Language WindOWing through CORPORA
Visualización del lenguaje a través de corpus

Isabel Moskowich
Begoña Crespo
Inés Lareo
Paula Lojo
Eds.

Language Windowing through Corpora.

Visualización del lenguaje a través de corpus

Part II

L-Z

Editors

Isabel Moskowich-Spiegel Fandiño

Begoña Crespo García

Inés Lareo Martín

Paula Lojo Sandino

Universidade da Coruña

A Coruña 2010

ISBN: 978-84-9749-401-4

Cover designed by Inés Lareo and Alejandro González

Contents/Contenidos Part II

Foreword

Though Corpus Linguistics, both as a methodology and as a branch of linguistics itself, has been among us for the last forty years, its development, mainly in the Anglophone world, has had a repercussion on the rest of the community of linguists. In Spain, for instance, the recently created association of Corpus Linguistics (AELINCO) testifies to this. The collection of essays we are presenting here are just a mere sample of the interest the topics relating to Corpus Linguistics have arisen everywhere.

Such different topics as those related to Computational Linguistics found in "Obtaining computational resources for languages with scarce resources from closely related computationally-developed languages. The Galician and Portuguese case" or "Corpus-Based Modelling of Lexical Changes in Manic Depression Disorders: The Case of Edgar Allan Poe" belonging to the field of Corpus and Literary Studies can be found in the ensuing pages. Almost all research areas can nowadays be investigated using Corpus Linguistics as a valid methodology. This is reason why *Language Windowing through Corpora* gathers papers dealing with discourse, variation and change, grammatical studies, lexicology and lexicography, corpus design, contrastive analyses, language acquisition and learning or translation.

This work's title aims at reflecting not only the great variety of topics gathered in it but also the worldwide interest awaken by the computer processing of language. In fact, researchers from many different institutions all over the world have contributed to this book. Apart from the twenty-two Spanish Universities, people from other Higher Education Institutions have authored and co-authored the essays contained here, namely, Russia, Venezuela, Brazil, UK, Finland, Portugal, Poland, Austria, Mexico, Thailand, Iran, the Netherlands, Belgium, Japan, Turkey, China, Italy, Malaysia, Romania and Sweden. All these essays have been alphabetically arranged, by the names of their authors, in two parts. Part 1 contains the papers by authors from A to K and Part 2, those of authors from L to Z.

Our special thanks to all the referees who carried out the selection of papers and to the contributors to this volume for giving us the opportunity to make a patchwork of different views and perspectives of what is being currently done in the field. Our thanks to Ms Agnieszka Kozera for her work as an assistant to the editors. We do hope the contents of these essays are illuminating for readers who may excuse all the mistakes and misprints that might remain after a hard editorial work.

The Editors

THAI-NEST: A framework for Thai named entity tagging specification and tools

THANARUK THEERAMUNKONG, MONTHIKA BORIBOON, CHOOCHART HARUECHAIYASAK,
NICHNAN KITTIPHATTANABAWON, KRIT KOSAWAT, CHUTAMANEE ONSUWAN,
ISSARIYAPOL SIRIWAT, THAWATCHAI SUWANAPONG, AND NATTAPONG TONGTEP
Thammasat University
National Electronics and Computer Technology Center

Abstract
A THAI-NEST framework is presented for a construction of Thai news corpus with named entity (NE) tagging process. Three main components of the framework are corpus tagging specification, tagging process, and tagging tools. To be in line with the Text Encoding Initiative (TEI) standardization, a tagging specification is developed by taking into account some characteristics of Thai NEs, including proper nouns, expressions of date, time, and quantity, and other extended named entities. The developed specification includes a tag set and its tagging schema. A set of tagging tools is designed and implemented with an effective GUI. The tool set supports two tagging levels of NE type and NE structure. Results and statistics of our ongoing corpus construction are reported.
Keywords: Named Entity, Thai Language, News Corpus, Language Resource

Resumen
Se presenta el marco THAI-NEST para construir un corpus de noticias tailandesas mediante un proceso de etiquetaje de entidades nombradas (NE). Los tres componentes principales del marco son la especificación del etiquetaje del corpus, el proceso de etiquetaje y las herramientas de etiquetaje. Para seguir la línea de la estandarización de la Text Encoding Initiative (TEI), se desarrolla una especificación de etiquetaje teniendo en cuenta algunas características de entidades nombradas tailandesas, incluyendo nombres propios, expresiones de fecha, hora y cantidad, así como otras entidades nombradas. La especificación desarrollada incluye un conjunto de etiquetas y su esquema de etiquetaje. Se diseña un conjunto de herramientas de etiquetaje que se implementa con un GUI efectivo. El conjunto de herramientas admite dos niveles de tipo y estructura de entidades nombradas. Se informará de los resultados y estadísticas de nuestro corpus en construcción.
Palabras clave: Entidad Nombrada, Lengua Tailandesa, Corpus de Noticias, Recurso Lingüístico

1. INTRODUCTION[1]

Named Entity Recognition (NER) is considered one of the fundamental tasks in NLP. In Thai NLP community, however, the topic of NE related tasks is not as widely discussed as in other languages. The early Thai NER had some limitations. Firstly, it was specific to a very few

[1] This research is supported by the National Electronics and Computer Technology Center under the project code NT-B-22-KE-38-52-01, and partially supported by National Research Council of Thailand (NRCT) via Thammasat University as well as Thailand Research Fund under the project number BRG5080013.

areas such as agricultural (Kawtrakul *et al.*, 2001) and political news (Chanlekha & Kawtrakul, 2004). Secondly, only a small class of NE types (person name, organization, and place) was mentioned. Lastly, evaluation tasks were usually performed on small-size corpora. With this in mind, the main goal of this project is to design a new framework for constructing a large-scale NE corpus with a larger set of NE types from various domains.

The proposed framework (THAI-NEST) includes the following components: (1) a specification for Thai NE tag set, (2) a tagging process, and (3) tagging tools. Our tag set specification was adapted from the TEI guidelines (Barnard & Ide, 1997) to suit the Thai language characteristics. The tag set followed some specifications proposed by TEI guidelines with additional modifications. Our NE tag set includes person name, organization name, place name, date and time expression, quantitative expression. In addition, we also considered an annotation of other named entities previously proposed by the Sekine's Extended Named Entity Hierarchy (Sekine, 2007).

The tagging process consists of three main steps: (1) news collection, (2) news article metadata and structure tagging, and (3) NE tagging and verification. Also due to a very short project time frame of one year, our tagging process was designed such that many steps can be carried out in a pipeline and parallel processing manner. For example, the news article structure tagging and the NE tagging can be performed as a pipeline process. Moreover, the tagging of different NE types can be done in parallel. In this paper, we will discuss and share our experience in designing the framework to allow the maximum resource allocation. With 10,000 news articles being annotated, our corpus will be the largest Thai NE corpus to date.

The remainder of paper is organized as follows. The related work in NE related tasks, i.e., corpus construction and NER, is given in the next section with special focus on Thai language. In Section 3, we present the proposed tagging framework which includes tag set design, tagging process and tools. A flowchart diagram with some tagging examples will be given to illustrate the proposed framework. Section 4 provides a summary of current corpus statistics. A summary with discussion is presented in the last section.

2. Related Work

Named entities recognition (NER) is one of the most extensively studied topics in NLP. There have been many organized conferences and workshops to discuss related issues including corpus designs and algorithms for recognizing and extracting NEs. Early conferences include MUC (Message Understanding Conference) and CoNLL (Conference on

Computational Natural Language Learning). MUC, initiated and funded by the Defense Advanced Research Projects Agency (DARPA), is perhaps the first widely recognized conference which focused on designing and creating a large-scale corpus specifically for NE related tasks (Chinchor, 1998). There are three-main subtasks considered under MUC: (1) entity names (person, organization, and location), (2) temporal expressions (date and time) and (3) number expressions (monetary expression and percentage).

CoNLL, on the other hand, focuses on language-independent NER task (Sang *et al*., 2003). In addition to basic NE types, CoNLL corpus also includes miscellaneous (MISC) names belonging to different domains such as adjectives (e.g., Italian) and events (e.g., World Cup, Olympics). Another effort in NE tasks is the ACE (Automatic Content Extraction) conference organized by the National Institute of Standards and Technology (NIST) (Linguistic Data Consortium, 2008). Under ACE, entities are categorized into seven types: person (PER), organization (ORG), geo-political entity (GPE), location (LOC), facility (FAC), vehicle (VEH), and weapon (WEA). Each tag is allowed to have subtypes such as *PER.Individual* to denote individual person and *PER.Group* to denote a group of people. ACE aims to develop some automatic content extraction techniques from different text sources such as newswire, broadcast conversation, and weblogs.

Recently the NE related tasks are increasingly recognized and have been discussed among Asian NLP community. Kumano *et al*. (2003) constructed a cross-lingual Japanese-English broadcast news corpus of 1,100 article pairs with NE tags. Their goal was to acquire NE translation knowledge by utilizing NE extraction techniques. Takenobu *et al*., (2006) reported a collaboration among many countries in Asia to create a common standard for Asian language resources. Their project focuses on constructing a lexicon set with an upper-layer ontology. The NE related tasks have yet to be discussed. In IJCNLP 2008 (The Third International Joint Conference on Natural Language Processing), there were two workshops related to NER tasks: (1) Named Entity Recognition for South and South East Asian Languages (NERSSEAL) and (2) Asian Language Resources (ALR). There are many reported NE corpora and tools in many different Asian languages including Indian, Telugu, Bengali, and Tamil (Ekbal & Bandyopadhyay, 2008; Saha *et al*., 2008; Sangal *et al*., 2008).

As for NE related tasks in Thai, Tongtep and Theeramunkong (2008) have presented a pattern-based approach for named entity extraction from Thai news documents. This named entity extraction is later applied as preprocessing for relation extraction from Thai news documents in Tongtep and Theeramunkong (2009). During the most recent SNLP 2009 (The Eighth International Symposium on Natural Language Processing), there were some reports

on Thai NE related tasks. Lertcheva and Aroonmanakun (2009) applied the CRF (Conditional Random Fields) algorithm to construct an NER model for Thai language. The corpus is based on the BEST 2009 word segmentation corpus (Kosawat *et al*., 2009) and contains only 90,000 words. Only three named entities (person, organization and place) were considered. Suwanapong and Theeramunkong (2009) proposed a method based on the SVD (Singular Value Decomposition) algorithm to identify aliases in Thai sports news articles. Inyaem *et al*. (2009) presented a method based on domain-specific NE to extract terrorism events from Thai news articles. Sutheebanjerd and Premchaiswadi (2009) proposed a different method to extract NEs from Thai news articles. Their work only covered person names and their models were trained using only approximately 1,000 news articles. Tirasaroj and Aroonmanakun (2009) presented a study on linguistic structure of Thai product names from economic news articles.

In sum, previous work on Thai NE related tasks was very limited to a few NE types and also domain-specific to certain topics. There has not been any open large-scale NE corpus with multiple NE types for Thai language. Perhaps the most related works to ours are the development of "Simple Named Entity Guidelines for Thai" and "Time Annotation Guidelines for Less Commonly Taught Languages: Thai" carried out by a research group at the Linguistic Data Consortium (LDC) (Linguistic Data Consortium, 2006a; Linguistic Data Consortium, 2006b). The simple named entity guidelines are based on the MUC-7 NE Guidelines which cover three basic NE types of person, organization and location. The time annotation guidelines are based on the TIMEX2 standard which provides a tag set for tagging temporal expressions.

Compared with the guidelines above, our proposed NE tag set is designed based on the adaptation of TEI guidelines which provide rich details for the designed tag sets (Barnard & Ide, 1997). Some of the tags are modified and added to make them suitable for the Thai language characteristics. For the application of detection and recognition tasks, tag sets are modified. For example, unlike TEI guidelines which classify <*placeName*> into various geopolitical subtypes, <*district*>, <*settlement*>, <*region*>, <*country*>, and <*bloc*>, we propose <*prefix*>, <*infix*>, and <*suffix*> to be used as a feature set for training the NER model. Such design provides additional syntactic features to compensate for the lack of capitalization in Thai.

Our guidelines cover seven NE types along with the framework of process and tools to assist the tagging process. We use Thai news articles collected from the Web as the main resource for the corpus. Once completed, we plan to release the corpus to the Thai NLP community.

3. THE PROPOSED TAGGING FRAMEWORK

In this section, we describe the tagging specification, process, and tools. The details on the tag set design with a tagging example are given to illustrate our proposed framework.

3.1. Tag Set Design

Many single and multiple word NEs in Thai derive from existing lexicon. Thus, some components of NE may be confused with ordinary words. For example, in a sentence หัวหน้าชอบอ้างอภิสิทธิ์ (อภิสิทธิ์, a single word, has two meanings: the name of a person (NE) or a privilege). Regular dictionaries often do not include NEs because they are an open set, i.e., new NEs could be coined everyday. So, it is not reasonable to recognize NEs by mapping with common dictionary, except for some frequently used ones. A more practical approach is by using Machine Learning (ML) technique in which annotated corpora are employed to train a model. Features such as capitalization, word boundaries are used as contextual clues in Latin-based writing systems. However, these features do not exist in Thai as shown in the above example where written text appears as a long string of uniform connected symbols. Contextual understanding and background knowledge are required to identify Thai NEs.

NE-Type Level	NE-Structure Level	Description	Example
<persName>		■ Proper noun or proper noun phrase referring to people	<persName> <roleName>ดร.</roleName> <forename>อนารักษ์</forename> (<addname>ปิง</addname>) <surname>ธีระมั่นคง</surname> </persName>
	<roleName>	• Titles, honorific titles, ranks, kinship terms, etc. used before given name	
	<forename>	• First name	
	<middlename>	• Middle name (if any)	
	<surname>	• Family or last name	
	<addName>	• Additional or alias name	
<orgName>		■ Name of organization, institution, nation, etc.	<orgName> <component> <prefix>สาขาวิชา</prefix> ภาษาศาสตร์ </component> <component> <prefix>คณะ</prefix>ศิลปศาสตร์ </component> <component> <prefix>มหาวิทยาลัย</prefix> ธรรมศาสตร์ <infix>ศูนย์</infix>รังสิต </component> </orgName>
	<prefix>	• Set of words indicating types used before a given organization name	
	<infix>	• Set of words indicating sub-types used inside a given organization name	
	<suffix>	• Set of words indicating types used after a given organization name	
	<addName>	• Additional or alias name, metonyms of a given organization name	
	<component>	• Members of single contiguous organization sequence	
<placeName>		■ Name of place, location, geographical bodies, country, etc.	<placeName> <prefix>พระที่นั่ง</prefix>จักรี <suffix>มหาปราสาท</suffix> </placeName>
	<prefix>	• Set of words indicating types used before a given place name	
	<infix>	• Set of words indicating sub-types used inside a given place name	
	<suffix>	• Set of words indicating types used after a given place name	
	<addName>	• Additional or alias name, metonyms of a given place name	
	<component>	• Members of single contiguous place name sequence	
<date>		■ Date in any format	<date> <offset>ก่อน</offset> วันที่ 1 มีนาคม 2009 </date>
	<offset>	• Set of words (before or after date expression) including repositions, subordinating conjunctions	
<time>		■ Time of day in any format	<time>19.30 น. <offset>เป็นต้นไป</offset> </time>
	<offset>	• Set of words (before or after time expression) including repositions, subordinating conjunctions	
<measure>		■ Word or phrase referring to some quantity of an object or commodity	<measure> <offset>ประมาณ</offset> <quantity>30</quantity> <unit>คน</unit> </measure>
	<offset>	• Set of words (before or after quantitative expression) including adverbial and adjectival phrases	
	<quantity>	• Digit and characters signifying measurement, amount, value, and ordinal number	
	<unit>	• Noun classifiers	
<name>		■ Extended named entities excluding those that refer to the person name, organization name, and place name (e.g., food's name, disease, product's name, etc.)	<name>รถโตโยต้า</name>

Figure 1: NE-Types, structures, and examples

Like other languages, Thai NEs are also presented with clue words that are specific to some types of NE; for example, the following terms นาย (Mr.) นาง or (Mrs.) can

be used to indicate the beginning of a Person Name, while จำกัด (Ltd.) typically used to indicate the end of an Organization Name. Those clue words are tagged separately as NE internal structure. Their positions will later be learned by ML algorithms to better predict the NE types and boundaries. Unlike NEs which are open set and hard to recognize, clue words are expected to some extent to be a closed set. Since, at this stage, we develop our corpus manually, word segmentation is beyond the scope of this paper.

In our proposed framework, Thai NE tagging comprises two levels 1) NE-Type Level, and 2) NE-Structure Level. The types and structures of NEs are fully listed in Figure 1 along with tagging examples.

3.2. Tagging Process

A flowchart (shown in Figure 2) summarizes main steps involved in the tagging process.

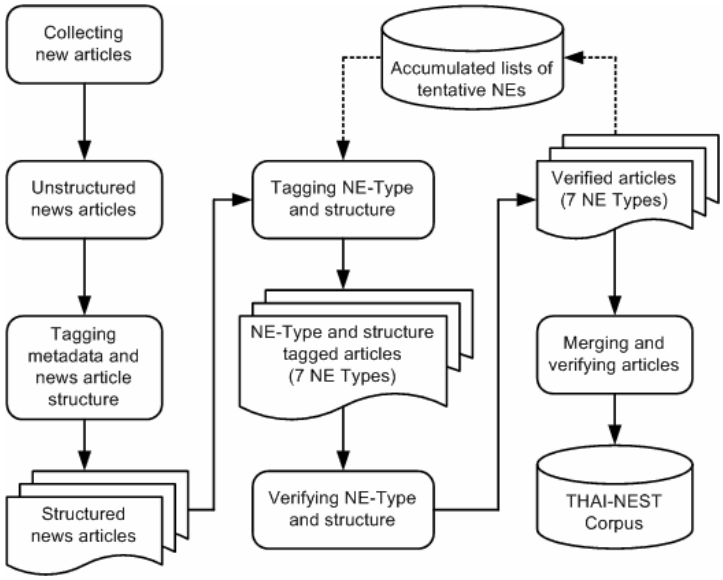

Figure 2: The tagging process

3.3. Collecting news articles

News articles during January to December 2009 have been collected from twenty-one Thai online newspaper publishers including seven major categories: crimes (CR), politics (PO), foreign affairs (FO), sports (SP), education (ED), entertainment (EN), and economic issues (EC). Over 300,000 news articles have been collected and imported to the next step using our developed Thai news structure tagging tool.

3.4. Tagging metadata and news article structure

From the collected news articles, we selected and balanced a set of 10,000 news articles in terms of news categories, publishers and time periods. Thai news structure tagging tool automatically converted the selected news articles into XML format with UTF-8 encoding. Then, metadata and news article structure were systematically assigned to the texts.

As shown in Figure 3, assigned metadata contains news title, author, publisher name, publisher URL, published date, news category, and news source URL. The news article structure includes headline, lead, and body of the news. Lastly, file names were given according to published date, news category, publisher, and file status (original, pending, submitted, and verified).

```
<?xml version="1.0"?>
<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="en">
  <teiHeader>
    <fileDesc>
      <titleStmt>
        <title>news title or headline</title>
        <author>journalist or publisher</author>
      </titleStmt>
      <publicationStmt>
        <publisher>news publisher</publisher>
        <pubPlace>publisher URL</pubPlace>
        <date>published date (yyyy-mm-dd hh:mm:ss)</date>
      </publicatonStmt>
      <notesStmt>
        <note>news category</note>
      </notesStmt>
      <sourceDesc>
        <bibl>news source URL</bibl>
      </sourceDesc>
    </fileDesc>
  </teiHeader>
  <news>
    <headline>news headline</headline>
    <lead>news lead</lead>
    <body>
      <p>news details</p>
      …
    </body>
  </news>
</TEI>
```

Figure 3: Thai news article metadata and structure

3.5. Tagging NE-Type and structure

Using our developed Thai NE Annotation Tool as shown in Figure 4, NE of seven types in news articles were tagged in parallel by linguists according to our NE annotation guidelines.

Therefore, one original new article produces seven separate files according to their tagged NEs.

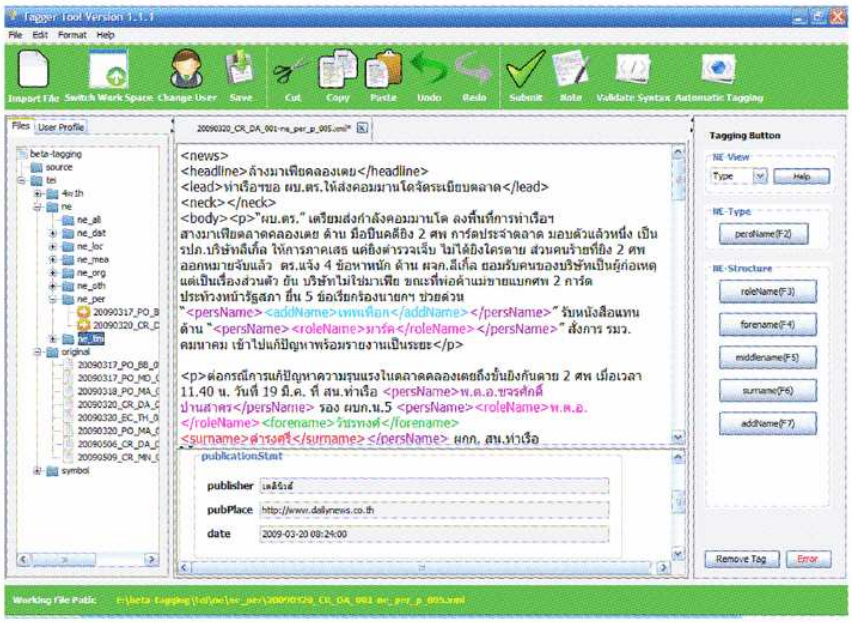

Figure 4: Snapshot of Thai NE Annotation Tool

To save time and reduce annotators' effort, the NE annotation tool was designed to incorporate several useful functions such as syntax validation, note/memo taking, log and status control, and customized GUI. Misspellings and other irregularities may be marked and kept separately. In addition, an automatic tagging function, which can be used to speed up the process, was implemented by using accumulated lists of tentative NEs either created manually by annotators or obtained from the upcoming verifying process.

3.6. Verifying NE-Type and structure

To ensure the validity and consistency of our tagging, NE of seven types was separately listed, double-checked, and corrected (if necessary). In this process, another tool – NE Tagging Verification Tool was introduced. As mentioned above, lists of each NE type can be exported to the automatic tagging function.

3.7. Merging and verifying articles

A designed function available in NE Tagging Verification Tool merges a set of separate files (from the same original file) into one file with completed NE tagging. There are three merging result status: finished, incomplete (more files are needed), and failed (addition or

deletion of lines and/or characters). Only files with finished status will proceed to our THAI-NEST corpus.

4. CORPUS STATISTICS

A summary of NE tagging statistics can be generated from the NE Tagging Verification Tool. The reports include (1) the number of files, (2) the total and distinct number of tagged NEs, and (3) file size, with respect to year, month, publisher, and news category. Moreover, the user notes, logs, and errors can be viewed separately. Figure 5 illustrates a snapshot of the system showing the total and distinct number of tagged NEs.

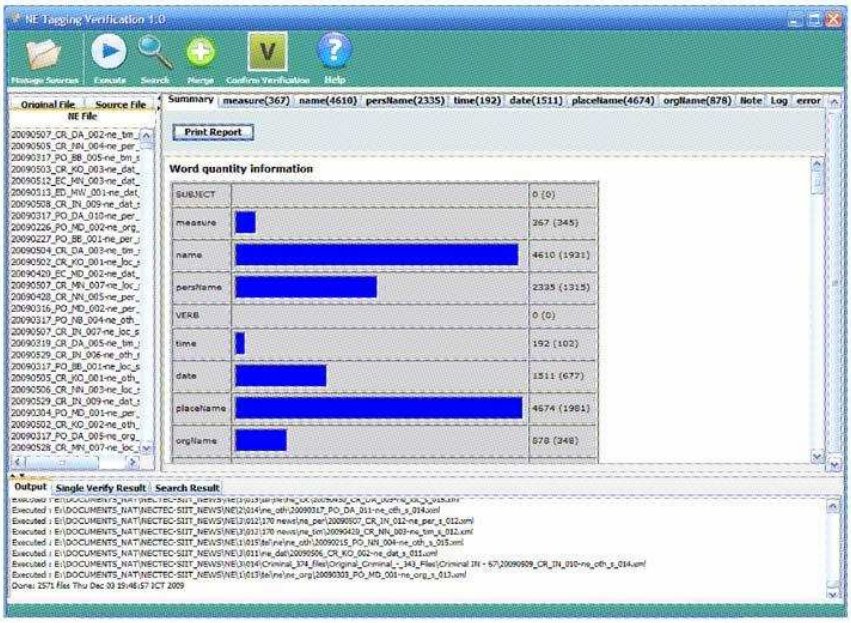

Figure 5: Snapshot of NE Statistics reported by NE Tagging Verification Tool

The current status of NE tagging process can be summarized in Table 1. To date, the numbers in the table are calculated from 7,520 tagged files.

Table 1: The number of NEs in the NE-Type level

NE-Type	#tokens	#unique tokens	#docs	#tokens/doc
<*persName*>	4,110	2,071	584	7.04
<*orgName*>	18,374	5,339	2,110	8.71
<*placeName*>	6,411	2,319	981	6.54
<*date*>	3,197	1,314	1,224	2.61
<*time*>	413	165	1,221	0.34
<*measure*>	3,068	2,348	339	9.05
<*name*>	9,482	3,842	1,061	8.94
Total	**45,055**	**17,398**	**7,520**	**43.22**

From Table 1, we observed that the largest average number of tokens per document (last column) is <*name*> – extended named entities, the smallest number is <*time*>. On average, the number of NEs found in each news article is 43.22. It is worth mentioning that the discrepancy in the number of tagged documents to date can be explained by the parallel process we have adopted (one annotator for each NE-type tag set). Due to the difference in annotators' experience and knowledge and task difficulty, the numbers of tagged documents varied among the NE types.

5. SUMMARY AND DISCUSSION

We proposed a new framework called THAI-NEST (THAI-Named Entities Specification and Tools) to support the NE corpus construction from Thai news articles. Since the project time frame is set within one year, the framework was carefully designed to allow efficient resource allocation and usage (i.e., time and humans). In this framework, three issues were considered: (1) a specification for Thai NE tag set, (2) a tagging process, and (3) tagging tools. Our tag set specification was adapted from the TEI guidelines to suit the Thai language characteristics. The tagging process includes the verification step which allows the tagged corpora of different NE types to be merged and checked for any error. To allow the maximum efficiency, the tagging tools were designed according to the proposed tagging process.

It is worth noting that one of the common problems found during the tagging process is difficulties in recognizing various jargons and metonym from different domains. Therefore, annotators should be assigned based on their familiarity with the domain. Another problem is ambiguous cases between organization and place names. Determining these two NE types requires deeper interpretation of contextual information and background knowledge. As a

final step of this project, we hope to review each individual tagging process and its shortcomings and to provide some suggestions for a creation of other Thai corpora.

As for future works, we plan to train and evaluate models using available machine learning techniques such as the CRFs. We will provide an initial technical report on training the NER model by using the corpus. To promote the NER task for Thai language, we plan to make the corpus publicly available. Another plan is to extend the use of this corpus by including another tag set of 4W1H (*Who*, *What*, *When*, *Where*, and *How many*). The 4W1H tag set can be partially transformed from the current NE tag set with further modification. For example, person names from the NE tag set can be mapped into *Who* of the 4W1H tag set. In addition, we will include a tag set of *subject* and *verb* to provide the *action* part of sentences. The newly designed 4W1H plus *subject* and *verb* tag set would be very useful for implementing a QA system from news articles.

REFERENCES

Barnard, D.T. & Ide, N.M. (1997). The text encoding initiative: flexible and extensible document encoding. *Journal of the American Society for Information Science*. 48(7), 622-628.

Chanlekha, H. & Kawtrakul, A. (2004). Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information, *Proceedings of the IJCNLP 2004*. (pp. 49–55).

Chinchor, N.A. (1998). Overview of Proceedings of the Seventh Message Understanding Conference (MUC-7)/MET-2, *Proceedings of the Seventh Message Understanding Conference (MUC-7)*. (pp. 5).

Ekbal, A. & Bandyopadhyay, S. (2008). Development of Bengali Named Entity Tagged Corpus and its Use in NER Systems, *Proceedings of the IJCNLP 2008 workshop on Asian Language Resources*. (pp. 1-8).

Inyaem, U., Meesad, P., & Haruechaiyasak, C. (2009). Named Entity Techniques for Terrorism Event Extraction and Classification, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 175-179).

Kawtrakul, A., Collier, N., Takeuchi, K., Ono, K., Suktarachan, M., Chanlekha, H., Waiyamai, K. (2001). Collaboration on Named Entity Discovery in Thai Agricultural

Texts, *Proceedings of the 8th International Workshop on Academic Information Networks and Systems*. (pp. 77-82).

Kosawat, K., Boriboon, M., Chootrakool, P., Chotimongkol, A., Klaithin, S., Kongyoung, S., Kriengket, K., Phaholphinyo, S., Purodakananda, S., Thanakulwarapas, T., & Wutiwiwatchai, C. (2009). BEST 2009: Thai Word Segmentation Software Contest, *Proceedings the 8th International Symposium on Natural Language Processing*. (pp. 83-88).

Kumano, T., Kashioka, H., Tanaka, H., & Fukusima, T. (2003). Construction and analysis of Japanese-English broadcast news corpus with named entity tags, *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition*. (pp. 17-24).

Lertcheva, N. & Aroonmanakun, W. (2009). A Linguistic Study of Product Names in Thai Economic News, *Proceedings the 8th International Symposium on Natural Language Processing*. (pp. 26-29).

Linguistic Data Consortium (LDC). (2008). ACE (Automatic Content Extraction) *English Annotation Guidelines for Entities Version 6.6 2008.06.13*.

Linguistic Data Consortium (LDC). (2006). Simple Named Entity Guidelines Version 6.4 Thai. Retrieved from *http://projects.ldc.upenn.edu/LCTL/Specifications/SimpleNamedEntityGuidelinesV6.4-Thai.pdf*

Linguistic Data Consortium (LDC). (2006). Time Annotation Guidelines for Less Commonly Taught Languages: Thai (Based upon the TIMEX2 Standard) Version 1.0. Retrieved from *http://www.ldc.upenn.edu/Projects/LCTL/Specifications/TimeAnnotation GuidelinesV1. 0-Thai.pdf.*

Saha, S.K., Sarkar, S. & Mitra, P. (2008). Gazetteer Preparation for Named Entity Recognition in Indian Languages, *Proceedings of the IJCNLP 2008 workshop on Asian Language Resources*. (pp. 9-16).

Sangal, R., Misra Sharma, D., & Kumar Singh, A., Eds, *Proceedings of the IJCNLP 2008 workshop on Named Entity Recognition for South and South East Asian Languages*.

Sekine, S. Sekine's Extended Named Entity Hierarchy. Retrieved from *http://nlp.cs.nyu.edu/ene/.*

Sutheebanjerd, P. & Premchaiswadi, W. (2009). Thai Personal Named Entity Extraction without Using Word Segmentation or POS Tagging, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 221-226).

Suwanapong, T. & Theeramunkong, T. (2009). Aliases Discovered in Thai Sports News Articles, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 63-66).

Takenobu, T., Calzolari, N., Huang, C. R., Prevot, L., Sornlertlamvanich, V., Monachini, M., YingJu, X., Kiyoaki, S., Charoenporn, T., Soria, C., & Hao, Y. (2006). Infrastructure for standardization of Asian language resources, *Proceedings of the COLING/ACL on Main conference poster sessions*. (pp. 827-834).

Tirasaroj, N. & Aroonmanakun, W. (2009). Thai Named Entity Recognition Based on Conditional Random Fields, *Proceedings of the 8th International Symposium on Natural Language Processing*. (pp. 216-220).

Tjong, K. S., Erik, F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition, *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003*. (pp. 142-147).

Tongtep, N. & Theeramunkong, T. (2008). Pattern-based Named Entity Extraction for Thai News Documents, *Proceedings of the 3rd International Conference on Knowledge, Information and Creativity Support Systems (KICSS'08)*. (pp. 82-89).

Tongtep, N. & Theeramunkong, T. (2009). A Feature-based Approach for Relation Extraction from Thai News Documents, *The Pacific Asia Workshop on Intelligence and Security Informatics 2009 (PAISI-09)*. (pp. 155-160).