Pattern-based Extraction of Named Entities in Thai News Documents

Nattapong Tongtep[1] and Thanaruk Theeramunkong[2]

Sirindhorn International Institute of Technology, Thammasat University
{nattapong[1], thanaruk[2]}@siit.tu.ac.th

Abstract

Named entity extraction is a nontrivial and challenging task for information extraction in Thai language since a Thai text has no word, phrase and sentence boundary. This paper proposes a pattern-based method to extract Thai named entities, such as person name, organization name, location, date and time, as well as action phrases from a text, without assistance of word segmentation and part-of-speech tagging. The experimental results show that the proposed method can detect named entities with approximately 68-100% correctness, using a large-scale Thai dictionary and a set of predefined pattern matching templates.

Keywords: Named Entity, Information Extraction, Pattern Classification

1. Introduction

Information extraction (IE) is a basic task to obtain valuable information from an unstructured text. As a main process in IE, named entity (NE) extraction/ classification detects a chunk of words which specifies a unique existence, such as a person, a time period and a location, and then assigns a type to such detected NE. Each NE usually specifies a main point in the text, which often relates to 4W (who, what, when and where) questions. Problems related to NE include NE boundary disambiguation, clueless NE detection, misspelled NE detection, and NE tagging (classification). So far, a large number of research works have been done to detect NEs in various languages. Part of the early NE research on English texts, a method to extract and recognize company names, was proposed by Rau in 1991 [1]. The method relied on heuristics with manually constructed rules.

Later, from 1991 to present, a series of Message Understanding Conferences (MUC) have attracted a lot of researchers to invent plentiful novel methods in extracting information, especially NEs, from a free text [2; 3; 4; 5]. NEs can be extracted from not only general documents but also web documents via search queries [6]. Besides English, there have been a number of works related to NE extraction in several languages, including Japanese [7; 8; 9; 10], Chinese [11; 12; 13; 14], Vietnamese [15], Hindi [16] and Korean [17].

NE is likely to appear synchronously in several news articles, whereas a common noun is less likely. It was reported that [22; 23] NEs can be recognized with 90% accuracy just by comparing time series distributions of a word in two newspapers, but the recall was not sufficient yet. One of the difficulties in extracting NEs in Thai language is that a Thai text has no explicit boundary for words, phrases or even sentences [18; 19; 20; 21]. Spaces are occasionally inserted between words or phrases within sentences, but there is no standard rule for using spaces. In Thai, a paragraph often contains chunks of phrases that do not together constitute sentences grammatically. Sometimes, the main subject, verb, or object can be omitted from a sentence or their positions, and it is still considered valid. The high language-structure ambiguity seriously hinders any automated process on Thai language. For Thai NE extraction, a number of works [22; 23] used clue words, spaces and discourse context to discover proper names in Thai, and then to assign a part-of-speech (POS) tag to each detected proper name. Some recent works [24; 25; 26] have studied compound nouns in news documents where compound nouns tend to be NEs (proper names). In [27], performance of Thai NE extraction using maximum entropy models with simple heuristic information to extract names of person, organization, and location was reported to be in the range of 75-90% F-measure.

In this paper, we propose a pattern-based approach for NE extraction, especially a person, a time period, a location and an event from Thai news documents. Techniques of longest word matching, longest pattern matching and pattern categorization are applied to construct a set of patterns in the form of regular expressions for retrieving our focused NEs. Moreover, we have introduced a concept of Thai Character Clusters (TCCs) [28] and stop words/phrases into pattern construction. By experiments, we also investigate a suitable sequence of applying regular expressions. A number of experiments are done on various types of news. The organization of the paper is as follows. Sections 2-3 describe the types of NEs including action phrases. Our pattern-based extraction is presented in Section 4. In Section 5, the experimental result and error analysis are discussed. A conclusion is given in Section 6.

2. Types of Named Entities

This section describes the types of NEs we set as targets of extraction from Thai news documents. They are date, time, person name and location. In our approach, each type of NEs can be viewed as a chunk of information.

2.1 Temporal Expression

Like most languages, Thai date and time expressions are a combination of characters, symbols and numbers. Thai numbers are of two types of characters; Roman and Thai. Yingseree et al. [29] classified time expressions into two types: absolute and relative time expressions. The absolute time expressions specifies a time point or time period that is unique in any situation while the relative time expression represents a time point/period that depends on the situation, such as geographical location and the previous time point. Absolute time expressions include the expressions of day of the month, day in week, month, year, era, time unit, and zodiac. The relative time expressions involve the expression of season and a preposition, such as “before”, “after”, and “two days before”. Some detailed date/time subtypes are:

Day of the month: A day of the month represented by an Arabic number (1-31), a Thai number (๑-๓๑) or a day of the month in words (หนึ่ง – สามสิบเอ็ด).

Day in Week: A day-in-week is represented by its full name (จันทร์ – อาทิตย์) (Monday - Sunday) or its abbreviation (จ. - อา.) (Mon. - Sun.).

Month: A month is represented by its full name (มกราคม – ธันวาคม) (January - December), its abbreviation (ม.ค. - ธ.ค.) (Jan. - Dec.), an Arabic number (1-12) or a Thai number (๑-๑๒).

Year: A year is represented by an Arabic numbers, a Thai number or a year in words (สองพันแปด) (two thousand eight).

Year in Chinese Zodiac: A 12-year cycle in Chinese Zodiac is represented by names of auspicious animals.

Era: Buddhist era (พุทธศักราช), Christian era (คริสต์ศักราช) and so on.

Time Unit: A unit is used for representing time such as seconds (วินาที).

Season: A name is represented a season such as summer (ฤดูร้อน).

Western Zodiac: An annual cycle of twelve stations along the ecliptic such as Capricorn (ราศีมังกร) and Aquarius (ราศีกุมภ์).

The full list of the names is shown in the Appendix.

2.2 Location Expression

Location expressions in NE extraction from news documents are varied in the dimensions of relativeness vs. absoluteness, and political vs. organizational vs. natural. In this work, we classify a location expression into four types as follows.

Organizational: A location where people work together under certain regulations such as institute, company, school, university, office, society, and club.

Political: A location defined by managerial or political reasons such as district, city, village, town, province, country, kingdom, and continent.

Natural: A geographical location pointing to either natural landscapes such as river, canal, and field, or artificial landscape such as road, lane, artificial canal, bridge, and buildings.

Relative: A location referred relatively using a prepositional marker such as "in a car," "on the table," and "opposite of the building".

Like English, a location expression in Thai language may be led by a preposition or a marker. Moreover, a series of location expressions may represent a certain location with increasing details.

2.3 Person Name Expression

A person name is a type of proper name that specifies a person. While a first name may precede a last name in some languages, it may be reversed in other languages. Like English, a person name may be led by a title, followed by last name and first name. Besides this, a foreigner's name can be represented in Thai by its spelling [30]. In several cases, a person name in Thai can be recognized by applying a set of patterns on a running Thai text. In this work, the following pattern is usually found in a news document and can be used to recognize a person name. In the pattern, (…) means optional, […] specifies mandatory, and | indicates exclusive disjunction.

(TITLE)[(FN)(MN)(LN) | (NN)]

Here, (TITLE) is a title of a person. In terms of computation, the title acts as a clue word for detecting a person name. FN, MN and LN denote a string presenting a first name, a middle name, and a last name, respectively. NN stands for a nickname. For example, “นายจอห์น” (Mr. John) can be analyzed as “(TITLE:นาย)(FN:จอห์น)”, “ดร.ซูซาน แครกเกอร์ ” (Dr. Susan Cracker) as “(TITLE:ดร.)(FN:ซูซาน)(LN:แครกเกอร์)”, “นายนิค” (Mr. Nick) as “(TITLE:นาย)(NN:นิค)”.

Table 1 Type of verbs in Thai

Types	Examples
Intransitive verb	น้อง*นอน* (My brother *sleeps*.)
Transitive verb	ฉัน*กิน*ข้าว (I *eat* rice.)
To-Be verb	เขา*เป็น*นักเรียน (He *is* a student.)
Auxiliary verb	นายดำ*จะ*ไปโรงเรียน (Dum *will* go to school.)
Infinitive verb	ฉันชอบไป*เที่ยว*กับเธอ (I like to *travel* with you.)

3. Action Phrases and Clue Phrases

An action expression refers to a verb, a series of verbs, or a verb phrase that presents an activity. It plays a major role in event extraction from a news document. Although an action expression is not counted as a type of NE, it is worth extracting it from a document, especially news documents, in order to obtain information related to an event. Table 1 displays the five common types of Thai verbs; i.e. transitive, intransitive, to-be, auxiliary, and infinitive verbs.

However, due to the fact that Thai language has no word boundary, detecting a verb, a series of verbs or a verb phrase from a news document may not be straight-forward. Especially, a string whose spelling is equivalent to a short verbal word in a dictionary may not be such a verbal word but just a part of a longer string which indicates another word. From this point of view, it seems better to focus on only a longer a verb phrase. Then one potential constraint is to handle a verb phrase that is longer than two syllables. For example, กระวีกระวาด (grà-wêek-rá-wâat: to hurry; four syllables) and ละลานตา (lá laan dtaa : be dazzled; three syllables).

We know each characteristic of each NE because of its pattern, but another problem is boundary correction. Shorter or longer boundary detection can be found in NE extraction. In order to solve this problem, we add a set of clue phrases between NE tagging sequences. These clue phrases are useful to reduce ambiguity of NE boundary, and then increase the accuracy of NE extraction, especially for location and person. From Table 3, we add person's position between time and location tagging sequence, question, conjunction, adverb and verb phrases with three or more syllables between location and person tagging sequence. Question phrases with three or more syllables such as เพราะเหตุใด (prór hèt dai : why; three syllables). Conjunction with three or more syllables such as เช่นเดียวกับ (chên dieow gàp : as well as; three syllables). Adverb phrases with three or more syllables such as ชั่วลูกชั่วหลาน (chûa lôok chûa lăan: forever; four syllables).

4. Pattern-based Extraction

Since we would like to extract NE such as person, date, time, location, and action phrases in Thai news documents, we assume that the combination of various existing solutions involved with this task can be another way to achieve both detection and extraction. Nowadays, information is available in various kinds of format, lists of words, and dictionaries which are great tools that can be used to process natural language. Longest word matching, longest pattern matching, pattern categorization, tagging sequence, three or more syllables of part-of-speech and TCCs algorithm (an unambiguous unit that is smaller than a word and cannot be further divided, to reduce the ambiguity of word boundary in Thai documents), combined with clue words, and word lists from dictionaries, are all techniques that we call pattern-based extraction.

There are four possible pattern boundaries

(SPC|TXT)(PTT)(SPC|TXT)

For example:

(SPC)(PTT)(SPC)
คนร้ายคือ นายไก่ อยู่บ้านเลขที่
(SPC)(PTT:นายไก่)(SPC)

(SPC)(PTT)(TXT)
ก่อนหน้านั้น นายไก่ได้แจ้งว่า
(SPC)(PTT:นายไก่)(TXT:ได้แจ้งว่า)

(TXT)(PTT)(SPC)
จับกุมนายไก่ ก่อนจะหลบหนี
(TXT:จับกุม)(PTT:นายไก่)(SPC)
(TXT)(PTT)(TXT)
จับกุมนายไก่ข้อหาฆ่าคนตาย

(TXT:จับกุม)(PTT:นายไก่)(TXT:ข้อหาฆ่าคนตาย)

Since we know that a clue word increases accuracy to detect NE, we can classify four types of patterns (PTT)

(CW|∅)(TIF)(CW|∅)

For example:

(CW)(TIF)(CW)
บริษัท ทีวีบูรพา จำกัด
(CW:บริษัท)(TIF: ทีวีบูรพา)(CW:จำกัด)

(CW)(TIF)(∅)
จังหวัดนครศรีธรรมราช
(CW:จังหวัด)(TIF:นครศรีธรรมราช)(∅)

(∅)(TIF)(CW)
ตลาดหลักทรัพย์แห่งประเทศไทย
(∅)(TIF:ตลาดหลักทรัพย์)(CW:แห่งประเทศไทย)

(∅)(TIF)(∅)
จีนี่ เรคคอร์ดส
(∅)(TIF:จีนี่ เรคคอร์ดส)(∅)

The SPC, PTT, TXT, TIF, CW and ∅ are space, pattern, text, tagging information, clue word and without clue word, respectively. TIF is composed of character, number, space, symbol, and marks ordered by its longest matching combination. We will consider TCCs algorithm in every connection between (TXT)(PTT) and (PTT)(TXT). The combinations of boundaries and types are 16 patterns:

(SPC|TXT)(CW|∅)(TIF)(CW|∅)(SPC|TXT)

For each NE type, level of ambiguity is different. NE type with (CW)(TIF)(CW) is less ambiguity than (∅)(TIF)(∅). The position of clue word in each NE type can be classified as shown in Table 2. In Table 2, we can conclude that person and action phrases have the most ambiguity found, because of no clue word for detection. When we consider a semantic level, date and time have less ambiguity than person and action, so we can construct tagging sequence starting with date and time and so on, as in Table 3.

Algorithm 1 shows how NEs can be extracted using our pattern-based method. Given a set of news documents *D* and a set of patterns *P*, the algorithm will output a set of NEs *E* grouped by their types. In this work we have nine entity types (*t*=9) and E_1-E_9 corresponding to a set of extracted NEs of type DATE, TIME, POSITION, LOCATION, QUESTION-PHRASE, CONJUNCTION-PHRASE, ADVERB-PHRASE, VERB-PHRASE, and PERSON, respectively. The function *PatternOrdering* will rearrange the set of patterns *P* with respect to their groups and length by the functions *GroupByPattern* and *Ordering ByLength*, and output a set of pattern sets *FPG={FPG$_1$,FPG$_2$,... FPGt}*, where *FPG$_i$* is the set of *type-i* patterns ordered by their length.

Algorithm 1: Pattern-based NE Extraction

```
$D = \{d_1, d_2, d_3,...,d_w\}$
# a set of w news documents

$P = \{p_1, p_2, p_3,\ldots, p_n\}$
# a set of n patterns

$E = \{E_1, E_2, E_3,...E_t\}$
# a set of extracted NEs from the document set D

# $E_i$ is a set of extracted NEs of type i
# t is the number of NE types being considered
```

```
main(){
  if(updated(P)){
    newP = PatternOrdering(P);
  }
  $E_i = \varnothing$;
  foreach $d_n \in D$ {
    foreach $newP_i \in newP$ {
      foreach $newP_{ij} \in newP_i$ {
        $LPG_{ij}$ = OrderingByLength($PPG_{ij}$);
        foreach $PPG_{ijk} \in PPG_{ij}$ {
          $ONE_{ijkn}$ = apply($PPG_{ijk}$,$d_n$);
          $E_i = E_i \cup \{ONE_{ijkn}\}$;
        }
      }
    }
  }
}
function PatternOrdering (P){
  PG = NETypeGrouping(P);
  FPG = $\varnothing$;
  foreach $PG_i \in PG$ {
    $PPG_i$ = GroupByPattern($PG_i$);
    $FPG_i = \varnothing$;
    foreach $PPG_{ij} \in PPG_i$ {
      $LPG_{ij}$ = OrderingByLength($PPG_{ij}$);
      $FPG_i = FPG_i \cup \{LPG_{ij}\}$;
    }
    $FPG = FPG \cup \{FPG_i\}$;
  }
  return FPG;
}
```

5. Experiment

Our domain for NE extraction is Thai news documents, collected from websites in these categories: crime, economic, foreign, politic and sport. News documents in each category are taken from Thai three news publishers: KomChad-Luek[1], DailyNews[2], and Manager[3]. There are several data owners contributing their data in digital format and publishing online. We easily collect the lists of word and various kinds of dictionaries as shown in Table 4. Words in the list of word and dictionaries are ordered by maximum longest word. After we combine the number of words from lists and dictionaries with pattern-based extraction, we have possible patterns for each NE as in Table 5.

Table 2 The position of clue word in each NE type. ‘Pre’ and ‘Suf’mean prefix and suffix, respectively.

(Pre,Suf)	DAT	TIM	LOC	PER	ACT
(+;+)	Y	Y	Y	N	N
(+;−)	Y	Y	Y	Y	Y
(−;+)	Y	Y	Y	N	N
(−;−)	Y	Y	Y	Y	Y

5.1 Implementation and Results

After we applied pattern-based extraction with all words in word lists and dictionaries to all Thai news documents, the results are shown in Table 6. The number of correct results for date, time, conjunction phrase with three or more syllables, and adverb phrase with three or more syllables are very high, but the number of correct results for persons is low. We know that the identification of a person’s surname and a person’s name without surname are very hard to find the exact boundary. The number of correct results in each news category is shown in Table 7. From the result, the identification of date and time are highly significant in every news category. Among all categories, we gain the highest accuracy for the ‘crime’ category. The performance order in detecting NE is crime, economics, politics, foreign and sports, respectively. When we investigated Thai sports news documents, we found that the written style for news in this category is quite different from our provided patterns, and a lot of complex sentences have been found. On the other hand, in the crime category, written structure and sentences are closer to our pattern-based extraction, which is Thai common usage, so the number of possible NEs are numerously detected in this category. From Thai three news publishers, Manager achieved the best

[1] KomChadLuek: http://www.komchadluek.net

[2] DailyNews: http://www.dailynews.co.th

[3] Manager: http://www.manager.co.th

correctness, followed by KomChadLuek and DailyNews.

Table 3 Tagging sequence

Tagging Sequence	Named Entity and Clue Phrase*
1	Date
2	Time
3	Position*
4	Location
5	Question phrase with three or more syllables*
6	Conjunction phrase with three or more syllables*
7	Adverb phrase with three or more syllables*
8	Verb phrase with three or more syllables
9	Person

5.2 Error Analysis

This section shows some types of errors found in NE extraction. As the first type, an extracted NE may be a part of a compound noun and it should not be extracted. Some examples are กระทรวงต่างประเทศ*เกาหลีใต้* (Ministry of Foreign Affairs of *South Korea*), โฆษกรัฐมนตรีกระทรวงต่างประเทศ*เกาหลีใต้*)Foreign Ministry spokesman, *South Korea*) and ทีมชาติ*ไอร์แลนด์*)*Ireland* National Team). A verb may be a part of project's name and it should not be extracted, for example โครงการ *รับประกันคุณภาพ* สินค้าและบริการ)The *Quality Assurance* Project of Products and Services), and รัฐมนตรี *ช่วยว่าการ*)*Vice* Minister). A person title abbreviation is incorrectly recognized such as น*สพ.*เซาเทิร์น เมโทรโปลิส (Southern Metropolis Newsletter, *นสพ.* stands for Newsletter but *สพ.* stands for Veterinarian). A person name without a surname such as เอกชัย *อยู่ที่โรงพยาบาล* (Ekachai *YooTeeRongPayaban*, YooTeeRongPayaban is an action phrase which means 'to stay in a hospital', but it is recognized as Ekachai's surname). Some pronouns may be recognized as person title such as ให้*คุณ*ฟรี ("to give *you* for free" is recognized as "to give to *Mr.* Free"). A person title may be recognized as a part of a verb such as ขอบ*คุณ* ประชาชน (Thank *you* citizen). Other errors are triggered by word boundary ambiguity, inappropriateness of patterns and misspelled words.

Table 4 List of dictionaries and word lists

NE Type &Phrase	Information Type	No. of Records
PER	Title	555
PER	Title Abbreviation	340
PER	Title Exception	198
PER	Position	244
LOC	Amphoe	924
LOC	Province	76
LOC	Thai government department	2449
LOC	Thai government organization	309
LOC	Common country name	192
LOC	Full country name	192
LOC	Capital city name	194
DAT-TIM	Time unit	26
DAT-TIM	Season	17
DAT-TIM	Number	30
DAT-TIM	Month	24
DAT-TIM	Era	16
DAT-TIM	Day	14
DAT-TIM	Year	24
PREP	Preposition for location	63
ACT	Verb phrase with three or more syllables	2767
ACT	Verb	13027
QUE	Question phrase with three or more syllables	6
CONJ	Conjunction phrase with three or more syllables	88
ADV	Adverb phrase with three or more syllables	912

6. Conclusion and Future Works

This paper presented a pattern-based approach in NE extraction from Thai news documents. The extracted NEs are person, organization, location, date and time, as well as action phrases from a text. Techniques of longest word matching, longest pattern matching and pattern categorization are applied to construct a set of patterns in the form of regular expressions for retrieving our focused NEs. Our pattern-based approach is possible to apply with news in other languages. By experiments, we found that with a suitable sequence of applying regular expression, we can gain up to a correctness of 68-100%, using a large-scale Thai dictionary and a set of predefined pattern matching templates. System performance and machine learning techniques with more news documents are topics for our future work. We also plan to apply our NE extraction model to other domain documents such as medical documents. Further investigation needs to be done in both syntax and semantic level.

Table 5 No. of patterns for each NE type and phrase

NE Type & Phrase	No. of patterns
DAT-TIM	350
LOC	9284
QUE	6
CONJ	88
ACT	2767
ADV	912
PER	2582
Total	15077

Table 6 Experimental result after extracting NEs and phrases with pattern extraction

Tagging	No. of Matched Instances	No. of Correct Instances	% Correct-ness
DAT-TIM	95	95	100.00
LOC	44	33	75.00
QUE	0	-	-
CONJ	18	18	100.00
ADV	25	25	100.00
ACT	83	70	84.33
PER	69	47	68.11

Table 7 The correctness in each news category (No. of correct instances / No. of matched instances)

Category	DAT-TIM	LOC	ACT	PER
Crime	24/24	27/33	18/18	25/32
Economics	19/19	4/5	26/35	7/10
Foreign	14/14	2/5	11/12	2/7
Politics	14/14	-	10/12	13/19
Sports	24/24	0/1	5/6	0/1

7. Acknowledgement

This work has been supported by the Thailand Research Fund (TRF), under Grant No. BRG50800013 and a TRF Royal Golden Jubilee Ph.D. Program, under Grant No.PHD/0034/2550, and was supported by the National Electronics and Computer Technology Center (NECTEC), Thailand, under Grant No. NT-B-22-I4-38-49-05.

8. References

[1] Rau, L. F., Extracting Company Names from Text, In Proceedings of the Seventh IEEE Conference on Artificial Intelligence Applications, February 24-28, pp. 29–32, 1991.

[2] Grishman, R. and Sundheim, B., Message Understanding Conference 6: A Brief History. In Proceedings of International Conference on Computational Linguistics in June 1996, 1996.

[3] Kozareva, Z., Ferrandez, O., Montoyo, A. Munoz, R., Suarez A., and Gomez. J., Combining Data-driven Systems for Improving Named Entity Recognition, Data & Knowledge Engineering, Vol. 61, No. 3, pp. 449–466, 2007.

[4] Etzioni, O., Cafarella, M., Downey, D., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., and Yates, A., Unsupervised Named Entity Extraction from the Web: An Experimental Study, Artificial Intelligence, Vol. 165, No. 1, pp. 91–134, 2005.

[5] Ananiadou, S., Friedman, C., and Tsujii, J., Introduction: Named Entity Recognition in Biomedicine, Journal of Biomedical Informatics, Vol. 37, No. 6, pp. 393–395, 2004.

[6] Paşca, M., Weakly-Supervised Discovery of Named Entities Using Web Search Queries, In Proceedings of the Sixteenth ACM Conference on

Information and Knowledge Management (CIKM'07), pp. 683–690, USA, 2007.

[7] Shinyama, Y., and Sekine, S., Named Entity Discovery Using Comparable News Articles, In Proceedings of the Twentieth International Conference on Computational Linguistics (COLING '04), pp. 848, USA, 2004.

[8] Utsuro, T., Sassano, M., and Uchimoto, K., Combining Outputs of Multiple Japanese Named Entity Chunkers by Stacking, In Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing (EMNLP'02), pp. 281–288, USA, 2002.

[9] Kumano, T., Kashioka, H., Tanaka, H., and Fukusima, T., Construction and Analysis of Japanese-English Broadcast News Corpus with Named Entity Tags, In Proceedings of the ACL 2003 Workshop on Multilingual and Mixed-Language Named Entity Recognition, pp. 17–24, USA, 2003.

[10] Isozaki, H., Japanese Named Entity Recognition based on A Simple Rule Generator and Decision Tree Learning, In Proceedings of the Thirtieth Annual Meeting on Association for Computational Linguistics (ACL'01), pp. 314–321, USA, 2001.

[11] Gao, J., Li, M., Wu, A., and Huang, C.N., Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach, Comput. Linguist., Vol. 31, No. 4, pp. 531–574, 2005.

[12] Wu, Y., Zhao, J., Xu, B., and Yu, H., Chinese Named Entity Recognition based on Multiple Features, In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05), pp. 427–434, USA, 2005.

[13] Fu, G., and Luke, K.K., Chinese Named Entity Recognition Using Lexicalized HMMs, SIGKDD Explor. Newsl., Vol. 7, No. 1, pp. 19–25, 2005.

[14] Ye, S., Chua, T.S., and Jimin, L., An Agent-based Approach to Chinese Named Entity Recognition, In Proceedings of the Nineteenth International Conference on Computational Linguistics, pp. 1–7, USA, 2002.

[15] Thao, P.T.X., Tri, T.Q., Dien, D., and Collier, N., Named Entity Recognition in Vietnamese using Classifier Voting, ACM Transactions on Asian Language Information Processing (TALIP), Vol. 6, No. 4, pp. 1–18, 2007.

[16] Li, W., and McCallum, A., Rapid Development of Hindi Named Entity Recognition Using Conditional Random Fields and Feature Induction, ACM Transactions on Asian Language Information Processing (TALIP), Vol. 2, No. 3, pp. 290–294, 2003.

[17] Chung, E., Hwang, Y.G., and Jang, M.G., Korean Named Entity Recognition using HMM and Co Training Model, In Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages, pp. 161–167, USA, 2003.

[18] Sornlertlamvanich, V., Potipiti, T., and Charoenporn, T., Automatic Corpus-based Thai Word Extraction with the C4.5 Learning Algorithm, In Proceedings of the Eighteenth International Conference on Computational Linguistics, pp. 802–807, 2000.

[19] Sukhahuta, R., and Smith, D., Information Extraction Strategies for Thai Documents, International Journal of Computer Processing of Oriental Language, Vol. 14, pp. 153–172, 2001.

[20] Narupiyakul, L., Thomas, C., Cercone, N., and Sirinaovakul, B.,

Thai Syllable-based Information Extraction using Hidden Markov Models, In Proceedings of the Fifth International Conference on Intelligent Text Processing and Computational Linguistics, Lecture Notes in Computer Science, Vol. 2945, pp. 537–546, 2004.

[21] Kawtrakul, A., Suktarachan, M., Varasai, P., and Chanlekha, H., A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling, In Proceedings of the Third Workshop on Asian Language Resources and International Standardization (COLING’ 02), pp. 1–8, USA, 2002.

[22] Kawinpanithan, A., and Aroonmanakun, W., A Computational Linguistic Study of Context Clues of Proper Names in Thai, Master’s Thesis, Chulalongkorn University, Bangkok, Thailand, 2003.

[23] Chaicharoen, N., and Aroonmanakun, W., Computerized Integrated Word Segmentation and Part-of-Speech Tagging of Thai, Master’s Thesis, Chulalongkorn University, Bangkok, Thailand, 2001.

[24] Muanpai, N., and Kawtrakul, A., Enhancing Thai Document Retrieval System Performance with Noun Phrase Analysis, In the Ninth National Computer Science and Engineering Conference (NCSEC 2005), 2005.

[25] Suwanno, N., Suzuki, Y., and Yamazaki, H., Extracting Thai Compound Nouns for Paragraph Extraction in Thai Text, In Proceedings of 2005 IEEE International Conference (IEEE NLP-KE’05), pp. 657–662, 2005.

[26] Kriengket, K., Kosawat, K., and Anchaleenukul, S., A Computational Linguistics Study of Compound Nouns in Thai, In Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP 2007), pp. 31–36, 2007.

[27] Chanlekha, H., and Kawtrakul, A., Thai Named Entity Extraction by Incorporating Maximum Entropy Model with Simple Heuristic Information, LNCS (LNAI), Vol. 3248, pp. 49–55., 2004.

[28] Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong, T., and Chinnan, W., Character Cluster based Thai Information Retrieval, In Proceedings of the Fifth International Workshop on Information Retrieval with Asian Languages (IRAL ’00), pp. 75–80, USA, 2000.

[29] Yingseree, C., Suktarachan, M., and Kawtrakul, A., Time Expression Normalization for Thai Language, In Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP 2007), pp. 177–182, 2007.

[30] Viriyayudhakorn, K., Prayoonsri, C., Silpasuwanchai, C., Nattee, C., and Theeramunkong, T., A Statistical Approach to Classify Nationality of Name, In Proceedings of the Seventh International Symposium on Natural Language Processing (SNLP 2007), pp. 7–11, 2007.

9.Appendix

Day of Week: A day-of -week is represented by its full name or its abbreviation

จันทร์, จ.	Monday, Mon
อังคาร, อ.	Tuesday, Tue
พุธ, พ.	Wednesday, Wed
พฤหัสบดี, พฤ.	Thursday, Thu
ศุกร์, ศ.	Friday, Fri
เสาร์, ส.	Saturday, Sat
อาทิตย์, อา	Sunday, Sun

Month: A month is represented by its full name, its abbreviation, Arabic numbers or Thai numbers

มกราคม, ม.ค., ๑	January, Jan, 1
กุมภาพันธ์, ก.พ., ๒	February, Feb, 2
มีนาคม. มี.ค., ๓	March, Mar, 3
เมษายน, เม.ย., ๔	April, Apr, 4
พฤษภาคม, พ.ค., ๕	May, May, 5
มิถุนายน, มิ.ย., ๖	June, Jun, 6
กรกฎาคม, ก.ค., ๗	July, Jul, 7
สิงหาคม, ส.ค., ๘	August, Aug, 8
กันยายน, ก.ย., ๙	September, Sep, 9
ตุลาคม, ต.ค., ๑๐	October, Oct, 10
พฤศจิกายน, พ.ย., ๑๑	November, Nov, 11
ธันวาคม, ธ.ค., ๑๒	December, Dec, 12

Year in Chinese Zodiac: A 12-year cycle in Chinese Zodiac is represented by names of auspicious animals

ปีชวด, ปีหนู	Year of the rat
ปีฉลู, ปีวัว	Year of the ox
ปีขาล, ปีเสือ	Year of the tiger
ปีเถาะ, ปีกระต่าย	Year of the rabbit
ปีมะโรง, ปีงูใหญ่	Year of the dragon
ปีมะเส็ง, ปีงูเล็ก	Year of the snake
ปีมะเมีย, ปีม้า	Year of the horse
ปีมะแม, ปีแพะ	Year of the ram
ปีวอก, ปีลิง	Year of the monkey
ปีระกา, ปีไก่	Year of the rooster
ปีจอ, ปีหมา	Year of the dog
ปีกุน, ปีหมู	Year of the boar

Era

พุทธศักราช, พ.ศ.	Buddhist Era, B.E.
คริสต์ศักราช, ค.ศ.	Christian Era, A.D.
มหาศักราช, ม.ศ.	Major era, Saka era, Shalivahana era
ฮิจญ์เราะหฺศักราช, ฮ. ศ.	Hijra era, H.E. Anno Hejira, A.H.
จุลศักราช, จ.ศ	Thai minor era
รัตนโกสินทรศก, ร.ศ.	Ratana Kosindra era

Time unit: A unit used for representing time

วินาที, วท.	second, sec.
นาที, นท.	minute, min.
ชั่วโมง, ยาม, ชม.	hour, hr.
นาฬิกา, (น.), โมง	o’ clock
โมงเช้า	o’ clock in the morning
โมงเย็น, ทุ่ม	o’ clock in the evening
วัน	day
คืน, ค่ำ, ราตรี	night
สัปดาห์, อาทิตย์	week
ปักษ์	fortnight
งวด	times
เดือน	month
ไตรมาส	three months
ฤดู	season
ปี	year
ศก	era
ยุค	age
ทศวรรษ	decade
ศตวรรษ	century
สหัสวรรษ	millennium

Season: A name represents a season

ฤดูร้อน, คิมหันตฤดู, หน้าร้อน	Summer
ฤดูฝน, วัสสานฤดู, หน้าฝน	Rainy
ฤดูใบไม้ร่วง, สารทฤดู, หน้าหนาว	Autumn
ฤดูหนาว, เหมันตฤดู	Winter
ฤดูใบไม้ผลิ, วสันตฤดู	Spring
ฤดูหมอก, สิสิรฤดู, หน้าหมอก	Fog
ฤดูน้ำค้าง	Dew

Western Zodiac: An annual cycle of twelve stations along the ecliptic

ราศีเมษ	Aries
ราศีพฤษภ	Taurus
ราศีเมถุน	Gemini
ราศีกรกฎ	Cancer
ราศีสิงห์	Leo
ราศีกันย์	Virgo
ราศีตุลย์	Libra
ราศีพิจิก	Scorpio
ราศีธนู	Sagittarius
ราศีมังกร	Capricorn
ราศีกุมภ์	Aquarius
ราศีมีน	Pisces