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:

$$(\mathcal{C}, \mathcal{I}, \mathcal{R}, \mathcal{F}, \mathcal{A})$$

where:

$\mathcal{C}$ 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 $\mathcal{C}$, that is, each $R \in \mathcal{R}$ is a product of n sets, R: ($C_1 \times C_2 \times ... \times C_n$). 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: ($C_1 \times C_2 \times ... \times C_{n-1} \mapsto C_n$). For example, the function `Pay` is function of the concepts `Price` and `Discount`, and returns a concept `FinalPrice`, that is `Pay: Price`×`Discount`↦`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.

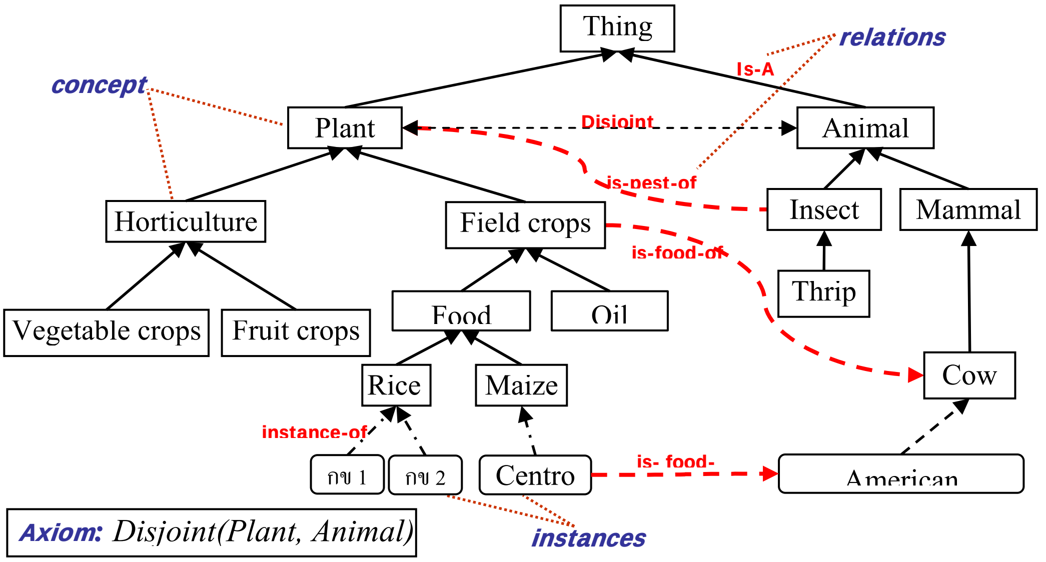


Figure 2 An example of ontology in the domain of agriculture

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.

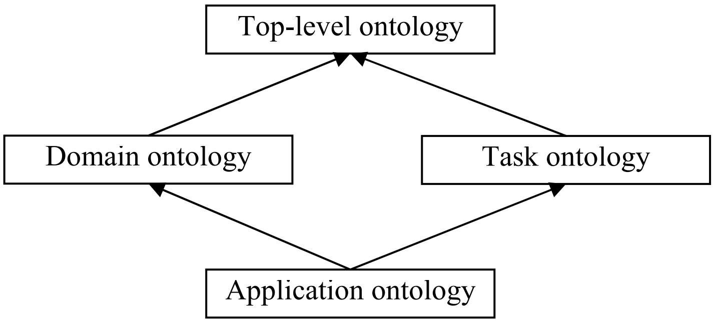


Figure 3 Categorization of ontologies classified by Guarino
Source: Guarino (1998)

- *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).

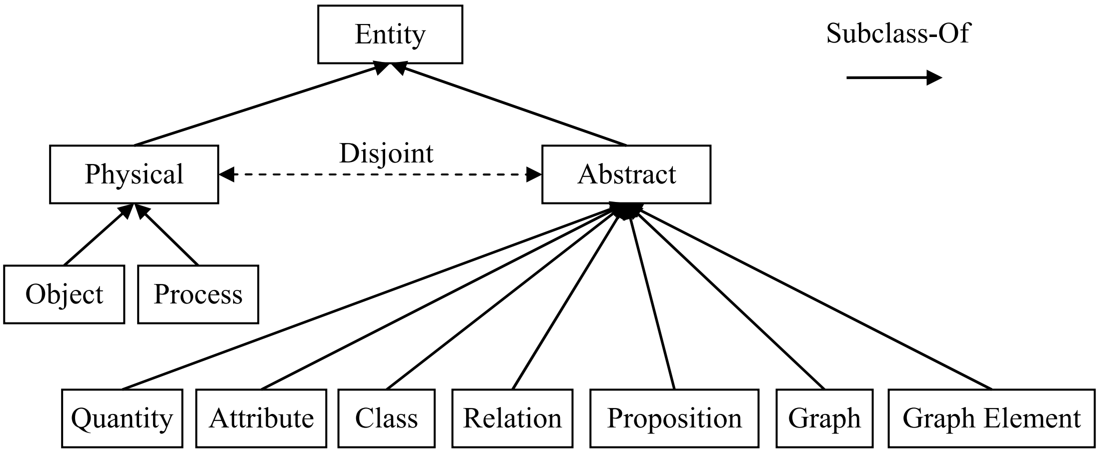


Figure 4 Structure of the first levels of SUMO

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).

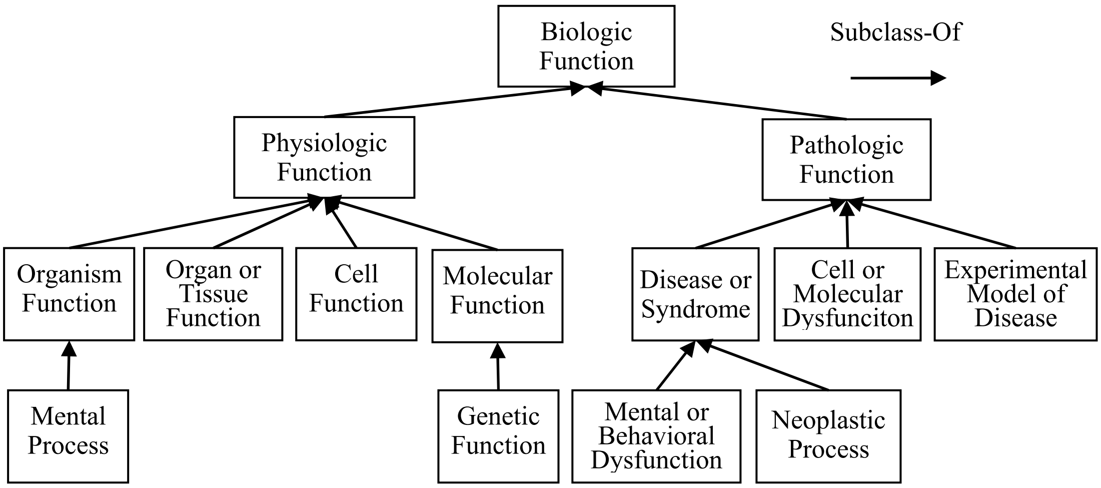


Figure 5 Partial view of the UML ontology

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).

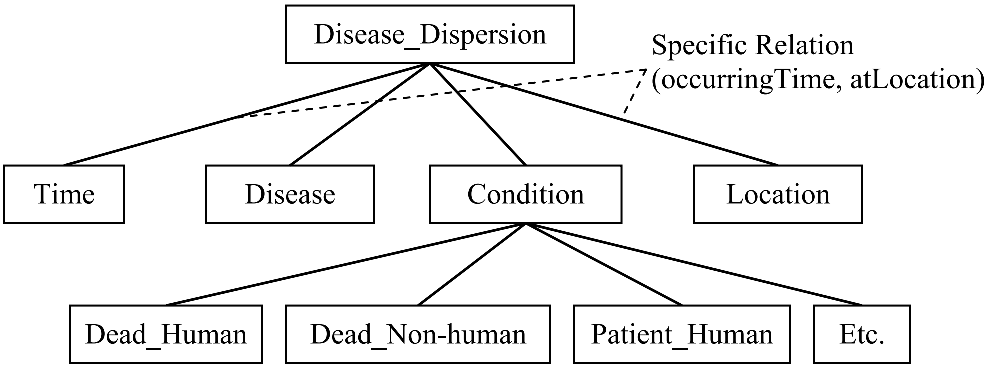


Figure 6 Task ontology about dispersion of disease
Source: Kawtrakul *et al.* (2007)

- 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.

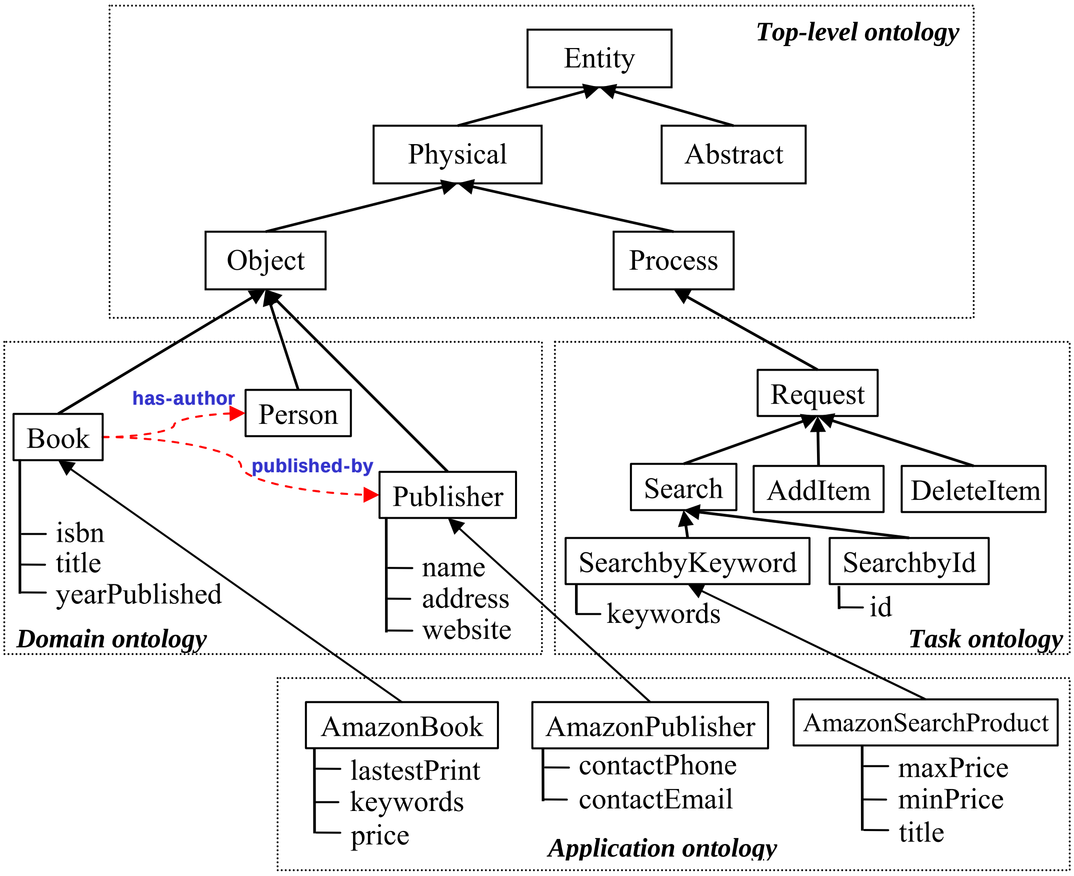


Figure 7 A simple ontology of Amazon web service for books

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.

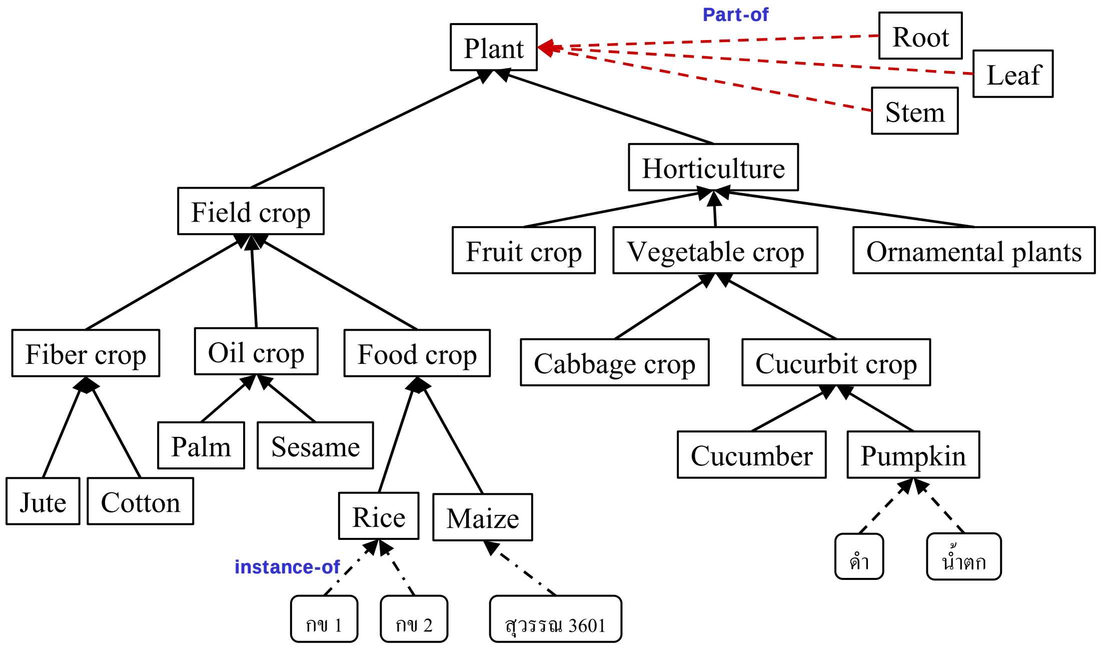


Figure 8 An example of lightweight ontology

- *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).

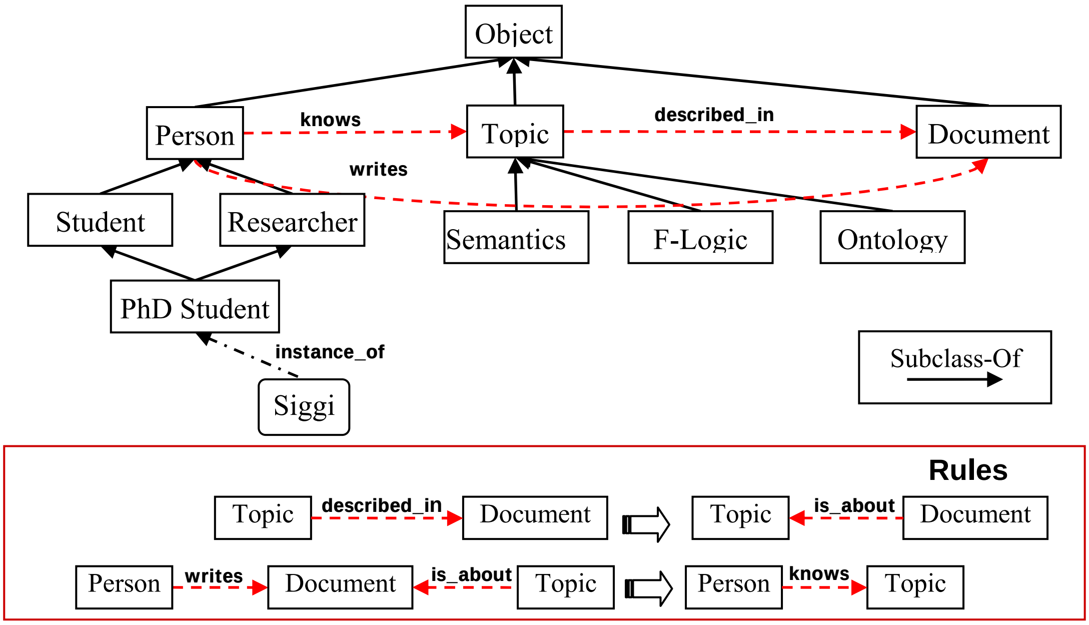


Figure 9 An example of heavyweight ontology

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).

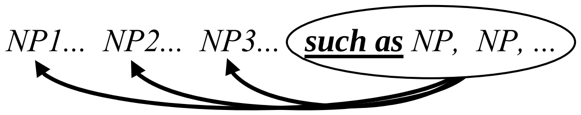


Figure 10 The cue phrase can modify any NP

(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:

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

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

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 ith 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, …, 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.

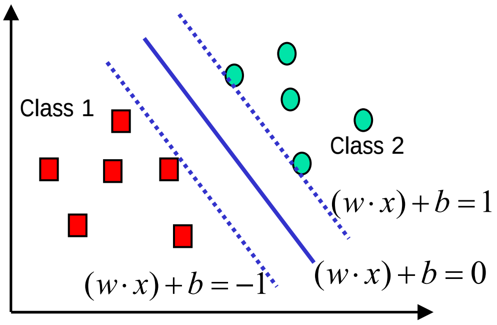


Figure 12 The optimal hyperplane separates circles from rectangles

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): $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2), \gamma > 0$ (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.

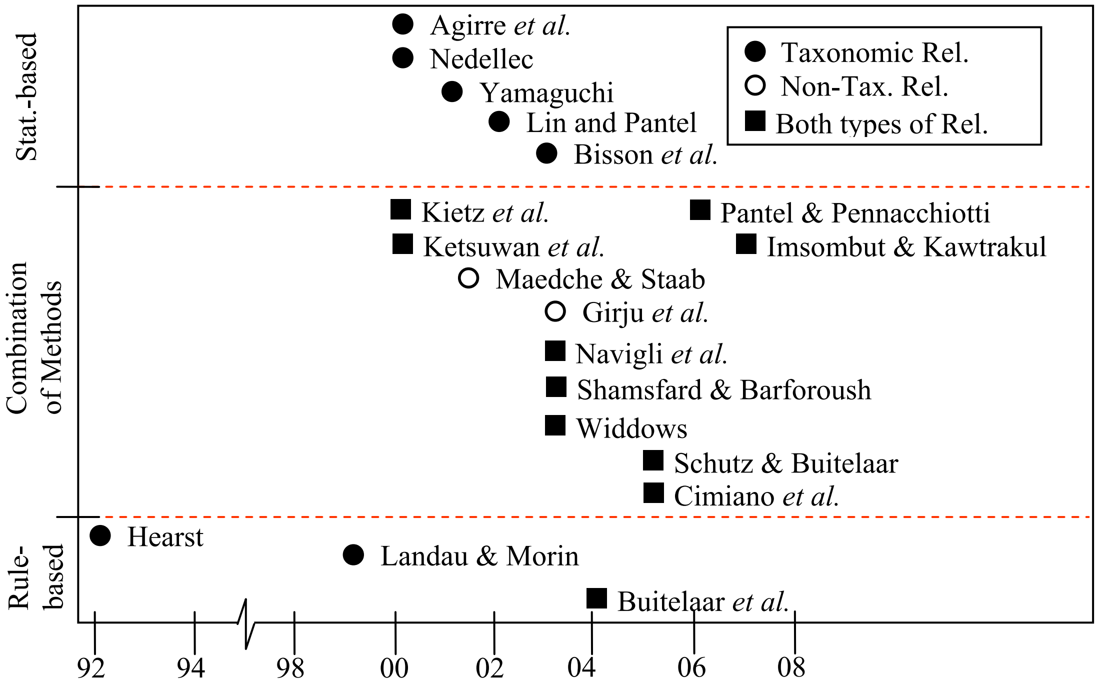


Figure 13 A research map of ontology learning

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.

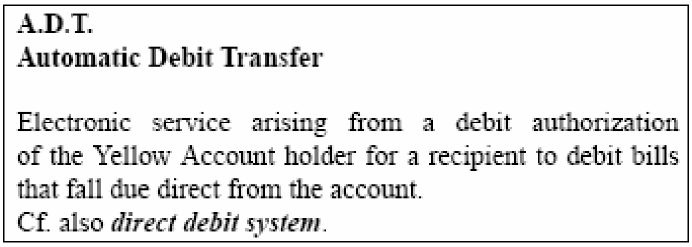
Approaches	Researches	Strengths	Weaknesses
Rule-based Approach	• Soergel (2004) • Wielinga (2001) • Clark *et al.* (2000)	Not complex	Need expert to define rules

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.

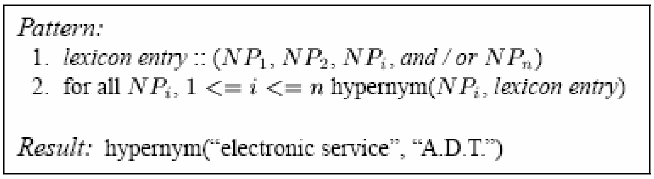
Kietz (2000) uses several heuristic rules to build taxonomy. The first example is the heuristic rule that matching pattern to texts as shown in Figure 14 and 15. From these Figures, the lexicon entry is *A.D.T.*, the NP is *Electronic service* and the system can extract the *hypernym(electronic service, 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* :: $(NP_1, NP_2, NP_i, \textit{and / or } NP_n)$
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 <u>pen</u> phuet phak chanit nueng /*
*(**Cabbage** <u>is</u> a kind of **vegetable**.).*
*(10) */kalam-pli <u>pen</u> rok-nao-le/*
(Cabbage <u>has disease as</u> Soft-rot.).
*(11) */kap-bai <u>pen</u> si-namtan/*
(*Leaf <u>is</u> 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.

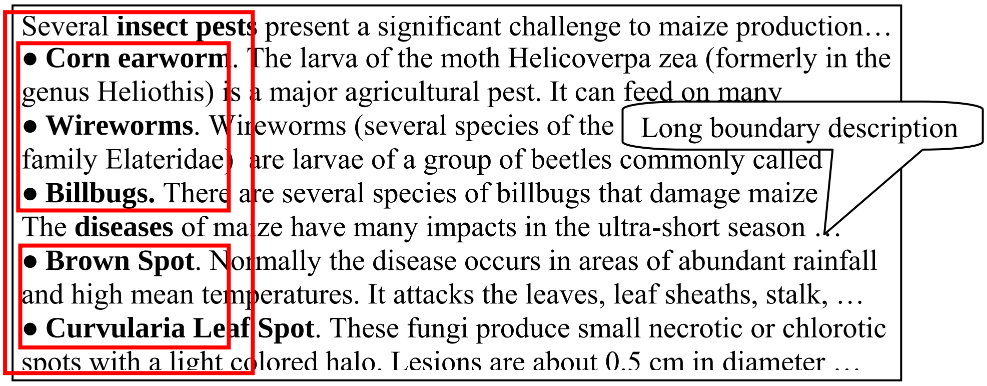
1.4 Problems of Item List Identification

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.

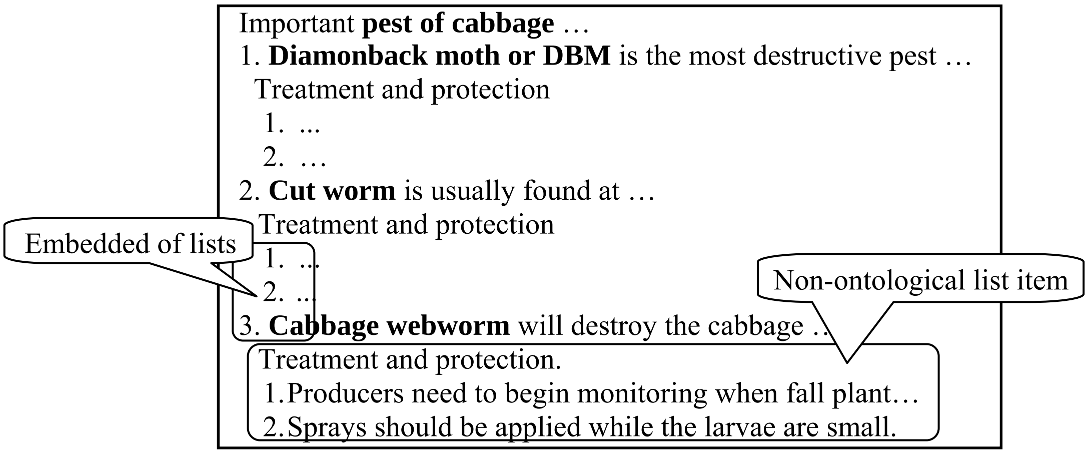
• *Embedded lists.* It frequently happens that a list contains another list, causing some identification problems. We solve this issue by detecting each list following the same bullet symbol or numbering order. Still, there is case which an embedded list may has a following number as the third item that 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.



Several **insect pests** present a significant challenge to maize production…
● **Corn earworm**. The larva of the moth Helicoverpa zea (formerly in the genus Heliothis) is a major agricultural pest. It can feed on many
● **Wireworms**. Wireworms (several species of the family Elateridae) are larvae of a group of beetles commonly called
● **Billbugs.** There are several species of billbugs that damage maize
The **diseases** of maize have many impacts in the ultra-short season …
● **Brown Spot**. Normally the disease occurs in areas of abundant rainfall and high mean temperatures. It attacks the leaves, leaf sheaths, stalk, …
● **Curvularia Leaf Spot**. These fungi produce small necrotic or chlorotic spots with a light colored halo. Lesions are about 0.5 cm in diameter …

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



Important **pest of cabbage** …

1. **Diamonback moth or DBM** is the most destructive pest …

   Treatment and protection

   1. ...
   2. …

2. **Cut worm** is usually found at …

   Treatment and protection

   1. ...
   2. ...

3. **Cabbage webworm** will destroy the cabbage …

   Treatment and protection.

   1. Producers need to begin monitoring when fall plant…
   2. Sprays should be applied while the larvae are small.

Embedded of lists
Non-ontological list item

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* <u>*daikae*</u> *phan* ***sacha, mercedes*** *lea* ***gabrielle****/*

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

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

(Last year a lot of roses have been imported from ***abroad*** <u>*such as*</u> ***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 [[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 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.

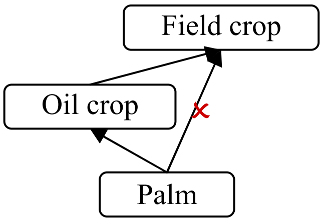
3.1 Term Mismatch

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,

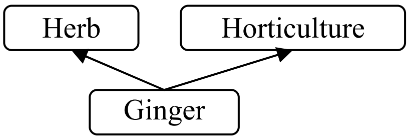


HYPONYM (Palm, Oil Crop),
HYPONYM (Oil Crop, Field Crop)
and *HYPONYM (Palm, Field Crop)*

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.

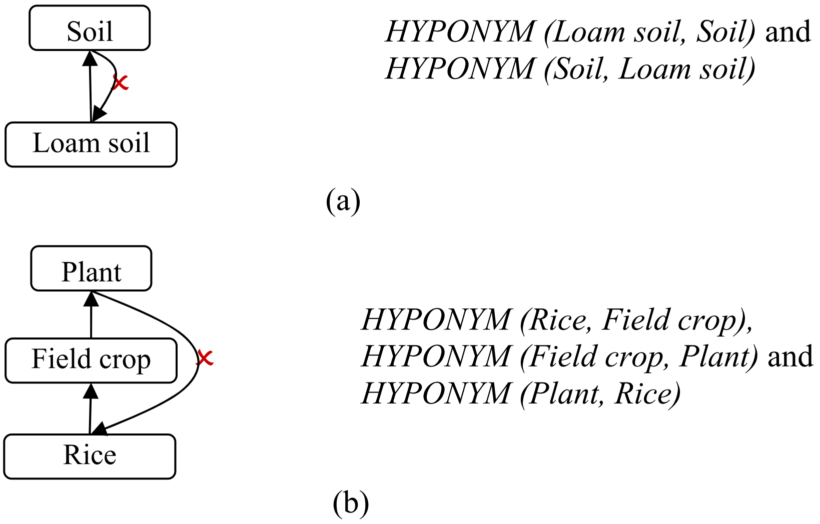


HYPONYM (Ginger, herb)
and *HYPONYM (Ginger, Horticulture)*

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).


Soil
Loam soil
HYPONYM (Loam soil, Soil) and
HYPONYM (Soil, Loam soil)
Plant
Field crop
Rice
HYPONYM (Rice, Field crop),
HYPONYM (Field crop, Plant) and
HYPONYM (Plant, Rice)

(a)

(b)

Figure 21 Examples of conflict relationship

(18) */din <u>pen</u> din-ruan/*

(Soil (that should be used) <u>is</u> loam soil.)

(19) */phuet thi pluk <u>pen</u> khao 50 poesen mai phon 50 poesen/*

(Plants that are cultivated <u>are</u> 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.