A Framework of NLP Based Information Tracking and Related Knowledge Organizing with Topic Maps

Asanee Kawtrakul[1], Chaiyakorn Yingsaeree[1], and Frederic Andres[2]

[1] NAiST Research Laboratory, Kasetsart University, Bangkok, Thailand
asanee.kawtrakul@nectec.or.th, chaikorn@gmail.com
[2] National Institute of Informatics, Tokyo, Japan
andres@nii.ac.jp

Abstract. This paper presents a computational framework for information extraction and aggregation which aims to integrate and organize the data/information resources that spread throughout the Internet in the manner that makes them useful for tracking events such as natural disaster, and disease dispersion. We introduce a simple statistical information extraction technique for summarizing the document into a predefined structure. We apply the topic maps approach as a semantic layer in aggregating and organizing the extracted information for smart access. In addition, this paper also carries out a case study on disease dispersion domain using the proposed framework.

1 Introduction

In order to monitor important events, such as disease dispersion, the occurrence of tsunami and terrorism connections, an operator and a decision maker need data, information and knowledge. Internet news and other online articles (e.g. wiki-like knowledge and web logs) are the good resources for these kinds of information which describe the world around us rapidly by talking about the update events, states of affairs, knowledge, people and experts who participate in. However, sources of these data are scattered across several locations and web sites with heterogeneous formats that offer a large volume of unstructured information. Moreover, the needed knowledge was too difficult to find since the traditional search engines return ranked retrieval lists that offer little or no information on semantic relationships among those scattered information, and, even if it was found, the located information often overload since there was no content digestion. Accordingly, the automatic extraction of information expressions, especially the spatial and temporal information of the events, in natural language text with question answering system has become more obvious as a system that strives for moving beyond information retrieval and simple database query.

However, one major problem that needs to be solved is the recognition of events which attempts to capture the richness of event-related information with their temporal and spatial information from unstructured text. Various advanced technologies including name ehintities recognition and related information extraction, which need natural language processing techniques, and other information technologies, such as GIS, are utilized to enable emerging of new methodologies for

Z. Kedad et al.(Eds.): NLDB 2007, LNCS 4592, pp. 272–283, 2007.

information extraction and aggregation with problem-solving solutions (e.g. the know-how from livestock experts from countries with experiences in handling bird flu situation). Ontology and Topic Map model are also applied for organizing related knowledge or related topics.

In this paper, we present a systematic attempt to provide a computational framework for information extraction and aggregation which aims to integrate and organize the data/information resources dispersed across web resources in a manner that makes them useful for tracking events such as natural disaster, and disease dispersion. The remainder of this paper is structured as follows: Section 2 describes the nontrivial problems in information tracking; Section 3 gives the conceptual framework for information collection, extraction and aggregation including the information service for different target user groups. Section 4 gives more details of the system process regarding the information extraction module. Section 5 discusses the knowledge service and visualization module. Finally, in Section 6, we conclude and discuss the next step and challenges.

2 Non-trivial Issues in Information Tracking

Lessons learned from special monitoring areas or areas that has past experiences with the interested events (e.g. the best practice for governments to handle bird flu situation), the collection of important events and their related information (e.g. virus transmission from one area to other locations and from livestock to humans) are important. However, collecting and extracting these data from the Internet have two main nontrivial problems: overload and scattered information, and salient information extraction from unstructured text.

2.1 Overloaded and Scattered Information

The knowledge applicable to an intended problem solving consists of data items and/or information that are organized and processed to convey understanding, experience, accumulated learning, and expertise. However, sources of these data are scattered across several locations and websites with heterogeneous formats. For example, the information about Bird Flu consisting of policy for controlling the events, disease infection management, and outbreak situation may appear in different websites as shown in Fig. 1. Consequently, collecting the needed information from scattered resources is very difficult since the semantic relations among those resources are not directly stated. Although we can gather those information, the collected information often overload since there is no content digestion. Accordingly, manually solving those problems will consume a lot of time and power, and the system that can collect, extract and organize those information automatically will definitely become a useful tool for knowledge construction and organization.

2.2 Salient Information Extraction from Unstructured Texts

In order to reduce time consumption for users to consume the information, only salient information must be extracted. As it happens, most of those information, such as time of the event, location that event occurred, and the detail of the event, are left

implicitly in the texts. For example: in the text in Fig. 2, the time expression "15 February" mentioned only "date and month" of the bird flu event but did not mention the 'year'. The patient and her condition (i.e. '37-year-old female', and 'died') was caused by bird flu which is written in the text as 'Avian influenza' and 'H5N1 avian influenza'. Accordingly, the essential component of computational model for event information capturing is the recognition of interested entities including time expression, such as 'yesterday', 'last Monday', and 'two days before', which becomes an important part in the development of more robust intelligent information system for event tracking.

Information extraction in traditional way extracts a set of related entities in the format of slot and filler, but the description of information in Thai text such as locations, patient's condition, and time expressions can not be limited to a set of related entities because of the problems of using zero anaphora [1]. Moreover, to activate the frame for filling the information, name entity classification must be robust as it has been shown in [2].

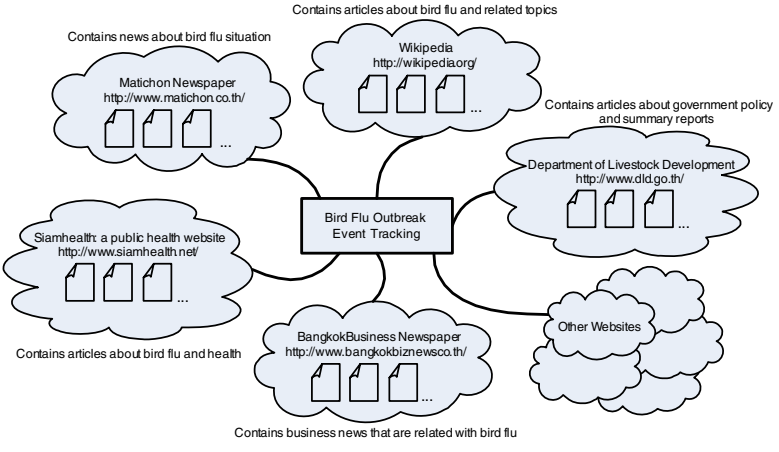


Fig. 1. The information required for tracking bird flu outbreak is scattered across the Internet

Avian influenza - situation in Egypt - update 5

16 February 2007

The Egyptian Ministry of Health and Population has confirmed the country's 13th death from H5N1 avian influenza. The 37-year-old female whose infection was announced on 15 February, died today.

Fig. 2. The example of the document containing bird flu outbreak situation

3 A Framework of Information Extraction for Event Tracking

A crucial first step in the automatic extraction of information from unstructured texts is the capacity to identify what events are being described and to make it explicit when these events occurred. Since the web consists of a large extent of unstructured

or semi-structured natural language text, several techniques (i.e., language engineering and knowledge engineering) are applied to information extracting and integrating. For language engineering, word segmentation [3], named entity recognition [2], shallow parsing [4], shallow anaphora resolution and discourse processing [2,5,6] are utilized. For knowledge engineering, the concept of frame for structuring the extracted information is applied. For ontological engineering, task-oriented ontology, ontology maintenance [7] and Topic Maps [8] model are applied for information aggregation and organizing for smart access. Fig. 3 overviews the system architecture which has been designed for event tracking and its related knowledge organization for aiding multi-users information service provision [7,9,10]. The framework consists of six main parts:

Information and Knowledge Extraction: To generate useful knowledge from collected documents, two important modules, information extraction and knowledge extraction, are utilized. Ontological topic maps are used as a knowledge base to facilitate the knowledge construction process. The information extraction and integration module is responsible for summarizing the document into a predefined frame-like/structured database, such as <disease name, dispersion location and time, status of patient's condition>. The knowledge extraction and generalization is responsible for extracting useful knowledge (e.g. general symptom of disease) from collected document. The extracted knowledge is represented as a structured knowledge and rules. The output of both modules is stored in RDF/OWL repository.

Distributed Information Collection: The information, both unstructured and semi-structured documents are gathered from many sources. Periodic web crawler and HTML Parser [11] are used to collect and organize related information. The domain specific parser [12] is used to extract and generate meta-data (e.g. title, author, and date) for interoperability between disparate and distributed information. The output of this stage is stored in the document warehouse.

Content-based Metadata Extraction: To organize the information scattered at several locations and websites, Textual Semantics Extraction [10] is used to create a semantic metadata for each document stored in the document warehouse. Guided by the ontology stored in Ontological Topic Map, the extraction process can be taught of as a process for assigning a topic to considered documents.

Knowledge Organization: After all required information and knowledge is generated, the Topic Map (ISO ISO13250) including topics and related associations is a proxy to access resource occurrences. The generation is done by combining the ontological topic map and the metadata extracted from content-based metadata extraction. The generated topic map is represented as a XTM document and, then, sent to the Knowledge Visualization module.

Knowledge Service: This module is responsible for creating response to users' query. The query processing is used to interact with the RDF/OWL Knowledge Repository, while inference engine is used to infer new knowledge that is not explicitly stored in the knowledge repository.

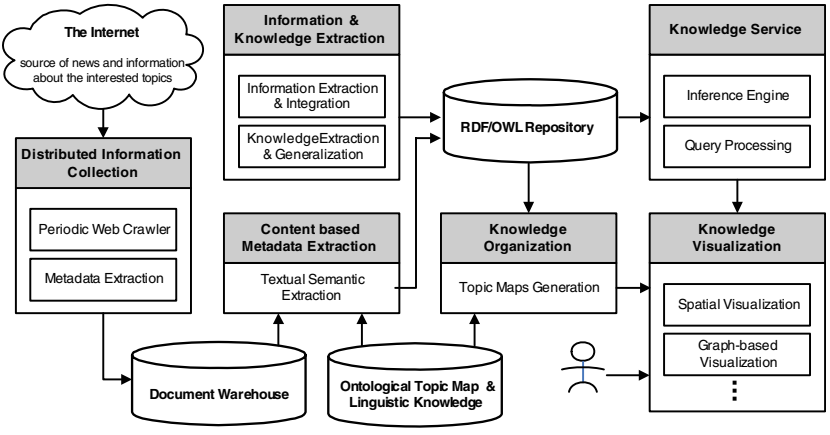


Fig. 3. The architecture of the proposed framework

Knowledge Visualization: After obtaining all required information from the previous module, the last step is to provide the means to help users consume that information in an efficient way. To do this, many visualization functions is provided. For example, Spatial Visualization can be used to visualize the information extracted from the Information Extraction module and Graph-based Visualization can be used to display hierarchal categorization in the topic maps in an interactive way [10].

Due to page limitation, this paper will focus in only Information Extraction module, Knowledge Service module and Knowledge Visualization module.

4 Information Extraction

The proposed model for extracting information from unstructured documents consists of three main components, namely Entity Recognition, Relation Extraction, and Output Generation, as illustrate in Fig. 4. The Entity Recognition module is responsible for locating and classifying atomic elements in the text into predefined categories such as the names of diseases, locations, and expressions of times. The Relation Extraction module is responsible for recognizing the relations between entities recognized by the Entity Recognition module. The output of this step is a graph representing relations among entities where a node in the graph represents an entity and the link between nodes represents the relationship of two entities. The Output Generation module is responsible for generating the n-tuple representing extracted information from the relation graph. The details of each module are described as followed.

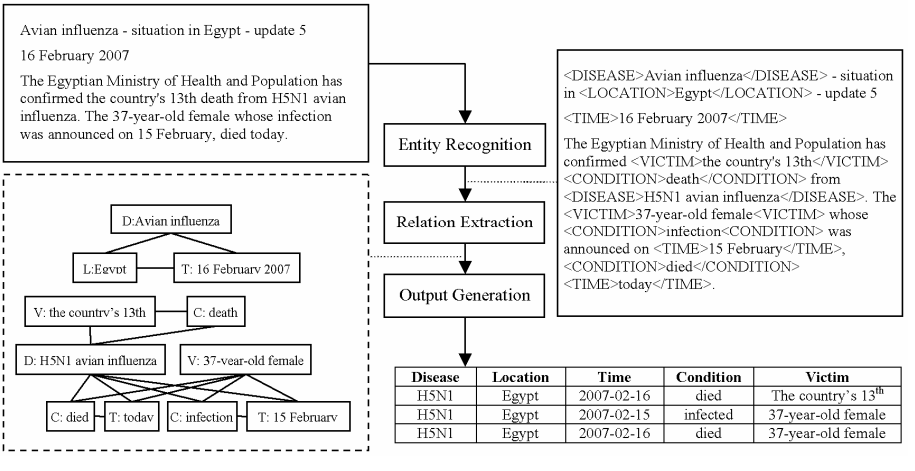
4.1 Entity Recognition

To recognize an entity in the text, the proposed system utilizes the work of H. Chanlekha and A. Kawtrakul [2] that extracts entity using maximum entropy [13],

heuristic information and dictionary. The extraction process consists of three steps. Firstly, the candidates of entity boundary are generated by using heuristic rules, dictionary, and statistic of word co-occurrence. Secondly, each generated candidate is then tested against the probability distribution modeled by using maximum entropy. The features used to model the probability distribution can be classified into four categories: Word Features, Lexical Features, Dictionary Features, and Blank Features as described in [2]. Finally, the undiscovered entity is extracted by matching the extracted entity against the rest of the document. The experiment with 135,000 words corpus, 110,000 words for training and 25,000 words for testing, shown that the precision, recall and f-score of the proposed method are 87.60%, 87.80%, 87.70% respectively.

4.2 Relation Extraction

To extract the relation amongst the extracted entities, the proposed system formulates the relation extraction problem as a classification problem. Each pair of extracted entity is tested against the probability distribution modeled by using maximum entropy to determine whether they are related or not. If they are related, the system will create an edge between the nodes representing those entities. The features used to model the probability distribution are solely based on the surface form of the word surrounding the considered entities; specifically, we use the word n-gram and the location relative to considered entities as features. The surrounding context is classified into three disjointed zone: prefix, infix, and suffix. The infix is further segmented into smaller chunks by limiting the number of words in each chunk. For example, to recognize the relation between VICTIM and CONDITION in the sentence "The [VICTIM] whose



Disease	Location	Time	Condition	Victim
H5N1	Egypt	2007-02-16	died	The country's 13th
H5N1	Egypt	2007-02-15	infected	37-year-old female
H5N1	Egypt	2007-02-16	died	37-year-old female

Fig. 4. Overview of the information extraction module

[CONDITION] was announced on ....", the prefix, infix and suffix in this context is 'the', 'whose', and 'was announced on ....' respectively.

To determine the best parameter of the system, we conduct the experiment with 257 documents, 232 documents for training and 25 documents for testing. We vary the n-gram parameter from 1 to 7 and set the number of words in each chunk as 3, 5, 7, 10, 13, and 15. The result is illustrated in Fig. 5. The legend in the figure is the number of words in each chunk; for example, WLLTY3 means that the number of words is set as 3. The evident shows that f-score is maximum when n-gram is 4 and number of words in each chunk is 7. The precision, recall and f-score at the maximum f-score are 58.59%, 32.68% and 41.96% respectively.

4.3 Output Generation

After obtaining a graph representing relations between extracted entities, the final step of information extraction is to transform the relation graph into the n-tuple representing extracted information. Heuristic information is employed to guide the transformation process. For example, to extract the information about disease outbreak (i.e. disease name, time, location, condition, and victim), the transformation process will starts by analyzing the entity of the type condition, since each n-tuple can contain only one piece of information about the condition. It then travels the graph to obtain all entities that are related to considered condition entity. After obtaining all related entities, the output n-tuple is generated by filtering all related entities using constrain imposed by the property of each slot. If the slot can contains only one entity, the entity that has the maximum probability will be chosen to fill the slot. In general, if the slot can contain up to n entities, the top-n entities will be selected. In addition, if there is no entity to fill the required slot, the mode (most frequent) of the entity of that slot will be used to fill instead. The time expression normalization using rule-based system and synonym resolution using ontology are also performed in this step to generalize the output n-tuple. The example of the input and output of the system are illustrated in Fig. 4.

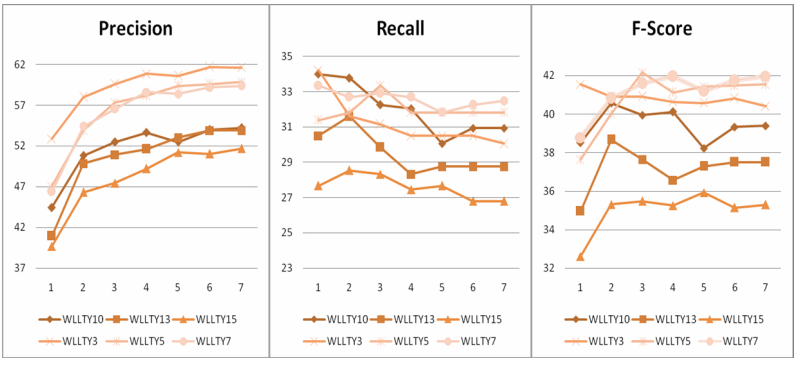


Fig. 5. Experimental results of relation extraction module

5 Knowledge Service and Visualization

After all required information and knowledge is generated and stored in the Ontological Topic Map and RDF/OWL Repository, users can consume those information and knowledge by using Knowledge Visualization module which combined the extracted information by interacting with Knowledge Service module and the topic maps model by interacting with Knowledge Organization module to generate visualizations that helps users consume the information in an efficient way. The details of Knowledge Service and Knowledge Visualization module are described as followed.

5.1 Knowledge Service

The Knowledge Service module is responsible for interacting with RDF/OWL Repository to generate the response to user's request. The framework currently supports four types of query. The detail of each query type is summarized in Table 1.

5.2 Knowledge Visualization

The Knowledge Visualization is responsible for representing the extracted information and knowledge in an efficient way. For example, we can create many visualization techniques that response to users who need concise and knowledge organization, such as Spatial Visualization and Graph-based Visualization as described below.

Spatial Visualization
The spatial-based visualization functions help users to visualize the extracted information (e.g. the bird flu outbreak situation extracted in Fig. 4.) using geographical information system, such as Google Earth. This kind of visualization allows the users to click on the map to get the outbreak situation of the area that they want. In addition, by viewing the information in the map users can see the spatial relations amongst the outbreak situations easier than without the map. The Google Earth integrated system for visualizing the extracted information about bird flu situation is shown in Fig. 6.

Graph-based Visualization
The graph-based visualization function is useful to show the global structure of topic maps and relations between different nodes in a 3D visual space. In a topic maps structure, various topics are associated to each other based on relationships. The graph viewer provides a better global understanding of the content by exploring through graph nodes. The kind of intuitive visualization of the Topic Maps allows browsing through all the topics and related relationships defined in the Topic Maps as shown in Fig. 7. The graph can be moved and restructured along its topological view according to the user's need.

Table 1. Detail of four query types supported by the Knowledge Service module

Query type	Description
Query by Object	A mechanism employed when users know the object but want to acquire more information/knowledge about it. The query example is as following: *SELECT qa_who,lblWho* *FROM {qa_who} ne:text {lblWho}* *WHERE (lblWho like "*เด็ก*")* *USING NAMESPACE ne = <http://naist.cpe.ku.ac.th/EventTracking#>*
Query by Relationship	A mechanism employed when users know the relation label. For example, user can access knowledge repository such as "ne:atLocation". The query example is as following: *SELECT disease,lblDisease,location,lblLocation* *FROM {disease} ne:text {lblDisease},* *{disease} ne:atLocation {location},* *{location} ne:text {lblLocation}* *WHERE (lblDisease like "*หวัดนก*") AND (lblLocation like "*เวียต*")* *USING NAMESPACE ne = <http://naist.cpe.ku.ac.th/EventTracking#>*
Query by Instance	A mechanism employed when users know the instance or some parts of instance label that can access to knowledge repository such as "ne:Disease-3-10". The query example is as following: *SELECT disease,lblDisease,location,lblLocation* *FROM {disease} ne:text {lblDisease},* *{disease} ne:atLocation {location},* *{location} ne:text {lblLocation}* *WHERE (lblDisease like "*หวัดนก*") AND (lblLocation like "*เวียต*")* *USING NAMESPACE ne = <http://naist.cpe.ku.ac.th/EventTracking#*
Knowledge Reasoning	A mechanism used for inferring new knowledge from existing information/knowledge by using inference engine provided by OWLIM plug-in. The custom designed rule set is required to create new knowledge from existing one. For example, to generate new knowledge about the region that have bird flu situation, one can custom rule sets as following: *RegionRules {* *Id : event_tracking* *Location <ne:locationOf> District* *District <ne:districtOf> Province* *Province <ne:ProvinceOf> Country* *Country <ne:CountryOf> Region* *...* *}* *SELECT disease, lblDisease, region, lblRegion* *FROM {location} ne: inRegionOf {region},* *{region} ne:text {lblRegion}* *WHERE (lblRegion like "เอเชียตะวันออกเฉียงใต้")* *USING NAMESPACE ne = <http://naist.cpe.ku.ac.th/EventTracking#>*

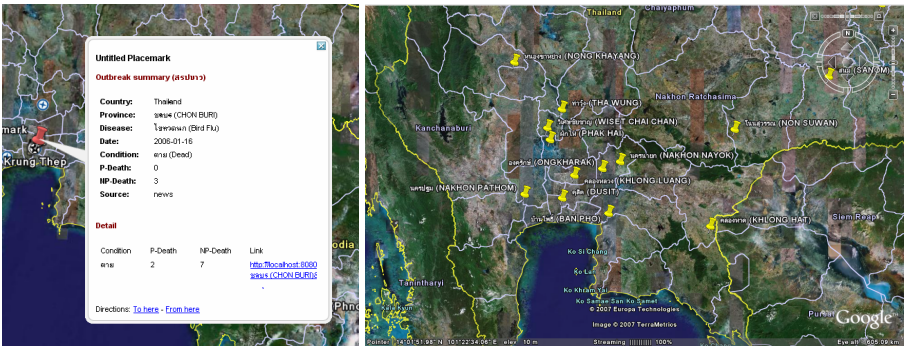


Fig. 6. Google Earth visualization for bird flu outbreak tracking

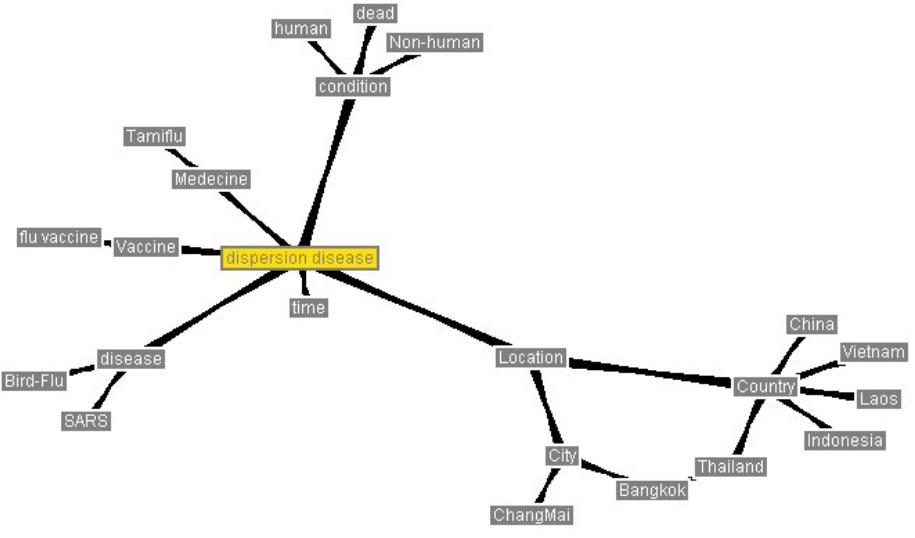


Fig. 7. Graph-based visualization of topic maps about dispersion disease

6 Related Work

The framework described in this paper is closely related to ProMED-PLUS [16], a system for the automatic "fact" extraction from plain-text reports about outbreaks of infectious epidemics around the world to database, and MiTAP [17], a prototype SARS detecting, monitoring and analyzing system. The difference between our framework and those systems is that we also emphasize on generating the semantic relations among the collected resources and organizing those information by using topic map model.

The proposed information extraction model that formulates the relation extraction problem as a classification problem is motivated by the work of J.Suzuki et. al. [19]

that proposed a HDAG kernel to solve many problems in natural language processing. The use of classification methods in information extraction is not new. Intuitively, one can view the information extraction problem as a problem of classifying a fragment of text into a predefined category which results in a simple information extraction system such as a system for extracting information from job advertisements [20] and business cards [21]. However, those techniques require the assumption that there should be only one set of information in each document, while our model could support more than one set of information.

7 Conclusion and Future Work

This paper presents a framework for extracting information and knowledge from unstructured documents that spread throughout the Internet by emphasizing on information extraction technique, event tracking and knowledge organizing. The work is going to develop the Textual Semantic Extraction for providing the automated topic maps construction process. This challenging work needs more complicate natural language processing with deeply semantic relations interpretation.

Acknowledgement

The work described in this paper has been supported by the grant of National Electronics and Computer Technology Center (NECTEC) No. NT-B-22-14-12-46-06, under the project “A Development of Information and Knowledge Extraction from Unstructured Thai Document”.

References

1. Kongwan, A., Kawtrakul, A.: Know-what: A Development of Object Property Extraction from Thai Texts and Query System. In: Proceedings of SNLP-2005, Bangkok, Thailand pp. 157–162,(2005)
2. Chanlekha, H., Kawtrakul, A.: Thai Named Entity Extraction by incorporating Maximum Entropy Model with Simple Heuristic Information. In: Proceedings of IJCNLP-2004, Hainan, China pp. 49–55 (2004)
3. Sudprasert, S., Kawtrakul, A.: Thai Word Segmentation based on Global and Local Unsupervised Learning. In: Proceedings of NCSEC-2003. Chonburi, Thailand (2003)
4. Satayamas, V., Thumkanon, C., Kawtrakul, A.: Bootstrap Cleaning and Quality Control for Thai Tree Bank Construction. In: Proceedings of NCSEC-2005, Bangkok Thailand pp. 849–860 (2005)
5. Grosz, B., Joshi, A., Weinstein, S.: Centering: A Framework for Modeling the Local Coherence of Discourse. Computational Linguistics 21, 203–225 (1995)
6. Chareonsuk, J., Sukvakree, Y., Kawtrakul, A.: Elementary Discourse unit Segmentation for Thai using Discourse Cue and Syntactic Information. In: Proceedings of SNLP. Bangkok, Thailand, pp. 85–90 (2005)

7. Kawtrakul, A., Suktarachan, M., Imsombut, A.: Automatic Thai Ontology Construction and Maintenance System. In: Proceedings of Ontolex Workshop on LREC, pp. 68–74 (2004)
8. Biezunski, M., Bryan, M., Newcomb, S.: ISO/IEC JTC1/SC34 (May 22, 2002) Available at http://www1.y12.doe.gov/capabilities/sgml/sc34/document/0322.htm
9. Kawtrakul, A., Imsombut, A., Thunyakijjanukit, A., Soergel, D., Liang, A., Sini, M., Johannsen, G., Keizer, K.: Automatic Term Relationship Cleaning and Refinement for AGROVOC. In: Proceedings of EFITA/WCCA. Vila Real, Portugal (2005)
10. Rajbhandari, S., Andres, F., Naito, M., Wuwongse, V.: Topic Management in Spatial-Temporal Multimedia Blog. In: the 1st IEEE International Conference on Digital Information Management (ICDIM 2006) Bangalore, India, December 6-8, 2006, pp. 81–88 (2006)
11. Thamvijit, D., Chanlekha, H., Sirigayon, C., Permpool, T., Kawtrakul, A.: Know-who: Person Information from Web Mining. In: Proceedings of NCSEC. Bangkok, Thailand, pp. 849–860 (2005)
12. Kawtrakul, A., Yingsaeree, C.: A Unified Framework for Automatic Metadata Extraction from Electronic Document. In: Proceedings of The International Advanced Digital Library Conference. Nagoya, Japan (2005)
13. Berger, A.L., Pietra, S.-A.D., Pietra, V.-J.D.: A maximum entropy approach to natural language processing. Computational Linguistics 22, 39–71 (1996)
14. Dean, M., Connolly, D., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F., Stein, L.A.: OWL Web Ontology Language Reference. W3C Recommendation (February 10, 2004) Available at http://www.w3.org/TR/owl-ref/
15. Garshol, L.M.: Living with Topic Maps and RDF (May 2003) Available at http://www.idealliance.org/papers/dx_xmle03/papers/02-03-06/02-03-06.html
16. Yangarber, R., Jokipii, L., Rauramo, A., Huttunen, S.: Information Extraction from Epidemiological Reports. In: Proceedings of HLT/EMNLP-2005. Canada (2005)
17. Damianos, L., Bayer, S., Chisholm, M.A., Henderson, J., Hirschman, L., Morgan, W., Ubaldino, M., Zarrella, J.: MiTAP for SARS detection. In: Proceedings of the Conference on Human Language Technology. Boston, USA, pp. 241–244 (2004)
18. Naito, M., Andres, F.: Application Framework Based on Topic Maps. In: TMRA 2005. Leipzig, Germany (2005) LNCS, vol. 3873, February 2006 pp. 42–52, DOI 10.1007/11676904_4, Charting the Topic Maps Research and Applications Landscape: First International Workshop on Topic Map Research and Applications, TMRA 2005, Leipzig, Germany, October 6-7, 2005, Revised Selected Papers Editors: Lutz Maicher, Jack Park ISBN: 3-540-32527-1
19. Suzuki, J., Sasaki, Y., Maeda, E.: Kernels for structured natural language data. In: Proceeding of NIPS 2003 (2003)
20. Zavrel, J., Berck, P., Lavrijssen, W.: Information extraction by text classification: Corpus mining for features. In: Proceedings of the workshop Information Extraction meets Corpus Linguistics. Athens, Greece (2000)
21. Kushmerick, N., Johnston, E., McGuinness, S.: Information extraction by text classification. In: Proceedings of IJCAI-2001 Workshop on Adaptive Text Extraction and Mining (2001)