A Case Study of NLP Technology for Cultural Information Management: M-Culture

Watchira Buranasing[1] Sineenat Tiekouw[1] Sapa Chanyachathchawan[1]

Thanate Muangthong[1] Pisan Taesuwan[1] Virach Sornlertlamvanich[1]

Thatsanee Chalernporn[1]

[1]National Electronics and Computer Technology Center, NSTDA, Patumthani, Thailand
{watchira.bur, sineenat.tie, sapa.cha, thanate.mua, pisan.tae, virach.sor, thatsanee.cha}@nectec.or.th

Abstract

In this paper, we present Thai cultural knowledge center or m-culture, which is a collaborative project of National Electronics and Computer Technology Center and Ministry of Culture, as a case study describing Thai word segmentation process based on Thai Character Cluster algorithm.

We continuously tested the process with more than 100,000 cultural data records for over 2 years and found that it is able to enhance the website service performance and search engine quality: reliable, precise, and highly effective.

According to the result mentioned above, the extracted keywords generally have superior quality than legacy method, which is formerly evaluated by experts from Ministry of Culture.

Keywords: Thai word segmentation, NLP tools, cultural information, m-culture, Thai Culture Knowledge Center

1 Introduction

Thai Cultural Knowledge Center Website[6] is a cultural archive project; it has been implemented through close cooperation between National Electronics and Computer Technology Center and Ministry of Culture, which is under the 2011 Memorandum of Understanding (MOU). In the first phase of the project was to develop technology base for collections and management of cultural data, which was established standards and guidelines for its acquisition, digitization, documentation, preservation, security and management. In the second phase, the project focuses on data integration involves combining data from several disparate sources, which are stored using various technologies and provide a unified view of the data. The content database associates with person, organization, place and artifact. There are more than 100,000 records since November 2010 to June 2013.

NLP (Natural Language Processing) is a subfield of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human languages. In general, there are the common tasks for Natural Language Processing, that are Automatic Summarization: produce a readable summary of text, Co-reference resolution: Given a body of text determine which words refer to the same objects, Discourse Analysis: discover the nature of discourse relationships between sentences, Machine Translation: translate written text in one language into written text in another, Named Entity Recognition: given a stream of text, determine which items map to proper names identify the type of text, Natural Language Generation: convert information from database into readable language, Part of Speech Tagging: in natural language text identify noun, verb, conjunction, etc., Question answering: given natural language question, generate natural language answer., Relationship Extraction: given natural language text determine relationship between named entities and Sentiment Analysis: extract subjective information from set of documents.

This paper purpose is to demonstrate how NLP technology can be utilized in cultural information management system. Our goal is to enhance Thai language information retrieval and Thai language knowledge management. The re-

mainder of the paper is organized as follows. Section 2 illustrates some related works. Section 3 gives an overview of system framework design. Section 4 shows system implementation of m-culture. Discussion and conclusion are in section 5.

2 **Related Work**

One of a related research developed by Rene Witte, Thomas Kappler, Ralf Krestel, and Peter C. Lockemann is Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management [1]. It shows the modern semantic technologies offer the means to make the heritage documents accessible by transforming them into a semantic knowledge base, that using techniques from natural language processing and semantic computing.

Moreover, Horacio Saggion, Emma Barker, Robert Gaizauskas and Jonathan Foster modified Integrating NLP Tools to Support Information Access to News Archives [2] and Joao Graca, Nuno J. Mamede, Joao D.Pereira developed NLP Tools Integration Using a Multi-Layered Repository[3]. They offer Natural Language Processing framework and integration for Supporting information access to archive.

3 System Framework Design

Existing information management system or content management system (CMS) could not process information in Thai language properly especially in information retrieval process. In our system, Apache Solr is selected for implementing information retrieval process, but the results are inadequacy. This is because of Thai language characteristics that have no explicit word boundary [4].

Describe about NLP from introduction paragraph2

According to the limitation described above, research team design a new input process and develop some NLP tools. The main objective is to increase the quality of raw data by combining the expertise of experts in cultural information for ministry of culture, THAILAND, and capability NLP technology. The new data input process workflow is shown in figure [1].

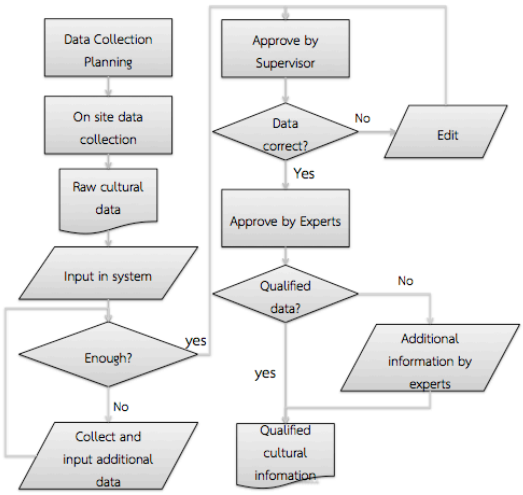

Figure [1]: New data input workflow

Information from new input workflow is concise and reliable. Research team developed word segmentation algorithm based on Thai Character Cluster algorithm [5]. The results from segmentation process are highly accurate. These results are tagged with their word position. Then part of speech is analyzed. All of this information is used as training data for keyword extraction engine from Apache Solr.

The original, unique aspects this system has

4 System implementation

M-culture was launched in October 2010 for internal test on Internet as show in figure [2].

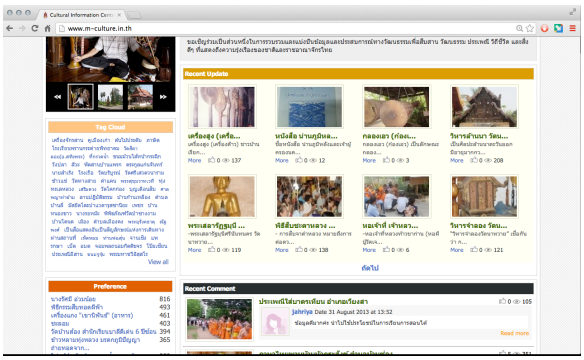

Figure [2]: M-culture on the web (http://www.m-culutre.in.th)

On first twelve months, keyword extraction process is based on Apache Solr algorithm and

user-input tags. Due to the large number of information, the number of tags is enormous and scattered as the number of tags and number of records ratio is higher than 0.6 in average. As the result, the result is unsatisfied by internal users and their supervisors particularly on Thai keywords. On the second year, research team focuses on improve quality of service including its search engine. In this phase, we implement new input process and implement our NLP tools into the system. Figure [3] shows the result from the modified search engine.

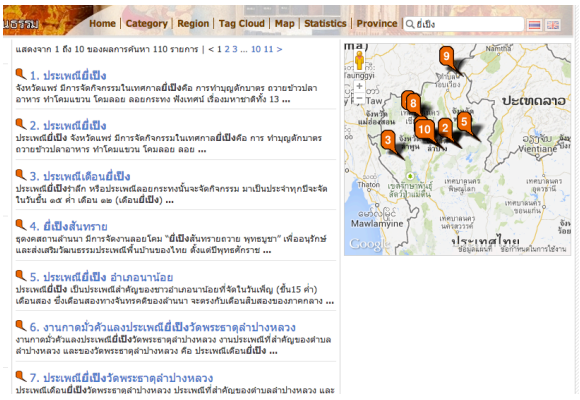

Figure [3]: Result from improve search engine

The research team implements "Character Cluster Based" word segmentation. As a result, word units are more precise. After that, each word will be annotated with part of speech (POS) and is trained by machine learning to get the qualified keywords. The results of improved search engine using lengthy Thai keywords are satisfied internal users and their supervisors. There are several assessments from ministry of culture's experts and investigators along entire development process. The overall quality of system including Thai search engine are improved from the original system.

5 Discussions and Conclusion

The m-culture launched for internal testing since October 2010 on URL http://www.m-culture.in.th. A total of 5,370 users from every province in Thailand have been testing the system since its launch. There are several subjective tests by investigators from ministry of culture about quality of system. From their evaluation, the quality of overall system and search engine quality is better than their previous system; especially Thai keywords extracted from our NLP tools are highly precise and pertinent. There is survey from public users show in figure [4]. In the future, we have plan to implement keyword semantic and ontology into system.

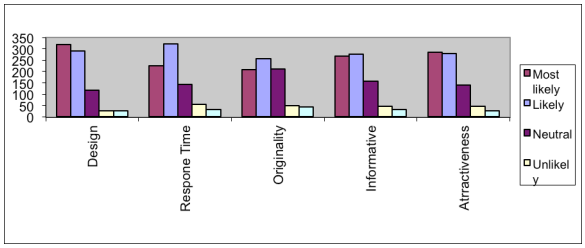

Figure [4]: survey from public users

6 Acknowledgments

We would like to express gratitude to Ministry of Culture, Department of Cultural Promotion, Department of Religion Affairs, Office of Contemporary Art and Culture for providing budget, information and other valuable resources.

7 References

[1] Rene Witte, Thomas Kappler, Ralf Krestel, and Peter C. Lockemann, *Integrating Wiki Systems, Natural Language Processing, and Semantic Technologies for Cultural Heritage Data Management*, Language Technology for Cultural Heritage ,page 213-230, 2011.

[2] Horacio Saggion, Emma Barker, Robert Gaizauskas and Jonathan Foster, *Integrating NLP Tools to Support Information Access to News Archives*, in Proceedings of the Fifth International Conference on Recent Advances in Natural Language Processing RANLP-2005, 2005.

[3] Joao Graca, Nuno J. Mamede, Joao D.Pereira, *NLP Tools Integration Using a Multi-Layered Repository*, in Proceedings of The fifth international conference on Language Resources and Evaluation, LREC 2006, 2006.

[4] Surapant Meknavin, Paisarn. Charoenpornsawat, Boonserm Kijsirikul. *Feature-based Thai word segmentation*, Proceedings of the Natural Language Processing Pacific Rim Symposium 1997

[5] Thanaruk Theeramunkong, Virac Sornlertlamvanich, Thanasan Tanhermhong, Wirat Chinnan. *Character Cluster Based Thai Information Retrieval,* Proceeding of the Fifth International Workshop on Information Retrieval with Asian Language, 2000, pp 75-80

[6] Ministry of Culture, Thailand. *Cultural Knowledge Center*, http://www.m-culture.in.th