International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Developing the Thai Regional Dialect Based on Semi-automatic Technique

Kunyanuth Kularbphettong

Abstract*: *A dialect is the specific language of a particular locality that has a unique style, both words and accents. It shows the identity culture and the way of life of people in each region of Thailand and it should be treated to continue to be a national heritage. This research presents the prototype of four regional dialect of Thai based on smart phone by using Ontology technique. The application is easily search and finds the term of dialect words for each region. To develop ontology, a semi-automatic approach was used to build domain ontology from spoken corpus in Thai dialect. The domain ontology is created to handle terms and relationships to describe the features of four Thai regional dialect words. Also, the longest matching approach was used to increase the capability of the system to Thai word segmentation. The empirical results showed that the prototype can efficiently provide satisfied information for users in information searching. System testing technique and questionnaires were needed to evaluate system work and user achievement. The results shown that user satisfaction, both experts and users, was fulfilled the system performances and easily found required information.

Index Terms*: *Thai regional dialect, Ontology, semi-automatic technique, word segmentation, mobile application.

I. INTRODUCTION

Thai language, the native and official language of Thailand, has many regional dialects. A regional dialect is a variety of a language spoken by people living in an area [1]. This kind of language variation is most noticeable, systematically different from other varieties in both structure and lexical features. The term of dialect is used to present the characteristics of grammar and vocabulary in case of pronunciation. Thai regional dialect is divided to four main regions: south, central, northeast and north. Each region has its own dialect and the spelling and meaning of the words does not match each other. With the rapid advancement of technology, the ubiquity of smart phones has affected significant changes the way of learning and searching information and it is potential for exploiting the benefits of these devices to educate and enhance learning Thai regional dialects.

Ontology is the well-defined blueprint of a conception to represent knowledge by identifying a set of representational terms [2]. Berners-Lee et al [3] proposed the concept of ontology that it is a collection of information described a theory of entity, of what types of things exist; Ontology is used to explain the meaning of interested thing and to categorize the documents in the interested data area. Also, ontology is considered as one of the pillars of the semantic web to be used to prepare structured vocabularies and allow human and intelligent agents to interpret term meaning. Even though semantic web focuses on the searching approaches of text matching, it supports to query the semantic among the meaning and relationships each object using the synonym, antonym, hypernym, hyponym correlation of words. However, the creation of the ontology is a very complicated process and it takes time consuming. To minimize an error and a tedious work of building ontology, semi-automatic approach was proposed to deploy with this project and also Thai word segmentation based on longest matching method was chosen to augment the efficiency of the research and OWL (Ontology Web Language), as a standard Semantic Web language used for the concrete manual ontology building, presents as structural knowledge to model Thai regional dialects knowledge with an ontology-based approach so as to search the information what exactly the meaning of words presents from Thai regional dialects. This work is separated as follow: the section 2 will be examined some related research, and then section 3 is described the research methodology and section 4 to the experiments and to the evaluation phase. The conclusion and future research are enacted in section 5.

II. RELATED WORKS

A literature reviews of relevant researches for exploration and required information shows that ontology has grown into the famous technique in various fields to handle a semantic framework for knowledge management. Ontology describes a content representing specific knowledge primitives (classes, relations, functions and constants) [4]. Ontology is usually adapted in information retrieval systems to make a comprehension for computer and machines. Humans use background knowledge in a great deal of information retrieval tasks [5]. The information retrieval system can use the concept of ontology to be useful in query data and information in the same kind of presumptions. McKone TE and Feng L [6] proposed ontology framework to present the structural framework of Health Risk Assessment (RsO) for organizing risk assessment information and methods.

Revised Manuscript Received on 30 July 2019.

Kunyanuth Kularbphettong* Faculty Department of Computer Science Program, of Science and Technology, Suan Sunandha Rajabhat University. Thailand.

Ontology is used to create meaningful content-oriented learning systems for learning through context related to learners' knowledge, Learning content and the learning domain for consideration the learner ability [7]. CRCTOL proposed the significant method of implementing by using the data mining techniques and ontology to analyze the full text content and the result found
that the proposed approach can provide accurate results in terms of analyzing concepts and relationships from complex structured sentences [8]. Also, Vietnamese ontology construction was proposed from a combination of methods based on statistics, lexical patterns and patterns of frequent sequences [9]. OntoGen is presented as a semi-automatic ontology editor that combines machine learning and text mining algorithms into user interfaces in order to reduce time and complexity when using ontology by naïve users [10].
Additionally, mobile application was used to provide management of learning environment [11]. There is much of research that indicated how to handle specifications for design of a mobile learning [12]. A knowledge base for many mobile applications can be utilized for the constructed ontology; such as information acquiring and recommendation systems [13]

III. RESEARCH METHODOLOGY

To develop the prototype, this project was divided the process of conducting re-search into 3 major steps as following: 1) analyzing the requirement of the system by studying Thai regional dialect information and creating the system framework; 2) designing and building the structure of ontology technique by setting the framework of the ontology based on the knowledge of Thai regional dialect information; 3) reviewing and evaluating the ontology and mobile application.

A. Analyzing the Requirement of the System

Rapid Application Development (RAD) is the rapid development approach that focuses on the reduction of cost and development time and RAD always applies to use in a small development team that has the knowledge and ability to the project in order to quickly develop the work system. RAD consists of 4 distinct phases; requirement planning; system design; development and cutover phase as shown in figure 1. In this research, RAD was employed to develop Thai regional dialect mobile application and to collect and analyze user requirements, the first step is to plan the requirement survey and assign duties and tasks within the system. The design of the user interface has been designed by participated users for ease of using. The system design uses the results from the requirement and planning to design and develop the application model by analyzing the necessary components of the system.

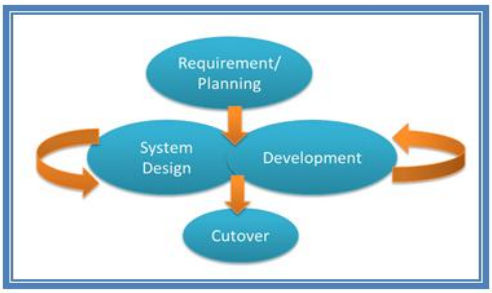

Fig. 1. Rapid Application Development Process [14]

In this model, it is designed to increase productivity and improves the precision of results by dividing the important components of the system into 3 parts: the user Interface part, the word segmentation part, and the ontology and database part. The user interface part consists of two main functions: providing information to users and displaying related data. Users enter information and the application sends the related data to the relevant processing section. The word segmentation part serves to cut sentences received from users using the longest matching technique and the ontology and database sections act to store and manage information and knowledge.

B. Designing and building the structure of ontology technique

In this research, there are a few of data and information related with Thai regional dialect on internet and mostly data and information are the document forms. Thai Wikipedia and other related web sites were used to gather 162 documents from Internet. The collecting keywords and indicating scope of data was gathered by manual process and then knowledge of Thai regional dialect was grouped to noun, pronoun, verb, preposition, conjunction, and adjective. Also, figure 2 displayed the system overview of this process. To design and create ontology, natural language processing language (NLP) was adapted to use in this research. However, with the limitation on natural language processing language, Thai language differs from English language because Thai language has no word boundary or delimiters between the multiple words and between sentences. In order to collect data, classification theory was also used as the basis for grouping and dividing information into primary classes, also each group of data was divided into sub-classes for analyzing the content, the extent of knowledge (domain) and the theory of cognitive systems. According to Reinberger's method [15], a parsed medical corpus was firstly refined on pattern matching and clustering algorithms to the classes of dependencies to build sets of semantically related words and establish semantic links between them.

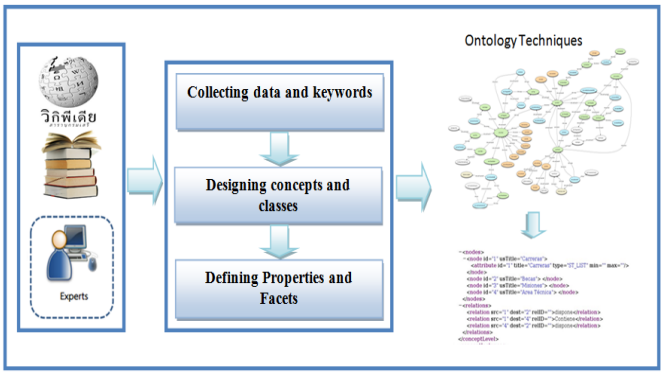

Fig. 2. The system overview of this process

The process of information extraction is an important process to extract information and to identify related information. It consists of the following steps:

1. Data preparation process to collect, clean, and consolidate data into a suitable format for use in analysis.
2. Word segmentation process to divide written text into meaningful units.
3. Data extraction process to extract keyword data from a set of documents.

4. Data grouping process to group a glossary dialect word of each region.

In this research, TF-IDF (Term Frequency-Inverse Document Frequency) was exploited to specify the weight of a word in a set of documents by calculating term frequency from documents.

$$W_{(f,d)} = TF_{(f,d)} \times IDF_{(f)} \quad (1)$$

$$IDF_{(f)} = log \frac{|D|}{|DF_{(f)}|} \quad (2)$$

Where TF (f,d) is the frequency of the feature (f) in (d), W(f,d) is the weight of a feature (f) in (d), | D | is the number of documents in the training data set (Training Set),| DF(f)| is the number of documents that feature (f) appears.

The results of the calculation are determined the features of keyword sets in each category and converted the feature to document-terms matrix as shown in Table 1.

TABLE I: THE DOCUMENT-TERMS MATRIX

Document	Features				Class
	$Feature_1$	$Feature_2$	…	$Feature_n$	
Doc_1	F_{11}	F_{12}	…	F_{1j}	A_1
Doc_2	F_{21}	F_{22}	…	F_{2j}	A_2
Doc_3	F_{31}	F_{32}	…	F_{3j}	A_3
…	…	…	…	…	…
Doc_n	F_{i1}	F_{i2}	…	F_{ij}	A_n

The data collected and gathered in ontology approach in this research was approved by experts and users in computer program, linguistic and cultural management and the ontology of this project is implemented in Protégé [16], a free open-source ontology editor and framework for building intelligent system. OWL (Web Ontology Language) is the language used for describing ontologies and defines the relationship between the data and knowledge. Also this ontology implemented for use in this system has been evaluated by experts and users to assess the validity of the ontology in the concept and relationship.

C. Reviewing and Evaluating the Ontology and Mobile Application

To discover the information from a set of documents, information extraction is the significant process that handles with the extraction of specified entities, events, and relationships from sources. In this research, feature extraction is the first step of pre-processing to recognize and classify key words. Term frequency-inverse document frequency (TF-IDF) is used to determine the relative frequency of words in a set of specified documents through an inverse proportion of the word over the entire document corpus. Accuracy, recall, precision and F-measure were used to evaluate the performance of text classification before building ontology.

$$precision = \frac{TP}{TP + FP} \quad (3)$$

$$recall = \frac{TP}{TP + FN} \quad (4)$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$F - measure = \frac{2 \times \text{precision} \times \text{Recall}}{\text{precision} + \text{Recall}} \quad (6)$$

Where TP represents the number of correctly classified documents, FP is all documents retrieved and accuracy is the percentage of documents correctly classified, recall is the percentage of relevant documents correctly retrieved (TP) with respect to all relevant documents (TP + FN) and F-Measure is consisted of a single measure Precision (P) and Recall (R).

To test the accuracy of the classification and relation-ship of ontology, experts evaluated with the prototype and questionnaires with consistent with the scope, the accuracy of the classification, consistent of terminology, the accuracy of the data relationships and the completeness of Thai regional dialect information [17].

Fig. 3. User Interface of this application

The architecture of this application is considered to be 3 subsystems including the user interface subsystem, the application subsystem, and data Storage subsystem. Figure 3 presents the user interface for this program. User enters information needed to be searched, then the system will cut the word and search in the storage section and show the result to the user. To test and evaluate the efficiency of the proposed system, Black box Testing is the basis of software testing to be used as a basis for analyzing and designing test cases to cover the user requirements. Black box testing is a significant test that it is not interested in the components within the system and look at the overall of the box. The advantage of black box testing is that it can be tested without the need for specific knowledge for programming languages to test and perform tests from the user's perspective. The test of the project was assessed in functional requirement testing, Function testing, Usability testing, Performance and Security testing. Functional Requirement test is a test that is interested in the function of the application that can meet the user requirements and Functional test is a test of the entire system whether it meets the requirement or not. Usability test was tested the user interface of the application that can access every menu or every screen and it is suitable for using in terms of convenience and ease of use. Performance test was used to test the number of users or the amount of data coming into the system as the user requirements. The trial process included participants who used and learned mobile application and after completion participants were asked to rate the quality of the application. The collected data were analyzed by the statistical means ($\bar{x}$) and standard deviation (S.D.).

IV. EXPERIMENTAL RESULTS

The experimental results of this research are separated by the research objectives into 2 issues: the accuracy of the classification and relationship of ontology, the usage of the system, and the user interface display.

A. The Accuracy of the Classification and Relationship of Ontology

To measure performance of classification, precision and recall were 87.8% and 82.95% respectively and the result of the accuracy of the classification and relation-ship was shown in table 2. Experts evaluated ontology with the prototype and questionnaires. Figure 4 was displayed the ontology of this project.

Fig. 4. The Ontology of this application

This is the strong evident to show that in terms of connection and information categories, the prototype worked very well with the average value of each topic through statistical methods against average (Mean) and standard deviation (SD) to find that the average is 4.45 and the standard deviation was 0.49, so the system.

TABLE II: THE RESULTS OF THE ACCURACY OF THE CLASSIFICATION AND RELATIONSHIP

	Experts	
	$\bar{x}$	SD
1. Consistent with the scope	4.75	0.44
2. The accuracy of the classification.	4.6	0.68
3. Consistent of terminology	4.4	0.75
4. The accuracy of the data relationships	4.3	0.92
5. The completeness of Tourism and Culture information.	4.2	0.83
Summary	4.45	0.49

B. The Result of the Qualities of the System

Black box testing and user satisfaction were used to test and evaluate the qualities of this application. Black box testing is a software testing technique that ignores the internal mechanisms of a system or component and focuses on the output that comes after the system response only by selecting input and execution conditions [18]. The results of Black box testing between experts and users was displayed in Table III.

TABLE III: THE RESULTS OF BLACK BOX TESTING

	Experts		Users	
	$\bar{x}$	SD	$\bar{x}$	SD
1.Function Requirement Test	4.52	0.59	4.55	0.54
2. Functional Test	4.50	0.60	4.60	0.58
3. Usability Test	4.70	0.48	4.11	0.44
4. Performance Test	4.48	0.66	4.17	0.60
5. Security Test	4.47	0.72	4.19	0.59

The results of data analysis to assess the satisfaction of the system users were expressed that experts and users were satisfied with the operation of the system as follows: means for experts and users were equal to 4.46 and 4.22 respectively

V. CONCLUSION

In this paper, the prototype of four regional dialect of Thai based on Android platform by using Ontology technique was proposed and this application can efficiently assist users to search and learn Thai regional dialect. Ontology is the heart of create knowledge for this project. Also, Term frequency-inverse document frequency (TF-IDF) is applied to determine the relative frequency of words in a set of specified documents and F-Measure is evaluated the related terms by precision (P) and Recall (R). The results from the development of prototype were found that the experimental group had significantly better performance in learning achievements both experts and users. The result from this research can be the fundamental information of further research.

ACKNOWLEDGMENT

The authors gratefully acknowledge the financial subsidy provided by Suan Sunandha Rajabhat University.

REFERENCES

1. Varieties of a language, Retrieved from http://media.openonline.com.cn/media_file/rm/dongshi2004/yyyyxgl/CHAPTER8/chapter8-2.htm#k, February, 2016.
2. T. R. Gruber. Towards principles for the design of ontologies used for knowledge sharing. In N. Guarino and R. Poli, editors, Formal Ontology in Conceptual Analysis and Knowledge Representation. Kluwer Academic Publishers, Deventer, The Netherlands, 1993.
3. Berners-Lee, T., Hendler, J. and Lissila, O. 2001. The Semantic Web. [Online]. Retrieved from: http://sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 .
4. Li, P. S., & Rui, M. S. (2005). Ontology-based learning content recommendation. Inter¬national Journal of Continuing Engineering Education and Life Long Learning, 15(3-6), 308-317.
5. Bowman, C.M., Danzing, P.B., Manber, U. and Schwartz, F., (1994). “Scalable Internet resources discovery: research problems and approaches”. Communications of the ACM, vol. 37, pp. 98-107
6. McKone TE, Feng L. “Building a Human Health Risk Assessment Ontology (RsO): A Proposed Framework.” Risk Anal. 2015 Nov;35(11):2087-101. doi: 10.1111/risa.12414. Epub 2015 May 15.
7. Zhiwen Yu1,2, Yuichi Nakamura1 , Seiie Jang2 , Shoji Kajita2 , and Kenji Mase2, “Ontology-Based Semantic Recommendation for Context-Aware E-Learning,”
8. Xing Jiang , Ah-Hwee Tan, Mining Ontological Knowledge from Domain-Specific Text Documents, Proceedings of the Fifth IEEE International Conference on Data Mining, p.665-668, November 27-30, 2005 [doi>10.1109/ICDM.2005.97]
9. Bao-An Nguyen, Don-Lin Yang., A Semi-Automatic Approach to Construct Vietnamese Ontology from Online Text, International review of research in open and distance learning, December 2012
10. B. Fortuna, M. Grobelnik, D. Mladenic, OntoGen: semi-automatic ontology editor, in: Proceedings of the 2007 Conference on Human Interface, 2007, pp. 309–318.
11. M. Luisa Sevillano-Garcia abd Estban Vazquez-Cano.,“The Impact of Digital Mobile Devices in Higher Education”, Educational Technology & Society, 18 (1), 106–118, 2015
12. M Sharples,”The design of personal mobile technologies for lifelong learning”, Computers & Education 34 (3), 177-193,2000
13. Kunyanuth Kularbphettong and Bundit Ngamkam. “A Recommendation System for Heritage-Tourism based on Mobile Application and Ontology Technique.” In: International Journal of Information Processing & Management 5.3 (2014).
14. Javatechig, “Rapid Application Development Model”, Retrieved from http://javatechig.com/se-concepts/rapid-application-development-model.
15. Reinberger M.-L., Spyns P., Pretorius A.J., & Daelemans W. (2004) Automatic Initiation of an Ontology. In Meersman, R. & Tari, Z. (eds.), CoopIS/DOA/ODBASE 2004, pp 600~617. Agia Napa, Cyprus. Berlin, Heidelberg: Springer-Verlag.
16. Protégé, Retrieved from: http://protege.stanford.edu/
17. Nisanart Tachapetpatboon and Kunyanuth Kularbphettong, “ONTOLOGY BASED KNOWLEDGE MANAGEMENT FOR CULTURAL TOURISM.”, Journal of Theoretical & Applied Information Technology, Vol. 75 Issue 3, p384-388. 5p.
18. Laurie Williams. “Testing Overview and Black-Box Testing Techniques”. , 2006.

AUTHORS PROFILE

Kunyanuth Kularbphettong is assistant professor of computer science program, Faculty of Science and Technology, Suan Sunandha Rajabhat University, Thailand. She received the Ph.D in information technology from King Mongkut's University of Technology North Bangkok, Thailand. Her research interests are IOT, data mining, machine learning, software applications and educational learning. (e-mail: kunyanuth.ku@ssru.ac.th).