Ontology-Based Terrorism Event Extraction

Uraiwan Inyaem
Faculty of Information Technology
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
uraiwaan@gmail.com

Choochart Haruechaiyasak
Human Language Technology Laboratory
National Electronics and Computer Technology Center
Pathumthani, Thailand
Choochart.Haruechaiyasak@nectec.or.th

Phayung Meesad
Faculty of Technical Education
King Mongkut's University of Technology North Bangkok
Bangkok, Thailand
pym@kmutnb.ac.th

Dat Tran
Faculty of Information Science and Engineering
University of Canberra
ACT, Australia
Dat.Tran@canberra.edu.au

Abstract— **The proliferations of terrorism news articles from thousands of different sources are now available on the Web. Summarization of such information is becoming increasingly important. The aim of this paper is to study and compare the linguistic feature methods that are appropriate for use in terrorism event extraction systems. The event extraction has a main function to named entity recognition and segments the terrorism events from news articles to display to the users. The research methodology in the paper compares many linguistic features techniques including the terrorism gazetteer, the terrorism ontology and terrorism grammar rule. The annotated entities are summarized into the three desired template events. The terrorism events are classified by using similarity measure based on Term Frequency-Inverse Document Frequency called TF-IDF-based event segmentation. Additionally, we use a finite state algorithm to learn these feature weights and also studied to emphasize the performance of the event extraction algorithms. The experimental results show that the terrorism ontology linguistic feature selection yielded the best performance with 85.15% for both precision and recall.**

Keywords; event extraction; terrorism ontology; feature selection; event segmentation;

I. INTRODUCTION

Thai people read news about terrorism events in the south of Thailand all the time from news online. News will typically include the title of the event, the date and time, the place, the occurrence event, and sometime a description. Unless the information is received as part of the decision support system for useful organization extracted, this data must be extracted into the decision support application manually. Similarly, finding information on a web page about the terrorism event must be entered manually into the decision support system. The extra time requires to do this data entry discourages users from recording the terrorism event all which reduces the usefulness of both the news and their decision support application.

This paper aims to examine partially automating this event extraction process. In our proposal, the user could select the entity of event and enters them automatically in the decision support system from the terrorism news articles. Grishman defines the task of information extraction that concerns the identification of instances of a particular class of events or relationships in a natural language text, and the extraction of the relevant arguments of the event of relationship as [1]. Most algorithms have been applied to English information extraction; relatively few have been made for Thai language. The top-ranked machine learning algorithms are suitable for the information extraction investigation, but may not be known to be the best Thai event extraction. This paper shows the terrorism ontology linguistic feature selection is the best approach, their results provide both precision and recall equal to 81.15%.

The rest of this paper is organized as follows. Section II gives related works on event extraction method and linguistic feature selection technique. Section III presents the framework for constructing the proposed Thai terrorism event extraction. Section IV describes performance measures in our empirical evaluation and outlines experiments and discussion detailing event extraction. Finally, Section V concludes the study and discusses future works.

II. RELATED WORKS

Related researchs in the event information extraction area fall under two main categories: a named entity recognizer and text summarizer to extract information. Angelo applied a named entity recognizer to first pre-process the data and then used an e-mail summarizer to extract attributes as [2]. However, we think this methodology is a difficult way to find the results. Goel and Wang used hand-coded pattern matcher to identify obvious attributes as [3]. When no matching pattern was found, a hidden Markov model as [4] was used. In their approach, the system produces fine accuracy.

The objective of our system is to study the named entity methodology including terrorism gazetteer, terrorism ontology and terrorism grammar rule. The tagged entities are summarized from each news article. The terrorism events are segmented into three categories by using similarity measure based on Term Frequency-Inverse Document Frequency called TF-IDF-based event segmentation as [5].

King Mongkut's University of Technology North Bangkok and
Rajamangala University of Technology Thanyaburi, Thailand

978-0-7695-3887-7/09 $26.00

Inyaem and et al. created training and testing corpus for Thai terrorism news article as [6]. The Thai terrorism news article corpus is manually tagged into three categories: (1) the occurrence event reporting article; (2) the find suspicious items reporting article; and (3) the arrest of suspects reporting article. The corpus is composed of 540 the occurrence events reporting article, 130 the find suspicious items reporting article and 80 the arrests of suspects reporting article.

Inyaem and et al. have defined three terrorism article patterns [6] as follows:

- *The occurrence events reporting article* is all news articles that contain occurrence terrorism event information clearly such as the date and time of occurrence event, terrorist, victim, tactic, weapon and place.
- *The find suspicious items reporting article* is all news articles which report the police to find the suspicious items in the area of three provinces such as Pattani, Yala, and Narathiwat as a part of the south of Thailand. The news article information contains the date and time of occurrence event, evidence and place.
- *The arrests of suspects reporting article* is all news articles which report the police to arrest the suspects in the area of three provinces such as Pattani, Yala, and Narathiwat as a part of the south of Thailand. The news article information contains the date and time of occurrence event, suspects, weapon and place.

III. PROPOSED SYSTEM FRAMEWORK

The architecture of the proposed Thai terrorism event extraction system is shown in Fig. 1. It can be viewed as a pipeline of processes that takes terrorism news articles corpus as input, and performs terrorism information event extraction process. Each step of the pipeline is discussed in more detail as follows.

A. Pre-processing

The terrorism news article is exported to a single text file in XML form. The file is tagged by hand and separated out into three separated files containing (1) occurrence event reporting, (2) found suspicious items reporting and (3) arrested suspects reporting respectively. The news article header and body are demarcated with XML tags such as <header> and <content>.

B. Linguistic Features Selection

The paper uses the linguistic features as the named entity type. Then, we compare and study linguistic features including terrorism gazetteer, terrorism ontology and terrorism grammar rule according to ANNIE rule-based recognizer as in [7]. The methods used in this study are described below.

1) Terrorism Gazetteer (TG): Firstly, the GATE graphic user interface tool as in [7] is used to create a terrorism gazetteer. The terrorism gazetteer consists of a set of lists complied into the finite state machines [8]. Its size is 36.5-kb. All entries have 19 types such as city, day, digit, district, evidence, the number of hunting, Level, month, occur, period, place, province, the number of terrorist, terrorist, title, the occupation of victim, victims, weapon and year.

Each list has attributes major type, minor type and language namely '*Amphur.lst:Location:Amphur:Thai'* This example sentence means *Amphur.lst* file is a list of terrorism gazetteer, *Location* attribute is major type, *Amphur* attribute is minor type and *Thai* is the Thai language consecutively. All attributes are used as input to Java Annotation Pattern Engine (JAPE) grammar [7]. The list entries may be entities or parts of entities, or they may contain context information for example *title* often indicate people.

2) Terrorism Ontology (TO): The Thai terrorism event extraction system also consists of terrorism ontology, which is developed by using GATE graphic user interface tool. The methodology starts with determining the scope of ontology for terrorism domain. The terrorism news article includes terrorist part, victim part, the date and time of occurrence event, place, evidence weapon, tactic and amount of victim and terrorist. Instead of reusing existing ontology but it is not possible in this Thai terrorism event extraction system so new ontology is created from Thai news articles. Then the expert domain analyzes the terms in terrorism domain by considering the overview of the specific domain. User defines and then implements the classes and instance hierarchy using GATE tool. Their properties and constraints are developed. The expert domain checks the final version again and then all instances are created in terrorism ontology system.

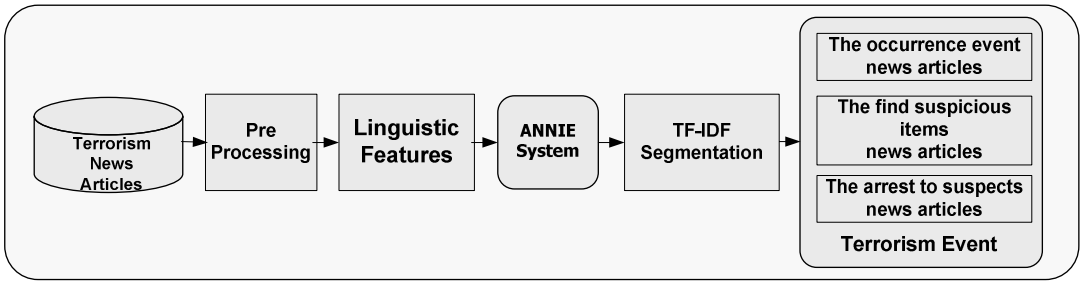

Figure 1. The proposed system framework of Thai terrorism event extraction

3) Terrorism Grammar Rule (TGR): The study uses Java Annotation Pattern Engine (JAPE) grammar in GATE tool for developing the terrorism grammar rule. JAPE provides the finite state transduction as in [8] over annotation based on regular expression. The JAPE phase runs sequentially and then constitutes a cascade of finite state transduction over annotations. The hand-coded rules are proposed for applied to annotation to identify named entities. The annotation comes from the format analysis, tokenizer and the gazetteer module. The Grammar rules can set their priorities based on pattern length, rule status and rule ordering. In this study, 19 grammar rules are created for the terrorism domain.

C. ANNIE System

This study uses the open-source called *A Nearly-New IE* (ANNIE). The ANNIE system is a part of General Architecture for Text Engineering (GATE) as in [7]. The ANNIE relies on finite state algorithms and the JAPE language. Their components form is a pipeline. The methodology start with the ANNIE system for named entity recognizing after the pre-processing and linguistic feature process respectively

D. TF-IDF-based Event Segmentation

In order to classify different terrorism news articles into the three categories described in previous section. The measure algorithm uses a similarity measure based on Term Frequency-Inverse Document Frequency (TF-IDF) as in [5]. This paper reproduced TF-IDF measure their method which defined in the paragraph as follow.

Let N be a news article. The study are trying to classify represented by a word-frequency vector F(N). Each component F(N, t) represents the frequency of token word t in news article N. This definition can compute the weight for a given category, C, as follows.

Let $TT(C,w)$ be the number of times that word w occurs in any news article. Some news belongs to category C divided by the total number of words in category C. Let $TT(Tr,w)$ be the number of times that word w occurs in any training news article.

Their news divided by the total number of words across all training news articles. Thus we can compute the *term frequency* as (1).

$$TF(C,w)\frac{TT(C,w)}{TT(Tr,w)} \tag{1}$$

Using this term frequency and the *document frequency* $DF(w)$ that define to be the number of categories in which the word w occurs at least once divided by the total number of categories, this can compute the weight, $W(C,w)$ as (2).

$$W(C,w)=\frac{TF(C,w)}{DF(w)^2} \tag{2}$$

Using this weight can compute a similarity measure, $SIM(N,C)$ for the given news article N to a given category C as in (3).

$$SIM(N,C)=\frac{\sum_{w\in N} C(N,w)W(C,w)}{\min\left(\sum_{w\in N} C(N,w),\sum_{w\in N} W(C,w)\right)} \tag{3}$$

When attempting to classify a news article, the category that produces the largest similarity score was selected.

E. Information Extraction

This task consists of the three subprocesses: (1) named entities recognition (2) computed the largest similarity score of each entity for news segmentation and (3) extracted the terrorism events. In order to extract event information from news articles, we compared the three linguistic feature techniques in named entities process. The proposed features were terrorism gazetteer, terrorism ontology, and terrorism grammar rules. The annotated entities of each news article were computed the largest similarity score using TF-IDF measure. The largest score of entity types were classified into the desired slot pattern. The terrorism event were extracted and stored in the corpus using hand-code file.

IV. Results

This part is divided into two parts; first part describes the performance measures for our experiments in detail. Second part describes the experimental result values and discussion respectively.

A. Performance measures

Our experiments are evaluated by comparing the summaries generated by human experts for the same test set of previously unseen texts. The comparison is performed using an automated scoring program that rates each system according to precision and recall measures as in [2].

Precision measures the reliability of the information extracted that shown in (4).

Recall measures the amount of the relevant information that the natural processing language system correctly extracts from the test dataset that shown in (5).

$$precision = \frac{\#\ correct\ slot\ fillers\ in\ output\ templates}{slot\ fillers\ in\ output\ templates} \tag{4}$$

$$recall = \frac{\#\ correct\ slot\ fillers\ in\ output\ templates}{slot\ fillers\ in\ answer\ keys} \tag{5}$$

B. *The experimental results and discussion*

The study was performed by using Thai news article corpus from previous work results as in [6]. The corpus consists of three news categories. Each category is further partitioned into training and test datasets. The number of the training documents is 900 news articles and the test dataset is 600 news articles. This task used the information extraction tool called ANNIE System of a part of GATE application to perform all the experiments in our work. In the experiments, we compare the linguistic features techniques using the default setting of ANNIE system with our proposed linguistic feature techniques. The performance measures for evaluating the event information extraction are precision and recall measure, as mentioned in previous section. Three algorithms: terrorism gazetteer, terrorism ontology, and terrorism grammar rule, were tested by using training set option. The results in terms of recall and precision measures were averaged across all training set experiments. The experimental results are summarized in Table I. The evaluating results of the event information extraction algorithms with terrorism ontology linguistic feature give higher performance,, than terrorism gazetteer and terrorism grammar rule linguistic features.

TABLE I. THE SUMMARY RESULTS FOR THE EXPERIMENT

Method	Recall	Precision
TG *	31.70	31.70
TG #	84.90	83.20
TO*	32.01	25.21
TO #	**85.15**	**85.15**
TGR*	32.18	26.56
TG R#	76.30	76.30

TG is Terrorism Gazetteer; TO is Terrorism Ontology; TGR is Terrorism Grammar Rules;
* is evaluated by default setting in ANNIE system
# is evaluated by several linguistic feature selections

V. CONCLUSION

In this paper report, the framework of Thai terrorism event extraction is proposed. The in the implementation, we compared several linguistic features selections to find the most suitable one for applying in the terrorism event extraction systems. The event extraction system has been developed to prove the concept. It has a main function to named entity and segment terrorism event from terrorism news article for displays to the users. Many linguistic features that are terrorism gazetteer, terrorism ontology, and terrorism grammar rule were studied and compared. The annotated entities were summarized into the three desired template event. The terrorism events were classified by using similarity measure called TF-IDF based event segmentation. The results from the experiments study showed that the terrorism ontology linguistic feature is the best approach with precision and recall equal to 85.15%.

ACKNOWLEDGMENT

This research is funded by KMUTNB (King Mongkut's University of Technology North Bangkok) in Thailand, RMUTT (Rajamangala University of Technology Thanyaburi) in Thailand and UC (University of Canberra) in Australia. The author would like to thank Asst.Prof.Dr.Phayung Meesad, Dr.Choochart Haruechaiyasak and Dr.Dat Tran for discussions and advice, and the anonymous reviewers for their helpful comments.

REFERENCES

[1] R. Grishman, "Information extraction: techniques and challenges," Springer-Verlag London, Vol.1299, pp.10-27, 1997.

[2] D. Angelo, "Automated email integration with personal information management applications," CLUK2004, in press.

[3] K. Goel, and P.C. Wang, "Automated extraction of event details from text snippets," unpublished.

[4] L. R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition," Proc. IEEE, 77(2). pp.257-286, 1989.

[5] J. A. Black, and N. Ranjan, "Automated event extraction from email," unpublished.

[6] U. Inyaem, P. Meesad, and C. Haruechaiyasak, "Domain knowledge based information filtering for terrorism news articles," NCCIT2008, in press.

[7] H. Conningham, K. Bontcheva, V. Tablan, and D. Maynard, "General architecture for text engineering, " unpublished.

[8] D. Maynard, "Finite State Transduction for information extraction and other tasks: ANNIE, JAPE-Part I," unpublished.