A State of the Art of Thai Language Resources
and
Thai Language Behavior Analysis and Modeling

Asanee Kawtrakul, Mukda Suktarachan, Patcharee Varasai, Hutchatai Chanlekha
Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand 10900.
E-mail: ak, mukda, pom, aim@vivaldi.cpe.ku.ac.th,

Abstract

As electronic communications is now increasing, the term Natural Language Processing should be considered in the broader aspect of Multi-Language processing system. Observation of the language behavior will provide a good basis for design of computational language model and also creating cost-effective solutions to the practical problems. In order to have a good language modeling, the language resources are necessary for the language behavior analysis.

This paper intended to express what we have and what we have done by the desire to make a bridge between the languages and to share and make maximal use of the existing lexica, corpus and the tools. Three main topics are, then, focussed: A State of the Art of Thai language Resources, Thai language behaviors and their computational models.

1. Introduction

As electronic communications are now increasing, the term Natural Language Processing should be considered in the broader aspect of Multi-Language processing system. An important phase in the system development process is requirement engineering, which can define as the process of analyzing the problems in a certain language. An essential part of the requirement engineering phase is computational language modeling which is an abstract representation of the behavior of the language. In order to have a good language model for creating cost-effective solutions to the practical problems, the language resources are necessary for the language behavior analysis.

This paper intended to express what we have and what we have done by the desire to make a bridge between the languages and to share and make maximal use of the existing lexica, corpus and the tools. Three main topics are, then, focussed:

• A State of the Art of Thai language Resources that will give an overview of what we have in Corpus, Lexicon and tools for corpus processing and analysis.

• Thai language behaviors (only in word and phrase level) analyzed from the varieties of corpus which consist of Lexicon growth, New word formation and Phrase/Sentence construction, and

• The computational models providing for those behaviors which consist of Unknown Word Extraction and Name Entities identification, New word generation and Noun phrase recognition.

The remainder of the paper is organized as follows. In section 2, we give the gateway of Thai language resources. Thai Language behaviors are discussed in section 3. In section 4, then, provides Thai Language Computational Modeling as a basis for creating cost-effective solutions to those practical problems.

2. A State of the Art of Thai Language Resource

This section gives a survey of a state of the art of Thai Language Resources consisting of Corpus, Lexicon and Tools. Here, we will present only the resources that open for public access.
2.1 Corpus

The existing Thai corpus is divided into 2 types; speech and text corpus developed by many Thai Universities. Thai Language Audio Resource Center of Thammasart University (ThaiARC) (http://thaiarc.ac.th) developed speech corpus aimed to provide digitized audio information for dissemination via Internet. The project pioneers the production and collection of various types of audio information and various styles of Thai speech, such as royal speeches, academic lectures, oral literature, etc.

For Text corpus, originally, the goal of the corpus collecting is used only inside the laboratory. Until 1996, National Electronics and Computer Technology Center (NECTEC) and Communications Research Laboratory (CRL) had a collaboration project with the purpose of preparing Thai language corpus from technical proceedings for language study and application research. It named ORCHID corpus (NECTEC, 1997). NAiST Corpus began in 1996 with the primary aim of collecting document from magazines for training and testing program in Written Production Assistance (Asanee, 1995). The existing corpus can be summarized as shown in Table 1.

| List | Corpus       | Type                        | Amount       | Status |
|------|--------------|-----------------------------|--------------|--------|
| NECTEC | Orchid Corpus | POS-Tagged Text             | 2,560,000 words | Online |
| Kasetsart Univ. | NAiST Corpus | Text                        | 60,511,974 words | Online |
| Thammasart Univ. | Thai ARC     | Digitized audio             | 4000 words++ | online |

Table 1: The List of Thai Corpus

2.2 Lexicon

There is a number of Thai lexicon which has been developed as shown in Table 2.

| Dictionary          | Type | Size (word) | status   | Web site                                           |
|---------------------|------|-------------|----------|----------------------------------------------------|
| Royal Institute Dictionary | Mono | 33,582      | Online   | http://rir3.royin.go.th/rtThdict/loo_kup.html     |
| Lexitron            | Bi   | 50,000      | Online   | http://www.links.nectec.or.th/lexit/lex_t.html    |
| NAiST Lexibase      | Mono | 15,000      | Online   | http://beethoven.cpe.ku.ac.th/                    |
| So Sethaputra       | Bi   | -48,000 Eng words -38,000 Thai words | Online | http://www.thaisoftware.co.th/                     |

From the table 2, Only Lexitron (from NECTEC) and NAiST Lexibase (from Kasetsart University) that were applied to NLP. NAiST Lexibase has been developed based on relational model for managing and maintaining easily in the future. It contains 15,000 word list with their syntax and the semantic concept information in the concept code.

2.3 Corpus and Language Analysis Tools

Corpus is not only the resource of Linguistic Knowledge but is used for training, improving and evaluating the NLP systems. The tools for corpus manipulation and knowledge acquisition become necessary.

NAiST Lab. has developed the toolkit for sharing via the Internet. It has been designed for corpus collecting, annotating, maintaining and analyzing. Additionally, it has been designed as the engine, which the end user could use with their data. (See a service on http://naist.cpe.ku.ac.th).
words but not being proper names or colloquial words. The unknown or new words occur much in the real document such as Proper names, Colloquial words, Abbreviations, and Foreign words.

The lexicon growth is observed from corpus size, 400,000, 2,154,700 and 60,511,974 words from Newspaper, Magazine and Agriculture text. We found that common word lists increased from 111,954 to 839,522 and 49,136,408 words according to the corpus size, while the unknown word lists increased from 288,046 to 1,315,178 and 11,375,566 words respectively as shown in table3.

Table 3: Lexicon-growth

| Size of Corpus/ words | Common words | Unknown words |
|-----------------------|--------------|---------------|
| 400,000               | 111,954      | 288,046       |
| 2,154,700             | 839,522      | 1,315,178     |
| 60,511,974            | 49,136,408   | 11,375,566    |

Regarding to 60,511,974 words corpus in the table 3, it composes of 35,127,012 words from Newspaper, 18,359,724 words from Magazine and 7,025,238 words from Agricultural Text. Unknown words occur in each category as shown in table 4.

Table 4: The Categories of Unknown words according to the various corpus genres

| Types of unknown word | Newspaper (words) | Magazine (words) | Agricultural Text (words) |
|-----------------------|-------------------|-----------------|---------------------------|
| Proper name           | 4,809,160         | 1,272,747       | 1,170,076                 |
| Spoken words          | 58,335            | 8,787           | 0                         |
| Abbreviation          | 70,109            | 43,056          | 0                         |
| Foreign words         | 304,519           | 239,107         | 3,399,670                 |
| Total                 | 5,242,125         | 1,563,697       | 4,569,746                 |

According to table 3 and 4, we could observe that not only unknown words increase but common words also increase and the main categories of increasing unknown word are proper names and foreign words. Consequently, a computational model of unknown word extraction and name entity identification has been developed and also of new word construction.

3.2 New Word Formation and Core Noun

Regarding to the growth of common word shown in table 3, we studied how the new words come from.

3.2.1 Basic Information about Thai

Thai words are multi-syllabic words which stringing together could form a new word. Since Thai has no inflection and no word delimiters, Thai morphological processing is mainly to recognize word boundaries instead of recognizing a lexical form from a surface form as in English. Let $C$ be a sequence of characters

$$C = c_1c_2c_3...c_n : n > = 1$$

Let $W$ be a sequence of words

$$W = w_1w_2w_3...w_m : m > = 1$$

Where $w_i = c_1i_1c_2i_2...c_ri_r : i_r > = 1, r > = 2$

Since Thai sentences are formed with a sequence of words with a stream of characters, i.e., $c_1c_2c_3...c_n$ mostly without explicit delimiters, the word boundary in “$c_1c_2c_3c_4c_5$” pattern as shown below could have two ambiguous forms. One is “$c_1c_2$” and “$c_3c_4c_5$”. The other one is “$c_1c_2c_3$” and “$c_4c_5$” (Kawtrakul, 1997)

![Figure 1: Word Boundary Ambiguity](image)

From the figure 1, if characters were grouped differently, the meaning of words will be changed too. For example, “$นําผู้$” can be grouped to “$นํา-ผู้$” (fold one's arms across the chest)” and “$นํา-ผู้$ (a clump of flower)”. From our corpus, we found that the sentence with 45 characters has 30 combinations of words sequence.

3.2.2 New word construction

Almost all-Thai new words are formed by means of compounding and nominalization, by using a set of prefixes.

3.2.2.1 Nominalization

Nominalization is a process by which a word can be formed as a noun by using prefixes added. Noun words formed by using prefixes “$การ$” and “$ความ$” are nouns which signal state or action. Words formed by using prefixes “$ผู$” “$ชาว$” “$นัก$” are nouns which signal human or profession. Prefix “$การ$” “$ความ$” are used in the process of forming a noun from verb or verb
phrase and sometimes from noun (Nominalization). นัก(ka:n) that co-occur with noun, represents the meaning about duty or function of noun it relates to. นัก(ka:n) that co-occur with verbs, always occur with action verbs. ชาว(ha:w) always co-occur with state verbs.

Prefix “ภ(ปʰ:u:) ““ภ(nak)” and “ภ(ตʰ:a:w)” are used in the process of new word formation. ภ(ปʰ:u:) and ภ(nak) co-occur with verb phrase. ภ(nak) sometimes can occur with a few fields of nouns, such as sport and music. So at the first time we kept words, which constructed from prefix “ภ(nak)” plus noun in the lexicon for solving the problem. Prefix “ภ(ตʰ:a:w)” can co-occur with noun only.

3.2.2.2 Compounding

Thai new words can, also, be combined to form compound nouns and are invented almost daily. They normally have at least two parts. The first part represents a pointed object or person such as นัก(man), นัก(pot), นัก(tail), นัก(plant). The second part identifies what kind of object or person it is, or what its purpose is like ขับ(drive a car), หุง(cooked), หาง(tiger), น้ำ(water). Table 5 shows the examples of compound noun in Thai.

Table 5: The Examples of Thai Compound Noun

| What or who | What type / what purpose |
|-------------|--------------------------|
| นัก(man)   | ขับ(drive a car), |
| นัก(pot)   | หุง(cooked), |
| นัก(tail)  | หาง(tiger), |
| นัก(plant) | น้ำ(water) |

Table 6 shows the patterns of compound noun.

Table 6: Compound noun pattern

| Compound noun structure | Examples | Meaning |
|-------------------------|----------|---------|
| n + n                   | นัก(plant) น้ำ(water) | Water Plant |
| n + v                   | นัก(room) นัก(sleep) | Bedroom |
| n + v                   | นัก(room) นัก(sleep) | Bedroom |
| n + v + n               | นัก(man) นัก(drive) นัก(car) | Driver |
| n + v + n               | นัก(child) นัก(hair) นัก(long) | A Long Hair Child |
| n + v + n               | นัก(human) นัก(leg) นัก(lame) | A Lame Man |
| n + n + n               | นัก(home) นัก(shape) นัก(THAI) | Thai Style House |
| n + v + v               | นัก(leaf) นัก(drive) นัก(ride) | Driving License |

From Table 6, it has shown that some compound nouns maintain some parts of those primitive word meanings but some changed to a new meaning. In this paper, we are only interested in compound noun grouping from primitive words which were changed the meaning to more abstract but still maintain some parts of those primitive word meanings, e.g. “นัก(brand) นัก(brand) นัก(cooker)” etc. The word “นัก” maintains its meaning which has a concept of human, but when it was compounded with “ขับ(car)” and “หุง(kitchen)”, their meanings have changed to the occupation by the word relation in the equivalent level. In case of compound noun that change a whole meaning such as “นัก(cooker)”, it will be kept in the lexicon.

3.2.2.3 Compound noun extraction problems

There are three non-trivial problems
- Compound Noun VS Sentence Distinction
- Compound Noun Boundary Ambiguity
- Core noun Detection

Compound Noun VS Sentence

Several NP structures have the same pattern as sentences. Since Thai language is flexible and has no word derivation, including to preposition in compound noun can be omitted, etc. This causes a compound noun having the same pattern as sentence. Thus, Thai NP analysis in IR system is more difficult than English. (See Figure 2)

**Table: Compound Noun**

| Sentence | นักกินผัก (birds eat fruit) |
|----------|-----------------------------|
| In Thai: | นักกินผัก (birds eat fruit) |
| Syntactic: | cn tv cn |
| Category: | Compound Noun |

**Figure 2: The comparison of noun phrase and sentence structure**

In figure 2, compound noun “นักกินผัก (birds eat fruit) (a dining table)” actually omit the preposition “สำหรับ (for)”, which is a relation that point to the purpose of the first noun “นัก(brand) table”.

The Compound Noun Boundary Ambiguity

After we have extracted noun phrase aiming for enhancing the IR system, we have to segment
that noun phrase into sub noun phrase or compound noun in order to specify the core noun as index and its modifier as subindex. For example, compound noun with “noun + noun + verb” structure:  พระ fingertip (child/N)มือ(hair/N)ตรี (long/V) etc. In this case, the second noun and verb have to be grouped firstly since it behaves similarly to a modifier by omitting the relative pronoun that represents its purpose, i.e., “who has”.

Another case of Compound Noun Boundary Ambiguity is word combination. Consider the sequence of words as the example of NP that composes of four words as follows:

NP = N1N2N3N4

There are 8 word combinations of compound noun as shown in figure 3.

\[
\begin{align*}
N_1 + N_2 + N_3 + N_4 \\
N_1 + N_3 + N_4 \\
N_1 + N_4 \\
N_2 + N_3 + N_4 \\
N_2 + N_4 \\
N_3 + N_4 \\
N_1 + N_2 + N_3 + N_4 \\
N_1 + N_2 + N_3 \\
N_1 + N_2 + N_4 \\
N_1 + N_3 + N_4 \\
N_2 + N_3 + N_4 \\
\end{align*}
\]

**Table 7**: The examples of core noun in NP

| Noun phrase(NP) | Core noun |
|------------------|-----------|
| W1 W2 โครงสร้ า + ประโยค | W1 located at the initial position |
| W1 W2 ข้อ + วิธี | W2 located at the final position |
| W1 W2 ป่าม + ต้นไม้ | W3 located at the second position |
| W1 W2 W3 ผล + มะละกอ + ดิบ | W2 located at the second position |

As mentioned above, the models of New Word Generation and Noun Phrase Recognition become one of the interesting works in Thai processing.

### 3.3 Phrase and Sentence Construction

Next, we will indicate the main problems that influence to MT, IE and IR system. These are constituent movement, zero anaphora and iterative relative clause.

#### 3.3.1 Constituent Movement

Constituent is the relationship between lexicon units, which are parts of a larger unit. Constituency is usually shown by a tree diagram or by square brackets:

Ex. [[การประชุมคณะกรรมการ] [อย่าง่ายเยี่ย]]

[[meeting committee] [very smoothly]].

Constituent acts as a chunk that can be moved together and it often occurs in Thai language (see Fig. 4). The constituents can be moved to the front, the middle or the end of the sentence.

#### Figure 3: Patterns of noun phrase analysis

In figure 3, word string has to be grouped correctly for the correct meaning.

The ambiguity of noun phrase boundary has also directly effected the efficiency of text retrieval.

### Core Noun detection

Due to the Information Retrieval, a head or core of noun phrase detection is necessary. In this paper, core noun refers to the most important and specific word that the information retrieval and extraction can directly retrieve or extract without over generating candidate words. However, by the observation, the core of noun phrase needs not to be the initial words. Some of them are at the final position and some have word relation in the equivalent level (As shown in Table 7).

**Table 8**: Categories of anaphora

| Reference                  | - Personal Reference, Demonstrative Reference, Comparative Reference |
|----------------------------|---------------------------------------------------------------------|
| Substitution               | - Nominal Substitution, Verbal Substitution, Causal Substitution    |
| Ellipsis                   | - Nominal Ellipsis, Verbal Ellipsis, Causal Ellipsis                |
| Conjunction                | - Additive, Adversative, Casual, Temporal                           |
| Lexical Cohesion           | - Reiteration(Repetition, Synonym or near Synonym, Superordinate, General word) - Collocation |
Observing from the corpus in: news, magazine and agricultural text, there are 4 types of anaphora. Ellipsis or zero anaphora was found most frequently in Thai documents and other anaphora happened as shown in table 9.

Table 9: Types of reference

| Type of Anaphora          | Magazine | News | Agriculture |
|---------------------------|----------|------|-------------|
| Zero anaphora             | 49.88%   | 52.38% | 50.04%      |
| Repetition                | 32.04%   | 27.78% | 34.49%      |
| Personal reference        | 12.18%   | 12.70% | 1.87%       |
| Nominal substitution      | 5.90%    | 6.08%  | 13.60%      |

Zero anaphora is the use of a gap, in a phrase or clause that has an anaphoric function similar to a pro-form. It is often described as “referring back” to an expression that supplies the information necessary for interpreting the gap.

The following is a sentence that illustrates zero anaphora:

มีถนนสองสายที่ต้องไปตรงแต่แคบและกว้างแต่คดเคี้ยว

There are two roads to eternity, straight but narrow, and broad but crooked.

In this sentence, the gaps in straight but narrow [gap], and broad but crooked [gap] have a zero anaphoric relationship to two roads to eternity.

Table 10 also shows the occurrence of zero anaphora in various parts of a sentence.

Table 10: Position of reference in sentences

| Position               | Frequency |
|------------------------|-----------|
| Subject                | 49.88%    |
| Object                 | 32.04%    |
| Possessive Pronoun     | 12.18%    |
| Following a Preposition | 5.90%    |

It is noticeable that zero anaphora in the position of the subject occurs with high frequency (49.88%). It shows that in Thai language, the position of subject is the most commonly replaced.

3.3.3 Iterative Relative Clause

Thai relative pronouns “ที่” (thi)  “ซึ่ง” (sung) and “อัน” usually refers to things, but it can also refer to event in general.

The relative pronoun is sometimes omitted because it makes the sentence more efficient and elegant.

The book that you ordered from that shop arrived two days later.

Sometimes relative pronoun refers to an event that takes place repeatedly in a phrase.

Although a sentence, which has several clauses inside, will be grammatical, but it is not a good style in writing and always causes a problem for parser and noun phrase recognition.

4. The Computational Model

The computational models in word and phrase level are developed according to the phenomena mentioned in section 3.

4.1 Unknown Word Extraction

Unknown word extraction model composes of 2 submodules: unknown word recognition and name entity identification.

4.1.1 Unknown word recognition

The hybrid model approach has been used for unknown word recognition. The approach is the combination of a statistical model and a set of context based rules. A statistical model is used to identify unknown word’s boundary. The set of context based rules, then, will be used to extract the unknown word’s semantic concept. If the unknown word has no context, a set of unknown word information, which has defined through corpus analysis, will be generated and the best one will be selected, as its semantic concept, by using the semantic tagging model. Unknown word recognition process is shown in figure 6.
4.1.2 Name Entity Identification

After unknown words have been extracted, Named Entity (NE) Identification will define the category of unknown word. The model based on heuristic rules and mutual information. Mutual information or statistical analysis of word collocation is used to solve boundary ambiguity when names were composed with known and unknown words. We use Knowledge based such as list of known name (such as country names), clue word list (such as person’s title) to support the heuristic rules. Using clue word or common noun that precedes the name can specify NE categorization. Based on the case grammar, NE categories can also defined. Moreover, the lists of the names from predefined NE Ontology can be used for predicting category too. The overview of our system is shown in figure 7. More detail see (Chanlekha, H. et al, 2002)

4.2 New Word Generation

Word formation is proposed to reduce the lexicon size by constructing new words or compound noun from the existing words. Based on word formation rules and common dictionary, the shallow parser will extract, a set of candidate compound nouns. Then probabilistic approach based on syntactic structure and statistical data is used to solve the problem of over- and under-generation of new word construction and prune the erroneous of compound noun from the candidate set. The process of new word construction is shown in figure 8. See more detail in (Pengphon, N. et al, 2002)

4.3 Noun Phrase Recognition

Entities or concepts are usually described by noun phrases. This indicates that text chunks like noun phrases play an important role in human language processing. In order to analyze NP, both statistical and linguistic data are used. The model of NP analysis system is shown in figure 9. More detail see (Pengphon, N. et al, 2002)

The first step is morphological analysis for
word segmentation and POS tagging. At the second step, the compound word is grouped into one word by using word formation module (see 4.2). The third step, statistical-based technique is used to identify phrase boundary. This step was provided for identifying the phrase boundary by using NP rules. Next step is Noun Phrase Segmentation. The condition of noun phrase segmentation is shown in figure 10.

**Figure 10** Noun phrase Segmentation

After noun phrase is correctly detected, the relation in noun phrases will be extracted. There are 2 types of relation: head-head noun phrase and head-modifier noun phrase. The process is based on statistical techniques by considering the frequency \( f_i \) of each word \( w_i \) in the document (See figure 11).

**Figure 11** Noun phrase relation

5. Conclusion

The computational language models for Thai in word and phrase level, consisting of Unknown Word Extraction and Name Entities identification, New word generation and Noun phrase recognition, are studied on the basis of their behavior analysis from the varieties of corpus. We expected that it could create cost-effective solutions to the practical problems in the application developments especially in Thai Information Retrieval and Information extraction system. We also give the gateway to access Thai language resources with hoping that it could be the bridge of the international collaboration for developing Multi-Language Processing applications.

Acknowledgement

This project has been supported by NECTEC.

Reference

[1] Bourigault, D. “Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases”. Proc. COLING 1992, 1992.
[2] Chen Kuang-hua and Chen Hsin-His, “Extracting Noun Phrases from large-scale Texts: A Hybrid Approach and Its Automatic Evaluation”, Proc. of the 32nd ACL Annual Meeting, 1994.
[3] Chanlekha, H. et al, “ Statistical and Heuristic Rule Based Model for Thai Named Entity Recognition”, Proc. of SNLP 2002, 2002.
[4] G. Salton, “Automatic Text Processing. The Transformation, Analysis, and Retrieval of Information by Computer”, Singapore: Addison-Wesley Publishing Company, 1989.
[5] Halliday, M.A.K and Hasan, Rugaiya. “Cohesion in English”. Longman Group, London, 1976.
[6] Kawtrakul, A. et al., “Automatic Thai Unknown Word Recognition”, Proc.of the Natural Language Processing Pacific Rim Symposium, Phuket, 1997.
[7] Kawtrakul, A. et al., “Backward Transliteration for Thai Document Retrieval”, Proc.of The1998 IEEE Asia-Pacific Conference on Circuits and Systems, Chiangmai, 1998.
[8] Kawtrakul, A. et.al., “Toward Automatic Multilevel Indexing for Thai Text retrieval System”, In Proceedings of The 1998 IEEE Asia-Pacific Conference on Circuits and Systems, Chiangmai, 1998.
[9] Kawtrakul, A. et.al., “A Lexibase Model for Writing Production Assistant System” Proc.SNLP'95, 1995.
[10] Kawtrakul, A. “Anaphora Resolution Based On Context Model Approach In Database-Oriented Discourse”. A Doctoral Thesis to The Department of Information Engineering, School of Engineering, Nagoya University, Japan, 1991.
[11] Pengphon, N. et al, “Word Formation Approach and Noun Phrase Analysis for Thai” , Proc. of SNLP 2002, 2002.
[12] Sornlertlamvanich, V. et.al., “ORCHID: THAI Part of Speech Tagged Corpus. Technical Report of NECTEC, 1997.
[13] WEBSITE : http://thaiarc.ku.ac.th