A Semantic Framework for Extracting Taxonomic Relations from Text Corpus

Phuoc Thi Hong Doan, Ngamnij Arch-int, and Somjit Arch-int
Department of Computer Science, Khon Kaen University, Thailand

Abstract: *Nowadays, ontologies have been exploited in many current applications due to the abilities in representing knowledge and inferring new knowledge. However, the manual construction of ontologies is tedious and time-consuming. Therefore, the automated ontology construction from text has been investigated. The extraction of taxonomic relations between concepts is a crucial step in constructing domain ontologies. To obtain taxonomic relations from a text corpus, especially when the data is deficient, the approach of using the web as a source of collective knowledge (a.k.a web-based approach) is usually applied. The important challenge of this approach is how to collect relevant knowledge from a large amount of web pages. To overcome this issue, we propose a framework that combines Word Sense Disambiguation (WSD) and web approach to extract taxonomic relations from a domain-text corpus. This framework consists of two main stages: concept extraction and taxonomic-relation extraction. Concepts acquired from the concept-extraction stage are disambiguated through WSD module and passed to stage of extraction taxonomic relations afterward. To evaluate the efficiency of the proposed framework, we conduct experiments on datasets about two domains of tourism and sport. The obtained results show that the proposed method is efficient in corpora which are insufficient or have no training data. Besides, the proposed method outperforms the state of the art method in corpora having high WSD results.*

Keywords: *Taxonomic relation, ontology construction, word sense disambiguation, knowledge acquisition.*

Received September 22, 2017; accepted October 28, 2018
https://doi.org/10.34028/iajit/17/3/6

1. Introduction

Ontologies play a vital role in knowledge management and semantic web due to the abilities in representing knowledge as well as inferring new knowledge from available information. A sufficient ontology may contain four components, including concepts, relations between concepts, instances, and axioms [26]. Among the relations between concepts, taxonomic relations (hierarchical relations) are the backbone component of an ontology [30, 38]. Taxonomic relations in an ontology are represented by the hierarchical arrangement of concepts based on hypernym/hyponym (i.e., broader/narrower) relations. The building of the taxonomic relations, hence, is considered as the first step in the ontology construction and corresponds to identifying hypernym/hyponym relation between the concepts [33, 34, 40].

In general, the taxonomic relations are constructed by domain experts and knowledge engineers. However, this task tends to be tedious, time consuming, and biased [15]. Additionally, due to the development of storage technology, a large amount of data, which describe domain information, is stored as text (e.g., a website that contains text documents about tourism domain is http://www.lonelyplanet.com). Therefore, the automatic extraction of taxonomic relations from text corpora has been investigated [6, 22, 29, 33].

In the literature, there are several approaches used to solve this issue. In fact, typical solution is a pattern-based approach, which was first proposed by Hearst [16]. In this method, the hypernyms in the text corpus are identified through the manual definition of a set of lexical syntactic patterns that frequently occur in the text. To improve the quality of the obtained results, some studies extend Hearst's approach [5, 6, 17, 35]. Most of them combine the Hearst's lexical patterns with Natural Language Processing (NLP) techniques and machine learning methods. However, these studies are limited due to the usage of the lexical patterns only from a text corpus. This corpus is used as input data for hierarchy construction, in which, the number of instances of lexical patterns is insufficient. Another class of approach is to use hierarchical clustering techniques or subsumption methods to extract taxonomic relations [4, 9, 17, 22]. Based on the statistic measurements, the taxonomic relations are produced by arranging the concepts in a hierarchy. Nevertheless, hierarchical clustering methods face a difficulty in assigning appropriate labels for intermediate nodes in a hierarchical tree [37].

Note that, most of the studies mentioned above rely only on the lexical patterns and the statistic measurements between concepts acquired from the text corpus. Therefore, when the text corpus is insufficient, these methods are ineffective.

An alternative approach for extracting taxonomic relations, called Web based approach, aims to leverage the large amount of information in web pages as a source of collective knowledge. A key challenge of this approach is how to acquire the relevant knowledge from the large amount of web pages. Ciamiano and Staab [8] propose a method to tackle this issue by using the results of web queries in search engines. Some studies, described by Sang [34] and Ortega-Mendoza *et al*. [27], attempt to improve the method of Ciamiano by adding the hypernym/ hypornym patterns into web queries. For example, with the pattern "such as", the query can be "hypernym such as" or "such as hyponym". Recently, Rios-Alvarado *et al*. [33] also propose a method for learning the concept hierarchies by identifying the hypernym relations between the concepts extracted from text. To obtain the results closed to the domain of the corpus, the authors improve Sang's method [34] by adding contextual and supervised information into the queries. The contextual information is given by nouns in the corpus with higher frequencies. Meanwhile, supervised information is the nouns extracted from the gloss of a synset[1] in WordNet that define the meaning of a considered term.

Nevertheless, as presented in [33], only the first sense in WordNet of a noun is considered as supervised information. As a result, with ambiguous nouns (i.e., nouns that have multiple senses), the final results are faulty because of the inaccurate data acquired from the web pages. For example, the word "park" has six senses in WordNet. However, the predominant sense of this word in the tourism domain is the first sense and a corresponding hypernym candidate is the word "area". Meanwhile, the right sense of this word in the sport domain is the third one and the word "stadium" is one of the hypernym candidates. As such, these methods cannot be applied in a variety of real-world applications.

In this paper, we propose a semantic framework that can overcome the above mentioned issues by using the Web based approach and Word Sense Disambiguation (WSD) to extract taxonomic relations from a text corpus. In particular, WSD algorithm aims at obtaining the contents of web pages via the queries that are related to interested domain. These web pages are considered as knowledge resources and are used to acquire hypernym candidates of concepts. Moreover, the WSD algorithm is also used to identify the grounded concepts from terms having the same synset as well as to evaluate whether two terms have a hypernym/hyponym relation in WordNet or not.

The contribution of this paper is twofold. First, we propose a new framework that applies word sense disambiguation to extract taxonomic relations from insufficient text corpora. By using the WSD module, most significant data from a large amount of web pages is obtained. Hence, noisy information in the hypernym-candidate extraction is reduced. Second, we introduce a WSD method by modifying knowledge-based methods such as Lesk algorithm [20] and the extension of Lesk algorithm [1]. For the proposed method, although the complexity of time is reduced to k times (k is the number of contexts in corpus that contain an ambiguous word), the accuracy of the obtained results is similar to the original methods.

The remainder of the paper is constructed as follows: In section 2, we present various approaches of previous studies for extracting taxonomic relations. In section 3, we describe the proposed method in detail. In section 4, we show the experiments and compare the results with a previous method. Finally, in section 5, we conclude the paper.

2. Related Work

Many existing studies have investigated the automatic extraction of taxonomic relations from text. To deal with this, there are several approaches which are classified as follows:

First, pattern-based approach that use lexical-syntactic patterns to extract taxonomic relations. One of pioneering studies is proposed by Hearst [16]. In Hearst's method, the hypernyms from the text are identified through the manual definition a set of the lexical syntactic patterns (e.g., "*X such as Y"*, where *X* is a hypernym of *Y*). However, as stated in [21], the accuracy of the results from the different corpora are completely dissimilar. For instance, 52% of the relations were extracted in Grolier's encyclopedia, but there was only 28% of accuracy on a different corpus (Lord of the Rings). To improve the accuracy of automatic methods for extraction of is-a relations from text, Cederberg and Widdows [5] use Lattent Semantic Analysis (LSA) and apply a graph-based model of noun-noun similarity to filter correct hypernym relations. Snow *et at*. [35] automatically extracted Hearst's defined patterns by combining supervised learning with lexical patterns. In another study, the Formal Concept Analysis (FCA) technique is used to group the words into a taxonomy based their attributes [6]. To identify these attributes of words, lexical knowledge dependency parsing instead of searching for patterns is applied. For example, the words such as "hotel", "apartment", "excursion", "car" and "bike" have the common attributes.

The second approach is use hierarchical clustering techniques or subsumption method to extract taxonomic relations [4, 9, 22, 39]. Specifically, Caraballo [4] cluster nouns into hierarchy using data on conjunction and appositive that appear in the Wall Street journal. Hierarchical clustering method of De Knijff *et al*. [9], meanwhile, base on two similar measures of concepts, such as document co-occurrence similarity and window-based similarity. Besides, the subsumption method uses concept co-

[1] Synset is a set of synonyms that share a common meaning. Each synset contains one or more lemmas which represent a specific sense of a specific word.

occurrences in different documents to build a domain taxonomy. This method base on the idea that a concept *X* subsumes concept *Y* if documents that contain *Y* are a subset of the documents that contain *X*. Hence, a concept may have many potential subsumers (i.e., potential hypernyms). To obtain a unique subsumer of a concept from potential subsumers, the authors in [9, 22] propose a score of potential subsumer by adding a weight that describes the relationships between the concept with its ancestors. The potential subsumer which have highest score will be selected as a subsumer of the concept.

The third, web-based approach uses the web as a source of collective knowledge to extract taxonomic relations. In more detail, studies in [27, 34] applied pattern-based methods for collecting the hypernym relations from the web. As reported in [34], to obtain the relevant to data from web, the information hypernym/hypornym is added into the query. For instance, for the pattern "such as", the formula of the query will be "hypernym such as" or "such as hyponym". Meanwhile, authors in [27] proposed a method consisting following steps. First, using a small set seed instances (e.g., pair of hyponym-hypernym such as apple–fruit), to discover a set of lexical patterns from the web. Then, a set of candidates of hyponym-hypernym are extracted on a target document collection by applying the patterns discovered in the previous stage. Finally, the confidence of extracting instances is assessed by evaluation function. Recently, the learning concept hierarchies from a specific domain corpus, proposed by Rios-Alvarado *et al*. [33], includes the two stages. In the first stage, the terms representing concepts are grouped into the clusters (topics) using clustering techniques. After that, relying on the web approach, the concept hierarchy of each topic is constructed. To obtain the results that close to the domain of the corpus, the authors proposed the way to construct a query set by adding information to the query such as contextual and supervised information.

3. Methodology

In this paper, we focus on extracting taxonomic relations from text corpus, which may be insufficient as well as have no training data. Hence, based on the approaches mentioned above, we propose a new framework by applying the web-based approach. In this framework, the WSD algorithm is used to produce the correct senses of concepts before their taxonomic relations are acquired from web pages. This is an improvement over previous web-based methods [8, 33, 34]. Specifically, the proposed framework for extracting taxonomic relations consists two main stages, including concept extraction and taxonomic-relation extraction, and it is divided into four major steps as shown in Figure 1. Each step can be described as follows:

3.1. Key-Term Extraction

Key terms in text documents are phrases consisting of one or more words (single word or multi-word)[2] that are used to describe events, concepts about a certain domain. For example, the term "accommodation" is one of the key terms of the corpus in the tourism domain. This is the first step of the taxonomic relation extraction because the taxonomic relations derived from the irrelevant concepts will be insignificant. By using NLP techniques and statistic methods, the key terms in domain corpus are extracted. The key-term extraction step consists of three sub steps as described in the following sections:

3.1.1. Term Extraction

The term extraction is considered as acquiring nouns and noun phrases in the text corpus. The process of extracting terms was conducted as in [12], consisting of the following steps. First, using Stanford POS tagger tool,[3] nouns and noun phrases are extracted by using the linguistic pattern $(JJ)^{*}(NN)^{+}$ or $(JJ)^{*}(NN)^{*}(NNS)^{+}$ (where, "JJ" is an adjective, "NN" and "NNS" means a singular noun and plural noun respectively, "*": zero or more time occurrences, "+": one or more time occurrences). After that, plural nouns are converted to singular nouns and multi-word terms are enriched by finding single nouns and all the combinations of adjacent words that belong to the phrase, where each combination is only a noun phrase. For example, terms created from the given term "underground metro system" include "underground metro", "metro system", "metro", and "system". Finally, we remove the terms which are stop words. Moreover, the single terms whose lengths are less than 2 and the multi-word terms whose number of words are greater than 5 are also discarded.

These extracted terms are used as inputs of term-filtering step and term-ranking step to obtain essential terms of domain corpus (key-terms).

3.1.2. Term Filtering

The extracted terms are filtered through measures of their features (e.g., the relevance of terms with domain corpus, the frequency of appearance of terms in corpus, and the lexical cohesion of words appearing in a multi-word term). In particular, we use two measures Domain Pertinence (DP) and Domain Consensus (DC) described as [9, 22]. Besides, Lexical Cohesion (LC) measure is used to filter significant multi-word terms. These measures are described in detail in Equations (1)-(5). Since the input datasets have no corresponding glossaries or training data, we use contrast corpora to filter the terms that relevant to

[2] In this paper, two concepts *word* and *term* are used interchangeably.

[3] http://nlp.stanford.edu/software/tagger.html

the domain corpus. The contrast corpus contains many concepts that are different from the domain corpus. For example, a corpus containing text documents in sport (football) domain is the contrast corpus of corpus in tourism domain.

DP is a measure that specifies whether a term is relevant to the domain of a certain corpus by using the contrast corpora. This measure is defined as follows:

$$DP_{D_i}(t) = \frac{freq(t, D_i)}{\max_j \ freq(t, D_j)} \qquad (1)$$

Where, *freq(t, D$_i$)* and *freq(t, D$_j$)* are the number of times that a term *t* appears in the domain corpus D_i and the contrast corpora D_j, respectively. This proportion in Equation (1) shows that a term has a high DP value if it appears more frequently in the domain corpus and less frequently in the contrast corpora. It means that, a term may be relevant to domain corpus if its DP value is high.

LC measure is used to determine how well the combination of individual words in a multi-word term (i.e., compound term). For example, “national park” is a significant multi-word term, meanwhile “first university” is not. We apply the formula that is based on not only the word occurrence but also the word association to compute the cohesion of individual words in a multi-word term as in [12]. This Equation is described as follows:

$$LC_{D_i}(t) = \begin{cases} PMI(t, D_i), & if \ len(t) = 2 \\ \min_i (LC_{D_i}(t_i)), & otherwise \end{cases} \qquad (2)$$

Where, *len(t)* is the number of words in compound term *t* and t_i is a sub-phrase of term *t*.

For the compound term *t* that has two individual words w_1 and w_2 (*len(t)*=2), the *LC* measure of term *t* in corpus D_i, denoted by *PMI(t,D$_i$)*, will be scored as follows:

$$PMI(t, D_i) = (\frac{p(w_1 w_2, D_i)}{p(w_1 w_2, D_i)+1})(\frac{\min(p(w_1, D_i), p(w_2, D_i))}{\min(p(w_1, D_i), p(w_2, D_i))+1}) \qquad (3)$$

Where, *p(w$_1$w$_2$,D$_i$)* is the simultaneous occurrence probability of w_1 and w_2 in corpus D_i, *p(w$_1$,D$_i$)* and *p(w$_2$,D$_i$)* are the probabilities of w_1 and w_2 appearing in corpus D_i.

In case that the amount of individual words in a compound term *t* is greater than two (*len(t)>2*), the *LC* measure of term *t* in corpus D_i is calculated by a recursive formula. Concretely, the minimum of the *LC* measures of sub-phrases t_i in compound term *t* is obtained recursively until the sub-phrases, whose phrase lengths equal to two words, are encountered. The compound terms having *LC* measures greater than zero are selected.

DC is a measure used to identify the importance of a term through the frequency of appearance of it in the corpus. The terms, which have high frequency and appear in many documents of domain corpus, are considered as significant terms. To calculate the frequency of terms in domain documents, we used the DC described as Equation (4). The terms having DC measure less than zero will be discarded.

$$DC_{D_i}(t) = -\sum_{d_k \in D_i} n\_freq(t, d_k) \times \log(n\_freq(t, d_k)) \qquad (4)$$

Where, *n_freq(t, d$_k$)* is the normalized frequency of term *t* in document d_k ($d_k \in D_i$) and *n_freq(t, d$_k$)* is calculated as a proportion of the frequency of term *t* in document d_k and the maximum frequency of term *t* in any documents in the domain corpus and defined as the Equation (5).

$$n\_freq(t, d_k) = \frac{freq(t, d_k)}{\max_{d \in D_i}(freq(t, d))} \qquad (5)$$

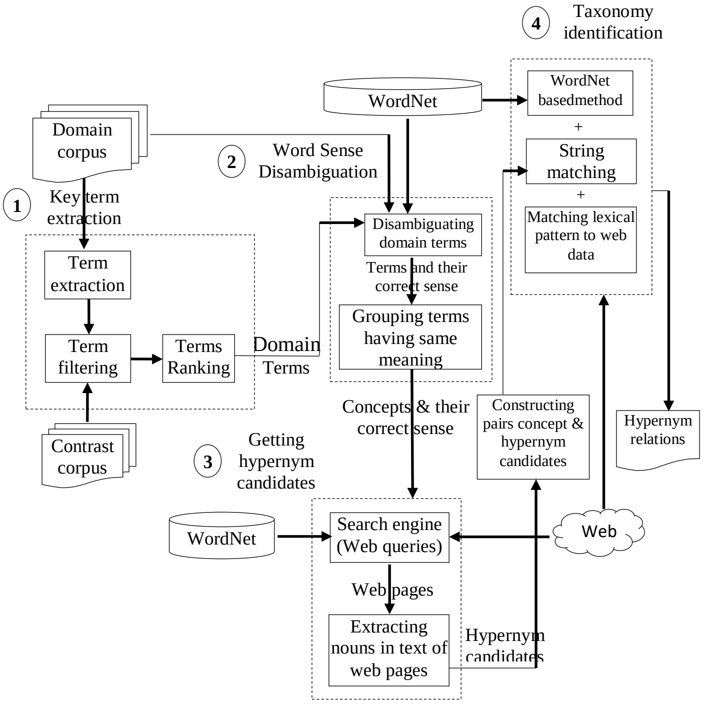
4
Taxonomy identification
WordNet
WordNet basedmethod
+
String matching
+
Matching lexical pattern to web data
Domain corpus
2
Word Sense Disambiguation
1
Key term extraction
Disambiguating domain terms
Terms and their correct sense
Grouping terms having same meaning
Term extraction
Term filtering
Terms Ranking
Domain Terms
Concepts & their correct sense
Constructing pairs concept & hypernym candidates
Hypernym relations
Contrast corpus
3
Getting hypernym candidates
WordNet
Search engine (Web queries)
Web
Web pages
Extracting nouns in text of web pages
Hypernym candidates

Figure 1. The framework of extracting taxonomic relations from a text corpus.

3.1.3. Term Ranking

After performing the above steps, each term *t* will be assigned a weight to determine whether it is a significant and relevant term of the domain corpus D_i or not. The weight of each term is computed based on the combination of aforementioned two measurements DP and DC and described as Equation (6):

$$weight(t, D_i) = \alpha \frac{DP_{D_i}(t)}{\max_{t \in D_i} DP_{D_i}(t)} + \beta \frac{DC_{D_i}(t)}{\max_{t \in D_i} DC_{D_i}(t)} \qquad (6)$$

Where, α and β are the weight values for domain pertinence measure and domain consensus measure, respectively (0<α,β<1 and α+β=1).

The terms with the highest weight are selected as key terms. The correct senses of these key terms would be produced by using WSD algorithm described in 3.2.

3.2. Word Sense Disambiguation

WSD is a technique that automatically determines the correct meaning of the ambiguous words in a particular context. For example, the word "accommodation" has five senses, but the predominant sense of this word in tourism domain is relevant to the. Due to the ability of semantic understanding at the lexical level, WSD has been used in many fields such as NLP, machine translation or information retrieval. In this paper, WSD is mainly used to obtain correct senses of concepts in specific domain corpus.

There are three main approaches for word sense disambiguation: supervised WSD, unsupervised WSD and knowledge-based WSD [24, 32]. Specifically, the supervised methods use machine-learning techniques to produce a classifier from labeled training sets. The obtained results from these systems have high accuracy [31, 36]. However, they need a large amount of training data. Even though there are some available training data (e.g., SemCor) but they are only useful for more general rather than specific domain fields. Besides, in practice, it is almost unfeasible to have hand tagging samples, especially for every new specific domain because the manual building is costly. Whereas, unsupervised methods do not utilize any manual sense-tagged corpus. These methods based on the idea that the sense of a word may be similar to the sense of the neighbouring words. By using clustering techniques, words in domain documents, which have the similar senses, are grouped into a cluster [28] (i.e., words are determined whether they belong to the same sense or not). Consequently, unsupervised WSD methods may not explicitly assign a sense label to a target word. An alternative approach for explicitly identifying correct senses of the words in domain corpus, which do not use any manually sense-tagged corpus, is knowledge-based approach. With this approach, available lexical resources (e.g., WordNet [13] and context (e.g., sentence or document containing ambiguous word) are used as knowledge bases [1, 14, 20].

To obtain correct senses of key terms in a special domain corpus, which have no training data sets, the knowledge-based approach is utilized. A well-known method of knowledge-based approach, called gloss overlap, is proposed by Lesk [20]. This method relies on the calculation of the words overlap between the sense definitions of the given words. Particularly, to identify the correct sense among the possibility senses of a target ambiguous word w, the words in common between the i^th sense's definition (gloss) and the context of the word w is computed as the Equation (7). For example, with a given context: "Thus studies of food, drink, accommodation and other forms of consumption have important implications for understanding the embodied performances of hospitality." and the definition of fifth sense of word "accommodation" in WordNet: "the act of providisng something (lodging or seat or food) to meet a need", we can see that the word in common is "food".

$$Score(s_i \in S)=|context(w)\cap gloss(s_i)| \qquad (7)$$

Where, *context(w)* is a bag of all words in a context window around the word *w*. *S* is the set senses *s* of the word *w*.

The sense whose score is highest will be selected as a correct sense of the word *w*.

However, the glosses in an available online dictionary tend to be short (e.g., in WordNet, a gloss has only from 6 to 8 words) to provide a sufficient vocabulary for computing. To overcome this issue, Banerjee and Pedersen [1] introduced a measure of semantic relatedness that is the extended gloss overlap. It expands the glosses of the words using explicit relations provided in WordNet such as hypernym, hyponym or meronym. Assume the set of relations includes gloss, hypenym, hyponym, the score of the relation between the given word *A* and word *B* is computed as follows:

$$\begin{aligned} relate(A,B)=&Score\ |gloss(A)\cap gloss(B)| \\ &+ Score\ |gloss(A)\cap hype(B)| \\ &+ Score\ |hype(A)\cap gloss(B)| \qquad (8) \\ &+ Score\ |hype(A)\cap hype(B)| \\ &+ Score\ |hypo(A)\cap hypo(B)| \end{aligned}$$

Where, *hype* and *hypo* are contractions of hypenym and hyponym respectively.

In particular, given an ambiguous word *w*, the correct sense of *w* is the highest value of similarity measures between the context of word *w* and its senses, described as Equation (9).

$$Sense(w)=\max_{s_{w,k}\in S_w}(Sim(s_{w,k},C)) \qquad (9)$$

Where, $s_{w,k}$ is the k^{th} sense of word *w*; S_w is a set of possible senses of word *w; C* is the context of word *w*; and $Sim(s_{w,k},C)$ is a similarity measure of the context with the k^{th} sense of *w*, calculated as follows:

$$Sim(s_{w,k},C)=\sum_{i=1}^{|C|}\sum_{j}^{m} relate(s_{w,k},s_{c_i,j}) \qquad (10)$$

Where, $s_{c_i,j}$ is the j^th sense of word $c_i \in C$; *m* is the number of senses of each word c_i; and $|C|$ is the size of the context window (i.e., the number of words surrounding *w*).

According to Banerjee and Pedersen [1], the process of finding the correct sense of an ambiguous word *w* was conducted on each document of corpus containing word *w* and consists of the following steps: First, they identified the context of word *w* (e.g., K. Meijer *et al*. [22] used a document as the context. Meanwhile, according to Jiang and Tan [17], a sentence is used as a context). After that, words (such as verbs, nouns, adjectives) in context window, surrounding the word *w*, were used to calculate the similarity between the senses and the context of the word *w* as in Equation (10). Finally, the sense with highest similarity measure, computed as in Equation (9), will be a correct sense of a document. This process is iterated in all contexts of domain corpus that contain word *w*. The sense with the most frequently occurring was selected as the correct sense of the word *w* in the corpus.

Nevertheless, in the case of large corpus, this approach is disadvantageous because the algorithm has high complexity. In fact, given an ambiguous word *w*, assume that *K* is the number of the contexts in the domain corpus containing word *w*, *N* is the size of a context window and *M* is the number of the senses on average per word *w*, the time complexity will be $K*M^N$.

We applied the method of Banerjee and Pedersen [1] to disambiguate word sense. However, to reduce the time complexity, we modified Banerjee's method by using the different way in getting the context of an ambiguous word in domain corpus. In fact, some previous studies show that the information about the domain of the corpus, called the context, has a strong influence on the distribution of sense of the words appearing in this corpus (i.e., the sense of a word depends on the context containing it) [19]. Hence, with the observation nouns appearing around an ambiguous word in the context play important role in identifying the relatedness between the word and its sense (e.g., a given ambiguous word "accommodation", the nouns such as "lodging", "food" have the highest frequencies in contexts containing the word "accommodation"). Thus, instead of considering words such as verbs, nouns, adjectives appearing around in a context window of each document as the above mentioned methods, we only focused on the nouns frequently occurring in contexts of all documents of domain corpus. The process of obtaining nouns in context consists of following steps. First, the sentences, containing the target ambiguous word *w*, are extracted from documents in the domain corpus. After that, we take a list of nouns from these sentences. Nouns, which are stopwords, are removed. Finally, these nouns are ranked according to their frequency of occurrence. The list of nouns with higher frequencies will be selected to represent the context of the ambiguous word *w*.

In the proposed method, we only consider *N* nouns with highest frequencies as a context. As a result, the time complexity in this case is $K+M^N$.

3.3. Taxonomic Relation Extraction

Based on the correct senses acquired from the WSD algorithm, grounded concepts are automatically derived from the corresponding key terms. Particularly, key terms that share the same meaning in WordNet are grouped and labeled with the name of the term, which has the highest frequent occurrence. These concepts are input data of taxonomic-relation extraction.

The process of extracting the taxonomic relations between the concepts is conducted by using web based approach that includes two steps:

1. Extracting hypernym candidates.
2. Identifying the right concept hierarchies.

However, to avoid the knowledge acquisition bottleneck as [34] and the noise data from web pages [33], we add the correct information sense of the concepts into web queries that aim at acquiring the accurate hypernym candidates.

3.3.1. Hypernym-Candidate Extraction

The process of obtaining the hypernym candidates of a concept consists of following steps:

First, the web queries are constructed based on the combination of lexical patterns and semantic information or contextual information. Similar to the previous studies [16, 33, 35], the set of lexical patterns that is used for identifying the hypernym relations of the concept *A* and concept *B* (assume that *A* is a hyponym of *B* or *B* is a hypernym of *A*) are shown in Table 1. The semantic information is nouns extracted in the gloss of the correct synset instead of the first synset in WordNet. Concretely, for the concepts contained in WordNet, we conduct to obtain the correct sense through the WSD algorithm in 3.2. By using NLP techniques such as tagging, lemmatization and stop-word removing, nouns from gloss definition corresponding to the correct sense are identified and added to queries. Besides, the context information, which includes nouns not only have high frequency of appearance, but also has high domain measurement, is also added to queries. These context information is useful for acquiring data, which is relevant to domain corpus, especially for concepts disappear in WordNet.

Next, by using search engines, web pages corresponding to the patterns, are obtained from queries.

Finally, hypernym-candidate extraction from these web pages is conducted. In more detail, we extract text data from web pages and filter sentences in this text data that contain the concept and corresponding patterns. Applying NLP techniques as above, the nouns and noun phrases in these sentences are acquired. The nouns or noun phrases appearing after patterns such as "is a", "and other", "or other", are considered as hypernym candidates. Meanwhile, for the remainder patterns in Table 1, the hypernym candidates are nouns or noun phrases that occur before the ppatterns. We found that, nouns or noun phrases having high appearance frequency, are significant hypernym candidates. Therefore, nouns or noun phrases with high appearance frequency are selected as hypernym candidates.

The taxonomic relations between these hypernym candidates and their corresponding concept are determined through taxonomic relation identification, presented in 3.3.2.

Table 1. Lexical patterns.

Lexical patterns for identifying hypernym relations between two concepts *A* and *B* (assume *B* is a hypernym of *A*)
A, *is a* B
A, *and other* B
A, *or other* B
B, *such as* A
B, *including* A
B, *especially* A
B, *called* A
B, *particularly* A
B, *for example* A
B, *among which* A

3.3.2. Taxonomic Relation Identification

In this section, we present the way to determine the taxonomic relations between two given terms. One of the methods to do it is to use a set of the hypernyms of a term in a lexical database WordNet, for example, in WordNet, the term "traveler" is a hypenym of the term "tourist". In particular, it is considered two given terms t_1, t_2 and their correct senses, function *is Hype WN(t_1,t_2)* is defined to confirm whether there is a taxonomic relation between two term t_1 and t_2 or not. The function *is Hype WN(t_1,t_2)* return *true* if t_2 is a hypernym of t_1 in WordNet and return *false* in otherwise.

Nevertheless, the coverage of WordNet is limited in many regions because it was created manually and difficult to adapt to the rapidly changing of domains. Thus, two other measures such as the string matching [17] and lexical-syntactic pattern matching [33, 34], are also applied.

String matching method is a simple way to determine the taxonomic relation between two terms based on the following observations:

- *Observation* 1: Given two terms t_1, t_2, denoted *wx* and *x* respectively (i.e., term t_2 is the head of term t_1), term t_2 may be a hypernym of term t_1 [17]. For example, term "holiday" is a hypernym of "public holiday" due to the former term is the head (or substring) of the latter term.
- *Observation* 2: Given two terms are compound nouns t_1 and t_2, in which their first words (from the right to left) are similar or semantic equal, a taxonomic relation can be induced from these terms if at least a word (or a compound word) in this term is a direct hypernym/hyponym or the inheritable hypernym/ hyponym of an another term [25]. For instance, given two terms are "car service" and "transport service", because of the word "car" is an inheritable hyponym "transport", hence the term "transport service" is a hypernym of a term "car service".

Hence, using string matching method, the function *SM(t_1,t_2)* is defined to identify the taxonomy relation of two term t_1 and t_2. The function *SM(t_1,t_2)* returns *true* if t_1 and t_2 satisfy observation 1 or observation 2 and returns *false* in otherwise.

However, not all taxonomy relations will be induced from the pairs of terms guaranteed two above measures. Therefore, a method that matches lexical-syntactic patterns to the Web data is applied to determine the taxonomic relation between two arbitrary terms. As presented in [33, 35], with given two terms t_1 and t_2 (assume that t_2 is a hypernym of term t_1), the process of the determining the taxonomic relationship between these terms consists following steps: First, the corresponding web query of each pattern *p*, used to produce the hypernym t_2, is created (e.g., "*t_1 is a t_2*" or "*t_2 such as t_1*"). Next, by using a Web search engine, the obtained result of the query, corresponding to the pattern *p*, is the number of hits (i.e., the number of web pages) containing pattern *p* and is denoted *Hit_pat (t_1, t_2, p)*. For example, with pattern "*is a*", *Hit_pat (t_1,t_2, "is a")* is the number of web pages that contain a string "*t_1 is a t_2*". The taxonomic relation between t_1 and t_2 was evaluated using web-based method [8, 33] that is described in the Equation (11).

$$ScoreWeb(t_1,t_2)=\frac{\sum_{p\in P} Hit\_pat(t_1,t_2,p)}{Hits(t_2)} \qquad (11)$$

Where, *P* is a set of patterns *p* and *Hits(t_2)* is the number of web pages that contain the term t_2.

In particular, the process of identifying the correct hypernyms from the list of hypernym candidates of a concept is described as Algorithm 1.

Algorithm 1. Identifying the list of the correct hypernyms of a concept

```
Input: Concept c, listHC (listHC is a set of hypernym candidates of concept c)
Output: The correct hypernyms of concept c.

Step 1. listCorrectHC=∅, listTempHC= ∅;
(-listTempHC is a list that contains hypernym candidates hc and measurements score, where, the score is the web-based measure between the concept c and the hypernym candidate hc that is computed by formula (11).
- listCorrectHC is a list of correct hypernyms of concept c. )
Step 2.

for (hc∈ listHC)
    if (isHypeWN(c,hc)=true or SM(c,hc)=true)
        // add hc to the top of the list listCorectHC
        add(listCorectHC, hc);
    else
        score=ScoreWeb(c,hc);
        // add hc and score into listTempHC
        add(listTempHC, hc, score);
Step 3. Rank listTempHC in descending order of score;
Step 4. Add listTempHC to the end of list listCorectHC;
Step 5. Return top elements in listCorectHC;
```

4. Experiments and Results

4.1. Experiments

Now, we turn to present the implementation of the proposed framework through an application with the input corpora is text documents in tourism and sport domains. The application is written in Java. It uses Stanford POS-tagger as a tool to extract phase. It also uses WordNet, an available lexical database provided by [13, 23], to disambiguate word senses. The experiments have also been performed through two datasets[4] Lonely Planet and SmartWeb consisting 1,801 files about tourism domain and 3,542 files about sport domain, respectively. The following are the results in the four-step process of the experiments:

First, in the key term extraction step, single terms and multi-word terms were extracted from the Lonely Planet and SmartWeb datasets. Specifically, in the Lonely Planet dataset, there are 38,446 extracted terms consisting of 8,278 single terms and 30,168 multi-word terms. By using lexical cohesion measurement, 1,783 significant multi-word terms were obtained from 30,168 multi-word terms. For the SmartWeb dataset, there are 37,634 terms, including 5,942 single terms and 31,692 multi-word terms. The significant multi-terms are 5,830 among 37,634 multi-word terms. Next, all of the terms are domain filtered using the contrast domain datasets. Concretely, the SmartWeb dataset is used as the contrast domain dataset for the Lonely Planet dataset. On the contrary, the Lonely Planet dataset is considered the contrast domain dataset of SmartWeb dataset. Finally, each filtered term is assigned a weight, which combines two measurements domain pertinence DP and domain consensus DC as in Equations (1) and (4), to determine whether this term will be selected as key term or not. The obtained results depend on the values of two parameters α and β, for instances, with α from 0.3 to 0.4 and β from 0.7 down to 0.6 respectively, the number of terms that relevant to the domain corpus is largest. As a result, 200 terms selected out of 3,583 terms of Lonely Planet dataset and 8,918 terms of SmartWeb dataset respectively are considered as key terms.

Second, in the word sense disambiguation step, correct senses of terms extracted in the first step were obtained. In more detail, the correct senses of terms are acquired from a semantic lexical dictionary WordNet. By using manual evaluation from experts, the accuracy (precision) of the obtained outcomes for ambiguous concepts in the two domains of tourism and sport is 74.1% and 65.8%, respectively.

Based on the correct senses, grounded concepts are automatically derived from the key terms in the corresponding data sets. In more detail, key terms that their labels share the same meaning in WordNet are grouped and the term, whose label frequently appear in the corpus, are selected to represent a corresponding concept. For example, in sport domain, the correct sense of term “coach” is the first sense, and synonym of this term is “manager”. Hence, these terms are grouped into a concept represented by a term with highest frequency appearance, such as “coach”. In addition, the automatic derived concepts would be selected again by human experts. The inducing taxonomic relations between these selected concepts are conducted on the third-step and the fourth-step.

In the third step, the hypernym candidates of concepts are acquired from web sources. Concretely, we build web queries by combining the lexical patterns as in Table 1 with the semantic information and the context information of the concept. The hypernym candidates of a concept are nouns or noun phrases with highest occurrence in the contexts of the obtained web pages.

To illustrate this performance, it is considered the concept “accommodation” of the Lonely Planet dataset. Through WSD module, the correct sense of this concept is the fifth sense in WordNet whose glossary definition is “the act of providing something (lodging or seat or food) to meet a need”. The semantic information is the nouns “act”, “something”, “lodging”, “seat”, “food”. The context information is nouns, having the highest domain measure, such as “beach”, “attraction”, “festival”, “coast”. Both of the above information is used for constructing its queries. For instance, a query of concept “accommodation” corresponding one of patterns in Table 1 is that “accommodation is a”+act+ something+lodging+ seat+ food. The hypernym candidates of the concept “accommodation” are nouns “service”, “facility” and “travel”.

Fourth, in the taxonomy-identification step, the taxonomic relations of the pairs of the concepts and their hypernym candidates are determined.

4.2. Results and Evaluation Methods

4.2.1. Evaluation Methods

There are many methods for evaluating the automatic construction of ontologies, in which, the gold standard and human evaluation approaches are usually applied. For the gold standard evaluation method, a constructed ontology is compared with a predefined ontology, built manually from scratch by domain experts (i.e., gold-standard/reference ontology) . The similarity between the ontologies is determined by comparing this one with the other one at two different levels: lexical (vocabulary) and conceptual (hierarchy). In fact, the comparison on a lexical level of two ontologies identifies the similarity between the lexicons (i.e., a set of terms/labels represented concepts). Meanwhile, taxonomic and other relations in ontologies are compared at the conceptual level [2,

[4] These datasets and the corresponding gold-standard ontologies are provided by authors in the paper [33].

10, 11, 22]. Nevertheless, in this research, we take into account the set of direct and non-direct is-a relations (taxonomic relations) that are produced by proposed method. Besides, concept hierarchies, used to compared, are not too deep. Therefore, we compare the taxonomic relations found by proposed method against the taxonomic relations of a gold-standard ontology at lexical layer through measures Precision, Recall, F-measure as presented in [2, 7, 17, 33]. In particular, the precision measure (*Prec*) is the proportion of the number of entities (concepts /taxonomic relations) extracted from the corpus that also appear in the golden standard ontology (CorrectExtrEnti) with the number of entities that extracted from the corpus (TotalExtrEnti). Recall measure (Rec) is the proportion of the extracted entities from the corpus that also appear in the golden standard ontology with the number of entities in the gold- standard ontology (GoldEnti). Both of these measures are described in Equations (12) and (13).

Additionally, most of the studies, which use the gold standard evaluation approach, usually assume that a gold-standard ontology describes the domain knowledge accurately because the quality of judging is influenced by the rightness of the gold-standard ontology [33]. However, in practice, the gold standard is built manually biased toward to experts [3], especially for taxonomic relations [18]. For instance, with above experiments, there are some extracted terms from the Lonely Planet dataset that relevant to tourism domain, but disappear in the gold-standard ontology such as “resort”, “destination”, “outdoor activity”. Hence, some taxonomic relations between these concepts will disappear in a gold standard.

In this paper, we use two types of the precision [18], such as prior precision and posterior precision to apply the evaluation on lexical level for the obtained results. The prior precision is similar to the precision (Prec) as in Equation (12). Meanwhile, posterior precision (Precpost), used for posterior evaluation by experts, is the proportion of the extracted entities from the corpus that also appear in the golden standard ontology (CorrectExtrEnti) or are correctly judged by human experts (CorrectEvalEnti) with the number of entities that extracted from the corpus (TotalExtrEnti) and showed in Equation (14). Besides, the comparison is performed by not only using string matching method, but also using a lexical resource WordNet. For example, the term “activity” disappears in the gold-standard ontology. However, with the second sense in WordNet, this term is a synonym of term “action”. Thus, “activity” is considered as a correct concept.

$$Prec = \frac{CorrectExtrEnti}{TotalExtrEnti} \tag{12}$$

$$Rec = \frac{CorrectExtrEnti}{GoldEnti} \tag{13}$$

$$Prec_{post} = \frac{CorrectExtrEnti + CorrectEvalEnti}{TotalExtrEnti} \tag{14}$$

4.2.2. Results

In this section, we present the outcomes of the proposed framework on two data sets (e.g., Lonely Planet and SmartWeb) corresponding to tourism and sport (football) domains. The obtained results are shown in Table 2. Besides, to determine the quality of the proposed method, we compare it with a state of the art method using web-based approach [33]. The outcomes of these two methods, which are also conducted on Lonely Planet dataset and SmartWeb dataset, are shown in Table 3. All the gold-standard ontologies of these datasets are taken from [33]. In more detail, there are 96 concepts and 103 hierarchical relations in the reference ontology of the Lonely Planet dataset. While, the reference ontology of SmartWeb dataset includes 359 concepts and 633 hierarchical relations. However, the reference ontologies are manually built and are independent of the domain corpus. Thus, not all concepts and relations defined in the reference ontology can be found in the domain corpus. Similarly, some significant concepts and relations appear in the domain corpus, but disappear in the reference ontology. For example, there are 78 terms out of 96 terms of reference ontology that appear in Lonely dataset and 206 terms out of 359 terms of reference ontology that appear in SmartWeb dataset. Since the process of extracting taxonomic relations consists of two main steps, so the evaluation is separated as follows:

- *Key-term extraction*: In both of two data sets, top 200 terms extracted by our framework have the highest precision and recall. Particularly, 44 terms, which were extracted from the Lonely Planet dataset, are found in the reference ontology. Also, for the SmartWeb dataset, 62 extracted terms are found in the reference ontology. Besides, hypernyms of key terms, which are obtained from taxonomy-extraction step and found in the reference ontologies, would be considered key terms. For example, in the SmartWeb dataset, the term “person” is found in the reference ontology. However, this term disappears in the set of extracted terms, but it is derived from taxonomy-extraction step. Thus, after the step of taxonomic-relation extraction, the accuracy of term extraction is scored again through prior precision and recall measures. Furthermore, according to human experts, there are many extracted terms that disappear in the reference ontology, but are also considered as key terms. Hence, the posterior precision is used to assess all of the extracted key terms. The obtained results are shown in Table 2.
- *Taxonomic relation extraction*: In this step, the

matching methods are used to evaluate the accuracy of the obtained results. In particular, the extracted taxonomic relations which can be directly matched with the taxonomic relations of reference ontology are assessed as the correct taxonomic relations. For example, in the SmartWeb dataset, term "official" is a subclass of "person" in the reference ontology, hence, there is a valid taxonomic relation the relationship between term "official" and its hypernym "person", denoted *rel(official, person)*. Another matching method is that the taxonomic relations, which can be derived from the reference ontology, are also evaluated as the correct taxonomic relations. For example, in the Lonely Planet dataset, "coast" is a subclass of "area", which can be inferred from the relations "coast" is a subclass of "nature" and "nature" is a subclass of "area". Thus, there is a valid taxonomic relation *rel(coast, area)*. By using these matching methods, the valid taxonomic relations of the datasets are produced. However, there are many correct taxonomic relations that are evaluated by human experts, but are not found in the reference ontologies. For example, the valid taxonomic relations in the Lonely Planet dataset, such as *rel(beach, destination)* or *rel(outdoor activity, activity*) are not found in the reference ontology. Hence, we also use priori precision, posterior precision and recall measures to evaluate the taxonomic relations. The obtained results of the extracting taxonomic relations from proposed framework are shown in Table 2.

Table 2. Results of proposed method with respect to the datasets.

Dataset	Task	Priori Precision	Posterior Precision	Recall
Lonely Planet	Concept extraction	22.5%	51%	56%
	Taxonomy extraction	27.8%	65%	35%
SmartWeb	Concept extraction	31%	47%	32%
	Taxonomy extraction	52%	70%	30%

As presented in Table 2, the recall and the priori precision measures in the taxonomy-extraction task of both data sets are rather low. The main reason is that reference ontologies are usually constructed manually. In fact, it is difficult to extract automatically some taxonomic relations that similar with those of the reference ontology. For example, in reference ontology of SmartWeb dataset, the hypernym (direct-super class) of concepts "free kick", "injury", "penalty" is "In Game Event", and concept "event" is the hypernym of concept "In Game Event". However, the extracted hypernyms of "free kick", "injury", "penalty" are "kick", "accident" and "punishment", respectively. In addition, in WordNet, concept "event" is an indirect super class of these hypernyms. Thus, these taxonomy relations are considered the correct taxonomy relations by human experts. This is the explanation why the posterior precision is higher than priori precision.

The obtained results by using the method of Rios-Alvarado *et al.* [33] and proposed method in two data sets, are shown in Table 3.

For Lonely Planet dataset, the precision of concept extraction of the Rios-Alvarado's approach (67%) is better than the precision of our method (51%). In contrast, for the SmartWeb dataset, the precision of our method (47%) are better than Rios-Alvarado's approach (44%). It may be attributed to the following reasons: First, as mentioned above, there are many concepts that are significant in domain, but disappear in the reference. Second, many terms are important concepts of a domain, but they occur with rather low frequency in the domain-text corpus. Hence, weights that determine the relevance of these terms with domain are also low. As a result, these terms are omitted.

In the step of taxonomic-relation extraction, we can see that the precision of the proposed method (65%) is better than Rios-Alvarado's method (53%) in the Lonely Planet dataset. Meanwhile, for SmartWeb dataset, the precision of Rios-Alvarado's method (77%) is better than the proposed method (70%). One of the main reasons is that the accuracy of results produced by the WSD algorithm in Lonely Planet dataset is higher than in SmartWeb dataset. For Lonely Planet dataset, hence, noisy information in extracting hyepernym candidates is reduced. Besides, in SmartWeb dataset, there are several terms (e.g., "header", "cross") that have no forms as patterns in Table 1 for obtaining additional information from the web pages. Consequently, it is difficult to find hypernym candidates of these terms.

Table 3. Performance comparison of the methods.

Dataset	Task	Proposed method Precision	Rios-Alvarado's precision
Lonely Planet	Concept extraction	51%	67%
	Taxonomy sextraction	65%	53%
SmartWeb	Concept extraction	47%	44%
	Taxonomy extraction	70%	77%

5. Conclusions

In this paper, we proposed a framework to extract taxonomic relations between concepts from the text corpora by leveraging the large amount of information from web sources. Unlike previous studies, the WSD salgorithm in this paper was applied to increase the accuracy in the processing of extracting the hypernym candidates from web pages that usually have noisy information. In addition, relying on the correct senses acquired from the WSD algorithm, the terms having the same meaning, called equivalent terms, are also obtained through a lexical resource. These equivalent terms are useful for identifying the grounded concepts in concept-extraction stage.

Besides, we offered a variant of WSD algorithm in the previous study [1] for disambiguating polysemous words. The quality of the obtained results is the similar to the previous studies [24], but the time complexity is reduced.

The obtained results from the experiments show this method is especially useful with corpora which are insufficient as well as have no training data. In case the datasets that have high WSD results, our method outperforms the state of the art method.

However, in case the domain corpora having many specific terms that disappear in the lexical resources, the quality of the outcome is not good because the senses of ambiguous concepts were taken from a certain lexical resource. In addition, there is a little information in the specific domains shared on the web pages. Hence, we found that the proposed method is efficient for datasets with popular domains (e.g., tourism, sport).

Acknowledgments

This work was supported by the scholarship from Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand.

References

[1] Banerjee S. and Pedersen T., "Extended Gloss Overlaps As A Measure of Semantic Relatedness," *in Proceedings of the 18th international joint Conference on Artificial Intelligence*, Acapulco, pp. 805-810, 2003.

[2] Bordea G., Lefever E., and Buitelaar P., "SemEval-2016 Task 13: Taxonomy Extraction Evaluation (TExEval-2)," *in Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, pp. 1081-1091, 2016.

[3] Brank J., Grobelnik M., and Mladenić D., "A Survey of Ontology Evaluation Techniques," *in Proceedings of the Conference on Data Mining and Data Warehouses*, Ljubljana, pp. 166-169, 2005.

[4] Caraballo S., "Automatic Construction of A Hypernym-Labeled Noun Hierarchy from Text," *in Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, Maryland, pp. 120-126, 1999.

[5] Cederberg S. and Widdows D., "Using LSA and Noun Coordination Information to Improve The Precision and Recall of Automatic Hyponymy Extraction," *in Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL*, Edmonton, pp. 111-118, 2003.

[6] Cimiano P., Hotho A., and Staab S., "Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis, " *Journal of Artificial Intelligence Research*, vol. 24, pp. 305-339, 2005.

[7] Cimiano P., Pivk A., Schmidt-Thieme L., and Staab S., *Ontology Learning from Text: Methods, Evaluation and Applications*, IOS Press, 2005.

[8] Cimiano P. and Staab S., "Learning by Googling," *SIGKDD Explorations*, vol. 6, no. 2, pp. 24-33, 2004.

[9] De Knijff J., Frasincar F., and Hogenboom F., "Domain Taxonomy Learning From Text: The Subsumption Method Versus Hierarchical Clustering," *Data and Knowledge Engineering*, vol. 83, pp. 54-69, 2013.

[10] Dellschaft K. and Staab S., "On how to Perform A Gold Standard Based Evaluation of Ontology Learning," *in Proceedings of International Semantic Web Conference*, Athens, pp. 228-241, 2006.

[11] Dietz E., Vandic D., and Frasincar F., "TaxoLearn: A Semantic Approach to Domain Taxonomy Learning," *in Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, Macau, pp. 58-65, 2012.

[12] Doan P., Arch-int N., and Arch-int S., "Improving key Concept Extraction Using Word Association Measurement," *in Proceedings of the 7th International Conference on Information Technology and Electrical Engineering*, Chiang, pp. 403-407, 2015.

[13] Fellbaum C., *WordNet: An Electronic Lexical Database*, Camb. MA MIT Press, 1998.

[14] Hadni M., Alaoui S., and Lachkar A., "Word Sense Disambiguation for Arabic Text Categorization," *The International Arab Journal of Information Technology*, vol.13, no. 1, pp. 215-222, 2016.

[15] Hazman M., El-Beltagy S., and Rafea A., "A Survey of Ontology Learning Approaches," *International Journal of Computer Applications*, vol. 22, no. 9, pp. 36-43, 2011.

[16] Hearst M., "Automatic Acquisition of Hyponyms from Large Text Corpora," *in Proceedings of the 15th International Conference on Computational Linguistics*, Nantes, pp. 539-545, 1992.

[17] Jiang X. and Tan A., "CRCTOL: A Semantic-Based Domain Ontology Learning System," *Journal of the American Society for Information Science and Technology,* vol. 61, no. 1, pp. 150-168, 2010.

[18] Kavalec M. and Svaték V., "A Study on Automated Relation Labelling in Ontology Learning," *Ontology Learning from Text: Methods, Evaluation and Applications*, no. 123, pp. 44-58, 2005.

[19] Koeling R., McCarthy D., and Carroll J., "Domain-Specific Sense Distributions and Predominant Sense Acquisition," *in Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, Vancouver, pp. 419-426, 2005.

[20] Lesk M., "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell A Pine Cone from an Ice Cream Cone," *in Proceedings of the 5th Annual International Conference on Systems Documentation*, Toronto, Canada, pp. 24-26, 1986.

[21] Medelyan O., Witten H., Divoli A., and Broekstra J., "Automatic Construction of Lexicons, Taxonomies, Ontologies, and other Knowledge Structures," *Data Mining and Knowledge Discovery*, vol. 3, no. 4, pp. 257-279, 2013.

[22] Meijer K., Frasincar F., and Hogenboom F., "A Semantic Approach for Extracting Domain Taxonomies from Text," *Decision Support Systems*, vol. 62, pp. 78-93, 2014.

[23] Miller G., "WordNet: A Lexical Database for English, " *Communications of the ACM*, vol. 38, no. 11, pp. 39-41, 1995.

[24] Navigli R., "Word Sense Disambiguation: A Survey," *ACM Computing Surveys*, vol. 41, no. 2, pp. 1-69, 2009.

[25] Navigli R. and Velardi P., "Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites," *Computational Linguistics*, vol. 30, no. 2, pp. 151-179, 2004.

[26] Noy N. and McGuinness D., "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[27] Ortega-Mendoza R., Villaseñor-Pineda L., and Montes-y-Gómez M., "Using Lexical Patterns for Extracting Hyponyms from The Web," *in Proceedings of the Mexican International Conference on Artificial Intelligence*, Aguascalientes, pp. 904-911, 2007.

[28] Pantel P. and Lin D., "Discovering Word Senses from Text," *in Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, pp. 613-619, 2002.

[29] Paukkeri M., García-Plaza P., Fresno V., Unanue R., and Honkela T., "Learning a Taxonomy from A Set of Text Documents," *Applied Soft Computing*, vol. 12, no. 3, pp. 1138-1148, 2012.

[30] Petasis G., Karkaletsis V., Paliouras G., Krithara A., and Zavitsanos E., *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution*, Springer Berlin Heidelberg, 2011.

[31] Pradhan S., Loper E., Dligach D., and Palmer M., "SemEval-2007 task 17: English Lexical Sample, SRL and All Words," *in Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, pp. 87-92, 2007.

[32] Ranjan Pal A. and Saha D., "Word Sense Disambiguation: A Survey," *International Journal of Control Theory and Computer Modeling*, vol. 5, no. 3, pp. 1-16, 2015.

[33] Rios-Alvarado A., Lopez-Arevalo I., and Sosa-Sosa V., "Learning Concept Hierarchies from Textual Resources for Ontologies Construction," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5907-5915, 2013.

[34] Sang E., "Extracting Hypernym Pairs from The Web," *in Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Prague, pp. 165-168, 2007.

[35] Snow R., Jurafsky D., and Ng A., "Learning-Syntactic-Patterns-For-Automatic-Hypernym-Discovery," *in Proceedings of the 17th International Conference on Neural Information Processing Systems*, Vancouver, pp. 1297-1304, 2004.

[36] Snyder B. and Palmer M., "The English all-Words Task, " *in Proceedings of SENSEVAL-3, the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, pp. 41-43, 2004.

[37] Tu D., Chen L., and Chen G., "Automatic Multi-Way Domain Concept Hierarchy Construction from Customer Reviews, " *Neurocomputing*, vol. 147, no. 1, pp. 472-484, 2015.

[38] Wong W., Liu W., and Bennamoun M., "Ontology Learning from Text: A Look Back and Into The Future," *ACM Computing Surveys*, vol. 44, no. 4, pp. 1-36, 2012.

[39] Yamane J., Takatani T., Yamada H., Miwa M., Sasaki Y., "Distributional Hypernym Generation by Jointly Learning Clusters and Projections," *in Proceedings of the 26th International Conference on Computational Linguistics*, Osaka, pp. 1871-1879, 2016.

[40] Zafar B., Qamar U., and Imran A., "A Domain-Independent Hybrid Approach for Automatic Taxonomy Induction," *in Proceedings of the 17th International Conference on Parallel and Distributed Computing, Applications and Technologies*, Guangzhou, pp. 372-375, 2016.

Phuoc Thi Hong Doan received the M.S degree from Hue University of Sciences, Vietnam, in 2004. She is currently a Ph.D student in Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand. Her research interests are data mining, natural language processing and information retrieval.

Ngamnij Arch-int received the PhD degree in computer science from Chulalongkorn University, Thailand in 2003. She is currently an associate professor in the Department of Computer Science at Khon Kaen University, Thailand. Her research interests include the semantic web, web services, semantic web services, and heterogeneous information integration.

Somjit Arch-int received the PhD degree in computer science from the Asian Institute of Technology, Thailand in 2002. He is currently an associate professor in the Department of Computer Science at Khon Kaen University, Thailand. His previous experiences include the development of several industry systems and consulting activities. His research interests are business component-based software development, objectoriented metrics, ontology-based e-business modeling, knowledge-based representation, semantic information integration, data mining, and semantic Web. He is a member of the IEEE Computer Society.