Capturing Knowledge for Interpretation based on Surrounding Words of Consecutive Serial Verb Constructions

Pakawan Pugsee*, Wanchai Rivepiboon and Pattarasinee Bhattarakosol
Department of Computer Engineering, Faculty of Engineering,
Department of Mathematics and Computer Science, Faculty of Science
Chulalongkorn University, Bangkok 10330, Thailand.
*Author for correspondence; e-mail: pakawan.pugsee@gmail.com

ABSTRACT
Documents are written using a natural language, so there are various interpretations because of different understanding of words. Therefore, implementing a mechanism to interpret statements to have only one clear point is very significant to readers. However, there are a variety of formats of a sentence, and some sentences consist of serial verbs that affect to the meaning. Thus, knowledge discovering about serial verb constructions (SVCs) is significantly vital before understanding the entire sentence. This paper proposes mechanisms to unclose knowledge concerning to the consecutive serial verbs by focusing on determining relations between serial verbs among surrounding words. These relations and syntactic information are represented into the semantic binding patterns. Then, patterns of SVCs are generated from semantic binding patterns with semantic classes of surrounding words. Furthermore, the semantic interpretations of SVCs can be initially classified by partial semantic information of surrounding words.

Keywords: serial verb constructions, semantic analysis, natural language understanding.

1. INTRODUCTION
As the fact that documents usually are written in the natural language, understanding these documents means that readers must understand the syntax and semantic of the language. Therefore, developing software to interpret or capture context from a document refers to the development of syntactic and semantic interpreters. However, these developments are difficult since meaning of words changes according to their positions in the sentences and its surrounding or compounding words. Additionally, same meaning can be written in different styles. These varieties are called the semantic heterogeneity and the syntactic heterogeneity respectively.

Consider formats of sentences. It must consist of a subject, verbs, and an object, where some sentences might have other verbs extension called serial verbs. The serial
verb construction, also known as verb serialization, is a syntactic phenomenon common to many languages, such as African, Asian, and Guinean languages. Contrary to subordination, where one clause is embedded into another, verb serialization strings two verbs together in a sequence in which no verb is subordinated to the other [1]. Therefore, sentences with serial verbs are difficult to analyze because the meaning of a serial verb is composition of the meanings of its components. For examples, a Chinese serial verb construction “qu chifan” is composed of two verbs “qu” and “chifan”; a Thai serial verb “chai ok-baep” is composed of two verbs “chai” and “ok-baep”. The meaning of “qu chifan” is “go to eat” and “chai ok-baep” is “use to design”. The meanings of these two consecutive serial verbs show that the meaning of serial verbs can be discovered by the meaning of both verbs and the second verb also shows the purpose of the first verb. Thus, the serial verb analysis is more complicated than the single verb analysis.

According to the influence of serial verbs in a sentence, the serial verbs analysis is needed so that the semantic understanding can be completed. Moreover, two consecutive SVCs which is the form of basic serial verbs can be widely used like the common verb of sentences. Therefore, this paper proposed statement modeling, called the framework of discovering knowledge about two consecutive SVCs, to interpret sentences with serial verbs. The objective of the proposed framework is to automatically extract useful information of two consecutive serial verbs from different raw sentences on many documents. The outcome of these processes is a useful meaning representation for the serial verb construction information.

The details of this research will be described in the following sections. The knowledge about serial verb constructions and the sentence analysis is elaborated in the next section. The overview and details of this proposed method are described in knowledge discovering methods section. Additionally, the results of experiments based on the proposed framework are demonstrated in experiments and results section. The last two final sections are discussion and conclusion respectively.

2. BACKGROUND KNOWLEDGE

Definitions and descriptions of SVCs in [2][3][4][5] show that there are different types of SVCs. Nevertheless, they can be classified into two basic groups that are the basic SVCs called V-V sequence which consists of only two verbs, and the complex SVCs which consists of more than two verbs. To automatically dissolve knowledge about SVCs, the sentence analysis process is needed.

There are two main processes of the sentence analysis: the syntactic analysis, and the semantic analysis. However, according to SVCs found in Asian languages, there is no exact word boundary in the sentences. Therefore, the first main process for the sentence analysis is the word segmentation. The examples of those Asian languages are Chinese, Japanese, Vietnamese, Thai, and Lao. There are several word segmentation techniques, such as lexical techniques, statistical techniques, and machine learning techniques. The details of word segmentation are not mentioned in this article because there are some word segmentation tools [6][7][8] for separating words in sentences with each language, including tagging their parts of speech.

Most of syntactic analysis researches use grammars to analyze a sequence of words in
sentences and to form the structure tree representing sentence structures. Some Asian languages, e.g. Chinese and Thai, described in [9] and [10], are the Subject-Verb-Object (SVO) language similar to English [11]; the sentence of this language is made up from a noun phrase subject followed by a verb phrase predicate. To parse syntactic patterns of sentences, the phrase structure grammar (PSG) is very useful because the result of parsing sentence by PSG, called the phrase structure tree, represents clearly sentence structures. Sentences are separated into constituent parts namely phrasal categories such as noun phrases, verb phrases, adverb phrases or prepositional phrases using phrase structure rules of PSG. Also, all phrases in sentences are broken into lexical parts that are nouns, verbs, adverbs, prepositions or other part of speech. The PSG of each language is created from the natural language grammar of each language that contains different phrase structure rules.

The semantic analysis is the most important part for analyzing sentences to interpret meaning of sentences and find information about two consecutive serial verbs. The semantic analysis is defined as a method for eliciting information and representing knowledge in [12]. There are various methods to perform the semantic analysis, such as lexical semantics and computational semantics. Lexical semantics emphasize the role played by the special primitives words used to represent the meanings of all words in a sentence. In this approach [13], the meaning is constructed through a restatement of the expression in terms of linked primitive words.

From the method proposed by [2] and [3], it has shown that the surrounding words of SVCs related to the syntactic patterns of SVCs. Additionally, [5] also supported that there are some relationships between the syntactic patterns and semantics of SVCs. However, these methods exclude systematic bindings between syntax to semantics which affects to the meanings of SVCs. Thus, this research analyzes both syntactic structure and semantic structure of basic SVCs to generate semantic bindings and to determine semantic interpretation.

The research methodology focuses on only two consecutive basic SVCs because the basic SVCs are found in sentences more than the complex SVCs. Moreover, the purpose of this method is to capture knowledge automatically by collecting different raw sentences with two consecutive serial verbs on various documents. Furthermore, the semantic analysis in this research is based on lexical semantics and a lexicon providing some basic word information. According to [14], there are researches combining fields of some machine learning techniques and the natural language processing. Therefore, a neural network is applied to improve performance of the natural language processing in discovering knowledge from words.

![Figure 1. The knowledge capturing framework.](image-url)
3. KNOWLEDGE CAPTURING METHODOLOGY

This section describes details of the proposed method for dissolving knowledge about two consecutive serial verbs. This method is divided into two main parts that are the syntactic analysis of sentences, and the semantic analysis to determine semantic bindings of two consecutive serial verbs. The framework of an approach is shown in Figure 1.

According to Figure 1, the syntactic analysis of sentences composed of two modules while the semantic analysis composed of four modules. The sentence with SVCs will be analyzed and parsed in the syntactic analysis process. Then, the structure of sentences and the syntactic patterns of SVCs are created from the phrase structure tree of sentences. In this semantic analysis, relationships between SVCs and their related words are unclosed by designed sentence theorems. All relations are learned to generate semantic binding patterns of SVCs. The details of the syntactic analysis of sentences and the semantic analysis are described in Section 3.1 and Section 3.2 respectively.

3.1 Syntactic Analysis of Sentences

From the syntactic analysis of sentences in Figure 1, the sentence with SVCs will be segmented by a tool and be analyzed by PSG rules. Therefore, the structure of sentences and the syntactic patterns of SVCs can be determined by phrase structure trees.

**Word Segmentation**

According to the word segmentation process, there are many languages that have no boundary between words, such as Thai and Chinese languages. Therefore, the word segmentation process must be implemented as a tool [7] to segment individual words and to tag their parts of speech. In the word segmentation, each word will be considered individually, except the serial verb that contains more than one verb will be considered as one individual phrasal verb. Thus, words such as nouns, auxiliaries, adverbs, adjectives and prepositions will be considered independently.

As the fact that a statement can contain punctuation, determiners, and compound words, so, the outputs from the word segmentation process must be edited for the correctness and completeness of the results. Consequently, punctuations and some determiners will be eliminated after the word segmentation process; additionally, the compound words will be maintained and correcting rules for compound words are generated.

Syntactic Analysis and Phrase Structure Analysis

After finishing the word segmentation process, the next step for the content interpretation is the process of syntactic analysis and the phrase structure analysis. Considering the syntactic analysis process, the structure of a sentence must be defined by the phrase structure analysis. Moreover, the syntactic analysis and the phrase structure analysis processes use a PSG to analyze parts of sentences. Since a sentence can contain basic phrases that can represent meaning of sentences, these basic phrases can be classified as noun phrases, verb phrases, adverb phrases or prepositional phrases. The PSG describing the sentence structures of the language is defined in a form of a set of rules that describes possible combinations of basic phrases. Thus, the automatic analysis of syntactic structures of sentences can be performed.

Generally, the PSG contains only structures of phrases in sentences. The
sentence can be composed of different types of phrases, such as noun phrases, verb phrases, adverb phrases, or prepositional phrases. Unfortunately, this information in the general PSG is not capable for the semantic information capturing. Therefore, in this research, a new PSG [15] is defined by adding semantic information of basic sentence structures into the basic PSG. The results are basic sentence structures classified into four different categories using functions: subjects, verbs, objects, and adverbials. Therefore, the main parts of sentences that are subjects, verbs, or objects can be defined by traversing branches of phrase structure trees. In addition, the analysis of syntactic structures of serial verbs in sentences can be achieved easily.

As mentioned previously, the inputs of the syntactic analysis and the phrase structure analysis module are the individual words and their parts of speech. These inputs will be passed through the new PSG. During the classifying process by the PSG, the probability of phrase patterns computed from all sentences in the corpus (Equation 1) is applied and the phrase structure trees for the sentences are created.

\[ \text{Prob}(\text{rule}) = \frac{\text{Number of parsed sentences by rule}}{\text{Number of parsed sentences by the same non-terminal symbol rules of rule}} \]  \( \text{(1)} \)

The advantage in applying the probability of phrase patterns is that the only one phrase structure tree will be generated for a sentence. The probability will be applied only when there are a variety of alternatives for creating a phrase structure tree of the sentence.

The consequence of creating phrase structure trees for a language is all possible syntactic patterns of a sentence with SVCs.

\[ \begin{align*}
1 & \rightarrow <1> \\
2 & \rightarrow <2> \\
3 & \rightarrow <3> \\
4 & \rightarrow <4> \\
5 & \rightarrow <5> \\
6 & \rightarrow <6> \\
7 & \rightarrow <7> \\
8 & \rightarrow <8> \\
9 & \rightarrow <9> \\
10 & \rightarrow <10> \\
\end{align*} \]

<1> refer: noun phrase (the subject of a sentence)
<2> refer: verb phrase (the main verb of a sentence)
<3> refer: prepositional phrase
<4> refer: adverb phrase
<5> refer: noun phrase (the object of a sentence)
<6> refer: subordinating clause
<7> refer: coordinating clause
<8> refer: relative clause
<9> refer: clause (the subject of a sentence)
<10> refer: empty string

Figure 2. Sentence diagram that SVCs are found as main verbs of Thai sentences.
These syntactic patterns of sentences in the form of phrase structure trees are called as a sentence diagram of the language. The example of a sentence diagram of Thai language, which SVCs are found as main verbs, is shown in Figure 2. Additionally, the diagram in Figure 3 expresses these phrase patterns of clauses in the form of transitions that SVCs are found as verbs in Thai clauses.

Referring to Figure 2, each node (circle with a number) represents each state during parsing each part of a sentence. State transitions take place, when phrases or clauses are found, that are represented by an edge in the diagram. Each path, which expresses phrases or clauses found in the sentence, is each syntactic pattern of the sentence. Therefore, all paths in the diagram represent all possible phrases or clauses in the sentence that appear in sentences containing SVCs in the main clause.

For example, a Thai statement is “ซอฟต์แวร์ NECTECII สามารถใช้ออกแบบวงจรรวมได้จริง” (sopwae - NECTECII - samat - chaiokbaep - wongchonruam - daiching) means “The NECTECII software can be used to design an integrated circuit practically”. Hence, its syntactic sentence structure can be discovered by starting at state 1, traversing to state 2, 3, 8, 5 and ending at state 14. As a result, this Thai statement is composed of four phrases that are the noun phrase, the verb phrase, the noun phrase and the adverb phrase, respectively. The detail of parsing this Thai statement will explain in section 4.1 Experiment in syntactic analysis.

Referring to Figure 3, sentence diagram is created from all possible syntactic patterns of clauses that contained SVCs as verbs. Like the sentence, state transitions take place, when phrases or sub clauses are found in the clause, which are represented by an edge in the diagram. Each path in diagram show each syntactic pattern of the clause with SVCs.

**Figure 3.** Sentence diagram that SVCs are found as verbs of Thai clauses.
According to Figure 2 and Figure 3, the complete sentences or complete clauses in the research scope have to be separated into two parts that are subject and verb. The state 3 mean the sentence or the clause pass the state that are composed of subject part and verb part, so it can be ended by $<10>$ empty string or it can be followed by other parts of sentence or clauses such as object part or adverb part or others, etc. Therefore, every path in sentence diagram will traverse to parse state 3.

### 3.2 Semantic Analysis

This phase is an important process for a sentence interpretation because it will produce meaning of the sentence by considering the created phrase structure trees and embedded semantic information in trees. Since the meaning of a word in a phrase structure tree usually depends on its position, so, relationships among words and their surrounding words must be determined. Finally, the knowledge from the sentence with SVCs can be obtained.

The proposed mechanism for the semantic analysis of components is composed of four processes. The first and the second processes are finding words surrounding SVCs and analyzing relationships between SVCs and surrounding words; semantic relations of SVCs with other parts of sentence structures will be generated. The next process is to define classes of words related to SVCs. The last process is to analyze all features to determine semantic binding patterns and semantic interpretation of SVCs.

From the semantic analysis in Figure 1, finding words surrounding SVCs is responsible for dissolving phrases or words relating to SVCs. Then, the relationship analysis process will be performed. The relationships analysis is the process for obtaining context information through the phrases’ relationships based on the sentence theorems. Next, the classes of related words are defined to add semantic information about SVCs. Finally, all relations and classes of related words are analyzed and learned by a neural network to generate semantic binding patterns of SVCs. As a result, the semantic interpretations of SVCs also are interpreted and represented. Details of these modules are described in the following section.

#### Finding Words Surrounding SVCs

After receiving words and syntactic patterns, as called a phrase structure tree, from the syntactic analysis phase, words surrounding SVCs will be identified and discovered by traversing branches of phrase structure trees of sentences. For example, a subject of SVCs will be found in the noun phrase node occurring before the verb phrase containing SVCs node in a phrase structure tree, while an object of SVCs will be included in the noun phrase occurring after SVCs. Therefore, the important words that impact on SVCs should be found in those phrases that related to verb phrases with SVCs.

#### Analysis Relationships

Once the surrounding words are exposed, the next process is to find the relationships among these words and serial verbs. In order to obtain the relationships among these words and serial verbs, the sentence theorems will be applied. The sentence theorem refers to formats of all possible sentences of the language. Thus, each language will have different sentence theorems. Therefore, in the development of an automatic interpretation of a sentence, these sentence theorems will be implemented as an algorithm of the
After analyzing relations between SVCs and other phrases, the semantic binding patterns are produced by binding all phrases related to SVCs based on the syntactic structures and relationship analysis; these phrases are such as the subject phrases of SVCs, the object phrases of SVCs, or the modifier phrases of SVCs. The position of each phrase will have impact to the semantic interpretations of sentences, described in the following section. Examples of the binding patterns in Thai language are presented in Figure 4.

Figure 4. Examples of semantic binding patterns.

After finishing the first two processes mentioned above, the surrounding words and relations that related to SVCs will be disclosed from the sentence. Then, the surrounding words determined classes of words in defining the phase. The first noun of the noun phrase occurring before SVCs in a sentence will be the head noun to define classes. Some word information in a lexicon and some designed heuristic rules are used to define classes of head nouns in noun phrases, classes of adverb phrases, and classes of prepositional phrases. Domains or classes of nouns are classified because both relations are the modifier parts of SVCs.

Defining Classes of Words

After analyzing relations between SVCs and other phrases, the semantic binding patterns are produced by binding all phrases related to SVCs based on the syntactic structures and relationship analysis; these phrases are such as the subject phrases of SVCs, the object phrases of SVCs, or the modifier phrases of SVCs. The position of each phrase will have impact to the semantic interpretations of sentences, described in the following section. Examples of the binding patterns in Thai language are presented in Figure 4.

Figure 4. Examples of semantic binding patterns.

After analyzing relations between SVCs and other phrases, the semantic binding patterns are produced by binding all phrases related to SVCs based on the syntactic structures and relationship analysis; these phrases are such as the subject phrases of SVCs, the object phrases of SVCs, or the modifier phrases of SVCs. The position of each phrase will have impact to the semantic interpretations of sentences, described in the following section. Examples of the binding patterns in Thai language are presented in Figure 4.

After analyzing relations between SVCs and other phrases, the semantic binding patterns are produced by binding all phrases related to SVCs based on the syntactic structures and relationship analysis; these phrases are such as the subject phrases of SVCs, the object phrases of SVCs, or the modifier phrases of SVCs. The position of each phrase will have impact to the semantic interpretations of sentences, described in the following section. Examples of the binding patterns in Thai language are presented in Figure 4.

After analyzing relations between SVCs and other phrases, the semantic binding patterns are produced by binding all phrases related to SVCs based on the syntactic structures and relationship analysis; these phrases are such as the subject phrases of SVCs, the object phrases of SVCs, or the modifier phrases of SVCs. The position of each phrase will have impact to the semantic interpretations of sentences, described in the following section. Examples of the binding patterns in Thai language are presented in Figure 4.
into various classes, such as the person class referring to persons or organizations (e.g. users, employees, officers); the artifact class referring to artifacts or instruments (e.g. computers, programs, software, instruments); the data class referring to data or information, the activity class referring to activities or processes; and the abstract class referring to abstract nouns.

According to the semantic hierarchy of noun classes [16] and noun sets [17], the main characteristics of nouns will represent individual identity of noun classes. So the characteristics of each class are classified from only main characteristics of main classes from semantic hierarchy of noun classes [16] and noun sets [17]. In word sense disambiguation researches [18-19] showed that main fundamental characteristics of the noun class can express the required concepts for information of nouns to classify into classes, while sub characteristics add more extensions for information of nouns to declare specific meaning. The reason is that the most of clearly identities of nouns are elementary characteristics of noun classes.

For adverb and prepositional phrases, beginning words of phrases, that are an adverb and a preposition respectively, used to be keywords for classes. For example, while “at” and “in” are used to express a location of an action in a sentence, “with” is used to express that the action in a sentence occurs in cooperation with nouns following by the preposition.

3.3 The Semantic Interpretations of SVCs

After passing the semantic analysis phase, the semantic binding patterns are properly stated. These patterns with their individual semantic will be analyzed using a neural network method. The learning mechanism in this research is the probabilistic neural networks [20] run by a tool [21]. Additionally, this research applied two layers of the probabilistic neural networks because the neural network is appropriate for nonlinear classification and more effective than other techniques in learning for ambiguities [22][23]. Moreover, the training is easy and instantaneous.

This entire process tried to predict the semantic interpretations of SVCs, which are classified into groups. Each language will have different number of groups. However, every language generally contains two fundamental groups: “Purpose”, and “Complement” [2] [4] [5]. The purpose group refers to the group that the second verb in two consecutive SVCs indicates the purposive action of the first verb while the complement group refers to the group that the second verb in two consecutive SVCs is the complement of the first verb.

The meaning of SVCs in the purpose group is interpreted from the combination of the meanings of two action verbs in two consecutive SVCs; these two action verbs have equal status. On the other hand, the meaning of SVCs in the complement group is an action where the first verb in two consecutive SVCs cannot indicate any clear actions for readers without the second verb.

Referring to the sentence interpretation processes, the proposed two main processes are unavoidable. Then, the meaning of a sentence with SVCs can be elaborated automatically. This proposed method is verified and validated by running experiment on Thai language as described below.

4. EXPERIMENTS AND RESULTS

In this research, the experiment was separated into two main parts based on the semantic binding patterns analysis process, including semantic interpretations of SVCs. The first part is examination of phrases’
relationships in sentences where the semantic binding patterns are derived. The second part is the process to determine groups of semantic interpretation so meanings of sentences with SVCs can be expressed without human intervention.

The research conducts experiments on one of the Asian language, called ORCHID (Open linguistic Resources CHanelled toward InterDisciplinary research) [24], which is Thai language corpus. The ORCHID is an initiative aimed at building linguistic resources to support research in the natural language processing. This corpus is a collection of annual reports and research proposals produced by different universities and research organizations in Thailand. Therefore, all sentences with two consecutive SVCs are gathered from sentences in the ORCHID corpus. Consequently, the number of main sentences with two consecutive SVCs in the corpus is 262 sentences.

Additionally, the number of sentences that two consecutive SVCs are founded as verbs in clauses of sentences is 124 sentences. In order to confirm the semantic binding patterns and the group of semantic interpretations, an unseen data set which is not found in the ORCHID corpus is applied. This Thai corpus is provided by Department of Linguistics, Chulalongkorn University (http://ling.arts.chula.ac.th/ThaiConc/). There are 144 sentences which contain serial verbs. Within these 144 sentences, there are 110 sentences that two consecutive SVCs are found as verbs in the main clause of sentences. Additionally, there are also 34 sentences that two consecutive SVCs are found as verbs in clauses of sentences.

This experiment has implemented modules to perform the proposed knowledge capturing framework are showed in Figure 5.

Referring to Figure 5, the word segmentation in the syntactic analysis process was managed automatically by a Thai word segmentation tool called SWATH [7]. The outputs from the first process, words and their part of speech, are constructed to be phrase structure trees of sentences by parsing with the phrase structure grammars of Thai language derived by the author in the parser module. Then, sentence structures from phrase structure trees are analyzed to find relations among all phrases in the sentence by the relation analysis module using sentence theorems. The output from...
the relation analysis module will be sent to two different modules: the first learning module and class defining module. However, the process of the class defining module cannot proceed unless the first learning process finished.

According to the first learning process, semantic binding patterns are dissolved. Thus, these patterns represent phrases that related to SVCs. Additionally, keywords in these phrases are classified into semantic classes of words following the detail in define classes of words module. All information from previous processes is defined as features and will be input for the second learning module. The results of this learning are the semantic interpretations of SVCs.

The following sections describe the result of experiments running in Thai language separated into two sections. The first section shows the example of output from syntactic analysis and the second section contains two experiments which are conducted to study on semantic binding patterns and the semantic interpretation of SVCs.
4.1 EXPERIMENT IN SYNTACTIC ANALYSIS

For example, a Thai statement is “ซอฟตแวร์ NECTECII สามารถใช้ออกแบบวงจรรวมได้จริง” in section 3.1. The output from SWATH is shown in Figure 6 and the phrase structure tree, the output from the parser, is expressed in Figure 7.

Referring to Figure 7, this Thai statement begins with the subject part (sj) in form of the noun phrase (np). Then, the rest part of the sentence is parsed into three parts that are verb, object and adverbial (voa). The verb part is the verb phrase (vp) and the object is the noun phrase like the subject part, while the adverbial part is the adverb phrase (advp). Therefore, this sentence can be discovered by the one path of the sentence diagram in Figure 2 and steps of traversing the path are shown on the left hand side of Figure 7. This path starts at state 1 and traverse to state 2, 3, 8, and 5 by the edge <1> (noun phrase - subject), <2> (verb phrase), <5> (noun phrase - object), and <4> (adverb phrase), respectively. The path ends with state 14, after get the edge <10> (empty string).

4.2 EXPERIMENT IN SEMANTIC ANALYSIS

This sub-section describes two experiments running in Thai language. The objective of the first experiment is to conduct the study on semantic binding patterns of SVCs and the second experiment is to conduct the study on semantic interpretations. The results of these experiments indicate that the proposed method for automatic discover knowledge of SVCs to interpret sentences with SVCs is acceptable.

Experiment I

Experiment I is set to classify binding patterns in sentences that are connected to two consecutive SVCs. Each relation in a sentence will be identified as individual feature to learn the binding patterns. As a result, the binding patterns of main verbs in sentences are classified into 18 patterns and the binding patterns of verbs in clauses are classified into 13 patterns. Additionally, all data are separated into three folds. One fold is used to test the data sets, when the rest two folds are used as the training data sets. To test on unseen data, the training data is the same training data set of each test data fold.

As the fact that there are various relations in sentences, these relations are defined as features of sentences. The examples of features are relation “Subject” and “Object” that are noun phrases in the sentences occurring before and after SVCs. The other examples of features are relation “Modifyv1” and “Modifyv2” that are a prepositional phrase and an adverb phrase following SVCs.

However, in this experiment, the defined features are sentence structures and relations in sentences. So, the features are selected from related relations that are related to SVCs only. Thus, from the above example, the unselected relations are “Modifyn1” and “Modifyn2” because both relations are relative clauses in sentences that occur after noun phrases. In addition, the roles of this relative clause are modifying nouns or noun phrases that are subject and object of sentences.

The input of classification is a vector matrix storing all features of each sentence and the output is semantic binding patterns derived by the probabilistic neural networks. The performance of the method in Experiment I was computed as accuracy rates by applying the formula in Equation 2. The semantic binding patterns of SVCs between the main verb of sentences and the verb of clauses are tested independently
because of different priority and different roles in sentences. The result of experiment I is shown in Table 1 and Table 2 for the verbs of the main clause and the verb of the sub clause in sentences, respectively.

Moreover, the percent accuracy of all data sets and the percent accuracy average in experiment I are graphically represented in Figure 8.

Table 1. Result of experiment I for verbs of main clauses in sentences.

<table>
<thead>
<tr>
<th>SVCs' Roles</th>
<th>Results of each Data Set</th>
<th>Avg. of Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(No. of All Sentences = No. of Sentences in Test Set, No. of Sentences with Correct Binding Pattern, No. of Accuracy)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Main Clauses (corpus)</td>
<td>Test: 1st or unseen Training: 2nd + 3rd = 175 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training: 2nd or unseen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(87, 86, 98.85%)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Main Clauses (unseen)</td>
<td>Test: 2nd or unseen Training: 1st + 3rd = 175 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training: 1st + 2nd</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(88, 88, 100%)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Main Clauses (unseen)</td>
<td>Test: 3rd or unseen Training: 1st + 2nd = 174 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(110, 110, 100%)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2. Result of experiment I for verbs of sub clauses in sentences.

<table>
<thead>
<tr>
<th>SVCs' Roles</th>
<th>Results of each Data Set</th>
<th>Avg. of Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(No. of All Sentences = No. of Sentences in Test Set, No. of Sentences with Correct Binding Pattern, No. of Accuracy)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Sub Clauses (corpus)</td>
<td>Test: 1st or unseen Training: 2nd + 3rd = 83 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training: 2nd or unseen</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(41, 40, 97.56%)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Sub Clauses (unseen)</td>
<td>Test: 2nd or unseen Training: 1st + 3rd = 83 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training: 1st + 2nd</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(42, 41, 97.62%)</td>
<td></td>
</tr>
<tr>
<td>Verbs in Sub Clauses (unseen)</td>
<td>Test: 3rd or unseen Training: 1st + 2nd = 82 sentences</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(34, 34, 100%)</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{accuracy} = \frac{\text{Total number of sentences with correct binding pattern of SVCs}}{\text{Total number of all sentences}} \times 100 \%
\] (2)

According to Table 1 and Table 2, the result data is shown in the set of three values that are the total number of all sentences, the total number of sentences with correct binding pattern of SVCs, and the percent accuracy of classification for each data set, respectively. The number of sentences in training sets is identified in the column header, while the number of sentences in test sets is the number of all sentences in each data set. To plot graph in Figure 8, only the percent accuracy for each data set fold and the percent accuracy average are expressed.

According to the results in Figure 8, the accuracy of corrected binding patterns is nearly 100% because binding patterns of two consecutive SVCs can be classified by...
formula rules. In contrast, the unclassified patterns can occur because they were existed in the test data only. To test for unseen data, the accuracy of corrected binding patterns is 100% because all patterns of unseen sentences were found in ORCHID corpus.

The feature set for identifying semantic binding patterns are composed of sentence structures and relations of SVCs among other words, phrases or clauses of sentences which depend on language grammar. Therefore, the rules for classifying binding pattern can be generated by analyzing sentence structure and relation in sentence to find the same feature values for SVCs in the same semantic binding pattern. For example, the semantic binding patterns of the two consecutive serial verbs, which are the verb of the main clause in sentences, can be identified as the first pattern in Figure 4 ([[Subject: NP][Verb: SV][Object: NP]]), if the value of feature showing relationship between the subject and main verb (“Subject”), and the value of feature presenting relationship between main verb and the object (“Object”) are more than zero, including the feature indicating sentence structures like one of sentence structures in Figure 9.

**EXPERIMENT II**

Experiment II performs the classification of semantic interpretation of SVCs, and all data are separated into three folds as

| S: = <NP,SV,NP> |
| S: = <NP,SV,NP,RELCL> |
| S: = <NP,RELCL,SV,NP> |
| S: = <NP,RELCL,SV,NP,RELCL> |

**Figure 9.** Examples of sentence structures.

same as Experiment I. As mentioned previously that there are two groups of semantic interpretations: Purpose, and Complement. The examples of Thai two consecutive SVCs in the purpose group are “ชัย/chi-sa-daeng” (“chai(use)”, sadaeng(display))” and “เลือก/leuak len” (“leuak(select), len(play)”). The examples of Thai two consecutive SVCs in the complement group are “ต้องการวัด/tong-karn wat” (“tongkarn(want), wat(measure)”) and “พยายามหา/phayayam ha” (“phayayam (try)” “ha(find)”).

In this experiment, the defined features are sentences structure, relations in sentences, and classes of surrounding words. The sentences structure and
relations in sentences are the same described features in Experiment I. The added feature is classes of surrounding words that are related to SVCs. Moreover, there are different classes of surrounding words that are the head noun of noun phrases, the preposition in the prepositional phrase and the adverb in the adverb phrase. The examples of surrounding word classes are elaborated below.

If relation features are “Subject” and “Object”, the examples of added features are noun classes of a subject and an object, such as “Person”, “Artifacts” and “Data”. If the other relations are “Modifyv1” or “Modifyv2”, the examples of added features are prepositional classes, such as “Location”, “Source or Destination”, “Purpose”, and “Mean” or adverb classes, such as “Manner” and “Time or Frequency”.

The input of classification is a vector matrix containing all features of each sentence and the output is semantic binding patterns derived from the probabilistic neural networks. The experiment was practiced on two different feature sets. The first feature set is the sentence structure, relations in sentences and classes of surrounding words. The first verb of two consecutive SVCs is added to the second feature set. In addition, the performance of the method in Experiment II is also separated into two evaluations. The first evaluation was computed as accuracy rates by applying the formula in the Equation 3, and the second evaluation was computed as precision rates by applying the formula in the Equation 4. The results of experiment II are shown in Table 3, Table 5 and Table 6. The purpose of evaluation by the precision rate is to measure the correctness and the performance of the classification. In binary classification, only the accuracy rate may not be the adequate

Table 3. Result of experiment II in terms of accuracy rates.

<table>
<thead>
<tr>
<th>SVCs’ Roles</th>
<th>Results of each Data Set</th>
<th>Avg. of Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(No. of All Sentences = No. of Sentences in Test Set, No. of Sentences with Correct Binding Pattern, No. of Accuracy)</td>
<td></td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (corpus)</td>
<td>Test: 1st or unseen Training: 2nd + 3rd = 257 sentences</td>
<td>65.54%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (corpus)</td>
<td>Test: 2nd or unseen Training: 1st + 3rd = 257 sentences</td>
<td>84.20%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (unseen)</td>
<td>Test: 3rd or unseen Training: 1st + 2nd = 258 sentences</td>
<td>82.41%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (unseen)</td>
<td>(129, 79, 61.24%)</td>
<td>(129, 82, 63.56%)</td>
</tr>
<tr>
<td></td>
<td>(129, 112, 86.82%)</td>
<td>(129, 107, 82.94%)</td>
</tr>
<tr>
<td></td>
<td>(144, 97, 67.36%)</td>
<td>(144, 91, 63.19%)</td>
</tr>
<tr>
<td></td>
<td>(144, 120, 83.33%)</td>
<td>(144, 114, 79.17%)</td>
</tr>
</tbody>
</table>
measurement, when the number of cases in one group is much greater than the number of cases in another group. In addition, both correct and incorrect classifications are used to calculate the precision rate, while the accuracy rate is calculated by only the correct classification.

\[
\text{accuracy} = \frac{\text{Total number of sentences with correct interpretations of SVCs}}{\text{Total number of all sentences}} \times 100 \quad (3)
\]

\[
\text{precision} = \frac{\text{Total number of sentences with correct SVCs classified to this interpretation}}{\text{Total number of sentences with SVCs classified to this interpretation}} \times 100 \quad (4)
\]

According to Table 3, the result data is also shown in the set of three values for each test data sets. The first value is the total number of all sentences in test set, and the second value is the total number of sentences correctly classified for semantic interpretations, including the percent accuracy of classification as the last value. Like experiment I, only the accuracy rates of semantic interpretation classification for all data sets in experiment II is shown in Figure 10.

According to the accuracy rates in Figure 10, it has shown that the accuracy of classifying semantic interpretations by the second feature set is about 80%, while the accuracy of using the first feature set is only 65%. Thus, the second feature set can determine semantic interpretations of SVCs better than that of the first feature set. In addition, the results can show that the knowledge about SVCs, such as syntactic patterns and syntactic binding patterns, is the useful information to identify semantic interpretation of SVCs.

However, the classification of semantic interpretations by the first feature set in Figure 10 is lower than 70% that means if this method learns sufficient data from different writing styles, the accuracy of interpretation will be much efficiency. The reason for that is that when comparing the result of classification by the first feature set between the test data from corpus and the test data from unseen documents, there is the fluctuation in
accuracy rates from variation of writing styles. The classification accuracy of data from unseen documents is higher than that of data from corpus when training by the first training data set (the second fold and the third fold), but when training by the third data set (the first and the second fold), the accuracy of classifying the test data from unseen documents is lower than that of classifying the test data from corpus.

After evaluating the accuracy of semantic interpretation classification, the precision of that classification for each group is figured by values in the confusion matrix. A confusion matrix displays the number of correct and incorrect classifications of semantic interpretation compared with the actual interpretation. There are two classes of semantic interpretation of SVCs that are the purpose group and the complement group so the confusion matrix of this classification is the $2 \times 2$ matrix. The rows present the number of actual classifications in the test data and the columns present the number of predicted classifications. All confusion matrix of all data sets are shown in Table 4.

According to each confusion matrix in Table 4, the first column in the first row show the number of SVCs in the purpose class of semantic interpretation classified correctly, while the second column in the second row indicate the number of SVCs in the complement class identified correctly. The number of SVCs in the purpose class of semantic interpretation was incorrectly classified as the complement class locating in the second column of the first row, and the

<table>
<thead>
<tr>
<th>SVCs’ Roles</th>
<th>Results of each Data Set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test: 1st or unseen</td>
</tr>
<tr>
<td></td>
<td>Training: 2nd + 3rd</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (corpus)</td>
<td>$[[28,31]]$</td>
</tr>
<tr>
<td></td>
<td>$[[19,51]]$</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (corpus)</td>
<td>$[[47,12]]$</td>
</tr>
<tr>
<td></td>
<td>$[[5,65]]$</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (unseen)</td>
<td>$[[25,17]]$</td>
</tr>
<tr>
<td></td>
<td>$[[30,72]]$</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (unseen)</td>
<td>$[[30,12]]$</td>
</tr>
<tr>
<td></td>
<td>$[[12,90]]$</td>
</tr>
</tbody>
</table>

Table 4. All confusion matrix of semantic interpretation classification of all data sets.
Table 5. Results of experiment II in terms of precision rates of “Purpose”.

<table>
<thead>
<tr>
<th>SVCs’ Roles</th>
<th>Results of each Data Set</th>
<th>Avg. of Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(No. of Sentences Classified to “Purpose”, No. of Sentences with Correct SVCs Classified to “Purpose”, No. of Precision)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 1st or unseen Training: 2nd + 3rd</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 2nd or unseen Training: 1st + 3rd</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 3rd or unseen Training: 1st + 2nd</td>
<td></td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (corpus)</td>
<td>(47, 28, 59.57%)</td>
<td>65.05%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (corpus)</td>
<td>(52, 47, 90.38%)</td>
<td>89.74%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (unseen)</td>
<td>(55, 25, 45.45%)</td>
<td>43.78%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (unseen)</td>
<td>(42, 30, 71.43%)</td>
<td>68.89%</td>
</tr>
</tbody>
</table>

Table 6. Results of experiment II in terms of precision rates of “Complement”.

<table>
<thead>
<tr>
<th>SVCs’ Roles</th>
<th>Results of each Data Set</th>
<th>Avg. of Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(No. of Sentences Classified to “Complement”, No. of Sentences with Correct SVCs Classified to “Complement”, No. of Precision)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 1st or unseen Training: 2nd + 3rd</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 2nd or unseen Training: 1st + 3rd</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test: 3rd or unseen Training: 1st + 2nd</td>
<td></td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (corpus)</td>
<td>(82, 51, 62.20%)</td>
<td>65.78%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (corpus)</td>
<td>(77, 65, 84.42%)</td>
<td>81.24%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. (unseen)</td>
<td>(89, 72, 80.90%)</td>
<td>81.34%</td>
</tr>
<tr>
<td>Sentence struc. + Rel. + Word cl. + 1st verb (unseen)</td>
<td>(102, 90, 88.24%)</td>
<td>88.58%</td>
</tr>
</tbody>
</table>
number of SVCs in the complement class, incorrectly classified as the purpose class, is appeared in the first column of the second row. For the example confusion matrix of the first data set in Table 4, the test data is composed of 129 sentences that the actual semantic interpretations of SVCs are 59 in the purpose group and 70 in the complement group. The number of correct semantic interpretation classification for the purpose class and the complement class are 28 and 51, respectively. The number of incorrect semantic interpretation of SVCs is 31 for the purpose class and 19 for the complement class.

Referring to Equation 4, the total number of sentences with correct SVCs classified to “Purpose” interpretation is the number of correct semantic interpretation classification for the purpose class, and the total number of sentences with SVCs classified to “Purpose” interpretation is the number of correct semantic interpretation classification for the purpose class adding the number of incorrect semantic interpretation classification for the complement class. The result of precision evaluation for the semantic interpretation “Purpose” and “Complement” are shown in Table 5 and Table 6, respectively.

According to Table 5 and Table 6, the evaluation by the precision rate is
different from the measurement of the accuracy rate because the precision of classification is separately calculated for each class of semantic interpretations and the precision values depend on both correct and incorrect classification. The results of precision measurement are shown in the sets of three values that are the total number of sentences with SVCs classified to each group interpretation, the total number of sentences with correct SVCs classified to each group interpretation, and the percent of precision for each group interpretation, respectively. The graph of the percent precision and the average of precisions for each data set are shown in Figure 11 ("Purpose") and Figure 12 ("Complement").

Similarly to the accuracy rate, the precision rates of classifying both semantic interpretations by the second feature set are higher than those of classification by the first feature set. Thus, these results of semantic interpretations reveal that both of semantic interpretations of SVCs can be determined by the sentence structure (syntactic patterns), relations (semantic binding patterns), surrounding word classes, and the first verb of two consecutive SVCs.

According to Figure 11 and Figure 12, although all average of precision rates of "Purpose" is lower than those of "Complement", most of them are nearly 70% or higher. In some data sets, the precision rates of "Purpose" and of "Complement" in the unseen document are in different directions when comparing to the rates of the corpus. The cause of these results is the difference of the ratio between number of "Purpose" SVCs and "Complement" SVCs in the corpus and in the unseen documents. However, the result of unseen document confirms that the semantic interpretations of SVCs can be determined by the sentence structure, relations, surrounding word classes and the first verb of two consecutive SVCs.

5. DISCUSSION

Having information of syntactic and semantics of statements leads to the content knowledge of whole documents. Therefore, the efficient syntactic and semantic interpreters are the mandatory requirements for information capturing to automatically discovery knowledge. Unfortunately, semantic heterogeneity and the syntactic heterogeneity still have been problems of interpretation, especially serial verbs.

In the linguistic research, SVCs are separated into an independent component, or called as a single particle, before performing the semantic and the syntactic analysis [2] [5]. However, these methods have some gaps in the interpretation of words since the meaning also depends on locations of words and its surroundings. Therefore, this research considers one SVC as one object although this object is composed of many words. In addition, the segmented words from the segmentation tool [7] are improved to analyze and to interpret meaning clearly. Since one object can be a component of a phrase, relations among phrases in sentences can be easily identified. Consequently, the syntax patterns can be uncovered from phrase structure trees conforming to grammar rules [15]. Therefore, the syntax patterns can be easily derived by the proposed method when comparing to the other techniques.

As the fact that the locations and relations of words of SVCs are very important to the automated interpretation, this paper proposes the relation analysis process which has focused on syntactic pattern, and semantic binding patterns.
Then, to interpret the relationship between single verbs in SVCs, this study analyze syntactic pattern, relations and surrounding words covering more than considering technique[4]. Consequently, the knowledge of the SVCs is automatically discovered and captured. Moreover, this research also considers the meaning of the surrounding words of SVCs and combines them to display a clear content for readers much better than the legacy techniques.

6. CONCLUSIONS

In the present time, document interpretation is an important issue for every human’s activities. However, until now, there is no such software that can perform an automated interpretation of content from the context. Moreover, many languages contain SVCs which cause difficulty in the interpretation process of software since the locations and every surrounding words influence to meaning of the SVCs and the entire statement. Thus, understanding of SVCs is much complicated than the single verbs.

Since SVCs are an important influencer for content finding, this research proposed methods to analyze the syntactic patterns, semantic binding patterns and related words of SVCs in order to discover knowledge about SVCs. As the results of performing both syntactic and semantic analysis processes, the useful information to analyze and dissolve knowledge about SVCs can be captured.

Generally, SVCs are composed of more than one verb with various patterns. Thus, knowledge about SVCs is actually helpful to find the meaning of SVCs that will finally lead to the understanding of the natural statements. According to the consideration of all surrounding words and locations of SVCs, including relationships between phrases in a sentence with SVCs, the understanding of sentences with SVCs is obtained.

Although this research performed the experiments on Thai language, the results show that the proposed solution of the SVCs interpretation is acceptable (more than 70% accurate). The incorrect interpretation was caused by limitations of learning process because of lacking of variety of the learning data for SVCs. Therefore, whenever the learning data is completed, the interpretation will be completed. Thus, this indicates that the proposed methods have filled in the gap obtained from performing semantic and syntactic analysis separately. Since it automatically uncovered semantic binding patterns and the semantic interpretation, the knowledge about two consecutive SVCs is clearly elaborated with reasonable accuracy. The benefit of this method is that the valuable information can be captured from documents using translation tools, question-answering systems or plagiarism detection systems [25] to make benefits to users in various areas from their understanding, especially partners in business areas. Furthermore, semantic binding patterns and patterns of SVCs are useful for building language resources of SVCs.

REFERENCES


Joan Bresnan, and Peter Sells, 1997; 247-288.


[22] Gao S., Wu W., Lee C. and Chua T., A maximal figure-of-merit learning approach to text categorization, Proceedings of the 26th annual international ACM SIGIR conference on...
Research and development in information retrieval (SIGIR ’03), 2003: 175-181.


[24] National Electronics and Computer Technology Center (NECTEC), Open linguistic Resources CHaneld toward InterDisciplinary research (ORCHID), 2010, Online: http://www.hlt.nectec.or.th/orchid/