PAPER *Special Section/Issue on ******

Construction of Thai Lexicon from Existing Dictionaries and Texts on the Web

Thatsanee Charoenporn[†, ††], *non-member,* Canasai Kruengkrai[††], *non-member,*
Thanaruk Theeramunkong[†], *member* and Virach Sornlertlamvanich[††], *non-member*

SUMMARY A lexicon is an important linguistic resource needed for both shallow and deep language processing. Currently, there are few machine-readable Thai dictionaries available, and most of them do not satisfy the computational requirements. This paper presents the design of a Thai lexicon named the TCL's Computational Lexicon (TCLLEX) and proposes a method to construct a large-scale Thai lexicon by re-using two existing dictionaries and a large number of texts on the Internet. In addition to morphological, syntactic, semantic case role and logical information in the existing dictionaries, a sort of semantic constraint called selectional preference is automatically acquired by analyzing Thai texts on the web and then added into the lexicon. In the acquisition process of the selectional preferences, the so-called Bayesian Information Criterion (BIC) is applied as the measure in a tree cut model. The experiments are done to verify the feasibility and effectiveness of obtained selection preferences.

Key words: *Thai Lexicon Construction, Logical and Selectional Preferences Semantic Constraints, Semantics Constraint Acquisition, Tree Cut Model*

1. Introduction

Recently several attempts have been made to construct machine-readable dictionaries (lexicons) in several languages, such as English WordNet [1], English FrameNet [2], Japanese EDR Dictionary [3], Chinese HowNet [4], IPAL [5] and so forth. The WordNet [1] includes attribute relations and adjective/adverb classifications which are semantic features extracted from the relations between words, and designed for understanding the semantics. All words are uniformly defined, regardless of their class. The concept level is considered, and synonymy, hyponymy, entailment, antonymy, meronymy, and troponomy are examined. In the FrameNet [2], a collection of lexical entries are grouped by frame semantics. Each lexical entry represents an individual word sense, and is associated with semantic roles and some annotated sentences. Lexical entries with the same semantic roles are grouped into a "frame" and the semantic roles are called "frame elements". HowNet [4] is a Chinese ontology with a graph structure of word senses called "concepts", and each concept contains seven fields including lexical entries in Chinese, English gloss, POS tags for the word in Chinese and English, and a definition of the concept including its category and semantic relations. IPAL [5] is a Japanese lexicon including a number of basic Japanese verbs, adjectives, and nouns. Each entry shows representative examples, information about semantics, morphology, grammatical categories, case frames and idiomatic usage. EDR [3] is a large-scale Japanese lexicon. Each record of EDR consists of entry information, grammatical information, semantic information, and pragmatic and supplementary information. Only synonymy, hyponymy, and entailment are examined. Most of these lexicons include not only the definition of words (or terms) but also the relationship among them, known as ontology. Applications of these lexicons vary from shallow processing, such as information retrieval, to deep processing, e.g., machine translation and, as more recent work, semantic web.

Among few Thai machine-readable dictionaries, Multilingual Machine Translation (MMT) [6] and LEXiTRON [7] are dominant. The former was constructed for machine-translation purpose while the latter was created as a general-purpose electronic dictionary. Due to these different purposes, one may contain information not existing in the other. Moreover, currently these machine-readable dictionaries cope with only limited semantic aspects of word entries. To implement in more practical applications, a more expressive dictionary with various types of information is required.

In this paper, we propose an approach to construct a content-rich computational lexicon by reusing information from the two existing dictionaries, i.e., the MMT dictionary and LEXiTRON, and then enriching the result with additional semantic information extracted from texts on the web. Toward this end, we design the specification of the lexicon and propose a method to acquire semantic information in automatic and semi-automatic ways rather than by only manual annotation. A number of practical approaches are applied for such tasks. As a by-product, we

[†] Sirindhorn International Institute of Technology, Thammasat University, 131 Tiwanont Rd. Bangkadi, Pathumthani, Thailand
[††] Thai Computational Linguistics Lab., NICT Asia Research Center

obtain a methodology of word sense representation, in contrast to the descriptive manner in general lexicons.

The remainder of the paper is organized as follows. Section 2 describes the existing dictionaries and reference thesaurus, MMT dictionary and LEXiTRON. In section 3, the structure of our lexicon and its construction is proposed. The method to acquire constraints from corpora is illustrated in section 4. In section 5, experimental results are reported together with some discussion. Finally, section 6 gives a conclusion and some future works.

2. Existing Dictionaries and Thesaurus

In the past, two large-scale dictionaries, namely Multilingual Machine Translation (MMT) Dictionary [6] for language processing and LEXiTRON [7] for human use were developed. The MMT dictionary contains 68,860 entries with 53,759 unique words while LEXiTRON covers 40,844 entries with 35,192 unique words. By the National Electronics and Computer Technology Center (NECTEC), the MMT dictionary was originally constructed for the Multilingual Machine Translation project, which is a six-year (1987-1992) cooperative project between groups of research institutes from five countries, i.e. China, Indonesia, Japan, Malaysia, and Thailand, organized by the Center of the International Cooperation for Computerization (CICC) in Japan.

Designed and developed for using in analysis and generation modules in the multilingual machine translation system, the MMT dictionary then contains as much detailed information as needed for concept disambiguation and word selection. The details are:

1. There are 13 part-of-speech categories and 51 sub-categories. The sub-category is designed to support syntactic analysis and grammar rule representation.
2. There is a mapping information of syntax and semantic relations, syntax-semantic mapping, for determining the syntax and semantic structure.
3. AKO (A-Kind-Of) is provided for grouping a semantic class. The semantic class is used to constrain possible arguments that a predicate can take.
4. Word entries are selected in preference of word length. Rather than a short word, a longer word representing more complex concept is then registered as an entry, to reduce complication in rule writing for machine translation. Therefore, word entries contain some large compound words, phrases and sometimes sentences.

The MMT dictionary consists of three levels of linguistic information; morphological, syntactic and semantic information. It was created based on some linguistic theories, such as phrase structure grammar, case grammar and dependency grammar.

WORD HEADER	จ่าย 3cf151 PAY	% Thai word % ID linking to description % Equivalent English word
MOR	TYPE.{S}	% Word formation (single/compound word)

SYN	CAT.{V}
	SUBCAT.{VACT}
	VPPAT.{SUB+V+DOB}
		% Word Category
		% Grammatical sub category
		% Syntax of verb in sentence

SEM	MAPS.{SUB=AGT, DOB=OBJ}
	AKO.{2-2-8}
		% Mapping of case relations to syntactic structure
		% A-Kind-Of relations indicating position in semantic hierarchy

Fig. 1 The MMT entry for the verb จ่าย ('pay').

Figure 1 shows an example of the word entry จ่าย ('pay'), in the MMT dictionary. The word header contains a Thai word entry and the corresponding concept ID and an English equivalent. Since the MMT dictionary is a sort of word-concept dictionary, the Thai word entry is uniquely defined a concept ID while the English equivalent is defined as an index for reference. Apart from the word header, the entry consists of 3 types of information, including morphological, syntactic, and semantic information. The morphological information (MOR) provides the type of word formation, i.e., a single or compound word. This information is used in generation rule for appropriate word selection according to its context. The syntactic information (SYN) indicates the grammatical category (CAT) and subcategory (SUBCAT) of the entry and the verb pattern (VPPAT) in sentence formation. The semantic information (SEM) provides a set of mapping between syntactic and semantic relations. The mapping information is used to identify the semantic relation from the provided syntactic relation when constructing a semantic structure in the analysis process, and vice versa in the generation process. It contains two types of slots of MAPS and AKO. The MAPS is assigned to a verb and adjective, providing syntactic relations among the circumstance words within the same sentence to their thematic roles (case roles) in the deep structure. The AKO (a-kind-of) indicates the location of a word in the conceptual hierarchy. It is considered to be the logical information of the concept to its class. However, these sorts of semantic information in the MMT dictionary are inadequate for representing and discriminating word meanings, due to the lack of other semantic relations. For example, it was not specified which semantic class that an agent (AGT) or object (OBJ) of the verb 'pay' should be.

LEXiTRON is the first Thai-English corpus-based electronic dictionary, which can be accessed via the internet. LEXiTRON dictionary is designed and developed for human use. It is extracted from the MMT dictionary to add more human friendly information, as well as to reduce unnecessary information. As a result, LEXiTRON includes the following information.

1. Word entries that reflect human understanding units. Most word entries are shorter than those in the MMT dictionary.
2. Part-of-speech categories are simplified. Basically, the main categories are selected rather than the sub-categories. There are only 12 categories, i.e. Noun, Pronoun, Verb, Determiner, Classifier, Auxiliary, Conjunction, Ending, Interjection, Preposition, Question word, and Negation.
3. AKO is simplified to capture the group of meaning for ease of searching a word sense.

Figure 2 displays the word entry of the verb จ่าย 'pay'. Note that there is no syntactic and semantic information in the LEXiTRON.

WORD HEADER	จ่าย V pay; spend	% Thai word % Word category % Equivalent English word
SYN	ให้เงิน, จ่ายเงิน	% Synonym
ANT	รับ	% Antonym
DEF	ให้เงินเพื่อวัตถุประสงค์อย่างใดอย่างหนึ่ง	% Definition of Thai word
SAMPLE	ปีนี้บริษัทมีนโยบายจะจ่ายโบนัสเพิ่มขึ้นจากเดิมที่เคยได้ 3 เท่า	% Sample sentence

Fig. 2 The LEXiTRON entry for the verb จ่าย ('pay').

3. The Proposed Lexicon and Its Construction

Although the MMT dictionary and LEXiTRON are very useful for the Thai language processing, pros and cons of both depends on purpose of the design. Some problems occur when we consider combining these two dictionaries.

- Information in these two dictionaries is individually defined for different purpose. Consistency of the information in word entry is crucial.
- The semantic information that is useful for word sense disambiguation (i.e. semantic constraint) is not explicitly defined in both of these two dictionaries, for example, the possible semantic class of the agent (AGT) or the object (OBJ) of the verb 'pay'.

TCL's Computational Lexicon (TCLLEX) adopts the semantic constraints to generate a word frame. Word arguments and semantic groups are defined to indicate the meaning that can be differentiated from others. Based on the designed framework, the word entries are selected from LEXiTRON. The corresponding word information especially the syntax-semantic mapping and AKO are extracted from the MMT dictionary. As a result, 68,860 word entries of 13,781 verbs, 35,688 nouns are extracted and redefined in TCL lexicon.

Since the MMT dictionary includes much more information than LEXiTRON, the word entries taken from LEXiTRON are revised using information from the MMT dictionary. Some information comes from LEXiTRON. For instance, the synonym of LEXiTRON is selected for Equal in TCLLEX, such as "ให้เงิน","จ่ายเงิน", the synonyms of "จ่าย" (to pay). On the other hand, arguments in the MMT dictionary are relied for the selection of semantic constraints. Among 35,192 words in LEXiTRON and 53,759 words in MMT dictionary, there are 22,173 words co-existing in both dictionaries. The information from both dictionaries is selected to compose TCLLEX. The ones that exist in either dictionary will be added. As a result, the union set of both dictionaries will be generated as the TCLLEX. Moreover, TCLLEX relies on the word entries, part-of-speech category and synonym from LEXiTRON while it relies on part-of-speech sub-category, syntax-semantic mapping from the MMT dictionary. TCLLEX is then generated to cover both word syntax and semantic.

The major manual revision is to reduce the duplication of semantic unit. A lot of word semantic units in the MMT dictionary are prepared for case resolution in grammar rules in sentence analysis and generation. For some expressions that are hard for disambiguation, a new semantic unit is prepared with some special information for the selection criteria defined in the rules. In this case, several semantic units are deleted or merged to a closer one. A new semantic class is generated when the number of members is large enough. There is no obvious value for a new class. After revision, a new class is manually generated to balance the concept hierarchy.

However, the details of information in both dictionaries are not equal. We need manual edition for the entries from LEXiTRON in case that they are not properly found in the MMT dictionary. In addition, the information such as the conceptual class for semantic and logical constraints which can be extracted from web corpus by BIC and Agglomerative Merging algorithms needs intensive revision from lexicographers in selecting from the candidates. In case of homonyms, we left them as choices for human evaluators to select.

Table 1 illustrates the list of the entries of TCLLEX and their sources. In the table, '*' specifies the items that will be considered in the future.

The TCLLEX's structure consists of four types of information, including general, morphological, syntactic, and semantic information. The general information of an entry conveys its Thai entry, corresponding English word, entry definition, and example sentence. The morphological information indicates the type of word composition (TYPE), which is of two types: *single word* and *compound* *word*. A single word is a lexical unit that cannot be divided into smaller units, such as โทรศัพท์ (telephone). In contrast, a compound word is a combination of more than one single word. For example, the word โทรศัพท์มือถือ (mobile phone) consists of two lexical units: โทรศัพท์ (telephone), and มือถือ (mobile).

Table 1 The list of TCLLEX entries and their sources.

TCL	MMT	LEXiTRON	Corpus
General Information			
Thai entry	Header	Header	-
Corresponding English	Header	Header	-
Entry Definition	-	Definition	-
Example Sentence	-	Sample	-
Morphological Information			
Word-type	TYPE	-	-
Syntactic Information			
Category	CAT	Header	-
Sub-category	SUBCAT	-	-
Verb Pattern *(for verbs)*	VPPAT	-	-
Semantic Information			
Logical Constraints			
Is-a (ISA)	AKO	-	-
Equal (EQU)	-	Synonym	-
Not-Equal (NEQ)	-	Antonym	-
Part-Of (POF)	-	-	Corpus*
Whole-Of (WOF)	-	-	Corpus*
Semantic Constraints			
Syntactic-semantic Mapping	MAPS	-	-
Agent	-	-	Corpus
Object	-	-	Corpus
Instrument	-	-	Corpus*
Location	-	-	Corpus*
Time	-	-	Corpus*
…	…	…	…

The syntactic information contains information of the entry’s syntactic structure, i.e. grammatical categories (CAT) and subcategories (SUBCAT). In case of a verb, its verb pattern is also included. In TCLLEX, there are 11 categories with 44 subcategories defined. Each category is divided into subcategories according its sub characteristic, such as its occurring position, composition, and reference. For example, the category ‘determiner’ is divided into 9 subcategories by its occurring position, such as *after noun*, *after noun and classifier*, *between noun and classifier*, etc. There are totally 11 verb patterns, classified according to the position of the obligatory arguments and its contextual environment.

The semantic information provides a set of logical and semantic constraints, which is useful for discriminating word senses. The logical constraints can be attached to a word of any category type. They illustrate the logical relationship among word senses in the lexicon. The semantic constraints are attached to a verb or an adjective. They represent the relationship among thematic roles in a verb or adjective pattern.

There are five types of logical constraints proposed in the work. They are ISA (a-kind-of), EQU (synonym), NEQ (antonym), POF (meronym), and WOF (holonym). The ISA constraint indicates a-kind-of relation of words in the semantic hierarchy which is currently composed of 189 conceptual classes. There are 16 kinds of semantic relations, constructed by analyzing a set of Thai sentences by Thai linguists. They are Agent, Object, Patient, Experiencer, Result, Goal, Location, Commutative, Measure, Complement, Cause, Time, Source, Instrument, Manner, and Beneficiary. Syntax-semantic mapping information (MAPS) is a key for extracting the constraint arguments. BIC and Agglomerative Merging algorithms are finally used to provide the candidates of classes for each constraint.

Furthermore, we classify the constraints into two types: obligatory and optional. While the obligatory constraints should be filled as much as possible, the optional constraints can be left empty. In the logical constraints, there is only one obligatory constraint, i.e. ISA. It is also considered to be the core structure of the TCLLEX. In the semantic constraints, AGT and OBJ are the obligatory constraints. The remaining constraints are categorized to be the optional constraints.

Unfortunately, POF and WOF are not included in these dictionaries. They are analogous to meronyms and holonyms. Recently, several approaches [8], [9], [10] have been proposed to learn meronyms and holonyms from a corpus. However, the acquisition of POF and WOF is left as our future work. Finally, the mapping between syntactic and semantic structure (MAPS) can be obtained from MMT dictionary.

4. Constraint Acquisition from Corpora

This section describes a method to acquire semantic constraints automatically from texts on the web. As the first step, only AGT and OBJ, which are obligatory constraints for selectional preferences, are focused in this work.

4.1 Semantic Constraint Acquisition

In this work, we propose a method to identify selectional preferences of a verb, a kind of semantic constraints, by using a large number of documents (or texts) gathered from the Web. By the MAPS in the MMT dictionary, we know that the subject of the verb ‘จ่าย’ (pay) is the agent, but we do not know which semantic class (concept) of the agent should be. Typically, one may think that the subject of the verb ‘จ่าย’ (pay) prefers to be humans. By parsing through text corpora, we can obtain examples of context nouns that are considered to be the subjects of the verb. Furthermore, a method to create a set of semantic constraints from these examples using Bayesian Information Criteria is described.

4.2 Bayesian Information Criteria

In order to obtain an optimal set of selectional preferences for a given verb, a model selection technique called the Bayesian Information Criteria (BIC) [11] is applied. The BIC is a model selection based on Bayesian theory. The problem of model selection is to choose the best model among a set of candidate models $\mathbf{m_i} \in \mathcal{M}$. The BIC of a model $\mathbf{m_i}$ can be approximated as follows:

$$BIC(\mathbf{m_i}) = l_i(D) - \frac{p_i}{2}\log|D| \qquad (1)$$

where $l_i(D)$ is the log-likelihood of the data D according to the model $\mathbf{m_i}$ and p_i is the number of independent parameters. The BIC is independent of the prior and related to the minimum description length (MDL). The details of BIC criterion can be found in [12]. As the probabilistic model for the semantic hierarchy, the tree cut model [13] is adopted.

The tree cut model is introduced to characterize the probabilistic model of the semantic hierarchy. Let $\mathbf{m} = (\Gamma,\Theta)$ be the model, including a partition in the semantic hierarchy Γ, and a set of parameters Θ. Given the noun class $C \in \Gamma$, the verb $v \in V$, and the syntactic relationship $r \in R$, the sum of the conditional probability distribution of $P(C|v,r)$ must be 1 as follows.

$$\sum_{C\in\Gamma} P(C\,|\,v,r) = 1 \qquad (2)$$

Two main assumptions for estimating probabilities in this model are: (1) the probability of a class C can be calculated from all nouns in the class $n \in C$, each of which is estimated using the maximum likelihood estimation (MLE), and (2) after calculating the class probability, the probability of each noun in the class noun $n \in C$ is assumed to be uniform distribution.

$$P(C\,|\,v,r) = \frac{\sum_{n\in C} freq(n\,|\,v,r)}{|D|} \qquad (3)$$

$$P(n\,|\,v,r) = \frac{P(C)}{|C|} \qquad (4)$$

where $freq(n|v,r)$ is the frequency of the noun n co-occurring with the verb v and the syntactic relationship r, $|D|$ is the data size, i.e., the total frequency of all nouns, and $|C|$ is the number of classes in the current partition. Based on this, the log-likelihood of class C according to $\mathbf{m_i}$ is:

$$l_i(D) = \log \prod_{n\in C} P_i(n\,|\,v,r) = \sum_{n\in C} \log P_i(n\,|\,v,r) \qquad (5)$$

$$BIC(\mathbf{m_i}) = \sum_{C\in\Gamma} P_i(n\,|\,v,r) - \frac{p_i}{2}\log|D| \qquad (6)$$

where the number of parameter p_i is equivalent to the number of classes in Γ minus one, i.e. $|C|-1$. P_i is the probability distribution of the model i. Finally, the objective function is defined as follows.

$$\mathbf{m}^* = \arg\max_{\mathbf{m_i}\in M} BIC(\mathbf{m_i}) \qquad (7)$$

4.3 The Agglomerative Merging Algorithm

We now describe an iterative algorithm for selectional preference generalization. Our algorithm searches the appropriate levels of noun classes on the semantic hierarchy by performing agglomerative merging in a bottom-up manner. One may consider the behavior of the algorithm as a simplified agglomerative clustering algorithm. We assume that all nouns are pre-classified onto their hierarchical classes according to the semantic information indicated by AKO. As a result, the algorithm does not have to make any decision about assigning nouns to the most probable classes. What it has to do is to repeatedly merge subclasses into a single class if the structure of the semantic hierarchy improves. We consider this structure as a model for representing selectional preferences. The improvement of the model can be measured by using the BIC as described in the previous section. The more the BIC increases, the more the model improves. The agglomerative merging algorithm tries to increase the objective function value in Equation 7 at every step. Thus, the BIC is used to test the improvement of the model both locally and globally.

Our algorithm starts by initializing the region of noun classes on the semantic hierarchy. The input data are given in the form of the co-occurrence tuple, <*v, r, n, freq*>, where *v* is the verb, *r* is the syntactic relationship, *n* is the noun, and *freq* is the co-occurring frequency. The co-occurrence tuples can be obtained by extracting and analyzing the snippets. It then finds appropriate leaf nodes having the same AKO to merge up into the parent node. Focusing on this partition, the BIC is measured locally. If the BIC score of the parent node is not greater than the BIC score of the children nodes, the algorithm keeps the structure of leaf nodes as it is. Otherwise, the BIC is measured globally to guarantee the overall improvement. These processes are implemented by MERGECLASSES and AGGLOMERATIVEMERGING algorithms shown in Figure 4. The algorithm iterates until it cannot find leaf nodes to merge or there remains one class.

Figure 3 illustrates an example of how the algorithm works, (originally, Li and Abe, 1998) [13]. Given the verb *fly* with the syntactic relationship *subject*, the co-occurring nouns are: crow (2), eagle (2), bird (4), and bee (2), where numbers in the parentheses indicate the co-occurring frequency of nouns. Let us focus on Figure 3a, which is the initial semantic hierarchy of the data. The algorithm starts by finding possible leaf nodes to merge. Since the local BIC score increases, it further measures the global BIC score by comparing the overall structure. The global BIC score also increases, it decides to merge the children nodes into the parent node. Figure 3b performs the same process. In Figure 3c, since the local BIC score decreases, it is not necessary to measure the global BIC score. Finally, we obtain the generalized semantic hierarchy in Figure 3d, whose remaining leaf nodes are considered to be selectional preferences [14].

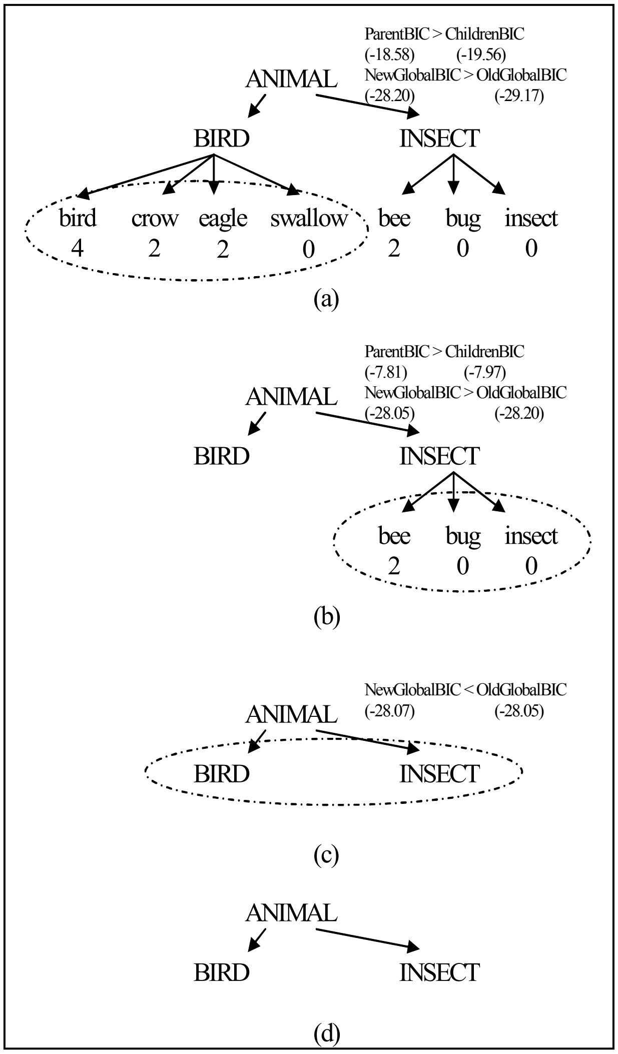

Fig. 3 An example of Agglomerative Merging

```
Algorithm 1: MERGECLASSES($\{c_i\}^k_{i=1}$)
  begin
    $c' \leftarrow \varnothing$ ;
    for $i$ from $i = 1, \ldots, k$ do
    $c' \leftarrow c' \cup c_i$;
    endFor
    if $BIC(c') > BIC(\{c_i\}^k_{i=1})$ then
      return $c'$;
    else
      return $\varnothing$;
    endif
  end
```

```
Algorithm 2: AGGLOMERATIVEMERGING
  input:  Semantic hierarchy $\Gamma$ containing a
          set of initial leaf nodes $c_i$, where
          $i = 1, \ldots, m$.
  output: Generalized $\Gamma$ with leaf nodes
          forming the optimal noun classes.
  begin
    repeat
      Find remaining nodes to merge, $\{c_i\}^k_{i=1}$;
      if k = 0 then
        break;
      endif

      $c'$ = MERGECLASSES($\{c_i\}^k_{i=1}$);
      if $c' \neq \varnothing$ ; then
        $\Psi = \Gamma \setminus \{c_i\}^k_{i=1} \cup c'$;
        if $BIC(\Psi) > BIC(\Gamma)$ then
          Re-distribute $P(n)$ for $n \in c'$
          according to Equation 3;
          DELETE($\Gamma$, $\{c_i\}^k_{i=1}$);
          APPEND($\Gamma$, $c'$);
          $m = m - k + 1$;
        endif
      endif
    until $m < 1$;
  end
```

Fig. 4 MERGECLASSES and AGGLOMERATIVEMERGING algorithms.

5. Experimental Methodology

5.1 Web Data Collection

As mentioned earlier, the Web is viewed as large and free corpora. Below we describe how to retrieve examples for selectional preference generalization through search engines. Common search engines usually return results, including a number of relevant links and their short descriptions. Since our objective is to extract the co-occurrence tuples, what we anticipate from the search engines is that, given a verb as a query, the returned short descriptions may contain the verb and its context. We refer to these short descriptions as *snippets*.

We implemented a simple web robot that sends the target verb to the search engines, and retrieves all the search results kept into a repository. Two major Thai search engines were used, including www.sansarn.com and www.siamguru.com. Then, we parsed HTML documents in the repository to extract only snippets. We obtained about 800-1000 snippets for each verb query. Each snippet contains 100-150 words on average.

The benefits of using the snippets from the search engines are two folds. Firstly, we can use the efficient search mechanism to get the context of the target word without implementing any string-pattern matching algorithms. Secondly, we obtain the large databases of the search engines, reflecting natural language usage in the society. One problem we faced is that the snippets are too heterogeneous. For example, since the descriptions of the web pages were produced from table data containing lists of items or bullets, the snippets did not contain grammatical features and were less meaningful. Consequently, we limited our web robot to crawl particularly on news sites, which are already categorized by both search engines. The search results from the news categories seem to contain more useful phrases having the target verb with its context.

5.2 Extracting Co-occurrence Tuples

Since we need the final input data of the algorithm in the form of the co-occurrence tuple, <*v, r, n, freq*>, linguistic tools for analyzing morphological and syntactic structure of Thai text are required. However, we only have a parts-of-speech tagger called Swath. A syntactic relationship *r* between a target verb *v* and its co-occurring noun *n* is manually assigned. In this section, we describe an approach that assists human subjects to do such task.

After retrieving snippets containing the target verb and its context, word segmentation and parts-of-speech tagging are performed using Swath. Note that Thai text has no explicit word boundaries like English text, so we have to segment it into meaningful tokens. We consider ±3 words of context around the target verb. This window size is enough to capture syntactic relationships. As the result, we obtain a set of tuples in the form of <*v, context relationship, n, freq*>.

We observe that the co-occurring frequencies have small different values. In order to filter out nouns which have insignificant dependence of the target verb, we measure dependence between words by using statistics taken from all the snippets. We apply the log likelihood ratio (LLR) [15] for selecting the most optimal nouns. Given the verb v and the noun n occurring within window z, a fast version of the LLR can be calculated as follows [16].

$$LLR_Z(v,n) = k_{11}\log\frac{k_{11}N}{Q_1R_1} + k_{12}\log\frac{k_{12}N}{Q_1R_2} + k_{21}\log\frac{k_{21}N}{Q_2R_1} + k_{22}\log\frac{k_{22}N}{Q_2R_2},$$

$$k_{11} = freq(v,n),$$

$$k_{12} = freq(v) - k_{11}$$

$$k_{21} = freq(n) - k_{11},$$

$$k_{22} = N - k_{11} - k_{12} - k_{21},$$

$$Q_1 = k_{11} + k_{12}, Q_2 = k_{21} + k_{22},$$

$$R_1 = k_{11} + k_{21}, R_2 = k_{12} + k_{22},$$

where *freq(v,n)* is the co-occurring frequency between *v* and *n*, *freq(v)* and *freq(n)* are frequencies of *v* and *n*, respectively. Only nouns with their LLR values greater than a pre-defined threshold are left. Table 2 shows the top 14 co-occurring nouns within window size +3 for a given verb ตรวจ 'check'. The second and third columns show their co-occurring frequencies and LLR values, respectively. The nouns within the window size -3 are considered in the similar way. Once the candidate nouns are produced, we ask human subjects to analyze and assign the most suitable syntactic relationships between the verb and candidate nouns. For example, from Table 2, we get co-occurrence tuples <ตรวจ 'check', obj, ร่างกาย 'body', 9>, <ตรวจ 'check', obj, หนังสือเดินทาง 'passport', 2>, and so on.

Word	Freq	LLR+3
ร่างกาย 'body'	9	24.6864
หนังสือเดินทาง 'passport'	2	6.4391
กล้ามเนื้อ 'muscle'	1	4.5825
พยาธิ 'worm'	1	4.5825
ป่าไม้ 'forest'	2	4.3856
คฤหาสน์ 'mansion'	2	4.3856
รถบรรทุก 'truck'	2	3.7537
บ้านพัก 'home'	2	2.8196
กระเป๋า 'bag'	2	2.8196
สุขภาพ 'health'	1	2.6056
ผลิตภัณฑ์ 'product'	1	1.4067
รถ 'car'	8	1.3848
โรงงาน 'factory'	1	1.0653
กะโหลก 'skull'	2	0.9742

Table 2 Co-occurring nouns of verb ตรวจ 'check' within window size +3.

5.3 Results and Discussion

Evaluation of selectional preference generalization is a difficult task. To this end, it requires a gold standard for checking the appropriateness of the acquired results. This gold standard can be produced by using the majority of the human agreements. At the present, there is no such gold standard for the Thai language. However, in order to observe the behavior of our algorithm, we selected Thai verbs, including ตรวจ 'check', สร้าง 'build', ซื้อ 'buy', and จ่าย 'pay' for evaluation. We considered two syntactic structures, including subject-verb and verb-direct object relationships. Tables 3 and 4 show some results of generalization.

Class	Prob.	Word Example
Subject of ตรวจ 'check'		
PEOPLE	1.00	ตำรวจ 'police'
Subject of สร้าง 'build'		
ABSTRACT_THING	0.69	สังคม 'society'
ORGANIZATION	0.04	รัฐบาล 'government'
PERSON	0.03	นักท่องเที่ยว 'tourist'
Subject of ซื้อ 'buy'		
PERSON	0.40	ชาวบ้าน 'villager'
CONSTRUCTION	0.35	โรงพยาบาล 'hospital'
ORGANIZATION	0.25	บริษัท 'company'
Subject of จ่าย 'pay'		
PERSON	0.54	นักเรียน 'student'
CONSTRUCTION	0.39	ธนาคาร 'bank'
CULTURAL_AB_THING	0.08	ประธาน 'chairman'

Table 3 Generalization results with subject-verb-relationship.

Class	Prob.	Word Example
Direct Object of ตรวจ 'check'		
ARTIFACT	0.34	รางวัล 'prize'
ABSTRACT_THING	0.22	เอกสาร 'document'
ANIMAL_PART	0.18	ร่างกาย 'body'
Direct Object of สร้าง 'build'		
ABSTRACT_THING	0.65	มาตรการ 'measure'
ARTIFACT	0.16	สะพาน 'bridge'
ATTRIBUTE	0.10	สถานการณ์ 'situation'
Direct Object of ซื้อ 'buy'		
ABSTRACT_THING	0.40	ธุรกิจ 'business'
ARTIFACT	0.27	ของที่ระลึก 'souvenir'
GRAIN	0.03	ข้าว 'rice'
Direct Object of จ่าย 'pay'		
IMMATERIAL_THING	0.31	ค่าเช่า 'rental fee'
SOCIAL_AB_THING	0.22	ค่า 'fee'
RESULT_OF_ACT	0.01	ดอกเบี้ย 'interest'

Table 4 Generalization results with verb-direct object relationship.

From these tables, it is noted that the result matches well with human intuition. For example, the subject of the verb ตรวจ 'check' falls into the class PEOPLE, which its children classes are PERSON and ORGANIZATION. The class ANIMAL_PART can be discovered to be the object of this verb. The computational time is very short, which is less than one second running on a personal computer with Pentium processor 2GHz and memory 512 KB. In addition, we observe that the noun sense ambiguity can lead to irrelevant results in some cases. For example, the noun โรงพยาบาล 'hospital' has two senses, which are categorized into two classes: CONSTRUCTION and ORGANIZATION. However, the class CONSTRUCTION is unlikely to be the subject of the verb ตรวจ 'check'. Since the tree cut model just deals with this problem by equally dividing the frequency of a noun among all the classes containing that noun, more sophisticated approach is needed for further improvement.

6. Conclusion

This paper presented the design of TCLLEX. Toward the construction of this large-scale Thai lexicon, information from the two existing Thai machine-readable dictionaries, called MMT dictionary and LEXiTRON was used. In addition to morphological, syntactic, semantic case role and logical information in the existing dictionaries, selectional preferences, a kind of semantic constraints, are automatically acquired by analyzing Thai texts on the web and then added into the lexicon. The Bayesian Information Criterion (BIC) is applied as the measure in a tree cut model to obtain the selectional preferences. In future work, we plan to explore approaches to extract other logical and semantic constraints. Additionally, we further study how the computational operations among word senses can be performed using the constructed lexicon.

Acknowledgments

This study is supported by Thailand Graduate Institute of Science and Technology (TGIST) of National Science and Technology Development Agency (NSTDA) under the grant number TGIST-TG-B-11-44-13-418D. The authors would like to thank Dr. Hitoshi Isahara, the leader of computational linguistic group of National Institute of Communications and Technology, Japan, for his kind support throughout this research. We also would like to thank the anonymous reviewers for giving a lot of useful comments.

References

[1] Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K., "Introduction to WordNet: An on-line lexical database", CSL Report 43, 1993.

[2] Baker, Collin F., Fillmore, Charles J., and Lowe, John B., “The Berkeley FrameNet project”, in Proceedings of the COLING-ACL, Montreal, Canada, 1998.

[3] EDR, “EDR Electronic Dictionary Technical Guide”, Japan Electronic Dictionary Research Institute, Ltd., 1990.

[4] Dong, Zhendong, and Dong, Qiang, “Hownet” [online], Available at http://www.keenage.com/zhiwang/e_zhiwang.html

[5] http://www.ipa.go.jp/STC/NIHONGO/IPAL/ipal.html

[6] Center of the International Cooperation for Computerization (CICC), “Thai Basic Dictionary: Technical Report”, 1995.

[7] Lexitron, Available at http://lexitron.nectec.or.th

[8] Berland, M. and Charniak, E., “Finding parts in very large corpora”, In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics. New Brunswick NJ, 1999.

[9] Sundblad, H., “Acquisition of hyponyms and meronyms from question corpora”, In Proceedings from the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering, Lyon, France, 2002.

[10] Girju, R., Badulescu, A., and Moldovan, D., “Learning semantic constraints for the automatic discovery of part-whole relations”, In Proceedings of the Human Language Technology Conference, Edmonton, Canada, 2003.

[11] Wasserman, L., “Bayesian model selection and model averaging”, Journal of Mathematical Psychology, 1999.

[12] Chickering, D.M., and Heckerman, D., “Efficient approximations for the marginal likelihood of Bayesian networks with hidden variables”, Machine Learning, 29, pages 181-212, 1997.

[13] Li, H., and Abe, N., “Generalizing case frames using a thesaurus and the MDL principle”, Computational Linguistics, 24(2): 217—244, 1998.

[14] Kruengkrai, C., Charoenporn, T., Sornlertlamvanich, V., Isahara,H., “Acquiring selectional preferences in a thai lexical database” In Proceedings of the 1st Joint Conference on Natural Language Processing (IJCNLP-04), China, 2004

[15] Dunning, T., “Accurate methods for the statistics of surprise and coincidence”, Computational Linguistics, 19(1): 61-74, 1994.

[16] Tanaka, T., “Measuring the similarity between compound nouns in different languages using non-parallel corpora”, In Proceedings of the 19th International Conference on Computational Linguistics, 2002.