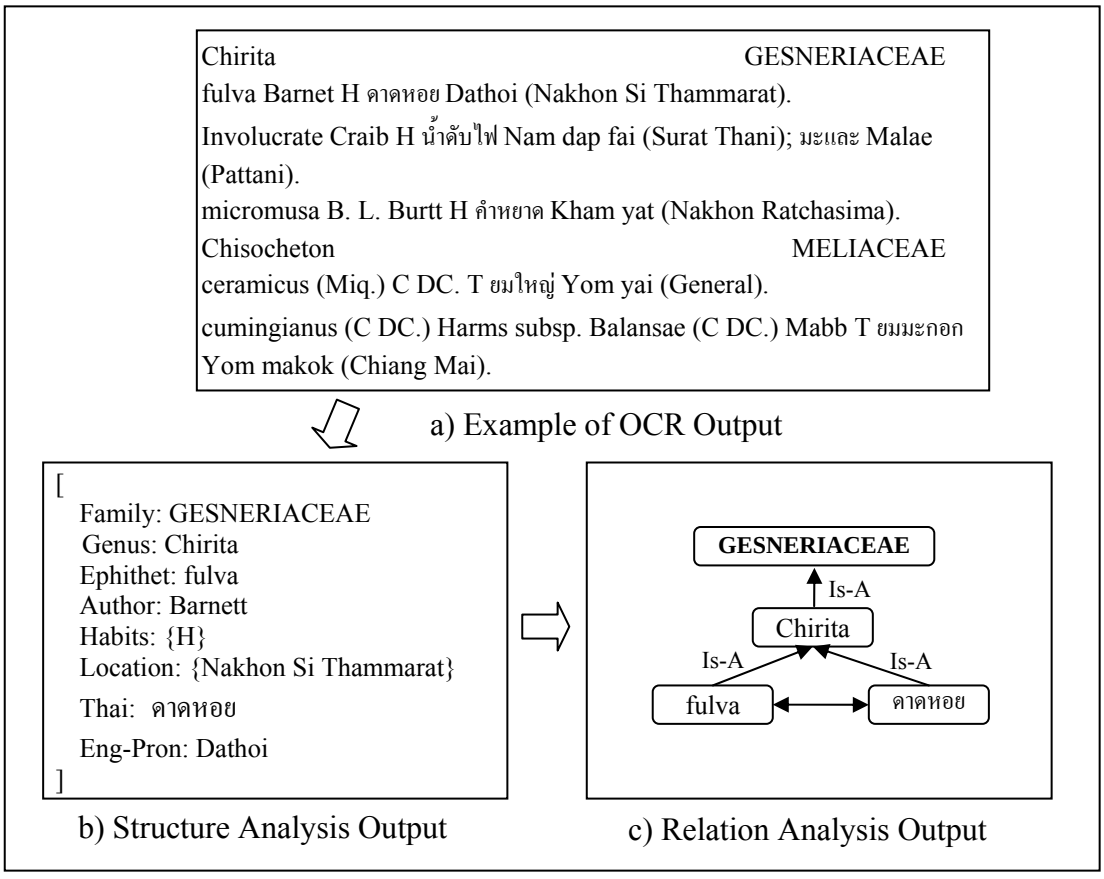


Figure 33 Dictionary based ontology extraction process

4. Ontology Integration and Reorganization

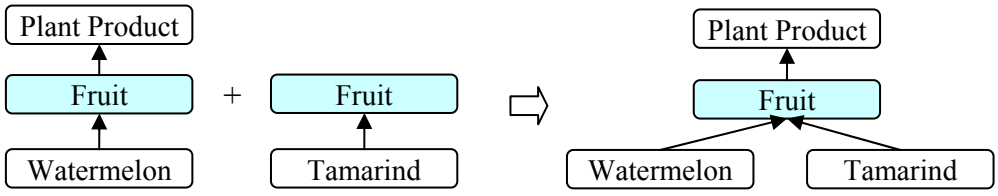
This section describes the details of the ontology integration and ontology reorganization process.

4.1 Ontology Integration

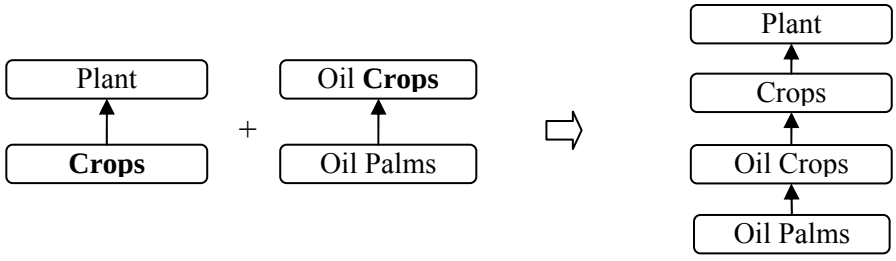
At this step, we united the related word/phrase pairs collected from our sources: Texts, some Dictionary and a Thesaurus. In order to integrate them, two heuristics are applied:

- *If the separated ontological trees have the same label nodes, then merge them.* Figure 34a shows the example of ontology integration by using this technique.

- *If the terms' head words match partially, then merge them.* The partially term's head words matching technique is shown in Figure 34b.



a) Same Label (Term Matching)



b) Partially Term's Head Words Matching

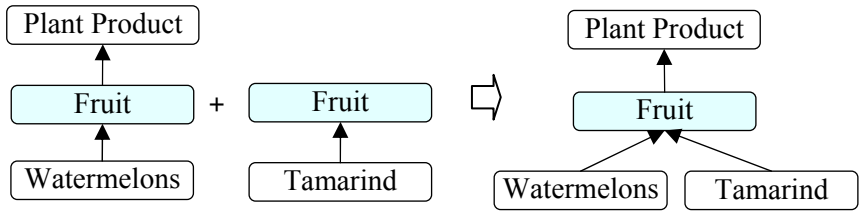
Figure 34 Techniques of ontology integration

There are two operations involved in this process of integration: Addition and Insertion. Figure 35 shows operations for ontology integration of a core-tree (left-hand-side tree) into a new ontological tree (right-hand-side tree), on the basis of information extracted from a dictionary or some raw texts.

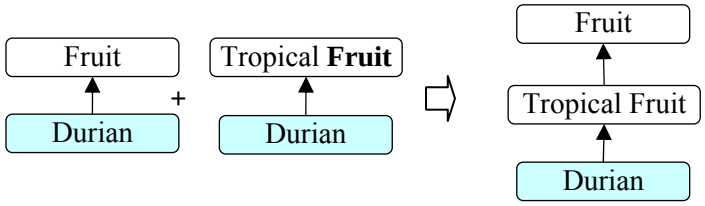
- *Addition*: A Child node will be added to the core tree, if the parent node has the same label or partially term's head word matching to the existing node in the core tree.

- *Insertion:* If the children nodes have the same label as the head word of the parent nodes then the new, more specific term will be inserted between two existing ontological terms.

The remaining terms that could not be integrated will be kept for the expert to be added later on, manually.



a) Addition



b) Insertion

Figure 35 Operations for ontology integration

However, the ontology integration process can make the problem of conflict relationship or inconsistent relationship since the sub-ontology tree may contain an incorrect relationship. The incorrect relationship can be caused by the problem of cue word ambiguity especially the cue word /pen/ as shown in the section of problems of ontology integration. We solve this problem by comparing the occurred frequency of each relation. The assumption of this solution is that the correct relationship has more frequency than the incorrect relationship. The relation that has less frequency is deleted. This process is beneficial for pruning the incorrect relationships.

4.2 Ontology Reorganization

When all nodes and relationships in the additional ontologies are added to the core tree completely, the ontology reorganizing operation will be processed respectively. There are three operations of ontology reorganizing: deleting, pruning and merging. Figure 36 shows the example of the process of these operations.

- *Deleting:* If there are duplicate relations, the system will delete the tree with less nodes.

- *Pruning:* Node, which does not have its own property and its children is the same set as its parent, should be deleted and its children should be transferred to under its immediate parent.

- *Merging:* If the two nodes or more than two have the common set of children nodes and these node's labels are similar then these nodes are merge to the new node and the common set of children will be transferred to the new node. The similarity of node's labels are compared by using edit distance technique (Levenshtein, 1966). Furthormore, the system will select the label that are the most frequency occurred in the corpus to be the concept label or concept representation of the new node.

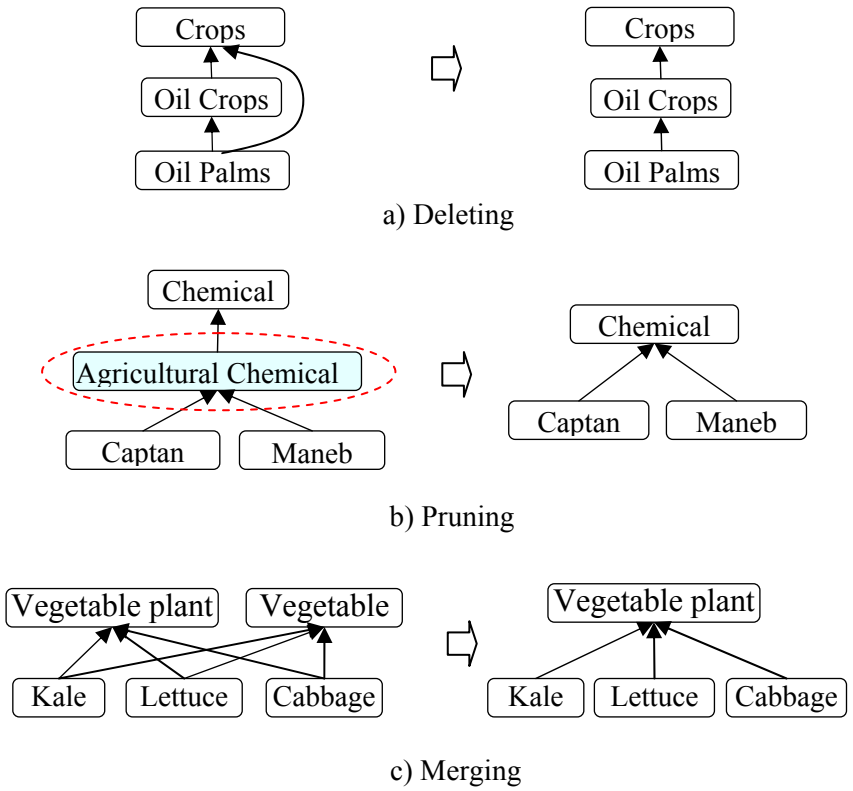


Figure 36 Operations for ontology reorganization

5. Verification Tool

Verification is required in order to ensure a high quality system, able to guide the expert to maintain the existing Ontology. This is why we have developed a user interface, allowing the expert to verify the quality of the output and to add related word pairs to the Ontological tree. Moreover, expert can delete node from the ontology tree if it is incorrect relationship.

Figure 37 shows the interface of the verification tool. The taxonomic relationships of the ontology are represented in the tree structure. When the node is selected, the tables in the right hand side will show the details of nodes that are the terms of this node and the semantic relationships of the node. The yellow icon show that the node and the relationship is correct and the red icon represent that this

relationship is not confirm by the expert then the user can verify this relationship with confirm or delete this node. Moreover, the user can search the node by enter the query word in the text box at the bottom right hand side of window.

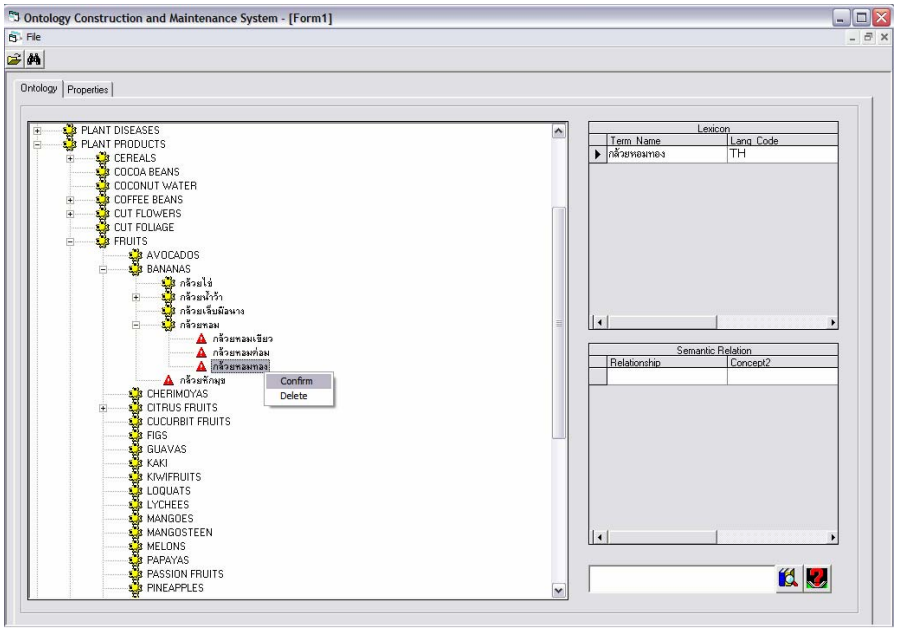


Figure 37 Ontology verification tool

RESULTS AND DISCUSSION

Evaluation Methods

The measurement of the system's performance was based on test cases in the domain of agriculture and divided according to the resources: text, thesaurus, dictionary and merging ontologies from all resources. To evaluate each methodology, we assumed that its inputs from the results of preprocessing are correct. To evaluate the results, the outputs of the system are compared with the results produced by the agreement of two experts of agriculture from Thai National AGRIS Center (2003). We use standard well-known measurements to assess the different approaches i.e. precision, recall and F1 variant of the popular F-measure. Precision (P) is the ratio of the number of extracted correct results to the total number of extracted results. Recall (R) is the ratio of the number of extracted correct results to the number of total correct results prepared by manually. Precision and Recall are given by standard formulas:

$$precision = \frac{\text{the number of extracted correct results}}{\text{the number of total extracted results}}$$

$$recall = \frac{\text{the number of extracted correct results}}{\text{the number of total correct results}}$$

The F-measure (F) combines the two parameters precision and recall. It was first introduced by Rijsbergen (1979). The standard formula is defined as:

$$F1 = \frac{(b^2 + 1) \cdot precision \cdot recall}{b^2 \cdot (precision + recall)}$$

b is a factor to quantify the value of precision and recall against each other. For the consequent test runs we use b = 1.

Results and Discussion

The experimental results of the system are summarized according to the resources: text, thesaurus dictionary. Moreover, the performances of the ontology integration and reorganization system are discussed in the end of this section.

1. Text-based ontology construction

The system's performance of this resource is evaluated according to the methodologies: cues-based ontology extraction and machine learning-based semantic relationship extraction from NPs.

1.1 Ontology Learning by using Cue

Concerning the calculation of the feature weights, the training corpus consists of 1,500 examples with positive and negative class. From the experiment by using several techniques: Information Gain, Information Gain Ratio and SVM (with linear kernel), for calculating the feature weights, we found that Information Gain and Information Gain Ratio give the similar results but Information Gain Ratio being better. SVM has a lower precision and it is difficult to set the proper kernel parameter. The weights of features from Information Gain Ratio are shown in Table 18.

Table 18 The Information Gain Ratio (or weight) of each feature.

	f_1	f_2	f_3	f_4	f_5
Information gain of lexico-syntactic pattern	0.001	**0.069**	**0.139**	0.014	0.029
Information gain of item list	0.050	0.190	0.045	0.043	**0.214**

From Table 15, we can conclude that the properties list feature (f_3) is the most important feature since it can prune the ambiguous cue words in patterns. Moreover, we found that the NE feature (f_2) is the next important feature for selecting the candidate term of the pattern-based ontology learning because NE usually occurs

in the agricultural document. We can conclude that this feature is appropriate for extracting specific domain term. If the candidate terms had the same NE class as the related term they should be selected. In addition, the co-occurrence feature (f_5) has the crucial role for selecting hypernym term of the list item terms because the hypernym and hyponym terms have more co-occurrence value than the other terms in the candidate term set. Conversely, the head word compatible feature (f_1) rarely occurs with the lexico-syntactic pattern and this feature usually occurs in the general domain documents. Hence, this feature is not significant for using to select the candidate term in the specific domain, as well as, the topic term feature (f_4) on the item lists.

In the evaluation process, we test the system with 3 aspects. First, we measure the performance of the system classified by the type of cues. Next, the system is evaluated based on two different test corpora i.e. technical documents and Thai encyclopedia. Finally, we compare our methodology with the Hearst's technique.

1.1.1 Evaluation based on the type of cues.

Based on our assumption that cues, lexico-syntactic patterns and item lists, could be used as the heuristic information for hinting the ontological relationship, we test the system based on each cue. As shown in Table 19, the results from using item lists are 0.83 of the precision, 0. 81 of the recall and 0.82 of the F-measure while the precision, recall and recall when using lexico-syntactic patterns as cue are 0.64, 0.69 and 0.66, respectively. The results of item lists are higher than the results from lexico-syntactic patterns since the item lists do not have the problem of anaphora and cue word ambiguity. However, the errors of item-list cue technique occur because some bullet lists are composed of two classes, for example, disease and pest. Then the system has the error in the detecting the paragraph that contains the hypernym term of the items in the second list.

Table 19 The evaluation results of the system classifying by the cue types.

Cues	P	R	F
Item list (a)	0.83	0.81	0.82
Lexico-syntactic pattern	0.64	0.69	0.66
Lexico-syntactic pattern with anaphora solving (b)	0.71	0.78	0.74
Both cues (a) + (b)	**0.74**	**0.78**	**0.76**

The important errors of pattern approach are caused by many sentences contained anaphora terms, then the system can not extract the correct ontological terms. The anaphoras causing these problems are the direct reference. They are definite NPs and zero anaphora. From observation, some of these anaphora terms can be solved by using heuristic rule by getting the subject of previous sentence. This method can increase the precision value by 7%, i.e. raising the level from 0.64 to 0.71. Even if including the anaphora resolution, the precision is not so high since it still has the problems of the ambiguity of the cue word /pen/ (be) that can be meant to other non-taxonomic relations i.e. being-state-of, being-status-of and made-of as shown in the example (18), (19) and (20), respectively.

*(18) */nai/(in) /raya/(period) /pen/(be) /tua-on/(young)…*
(Being in the young period)
*(19) */khao/(he) /pen/(be) /sot/(single)*
(He is single.)
*(20) */rongruean/(house) /pen/(be) /khrong-Lek/(steel structure)*
(House structure is made of steel.)

These problems can not be solved by keeping all non-taxonomic terms as a list like the property terms for pruning non-taxonomic relation of cue word /pen/ (be).

1.1.2 Evaluation based on different data set.

Dataset 1: Technical documents in the domain of agriculture. By testing with a dataset 1 (about 277,164 words), the system is able to extract about 2,043 concepts and 2,154 taxonomic relations when using both the lexico-syntactic patterns and the item list. The precision, recall and F-measure of the system when testing with this dataset are 0.75, 0.78 and 0.77, respectively.

Dataset 2: Thai encyclopedia in the topic of plant. The size of dataset 2 is about 25,476 words. By using both cues the system can extract about 224 concepts and 198 taxonomic relations. The precision, recall and F-measure of the system when testing with this dataset are 0.63, 0.72 and 0.67, respectively. The accuracy of the experiment with this dataset is less than the previous one because the documents in this genre contain a few ontological lists.

The evaluation results of the ontology extraction system classified by the data test set are shown in Table 20. From both data sets, the system can extract totally 2,228 concepts and 2,325 relations. The performances of the system, in total average, are 0.74 of the precision, 0.78 of the recall and 0.76 of the F-measure. As mentioned above, there are some ambiguities of the cue words /pen/ (be) that still remain in both datasets.

Table 20 The evaluation results of the system classifying by the data test set.

Data Set	Test Corpus size(words)	No. of concepts	No. of relations	P	R	F
DataSet1	277,164	2,043	2,154	0.75	0.78	0.77
DataSet2	25,476	224	198	0.63	0.72	0.67
Total	**302,640**	**2,228**	**2,325**			
average				**0.74**	**0.78**	**0.76**

1.1.3 Comparing with the Hearst's technique.

In this sub section, the system was evaluated by comparing with the Hearst's technique (Hearst, 1992), the most well known of pattern-based technique. Hearst's technique selected only terms that occur nearest the cue word of lexico-syntactic patterns to be the ontological terms. However, the ontology terms might occur far away from the cue word. This research then proposes the technique for selecting the correct ontology term even it was far away from the cue word. Figure 38 shows the comparison between applying Hearst's technique and our technique by varying the data test size. The experimental results show that our technique has the higher accuracy than Hearst's technique.

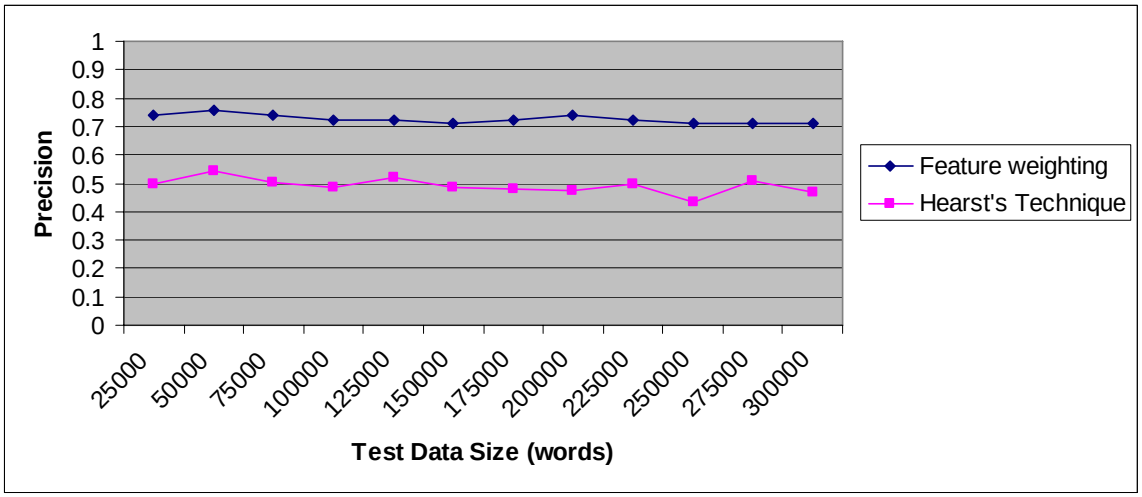


Figure 38 The accuracy of the systems classifying by the technique of selecting term.

1.2 Discovery of the semantic relations of nouns in NPs

In our experiment, we select 1055 pairs of head and modifier of NPs expressing 10 relationships. The 1055 pairs are compound noun that composed of two or three nouns. There are only 107 pairs that composed of three nouns then they need to be segmented or bracketed for identifying the head and the modifier. From the

experimental result, the precision of this bracketing process is 0.71 and the errors are caused by the incorrect pairs of words are occurred more than the right one. For instance, "[*/kak/(lees) /malet/(seed)*] */fai/(cotton) (cotton meal)*". Altogether, from 1055 pairs, the system can translate and disambiguate them, 347 pairs, by using a thesaurus to achieve a precision of 0.96 and a recall of 0.34. The remaining term pairs are processed by using a dictionary to obtain 420 pairs from 708 NPs. with a precision of 0.97 and a recall of 0.59. Table 21 shows the experimental results of the system for translating and selecting words' sense. When combining thesaurus- and dictionary-based translation, the precision and recall of the system are 0.97 and 0.73, respectively. We found that many terms that could not be processed by our system were domain specific terms absent from our knowledge resources.

Table 21 The experimental results for translating and selecting words' sense

Technique	precision	recall
Thesaurus-based Translation	0.96	0.34
Dictionary-based Translation	0.97	0.59
Thesaurus-based + Dictionary-based Translation	**0.97**	**0.73**

Concerning the learning of semantic relations, we did an experiment with 767-tagged examples for 10 semantic relations. The decision tree learning generated about 90 classification rules and the SVM learning system generated 10 learning models accordingly to the relations. The experimental results are shown in Table 22. Both learning systems show that the 'Topic' relation has the highest precision, as all NPs expressing this relation have the head word in the class 'data' or 'information'. The following examples show the NPs having the class 'data#1' as head word.

(21) */khomun/(data#1) /phum-akat/(weather#1)*
(weather data)

(22) */khomun/(data#1) /sathiti/(statistic#1)*
(statistical data)

Besides, the 'IS-A' relationship achieves the lowest precision. This is due to the fact that there is a problem of data sparseness then it is difficult to discovery the common semantic class of the head word or modifier word of NPs which embedded this relationship. The data sparseness is caused by a variety of heads and modifiers such as *plant* or *animal species* as follows:

(23) */khao/(rice#1) /bale/(barley#1)*
(Barley)
(24) */nok/(bird#2) /krachokthet/(ostrich#2)*
(Ostrich)
(25) */phlia/(aphid#1) /kai-chae /(bantam#1)*
(Durian psyllid)

Table 22 The evaluation results concerning the discovery of the semantic relation from NPs.

Relation	**SVM**		**Decision tree**	
	P	R	P	R
IS-A	0.77	0.70	0.61	0.75
container	0.70	0.54	0.70	0.54
location	0.82	0.80	0.75	0.76
made-of	0.82	0.82	0.76	0.85
purpose	0.83	0.67	0.65	0.57
possessor	0.85	0.86	0.86	0.74
property	0.85	0.76	0.88	0.74
source	0.88	0.82	0.80	0.67
part-whole	0.89	0.83	0.86	0.76
topic	1.00	0.90	1.00	0.90
Average	**0.84**	**0.77**	**0.79**	**0.73**

2. Thesaurus-based ontology construction

In this process, the AGROVOC thesaurus relationships are cleaned and refined before converting to ontological relationships. The BT/NT relationships are cleaned and converted to hypernym/hyponym relationship by using expert-defined rules, NP analysis and WordNet alignment techniques while the USE/UF relationships are cleaned and converted to synonym relationship by using only WordNet alignment technique. The RT relationships are refined to more specific relations by using expert-defined rules and training rules. We ran an experiment testing the training rules technique for refining the RT relationship by using 100 examples with 5 semantic relationships i.e. 'scientificNameOf' *(Malus <scientificNameOf> apples)*, 'usedToMake' *(almond <usedToMake> almond oil)*, 'subclassOf' *(avocados <subclassOf> tropical fruits)*, 'producedFrom' *(goat milk <producedFrom> goats)*, derivedFrom *(rice straw <derivedFrom> rice)*. It produced around 10 classification rules. The experimental results using these rules as well as expert-defined rules, noun phrase analysis, and WordNet Alignment are shown in Table 23.

Table 23 The experimental results classified by relationship

Relation-ship	No.	No. of refinement	Expert-defined rules		NP Analysis		WordNet Alignment		Training Rules	
			No.	P	No.	P	No.	P	No.	P
BT/NT	32176	21072	16587	1.00	2062	0.95	2423	0.95	**	**
USE/UF	21605	3553	-	-	-	-	3553	0.70	**	**
RT	27589	1420	622*	1.00*	-	-	-	-	798*	0.72*
Total	**81370**	**26045**	**17209**	**1.00**	**2062**	**0.95**	**5976**	**0.80**	**798***	**0.72***

Remarks: - indicates this technique can not revise this relationship, * indicates the experiment is run with some data, ** indicates the experiment is in initial state.

Based on an expert's review of a small sample of data, some initial rough estimates were made regarding the precision of the methods. The precision of the Expert-defined Rules technique was estimated to be around 1.00 and 0.95 correctness

for NP Analysis. The WordNet Alignment technique was estimated to be lower, about 0.94 precision, because some synonym relationships in WordNet should be replaced with the 'abbreviation_of' relationship. For example, in *AMP* <synonym> *Adenosine monophosphate*, <abbreviaton_of> should be used. The precision of the Training Rules technique was estimated to be about 0.72. Sources of error include ambiguity in concept classes used as arguments for a given rule, such as the following, 'If class X is *food#1* and class Y is *food#1*, and X RT Y, then X <*usedToMake*> Y' where, because X and Y belong to the same concept class, the system cannot distinguish between X and Y and may generate erroneous relationships, e.g., *pork* <*usedToMake*> *hams*, and *hams* <*usedToMake*> *pork*. These cases can be revised by using information from text such as the sentence '*pork is used to make ham*' and it needs the technique of verb analyzing for identifying the semantic relation.

3. Dictionary-based ontology construction

The experiment in this step, we can extract 37,110 terms with 21,620 relations from Thai Plant Name dictionary. By random checking, the accuracy of the system is 1.00 because this specific dictionary contains explicit taxonomic structures then it is not a complex task for extracting ontological terms and relationships.

4. Ontology Integration and Reorganization

The process of ontology integration is iteratively occurred when the system adds each extracted concepts and relationships to the core tree. The system integrated 1,452 relationships that extracted from corpus and dictionary to the core tree with term matching technique and 590 relationships with partially terms' head words matching technique. The accuracies of these techniques are 0.82 and 1.0 as shown in Table 24.

Table 24 The evaluation of ontology integration

Technique	Nb. of Rel.	Accuracy
Term matching technique	1,452	0.82
Partially terms' head words matching	590	1.0

Although these methods are looked simple, they give a promise results. The errors of term matching technique are caused of the terms have the same label with different meanings. For example, the term "/*kaew*/" can mean a flower or a variety of mango. Another example is /*chomphu*/ that can be '*rose apple*' or '*guava*' (called in the south of Thailand). This problem will be solved in the future work by defining the rules or constraints e.g. flower is a disjoint concept of fruit and not considering the local names in the merging process.

Moreover, we evaluate the accuracy of the ontology integration system according to the problem's categories as shown in Table 25.

Table 25 The evaluation of ontology reorganization system classified by the problem's categories

Problems Categories	Frequency	Accuracy
Term Mismatch	142	0.46
Redundancy in the class hierarchy	223	1.00
Conflict relationship	13	0.77

The accuracy of the system for merging the terms that have different names or have the problem of term mismatch is 0.46. The result shows that our proposed technique by using edit distance for solving the problem of term mismatch is work well only for transliterate terms e.g. */dai then em 45/ (Dithane M-45)* and */dai thaen em 45/ (Dithane M-45)*. However, this technique does not work with noun phrases that have the same semantic generated from different terms. These terms do not have the similarity when measured by using edit distance technique. For example,

(21) */ya/(medicine) /kha/(kill) /malaeng/(insect)*
(Insecticide)

(22) */sankhemi/(chemical) /kamchat/(eliminate) /malaeng/(insect)*
(Insecticide)

For solving this problem, the system needs the knowledge of lexicon meaning for mapping the synonym terms such as */kha/(kill)* has the same meaning as */kamchat/(eliminate)* and the techniques of paraphrase resolution are helpful. Concerning the problem of redundancy in the class hierarchy, the system can delete all redundancy relationships then the accuracy is 1.00. Moreover, the accuracy of conflict-relationship problem solving is 0.77. We used the statistical technique to solve this problem by deleting the relationship that has less frequency than another conflict relationship. However, some less frequent relationships could be used to imply the correct relationship e.g. the correct relation *HYPONYM(soil, loam soil*) has less frequently than the incorrect relation *HYPONYM(loam soil, soil)*. Some relationship like this example can be solved by analyzing the head word of NP i.e. *soil* is the head word of phrase *loam soil* then *soil* should be the hypernym of *loam soil*.

Table 26 shows the number of concepts and relationships, extracted from each resource, as well as the total number of concepts and relationships in the ontology resulting from the process of organization.

After a random check with 1,000 integrated terms, the organizing system's accuracy is 0.86 and the coverage is 0.90. The errors are due to the particular characteristics of the corpus extraction terms

Table 26 Experimental results classified by the resources

Source	Methodology	Nb. of Terms	Nb. of Rel.	Accuracy	Coverage
Text (90 docs.)	Ontology learning by using Cues	2,228	2,325	0.74	0.78
	Relation Extraction in Phrase Level	585	767	0.97*0.84 = 0.82	0.73*0.77 = 0.56
Thesaurus		33,450	26,045	0.87	0.87
Dictionary		37,110	21,620	1.00	1.00
3 Sources		**59,971**	**41,677**	**0.86**	**0.90**

When evaluating the system with three criteria, accuracy, coverage and portability, we can conclude as follows:

Accuracy: We evaluate the accuracy of the system by using the precision value. The precision of the ontology extraction system based on the thesaurus and dictionary is very good since they are structure sources. When considering only text-based ontology extraction, the average accuracy of the system for extracting both taxonomic and non-taxonomic relations is 0.76. Concerning the relation extraction in phrase level, it composed of two sequence steps that are 1) word translating from Thai to English and word disambiguating process and 2) semantic relationship learning, and then we multiply their precisions (i.e. 0.97*0.84) as the precision of this task that is 0.82 of the accuracy. The errors of text-based ontology construction are analyzed as previous mentioned i.e. cue word ambiguity and various semantic classes of head noun and modifier noun of NPs. Based on this analysis and proposed solutions, the performance of the system can be improved.

Coverage: The recalls are considered as the coverage of the system. The recall values of thesaurus- and dictionary-based ontology extraction system are equal to the precision values because these sources contain exactly number of terms and relationships. Hence, the number of total extracted results is equal to the number of total correct results then the recall value is equal to precision value. These two sources

give promise recalls. Concerning the text-based ontology extraction, the coverage of the system is 0.71. The recall of relation extraction in phrase level is very low because there are many terms that are domain specific terms and absent from our knowledge resources i.e. dictionary and thesaurus. Thus, these terms can not be translated. For solving this problem, we can apply NE classes for tagging semantic classes of these terms, e.g. /*namdokmai*/ is a kind of mango and its NE class is '*plant*', and the coverage of the system will be increased.

Portability: There are very difficult for evaluating this criterion as a quantity value. Accordingly, we evaluate this criterion by analyzing the competency of our techniques for applying to other domain and other language. Concerning text-based technique, our cues are very general, i.e. the cue word: /chen/(such as), /dai-kea/(i.e.), /pen/(be) and the item lists: bullet list and numbering list, and they can occur in other domains and other languages. By this reason, we can conclude that our proposed techniques can be applied to other domain and they can be adapted for other languages by changing some linguistic grammars e.g. NP patterns. In addition, the proposed technique for extracting ontology based on thesaurus can be directly applied to other thesaurus and the dictionary-based ontology extraction technique, proposed here, can be applied to other dictionaries that have explicit structure by modifying the rules of task-oriented parser in the process of structure analysis. For extracting ontology from other dictionaries that do not have the explicit structure of terms, the ontological elements can be extracted from words and their definitions by applying the approach of text-based ontology extraction.

CONCLUSION AND RECOMMENDATION

Conclusion

Ontology with a precise semantic is very important for improving information systems by and large, for automating reasoning, as well as for sharing and managing knowledge. In this work, we are interested in building an ontology that is both taxonomic and non-taxonomic; hence we included relations like meronyms, and functional relations. One of our aims is to solve the ontology development bottleneck problem by exploiting the enormous body of knowledge gathered over the years in various types of classification systems and thesauri. Then AGROVOC and a Plant dictionary have been used as a seed to transform these terms to ontological terms. Terms grow very fast in many domains. In the case of agriculture, this is illustrated by terms like, *plant species, disease name, chemical names, and pathogen names.* This is where text corpora become very useful, since they contain a lot of, frequently updated information. However, texts present also the hardest challenge, as they require much more work concerning the acquisition of ontological terms and mining their relationships.

We presented and evaluated here the learning methodologies for the automatic building of ontology that is composed of term and relation extraction. The ontology is constructed from several resources: text corpus, thesaurus and dictionary. In order to build ontology of Thai text, a shallow parser is used for candidate terms extraction, and cues-words: lexico-syntactic patterns and item list (numbering and bullet list) are used for relation extraction. Concerning the lexico-syntactic patterns, there are some problems of many candidate terms and cue word sense ambiguity, then the lexicon and the co-occurrence features of each candidate term are used to solve this problem. We also applied information gain ratio for weighting each feature to measure its relevance. This technique can be used to extract the hypernym term of the item lists from the set of candidate terms. One of the most important advantages of using cues is that it reduces the problems of concept and relation labeling which are the crucial problems of the research of ontological engineering.

Concerning with the thesaurus-based ontology construction, we propose three methodologies for data cleaning and semantic relationship refinement to solve the problem of producing well-defined semantics from poorly defined or underspecified semantics in a thesaurus. The system refines the semantic relationships though noun phrase analysis, WordNet alignment, and semantic relationship rules, some generated by experts and others generated from annotated examples by an inductive statistical machine learning system. Finally, the relationships were verified by the experts. Initial results are promising. Moreover, in order to extract ontological terms and relationships from a specific dictionary, a task oriented parser is used to build the ontological tree. Finally, all integrated sub-ontologies are integrated to the core-tree by using term matching technique and the ontology is reorganized for pruning the inconsistency relationships.

We consider our results quite good, given that the experiment is preliminary, but the vital limitation of our approach is that it works well only for documents that contain a lot of cues. Based on our error analysis the performance of the system can be improved and the methodologies can be extended to other sets of semantic relations and applied to other domain. However, concerning the process of pruning the redundancy relations, we analyze that these redundancy relations have the benefit for analyzing the using of words in text. For example, the relation *HYPONYM(animal, mammal)* and *HYPONYM(mammal ,cat)*. In our system, if we found the relation of *HYPONYM(animal, cat)*, the system will delete this redundancy relation. However, we found that texts sentence ‘cat is an animal.’ more than the sentence ‘cat is a mammal.’ then the redundancy relations have the useful for analyzing text and they can show the using of words in text.

Further works to complete the research are performing more tests on large corpora and evaluating the ontology by using in the real applications. Another research direction is to extract the semantic relation embedded in the sentence without the cues and to represent the confidence of relation by adding the frequency of the occurrence of each relation. Moreover, the techniques for translating Thai to English

and selecting words' sense can be applied for helping the task of Thai WordNet construction.

Recommendation

Although there are many studies in the field of ontology engineering, it still has many open problems. Special attention to improve the field must be given to the following tasks.

Axiom learning: The only report we found on learning axioms is by HASTI (Shamsfard and Barforoush, 2003), which learns some axioms in restricted circumstances. This system learned the explicit axioms from conditional sentences in texts. The implicit axioms from text need more attention.

Task ontology learning: Most of the methods for building ontology are focused on domain specific ontology. However, task ontology is another important knowledge for supporting the process of real world application. Hence, automatic construction of task ontology needs more research.

Evaluating ontology learning systems: Before using the ontology in applications, it should be evaluate the content in the aspects of consistency, completeness and conciseness (Gomez-Perez, 1996). Currently ontology learning systems (Gomez-Perez, 1999; Gruninger and Fox, 1995) are evaluated their results in specific domains. Special attention must be given to find formal, standard methods to evaluate the ontology learning systems.

Managing the evolution of the ontology: Ontology is a set of dynamic entities that evolves over time. While maintaining it throughout its life-cycle is an endless task, since new terms are created, others becoming obsolete. In the domain of agriculture, certain terms grow very fast, for example, *plant species, disease name, etc.,* while others tend to disappear. They have lived their time, or they have simply become obsolete with the appearance of new technologies. For example: *pineapple*

was an instance of an *economic plant* in the year *2002* while in *2004*, *oil palm* has become an economic plant instead as illustrated in Figure 39.

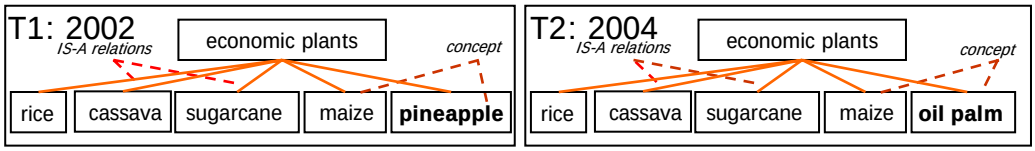


Figure 39 Examples of domain specific ontologies for an economic plant at different points in time

The management of ontology evolution and the relationships between different versions of the same ontology are crucial problems to be solved.

LITERATURE CITED

Agirre, E., O. Ansa, E. Hovy and D. Martinez. 2000. Enriching very large ontologies using the WWW. *In* **Proceedings of the Workshop on Ontology Construction of the European Conference of AI (ECAI-00)**.

Aldea, A., R. Banares-Alcantara, J. Bocio, J. Gramajo, D. Isern, A. Kokossis, L. Jimenez, A. Moreno and D. Riano. 2003. An ontology-based knowledge management platform, *In* **Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03).**

Aramaki, E., T. Imai, M. Kashiwagi, M. Kajino, K. Miyo and K. Ohe. 2007. Toward medical ontology using natural language processing. **Ontologies and Lexical Resources for Natural Language Processing**. Cambridge University Press.

Ayan, N. F. 1999. Using information gain as feature weight. *In* **Proceedings of the 8th Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN'99)**, Istanbul, Turkey.

Barker, K. and S. Szpakowicz. 1998. Semi-automatic recognition of noun modifier relationships, pp. 96-102. *In* **Proceedings of the 17th International Conference on Computational Linguistics**, Montréal.

Bodenreider, O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. **Nucleic Acids Res** 32(Database issue).

Borst, W.N. 1995. **Construction of Engineering Ontologies**. PhD Thesis. Centre for Telematica and Information Technology, University of Tweenty Enschede, The Netherlands.

Chanlekha, H. and A. Kawtrakul. 2004. Thai named entity extraction by incorporating Maximum Entropy Model with simple heuristic information. *In* **Proceedings of the IJCNLP' 2004**, Hainan Island, China.

Cimiano, P., L, Schmidt-Thieme, A, Pivk, S, Staab. 2004. Learning Taxonomic Relations from Heterogeneous Evidence. *In* **Proceedings of the 16th European Conference on Artificial Intelligence (ECAI 2004)**, Workshop on Ontology learning and population, Valencia, Spain.

Clark, P., J. Thompson, H. Holmback and L. Duncan. 2000. Exploiting a thesaurus based semantic net for knowledge-based search, pp. 988-995. *In* **Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence**.

Davies, J., A. Duke and A. Stonkus. 2002. OntoShare: using ontologies for knowledge sharing, *In* **Proceeding of WWW2002 Semantic Web workshop, 11th International WWW Conference WWW2002**.

Doan, A., J. Madhavan, P. Domingos, A. Halevy. 2004. Ontology matching: a machine learning approach, pp. 397-416. *In* S. Staab and R. Studer (eds.). **Handbook on Ontologies in Information Systems**. Springer-Velag. Invited paper.

Duch, W. and K, Grudzinski. 1999. Weighting and selection of features, pp. 32-36. *In* **Proceedings of Intelligent. Information Systems VIII Workshop**, Ustron, Poland.

Faure, D. and C. Nedellec. 1998. A corpus-based conceptual clustering method for verb frames and ontology acquisition. *In* **LREC workshop on adapting lexical and corpus resources to sublanguages and applications**.

Fensel, D., F. V. Harmelen, M. Klein and H. Akkermans. 2000. On-To-Knowledge: ontology-based tools for knowledge management. *In* **Proceedings of eBusiness and eWork 2000 (EMMSEC 2000)**.

Food and Agriculture Organization of the United State Nations (FAO). 2007. **AGROVOC Thesaurus**. Available Source: http://www.fao.org/aims/ag_intro.htm, April 26, 2007.

Girju, R., A. Badulescu and D. Moldovan. 2003. learning semantic constraints for the automatic discovery of part-whole relations. *In* **Proceedings of the Human Language Technology Conference**, Edmonton, Canada.

Gomez-Perez A. 1996. A framework to verify knowledge sharing technology. **Expert Systems with Application** 11(4): 519-529.

Gomez-Perez A. 1999. Evaluation of taxonomic knowledge on ontologies and knowledge-based systems. *In* **Proceeding of 12th Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW'99)**, Canada.

Gomez-Perez A. and D. Manzano-Macho. 2003. A survey of ontology learning methods and techniques. **Deliverable 1.5, OntoWeb Project**.

Gruber, T. R. 1993. A translation approach to portable ontology specifications. **Knowledge Acquisition** 5(2): 199–220.

Gruninger M., M.S. Fox. 1995. Methodology for the design and evaluation of ontologies. *In* **Proceeding of IJCAI95 Workshop on Basic Ontological Issues in Knowledge Sharing**.

Guarino, N., 1998. Formal ontologies and information systems. *In* **Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'98)**, Trento, Italy.

Hall, M.A. and Smith, L.A. 1998. Practical feature subset selection for machine learning, pp. 181-191. *In* **Proceedings of the 21st Australian Computer Science Conference**, Perth, Australia.

Hearst, M. 1992. Automatic acquisition of hyponyms from large text corpora, *In* **Proceedings of the 14th International Conference on Computational Linguistics**.

Imsombut A., M. Suktarachan, W. Yingsaree and A. Kawtrakul. 2004. Country report and activity from Thailand: ontology construction and maintenance system in agricultural domain. *In* **Proceeding of Workshop: The AFITA/WCCA2004, The Multilinguality in Agricultural Information Access Workshop**, Bangkok, Thailand.

Imsombut, A. and A. Kawtrakul. 2005. Semi-automatic semantic relations extraction from Thai noun phrases for ontology learning. *In* **Proceeding of the Sixth Symposium on Natural Language Processing 2005 (SNLP 2005)**, Chiang Rai, Thailand.

Imsombut, A. and A. Kawtrakul. 2007. Automatic building of an ontology on the basis of text corpora in Thai, To be appear in **Language Resources and Evaluation Journal special issue on Asian Language technology**, Springer.

Jacquemin, C. 2001. **Spotting and discovering terms through NLP**, MIT Press, Cambridge MA.

Jannink, J. 1999. Thesaurus entry extraction from an on-line dictionary. *In* **Proceedings of Fusion '99**, Sunnyvale CA.

Kang S. J. and J.H. Lee. 2001. Semi-automatic practical ontology construction by using a thesaurus, computational dictionaries, and large corpora, pp.45-52. *In* **Proceedings of ACL 2001 Workshop on Human Language Technology and Knowledge Management**, Toulouse, France.

Kasetsart University, **Thai Thesaurus**, Se-Education Public Company Limited, 1992

Kawtrakul, A. 2004. The development of resources on network for NLP. Technical Report. October.

Kawtrakul, A., A. Imsombut, A. Thunkijjanukit, D. Soergel, A. Liang, M. Sini, G. Johannsen, and J. Keizer. 2005. Automatic term relationship cleaning and refinement for AGROVOC. **AOS Workshop at EFITA 2005**, Vila Real, Portugal.

Kawtrakul A., M. Suktarachan and A. Imsombut. 2004. Automatic Thai ontology construction and maintenance system. *In* **Proceeding of OntoLex Workshop on LREC**, Lisbon, Portugal.

Kawtrakul, A. and P. Waewsawangwong. 2000. Multi-reature extraction for printed Thai character recognition. *In* **Proceeding of the Fourth Symposium on Natural Language Processing 2000 (SNLP2000)**, Chiang Mai, Thailand.

Ketsuwan, C., N. Pengphon and A. Kawtrakul. 2000. Automatic Thesaurus Extraction for Thai Text Retrieval Enhancement, *In* **Proceeding of WAINS 7 : E-Business for the new Millennium**, Bangkok, Thailand.

Kietz, J.U., A. Maedche and R. Volz. 2000. A method for semi-automatic ontology acquisition from a corporate intranet. *In* **Proceeding of Workshop Ontologies and Text, co-located with the 12th International Workshop on Knowledge Engineering and Knowledge Management (EKAW'2000)**, Juan-Les-Pins, France.

Laird, M. 2005. **Algorithm-SVM**. Avialable Source: http://search.cpan.org/dist/Algorithm-SVM/, April 26, 2007.

Landau M. F. and E. Morin. 1999. Extracting semantic relationships between terms: supervised vs. unsupervised methods, pp. 71-80. *In* **Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure**, Dagstuhl Castle, Germany.

Levenshtein, V. I.. 1966. Binary codes capable of correcting deletions, insertions and reversals. **Soviet Physics Doklady** 10(8): 707-710.

Li, S., Q. Lu and W. Li. 2007. Experiments of Ontology Construction with Formal Concept Analysis, **Ontologies and Lexical Resources for Natural Language Processing**, Cambridge University Press.

Lin, D. and P. Pantel. 2002. Concept discovery from text, pp. 577-583. *In* **Proceeding of the International Conference on Computational Linguistics**, Taipei, Taiwan.

Maedche, A. and S. Staab. 2001. Ontology learning for the semantic web. **IEEE Intelligent Systems** 16(2): 72–79.

Maedche, A., S. Staab, R. Studer, Y. Sure and R. Volz. 2002. SEAL - Tying up information integration and web site management by ontologies. **IEEE-CS Data Engineering Bulletin, Special Issue on Organizing and Discovering the Semantic Web**.

Mann, G. 2002. Fine-grained proper noun ontologies for question answering. *In* **SemaNet'02: Building and Using Semantic Networks**.

McGuinness, D. L., R. Fikes, J. Rice, and S. Wilder. 2000. An environment for merging and testing large ontologies. *In* **Proceedings of the Seventh**

International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Colorado, USA.

Miller, A.G. 1995. Wordnet: A lexical database for English. **Communications of the ACM** 38(11): 39-41.

Miller, G., C. Leacock, T. Randee, and R. Bunker. 1993. A semantic concordance, pp. 303-308. *In* **Proceedings of the 3rd DARPA Workshop on Human Language Technology**.

Mladenic, D. 1998. Feature subset selection in Text-Learning, p.95-100. *In* **Proceedings of the 10th European Conference on Machine Learning**.

Moldovan D. and R. Girju. 2001. An interactive tool for the rapid development of knowledge Bases. *In* **International Journal on Artificial Intelligence Tools (IJAIT)** 10(1-2).

Mori, T. 2002. Information gain ratio as term weight: the case of summarization of IR results, pp. 688-694, *In* **Proceedings of the 19th International Conference on Computational Linguistics (COLING 02)**, Taipei.

Nashvili, M. 2004. **Decision Trees**. Available Source: http://decisiontrees.net, April 26, 2007.

National Electronics and Computer Technology Center (NECTEC). 2007. **LEXiTRON**. Available Source: http://lexitron.nectec.or.th/, April 26, 2007.

Navigli, R., P. Velardi and A. Gangemi. 2003. Ontology learning and its application to automated terminology translation. **IEEE Intelligent Systems** 18(1).

Nedellec, C. 2000. Corpus-based learning of semantic relations by the ILP system, ASIUM. **Learning Language in Logic**. Lecture Notes in Computer Science 1925: 259-278.

Noy, N.F. and M.A. Musen. 2000. PROMPT: Algorithm and tool for automated ontology merging and alignment. *In* **Proceedings of AAAI-2000**, Austin, Texas.

Pantel, P., M. Pennacchiotti. 2006. Espresso: A Bootstrapping Algorithm for AutomaticallyHarvesting Semantic Relations. *In* **Proceedings of Conference on Computational Linguistics / Association for Computational Linguistics (COLING/ACL-06).** Sydney, Australia.

Pengphon, N., A. Kawtrakul and M. Suktarachan. 2002. Word formation approach to noun phrase analysis for Thai. *In* **Proceeding of the Fifth Symposium on Natural Language Processing 2002 (SNLP 2002)**, Prachuapkirikhan, Thailand.

Plas, L. and G. Bouma. 2007. Automatic acquisition of lexico-semantic knowledge for QA, **Ontologies and Lexical Resources for Natural Language Processing**. Cambridge University Press.

Pustejovsky, J., A. Rumshisky and J. Castano. 2007. Rendering semantic ontologies: automatic extensions to UMLS through corpus analysis. **Ontologies and Lexical Resources for Natural Language Processing**. Cambridge University Press.

Rijsbergen, C. J. V. 1979. **Information Retrieval**, 2nd edition. Department of Computer Science, University of Glasgow.

Roget, P. M. **Roget's Thesaurus of English Words and Phrases**. Longmans Green, London, 1962.

Schoening, J. 2003. **Standard Upper Ontology Working Group (SUO WG).** Available Source: http://suo.ieee.org/index.html, April 26, 2007.

Schutz, A., P. Buitelaar. 2005. RelExt: A Tool for Relation Extraction from Text in Ontology Extension. *In* **Proceedings of the 4th International Semantic Web Conference.** Galway, Ireland.

Scicluna, J., A. Polleres and D. Roman. 2005. **D14v0.2. Ontology-based Choreography and Orchestration of WSMO Services**. WSMO Working Draft. Available Source: http://www.wsmo.org/TR/d14/v0.2/20051008/, April 26, 2007.

Shamsfard, M. and A. A., Barforoush. 2004. Learning ontologies from natural language texts. **International Journal of Human-Computer Studies** 60(1): 17-63.

Soergel, D., B. Lauser, A. Liang, and F. Fisseha. 2004. Reengineering thesauri for new applications: The AGROVOC example. **Journal of Digital Information** 4(4). Article No. 257, 2004-03-17.

Srikant, R. and R. Agrawal. 1995. Mining generalized association rules, pp. 407-419. *In* **Proceedings of the 21th International Conference on Very Large Data Bases**.

Studer, R., V.R. Benjamins and D. Fensel. 1998. Knowledge engineering: principles and methods. **IEEE Transactions on Data and Knowledge Engineering**.

Stumme, G. and A. Maedche. 2001. FCA-MERGE: bottom-up merging of ontologies, pp. 225-230. ***In*** **Proceeding of 17th International Joint Conference on Artificial Intelligence (IJCAI '01)**, Seattle, WA, USA.

Sudprasert, S. and A. Kawtrakul. 2003 Thai word segmentation based on global and local unsupervised learning. *In* **Proceedings of the 7th National Science and Engineering Conference (NCSEC2003)**, Chonburi, Thailand.

Swartout, B., R. Patil, K. Knight, and T. Russ. 1996. Towards distributed use of large scale ontologies. In University of Calgary, editor, *In* **Proceedings of Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop (KAW)**, Alberta, Canada.

Thai National AGRIS Center. 2003. Thai National AGRIS Center. Available Source: http://thaiagris.lib.ku.ac.th/, April 26, 2007.

The J. Paul Getty Trust. 2007. **Art and Architecture Thesaurus Online**. Available Source: http://www.getty.edu/research/conducting_research/vocabularies/aat/, April 26, 2007.

The Library Development Division, Austin Texas State Library and Archives Commission. 2003. **Core References – Sources**. Available Source: http://www.tsl.state.tx.us/ld/pubs/ corereference/internal/chd.html, April 26, 2007.

The University of Waikato. 2007. **Weka 3: Data Mining Software in Java.** Available Source: http://www.cs.waikato.ac.nz/ml/weka/, April 26, 2007.

Thunkijjanukij, A., P. Lertpongwipoosana, T. Damrongson, P. Tongumpai, K. Phomphunjai, W. Srijankul, J. Suansawan, O. Kongpitak, S. Buddeewong. 2005. The development of Thai agricultural thesaurus, *In* **Proceeding of the 43th Kasetsart University Annual Conference**, Kasetsart University, Bangkok, Thailand.

Uschold, M. 1998. Knowledge level modeling: concepts and terminology. **Knowledge Engineering Review**, 1998.

Vanderwende, L. 1994. Algorithm for automatic interpretation of noun sequences, pp. 782-788. *In* **Proceeding of the 15th conference on Computational linguistics**.

Vapnik, V. 2000. **The Nature of Statistical Learning Theory**, 2nd ed., Springer.

Vapnik, V. 1998. **Statistical Learning Theory**. Wiley-Interscience, New York.

Vargas-Vera, M. and E. Motta. 2004. AQUA - Ontology-based Question Answering System. **Lecture Notes in Computer Science**, 2972:468-477.

Visser, P. R. S., D. M. Jones, T. J. M. Bench-Capon and M. J. R. Shave. 1997. An analysis of ontology mismatches: heterogeneity versus interoperability. **AAAI 1997 Spring Symposium on Ontological Engineering**, Stanford, USA.

Warawudhi, R. 2006. Thai noun phrase analysis for natural language processing. Master Thesis, Department of Linguistics, Kasetsart University.

Wielinga, B., A. Th, S. Wielemaker and J. Sandberg. 2001. From thesaurus to ontology, pp. 194-201. *In* **Proceedings of the International Conference on Knowledge Capture**. ACM Press, Victoria, Canada.

Yamaguchi, T. 2001. Acquiring conceptual relations from domain-specific texts. *In* **Proceedings of the IJCAI 2001 Workshop on Ontology Learning (OL'2001)**, Seattle, USA.

Zuckerman, E. L. 2005. **Clinician's Thesaurus, 6th Edition: The Guide to Conducting Interviews and Writing Psychological Reports**, The Guilford Press.

CIRRICULUM VITAE

NAME : Ms. Aurawan Imsombut

BIRTH DATE : December 24, 1976

BIRTH PLACE : Bangkok, Thailand

EDUCATION :

YEAR	INSTITUTE	DEGREE/DIPLOMA
1998	Kasetsart Univ.	B.Sc.(Computer Science)
2000	NIDA	M.Sc.(ISM.)

POSITION/TITLE : Lecturer

WORK PLACE : Faculty of Information Technology,
Dhurakij Pundit University