An Alternative Technique for Populating Thai Tourism Ontology from Texts Based on Machine Learning

Aurawan Imsombut
Faculty of Information Technology
Dhurakij Pundit University,
Bangkok, Thailand
aurawan.ims@dpu.ac.th

Chaloemphon Sirikayon
Faculty of Information Technology
Dhurakij Pundit University,
Bangkok, Thailand
chaloemphon.sir@dpu.ac.th

Abstract - **This paper proposes an alternative technique to perform ontology population by using natural language processing and machine learning techniques. This study conceptually considers the population task as classifying terms into ontological subcategories. The proposed technique adopts the recognition method named Conditional Random Fields (CRFs) to identify boundary of instances and define types of sub-concepts to generate relationships between instance-of and related concept. Also, the lexico-syntactic pattern is used to identify the relationships between instances. The experiments are conducted on Thai language documents in the tourism domain. The experimental results showed that the instances extraction step provided 77.62% and 70.87% of precision and recall measures, respectively, and relationships extraction step yielded 82.67% and 72.61% of recall measures.**

Keywords—Ontology Population; Tourism Ontology; Machine Learning; Conditional Random Fields (CRFs);

I. Introduction

Ontology is a set of terms or concepts and relationships between concepts that is collected for a domain-of-interest, for examples, tourism, medical, and agriculture domains. Ontology can be used in various applications that needs semantic meaning mechanism, such as information retrieval or recommendation systems. In general, performing ontology learning consists of several tasks including term extraction and normalization synonym identification, concept and instance recognition, and relation extraction [1]. Defining instances and instance's relationships into ontology can be called as Ontology Population. This task is an important step so as to expand knowledge in the ontology applicable to increasingly various applications. However, ontology population is considerable time consuming, and also needs experts' efforts and experiences. Thus, studies of automatic or semi-automatic ontology population are still needed.

This study focuses on instances of concepts relating to Attractions and Activities. Each concept is classified into sup-concepts. For example, Attraction concept consists of Cultural, Argo, Natural, and Shopping sub-concepts. These are most information searched by users and they are used for decision making. In general, a word representing for instance for each concept in the ontology is a specific name, says, Name Entity (NE). This NE is proposition used to identify things such as persons, organizations, locations [2]. However, NE in Thai language does not have orthographical information. For examples, use capital letters in the beginning of the sentence as used in English language, or special characters such as Kanji, Katakana as used in Japanese language. Moreover, it is rather ambiguous when considering NE between single words and common words, between multi word NE and common noun phrase. These result in challenging task to extract NE in Thai language. The relationships between instances are inherited relationship from concepts that the instances belong to. And, normally, the relationships are verbs that occur together with those Concepts.

Steps in this research begin with gathering tourism documents from related websites. After that, HTML tags in the documents are removed and then fed into natural language processing including word segmentation and part of speech tagging. Next step, Name Entity Recognition is performed. In this step, the features used are considering word, near co-occurred words, POS of considering word, and POS of near words of cue word list, and the number of the words occurring in the document. The pattern recognition technique called Conditional Random Fields (CRFs) is adopted herein in order to define the boundary of NE and the type of sub-concepts for creating the relationships between instance-of and related concepts. Next step, the results are post-processing for (1) validating the association of class results, (2) improving the results of NE boundary identification, and (3) define more corrected sub-concept types with heuristic rules. In the part of finding the relationships of instances, this study adopts lexico-syntactic pattern to identify them, which is able to represent the related co-occurred patterns. For example, important activity patterns "…<Attraction> and <Activity>, <Activity>, … etc." shows has Activity relationship between attraction and activity. (e.g., important activities at the Kok river area are raft, camp, and bush walk)

The rests of the paper are as follow: Section 2 reviews the related literatures. Section 3 presents Ontology population process. The experiments are presented in Section 4. Section 5 is the conclusion.

978-1-5090-0806-3/16/$31.00 copyright 2016 IEEE
ICIS 2016, June 26-29, 2016, Okayama, Japan

II. RELATED WORKS

There are many research studies concerning to ontology population (i.e., define and classify instances). Most of those studies are to apply NLP techniques with Information Extraction (IE) techniques and Machine Learning (ML) techniques.

Martinez et al. [3] proposed a combination of NLP and IE techniques by using GATE tools for extracting NE from restaurant and hotel corpus, and use heuristic algorithm for solving different kinds of ambiguities to populate the instances into tourism ontology. Faria et al. [4] presented another combination of NLP and IE to create rules for automatic population of ontologies from text. They study conducted on legal and tourism corpora.

Zhang et al. [5] applied NLP and ML techniques called Maximum Entropy to extract relationships between entities for the field of tourism. Nanba et al. [6] applied NLP and used CRF as ML in order to identify travel blog, and extract travel information relating to the relationships between location names and local products. Nevertheless, Carlson et al.[7], Giuliano and Gliozo [8], Cimiano et al. [9] and Etizioni et al. [10] applied NLP, IE and ML techniques to ontology population.

III. ONTOLOGY POPULATION PROCESS

In this section, we describe the ontology population process. First, the Thai Tourism Ontology used in this study is presented. Next, we describe the methodology for populating ontology from text based on machine learning technique.

A. Ontology

Fig. 1 shows an example of tourism ontology used in this study. This was improved from Thai Tourism Ontology [11], which is collected in owl file format. The ontology consisting of 4 classes: Activity, Attraction, Province and Accommodation. Each class has sub-concept relationship to their sub-class and contains different properties e.g. has Activity, has Attraction, has Accommodation.

B. Methodology

The Ontology population process in composed by four sequential phrases: Feature Extraction, Instance Extraction by CRFs, Post-processing of instance extraction and Relation Extraction as shown in Fig. 2.

Pre-processed Documents are documents from tourism websites that were removed HTML tags with HTML parser. The documents are fed into Natural Language Processing (i.e. word segmentation and part of speech tagging) by using developed own tools. Word segmentation uses longest matching and defines POS with Hidden Markov Model (HMM).

Feature Extraction is the step to extract important features that is used by system to learn to classify boundary, and identify types of noun-identified proposition. The characteristics are as follows:

Lexical & POS features

- Words and POS of current word
- Words and POS of 3 words before current word
- Words and POS of 3 words after current word

Dictionary features

- Is current word in the cue word list? (e.g. Temple, Park)
- Are previous n-words before current word in the cue word list? (e.g. Temple, Park)
- Are not the words in dictionary?
- Do the words appear in location dictionary?

Fig. 1. An example of Tourism Ontology

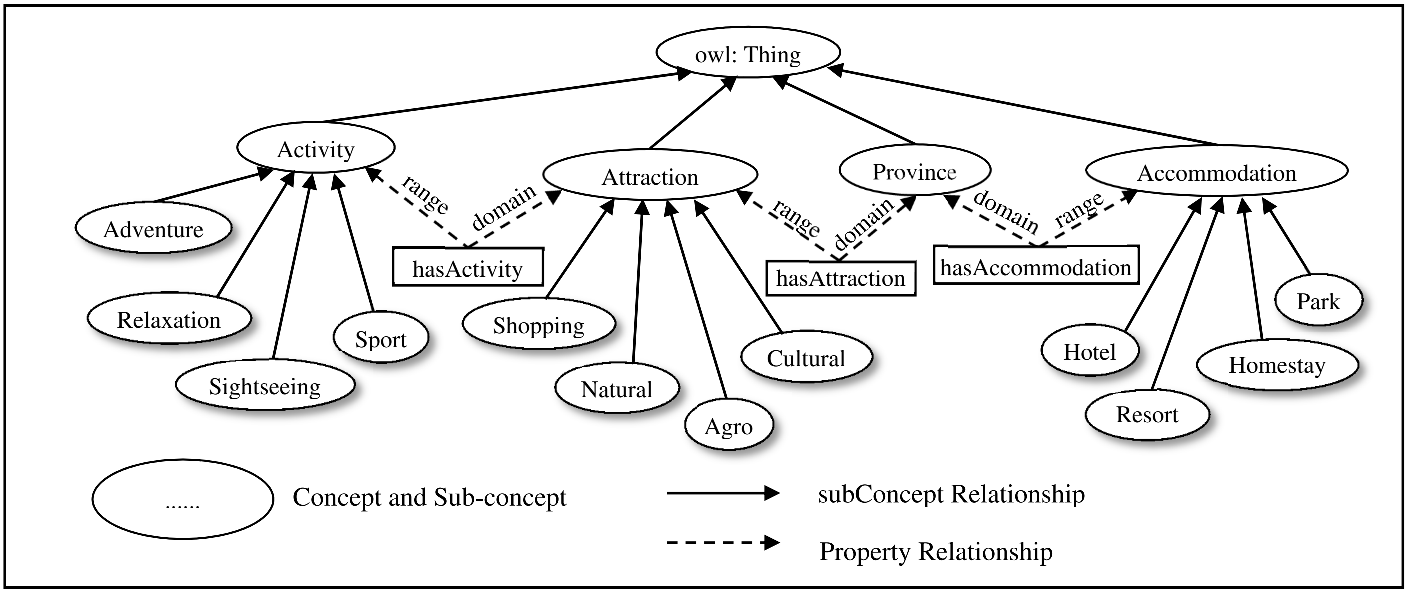
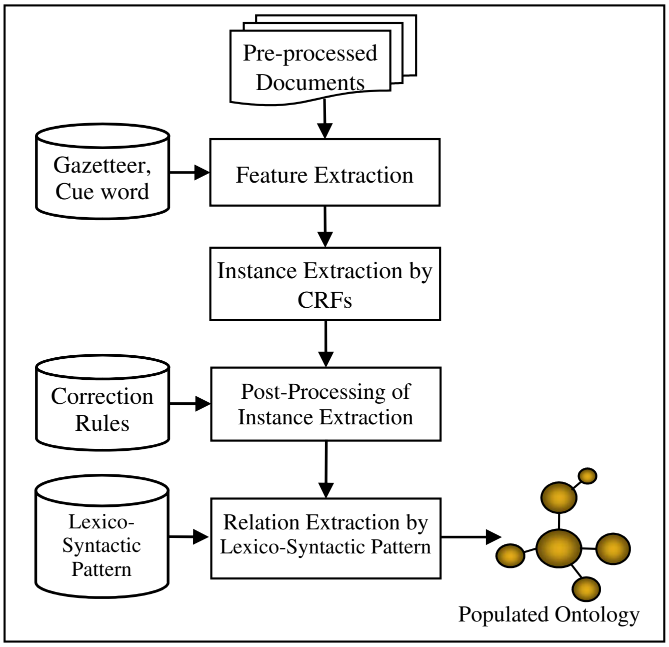


Fig. 2. Ontology population process



Repeated occurrence

- Do the words occurring before and after the considering word occur together with more than 3 times?

Instance Extraction by CRFs is step to extract noun-identified propositions. Noun-identified propositions are instances of concepts in ontology. This study identify boundary of NE and classify types of NE by recognition technique CRFs, a supervised learning which learns from class-labeled examples.

Conditional Random Fields (CRFs) [12] are undirected graphical models often used to predict sequences of labels for sequences of input samples such as natural language text. When applying CRFs to the named entity recognition problem an observation sequence is the token sequence in document and state sequence is its corresponding label sequence.

The conditional probability of a state sequence s=<s_1, s_2, ..., s_T> given an observation sequence o=<o_1, o_2, ..., o_T> defined as:

$$P(s|o) = \frac{1}{Z_o} exp(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t) \qquad (1)$$

Where $f_k(s_{t-1}, s_t, o, t)$ is a feature function and λ_k is a learned weight for each feature function. Z_o is a normalization factor over all state sequences defined as:

$$Z_o = \sum_s exp\,(\sum_{t=1}^{T} \sum_{k=1}^{K} \lambda_k f_k(s_{t-1}, s_t, o, t) \qquad (2)$$

In this study, we used CRF++ 0.58 implemented by Taku Kudo in the experiment for name entity recognition.

Each example or each word occurring in the document is labeled with observed class with entity boundary and type of subcategory. Entity boundary is used for identify boundary of word, and it is divided as follows:

- B: Begin, the beginning word of noun-identified proposition
- M: Middle, the middle word of the noun-identified proposition
- E: End, the ending word of the noun-identified proposition
- S: Single, the single noun-identified proposition
- O: Others, other words that are not the noun-identified propositions

And the types of sub category for classifying types of NE, that associates to subcategory of each concept in Thai tourism ontology. In this study, the interesting subcategories of Attraction concept composes of 4 groups, that is, Cultural, Agro, Natural and Shopping. An example of labeling answer for the sentence “Pa-Hom Pok mountain is tourist attraction that located in the area of Fa Hom Pok mountain national park.” is shown in Fig. 3.

Fig. 3. An example of labeling answer in a sentence

ดอย/B_ Natural ผ้าห่ม/M_Natural ปก/E_Natural เป็น/O สถานที่/O ท่องเที่ยว/O ที่/O ตั้ง/O อยู่/O ใน/O เขต/O อุทยาน /B_Natural แห่ง/M_Natural ชาติ/M_Natural ดอย/M_Natural ฟ้า/M_Natural ห่ม/M_Natural ปก/E_Natural

mountain/B_Natural Pa-Hom/M_Natural Pok/E_Natural is/O place/O travel/O that/O located/O is/O in/O area/O park/B_Natural of/M_Natural nation/M_Natural mountain/M_Natural Fa/M_Natural Hom/M_Natural Pok/E_Natural

The activity and province instances are classified by using dictionary.

Post-processing of instance extraction is the step to improve the results from identifying NE boundary. This step also defines the types of sub-concepts more accurately with heuristic rules, which, in turn, consists of 2 sup-processes.

1. The process of correcting entity boundary correction considers association in the occurring of boundary pattern of NE. If predicted subcategories have unassociated pattern to B-M_n-E, heuristic rule will be used to solve the correctness of correction rules. For example:

- if O-Mn-E then correct to B-Mn-E
- if B-On then correct to O-On
- if B-Mn-O then correct to B-Mn-E
- if B-Mn-O-E then correct to B-Mn-M-E

2. The process of correcting classified subcategory. After receiving the corrected identified boundary pattern, next process is to check whether created subcategories of each entity in the boundary are associated. If they are not associated enough, majority vote technique will be applied. Weight of subcategory B is larger than other subcategories.

Relation Extraction is to find relationships between instances by using lexico-syntactic pattern technique to identify relationships. Table I. shows the examples of lexico-syntactic pattern for hasActivity relationship. There are the co-occurred word patterns such as pattern “Interesting things for travelling in <Attraction> are <Activity>”. The sentence “Interesting things for travelling in Mon-Jam are camping, sightseeing, visiting temperate vegetable and fruit farms.” can be extracted three relationships as (hasActivity, Mon-Jam, Camping), (hasActivity, Mon-Jam, sightseeing), and (hasActivity, Mon-Jam, visiting temperate vegetable and fruit farms).

TABLE I. EXAMPLES OF LEXICO-SYNTACTIC PATTERN FOR HASACTIVITY RELATIONSHIP

lexico-syntactic pattern	Example Sentences
Tourists can <Activity> at <Attraction>	Tourists can go camping at Kun-Chang-Keint agro-stations area.
At <Attraction>, Tourists can <Activity>	At Small-Farm, the tourist can riding a horse, caravan site seeing, feeding animals, planning and others activities.
Besides <Activity> at <Attraction>, tourists also can <Activity>	Besides site seeing and taking photographs of Cheingmai-Grand Canyon, tourists also can take swimming and river cliff jumping.
Interesting things for travelling in <Attraction> are <Activity>	Interesting things for travelling in Mon-Jam are camping, sightseeing, visiting temperate vegetable and fruit farms.
<Attraction> has…which is suitable for <Activity>	Cheng-San Lake has shady atmosphere, which is suitable for canoeing.

IV. EXPERIMENTATION AND DISCUSSION

In the experiments, 100 Thai documents (or 40,000 words approximately) from the Thai tourism websites were used. The contents in the website were, for examples, attractions, accommodations, and activities. Those documents were pre-processed and were performed instance extraction and relation extraction. The preliminary results from the instance extraction process are shown in Table 2 and Table 3.

TABLE II. EXPERIMENTAL RESULT FOR INSTANCE EXTRACTION

Attraction	Precision	Recall	F-measure
Cultural	80.17%	74.62%	77.29%
Agro	66.67%	43.24%	52.46%
Natural	79.25%	87.50%	83.17%
Shopping	66.67%	53.33%	59.26%
All	77.62%	70.87%	74.09%

TABLE III. EXPERIMENTAL RESULT FOR RELATION EXTRACTION

Relationship	Precision	Recall	F-measure
hasActivity	**77.78%**	**67.31%**	**72.16%**
hasAttraction	**84.08%**	**74.16%**	**78.81%**
All	**82.67%**	**72.61%**	**77.31%**

The results of the cultural attraction extraction process showed the highest precision because most of their names were specific name such as temple names and people’ monument names. As a result, it was not difficult for the classification module to clarify them.

In contrast, the instances of agro attraction had long word names (i.e., the name of national park). Also, the shopping instances had larger variations of word names and showed ambiguous. These reasons resulted in small value of the recall values for agro attraction extraction. Their names were composed of common words. Then, the system often classified them as the O-class, not the NE name.

V. CONCLUSION

In conclusion, this study proposed Thai ontology population by using Conditional Random Fields (CRFs) to identify the instances of concepts with knowledge from dictionary and the cue word list. Furthermore, the post-processing of instance extraction is used as the process for improve the results from identifying NE boundary, and the lexico-syntactic pattern is used to identify the relationships between instances. The experiments were conducted on Thai language web documents in the tourism domain. Accordingly to these preliminary results, the approach could extract instances with acceptable results. For the future work, we will conduct the experiment on the instance extraction of other concepts and relationships extraction of longer-distance information.

REFERENCES

[1] Z. Zhang and F. Ciravegna, “Named Entity Recognition for Ontology Population using Background Knowledge from Wikipedia”, in Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances, IGI Global, 2011.

[2] Chinchor, Nancy. 1998. MUC-7 Named Entity Task Definition (Version 3.5). MUC-7. Fairfax, Virginia.

[3] Martinez, J. et al. 2011. Ontology Population: An Application for the e-tourism Domain. International Journal of Innovative Computing, Information and Control.

[4] Faria, C., Girardi, R. and Novais, P. 2012. Using Domain Specific Generated Rules for Automatic Ontology Population. Proceeding of 12th International Conference on Intelligent Systems Design and Applications.

[5] Zhang, Y., et al. 2009. Automatic Entity Relation Extraction for the Field of Tourism. Journal of Computational Information System.

[6] Nanba, H., et al. 2009. Automatic Compilation of Travel Information from Automatically Identified Travel Blogs. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and 4th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing (ACL-IJCNLP).

[7] Carlson, A., J. Betteridge, R. Wang, Jr. E. Hruschka and T. Mitchell. 2010. Coupled Semi-Supervised Learning for Information Extraction. In Proceedings of the third ACM International Conference on Web Search and Data Mining (WSDM ’10).

[8] Giuliano, C., and Gliozo, A. 2008. Instance-Based Ontology Population Exploiting Named Entity Substitution. In Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008).

[9] Cimiano P., Ladwig, G., and Staab. 2005. Gimme’ The Context: Context-driven Automatic Semantic Annotation with C-PANKOW. In Proceedings of the 14th World Wide Web Conference (WWW).

[10] Etizioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A. M., Shaked, T., Soderland, S., Weld, D., and Yates A., Web-scale information extraction in KnowItAll. In Proceedings of the 13th World Wide Web Conference (WWW).

[11] Kongthon, A., Kongyoung S., Sangkeettrakarn, C., Haruechaiyasak C., 2010. Thailand’s Tourism Information Service based on Semantic Search and Opinion Mining. Proceeding of International Technical Conference on Circuits/Systems, Computers and Communications.

[12] Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Sequence Data. Proceeding of 18thICML. San Francisco.