Semantic Relation Extraction for Extensive Service of a Cultural Database

Watchira Buranasing **Virach Sornlertlamvanich** **Thatsanee Charoenporn**

National Electronics and Computer Technology Center
Thailand Science Park, Klong Luang, Pathumthani 12120, Thailand
{watchira.buranasing, virach.sornlertlamvanich, thatsanee.charoenporn}@nectec.or.th

Abstract

Semantic relation extraction is a significant topic in natural language processing. It aims at discovering semantic relations between entities with various important applications such as knowledge acquisition, web and text mining, information retrieval and search engine, text classification and summarization.The propose of this research is to discovery semantics relation among a focused set of entities in a cultural archive. The approach is based on a set of relation templates which are determined by relation type and their arguments. The experimental performs a promising resource ranging between 88-100% of accuracy with moderate processing time per document.

Keywords: semantic relation extraction, cultural database, information extraction

I Introduction

Culture is a key dimension of the information society. The exchange of information on the world cultural heritage will help people from different cultures to understand each other better. The building of cultural archives is critical in ensuring the preservation of important and unique heritage that might otherwise be threatened by natural elements, war or other factors that might cause the artifact or site to diminish or disappear over time. A cultural archive derived from Thai Cultural Information Center Website, which is one of an important database for education, economy and society. The content database associates with person, organization, place and artifact. A size of database has been increasing in terms of volume of data from cultural specialist in 76 provinces of Thailand. There are more than 100,000 records uploaded in 8 months since November, 2010 to June, 2013 .

Relation extraction is one of the core topics in natural language processing. It is an established subfield of information extraction concerned with extracting related pairs of entities from text. The goal is to discover the relationships between pairs of entities in texts. It is a critical part in many applications such as factoid question answering, building knowledge bases and improving search engine relevance.

In general, there are feature based methods and kernel methods.[1] The feature based method explicitly extracts a variety of lexical, syntactic and semantic features for statistical learning, either generative or discriminative. In contrast, the kernel based method does not explicitly extract features; it designs kernel functions over the structured sentence representations (sequence, dependency or parse tree) to capture the similarities between different relation instances.

With in the decades of relation extraction research many approaches are developed and applied to a lot of tasks, but the Thai written does not have word boundary. Any significant text parsing technique usually requires the identification of word boundary. Therefore, in Thai language text, the word segmentation is a significant task requiring knowledge of the vocabulary and morphology of words in Thai language. For this reason, relation extraction from Thai text is not a trivial task.

This research proposed a method to extract relation instances from a cultural database with a practical application.

One of a related research developed by Dekang Lin and Patrick Pantel is DIRT – Discovery of

Inference Rules from Text [2]. It is a method to discover relation instances based on the outputs from dependency parsers. Such parsers and annotated training corpora are difficult to obtain in non-English languages.

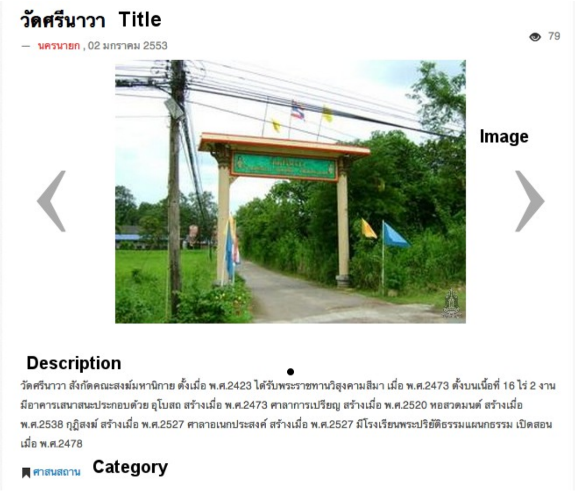
In addition, Eugene Agichtein et al. modified Snowball: A prototype system for extracting relations from large text collections[3], and Pantel and Pennacchiotti developed Espresso: Leveraging generic patterns for automatically harvesting semantic relations[4]. They are pattern-based approaches for semantic relation extraction. It seems to be more practical for languages with limited NLP resources.

The remainder of the paper is organized as follows. Section II gives an overview of the proposed semantic relation extraction for extensive service of a cultural database. Section III shows the experiment results. Section IV concludes and discusses some directions.

II Overview of the Proposed Semantic Relation Extraction for Extensive Service of a Cultural Database

A. Relation Template Design.

The content of each document from a cultural database including four components, there are images, title , description and category as shown in Fig.1



วัดศรีนาวา Title

— นครนายก , 02 มกราคม 2553

79

Image

Description

วัดศรีนาวา สังกัดคณะสงฆ์มหานิกาย ตั้งเมื่อ พ.ศ.2423 ได้รับพระราชทานวิสุงคามสีมา เมื่อ พ.ศ.2473 ตั้งบนเนื้อที่ 16 ไร่ 2 งาน มีอาคารเสนาสนะประกอบด้วย อุโบสถ สร้างเมื่อ พ.ศ.2473 ศาลาการเปรียญ สร้างเมื่อ พ.ศ.2520 หอสวดมนต์ สร้างเมื่อ พ.ศ.2538 กุฏิสงฆ์ สร้างเมื่อ พ.ศ.2527 ศาลาอเนกประสงค์ สร้างเมื่อ พ.ศ.2527 มีโรงเรียนพระปริยัติธรรมแผนกธรรม เปิดสอนเมื่อ พ.ศ.2478

ศาสนสถาน Category

Figure1. Document from cultural database including four components

The assumption of the cultural database are

1. There is only one main subject of relations in each documents.

2. The main subject belongs to only one cultural domain.

In addition to the above two assumptions, this research focuses on unary relation extraction, it was introduced by Hoffmann et al.,[5] and Chen et al.[6]. The method assumes that the subject of the relation is the title of document and each relation remains one argument to be extracted.

Table 1. RELATION TEMPLATE

Domain	Relation	Surface	Argument
Place	IsLocatedAt	ตั้งอยู่ที่ ตั้งที่ ที่ตั้ง	LOC (Location)
	IsBuiltIn	สร้างขึ้นใน สร้างใน สร้างขึ้นเมื่อ สร้างเมื่อ ตั้งขึ้นเมื่อ ตั้งเมื่อ ก่อตั้งเมื่อ	DATE
	IsBuiltBy	สร้างขึ้นโดย สร้างโดย ตั้งขึ้นโดย ตั้งโดย ก่อตั้งโดย	PER, ORG (Personal, Organization)
	HasOldName	เดิมชื่อ ชื่อเดิม	LOC,ORG (Location, Organization)
Person	MarriedWith	สมรสกับ	PER (Personal)
	HasFatherName	บิดาชื่อ	PER (Personal)
	HasMotherName	มารดาชื่อ	PER (Personal)
	HasOldName	เดิมชื่อ ชื่อเดิม	PER (Personal)
	HasBirthDate	เกิดเมื่อ	DATE
	BecomeMonkIn	อุปสมบทเมื่อ	DATE
Artifact	IsMadeBy	ผลิตขึ้นโดย ทำขึ้นโดย ผลงานโดย	PER, ORG (Personal, Organization)
	IsSoldAt	จำหน่ายที่	LOC, ORG (Location,Organization)

Table 1 shows the relation template, this focusing on three cultural domains, as shown in the first column, which are place, person and artifact. Based on these domains, the possible subject of the relations is a place, a human and a man-made object. Therefore, the set of relations corresponding to the subject, such as the subject is a place, consequently, the related information has to be *where* it is, *when* it was built and *who* built it.

The formal expressions for these relations are IsLocatedAt, IsBuiltIn and IsBuiltBy as shown in the second column.

The surface forms of the relations used for searching the relation texts are shown in the third column. Named entity types, associated with the main subject domain and their relations are shown in the forth column.

B. Searching Relation Texts

We use Apache SOLR4 for indexing and searching. Apache SOLR works well with English-language free text and it has a non english languages handling extension. For Thai text, there is ThaiWordFilterFactory module, it invokes the Java BreakIterator and specifies the locale to Thai.

The Java BreakIterator uses a simple dictionary-based method, which does not tolerate word boundary ambiguities and unknown words. This research focused on the method,which is process with Thai text in lower units called character clusters.[7] A character cluster functions as an inseparable unit which is larger than or equal to a character but smaller than or equal to a word.
We introduce Canasai's ThaiWordTokinizeFactorymodule [8] and plug it into Apache SOLR by replacing the default Whitespace Tokenizer Factory. The character cluster generator class is based on the spelling rules [9]. In the Thai language, the sentence boundary markers are not explicitly written. White spaces placing between text segments can function as word, phrase, clause or sentence boundaries. Our approach obtained a related text, it proceed as follows: After finding the position of the target relation surface, we look up at most ± 4 text segments to generate relation. This length should be enough for morphological analyzer and named entity recognizer.

C. Learning named entities

We control semantic drift of the target arguments using named entities. The named entity recognizer has been built from an annotated corpus.[10] According to the relation templates, this method trained the model with four named entity tags. The list of named entity tags are location (LOC), person name (PER), organization name (ORG) and date (DAT). Thai morphological analyzer is use to obtain word boundaries and POS tags. This work trained the morphological analyzer using ORCHID corpus[11] and TCL's lexicon [12]. In addition, we also built a special corpus from cultural database.

Subsequently, the corpus format is converted into IOB tagging format for named entity tagging. The final form of the corpus contains three columns, there are word, POS tag and named entity tag.

The sample size of this research as shown in Table 2.

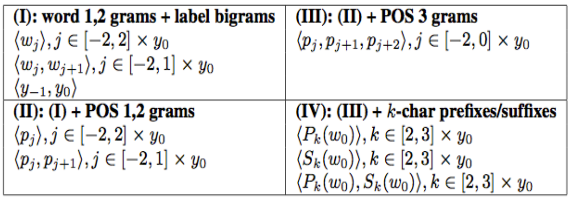
Table 2. THE SAMPLE SIZE

PER	ORG	LOC	DATE
33,231	20,398	8,585	2,783

The samples are split into two sets, 90% of the samples are used as training set , and 10% of the samples are used as test sets. We trained the named entity models using k-best MIRA-Margin Infused Relaxed Algorithm[13]. It sets k = 5, and sets the number of training iterations to 10. The method denotes the word by w , the k-character prefix and suffix of the word by $P_k(w)$ and $S_k(w)$, the POS tag by p and the NE tag by y.

The summarization of all feature combinations used in the experiments is shown in Table 3

Table 3. NAMED ENTITY FEATURES



(I): word 1,2 grams + label bigrams $\langle w_j \rangle, j \in [-2, 2] \times y_0$ $\langle w_j, w_{j+1} \rangle, j \in [-2, 1] \times y_0$ $\langle y_{-1}, y_0 \rangle$	**(III): (II) + POS 3 grams** $\langle p_j, p_{j+1}, p_{j+2} \rangle, j \in [-2, 0] \times y_0$
(II): (I) + POS 1,2 grams $\langle p_j \rangle, j \in [-2, 2] \times y_0$ $\langle p_j, p_{j+1} \rangle, j \in [-2, 1] \times y_0$	**(IV): (III) + k-char prefixes/suffixes** $\langle P_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle S_k(w_0) \rangle, k \in [2, 3] \times y_0$ $\langle P_k(w_0), S_k(w_0) \rangle, k \in [2, 3] \times y_0$

The baseline features (I) include word unigrams/bigrams and NE tag bigrams. Other features (II, III, IV) are applied to observe their effects.

The method used the conlleval perl script [14] for evaluation. Relation extraction based on all features (IV) gives best performance for named entity model as shown in Fig.2

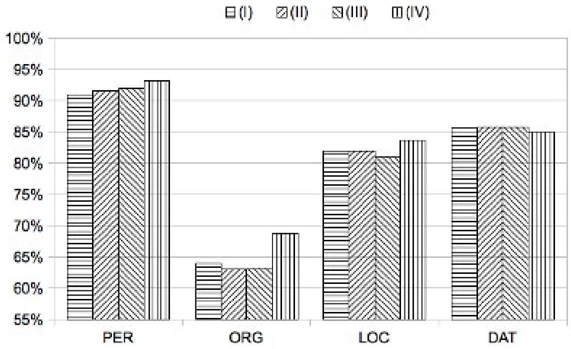


Figure 2 THE RESULT FROM NAMED ENTITY MODEL

D. Relation Extraction Enhancement

In fact, there are some of relations, such as IsLocatedDat, IsBuiltIN, IsBuiltBy, HasFatherName, HasMotherName, HasBirthDate must have only one named entity, that corresponds to the main subject of document. We improve the accuracy of relation extraction for these relations by focuses on distance between main subject and it's relation to determine the nearest position. For example as shown in Table 4, “SriNaVa Temple, that belonged to Maha Nikai was built in B.E. 2423. SriNaVa Temple has a Manor House consist of a buddhist sanctuary, which was built in B.E. 2473 and a sermon hall, which was built in B.E. 2520 ” (วัดศรีนาวา สังกัดคณะสงฆ์มหานิกาย ตั้งเมื่อ พ.ศ.2423 วัดศรีนาวามีอาคารเสนาสนะประกอบด้วย อุโบสถ สร้างเมื่อ พ.ศ.2473 ศาลาการเปรียญ สร้างเมื่อ พ.ศ.2520).

Table 4. THE DISTANCES BETWEEN SUBJECT AND RELATION INSTANCES.

Subject	Position	Relation	Distance
SriNaVa Temple วัดศรีนาวา	0	IsbuitIn ตั้งเมื่อ	95
	0	IsbuitIn สร้างเมื่อ	437
	0	IsbuitIn สร้างเมื่อ	521
SriNaVa Temple วัดศรีนาวา	315	IsbuitIn ตั้งเมื่อ	220
	315	IsbuitIn สร้างเมื่อ	122
	315	IsbuitIn สร้างเมื่อ	206
buddhist sanctuary อุโบสถ	418	IsbuitIn ตั้งเมื่อ	19
Sermon hall ศาลาการเปรียญ	481	IsbuitIn สร้างเมื่อ	40

Our approach discovered this sentences as shown in Table 5.

Table 5. THE RESULT FROM RELATION EXTRACTION ENHANCEMENT

Subject	Relation	Named Entity
SriNaVa Temple วัดศรีนาวา	IsbuitIn ตั้งเมื่อ	B.E.2423 พ.ศ. 2423
buddhist sanctuary อุโบสถ	IsbuitIn สร้างเมื่อ	B.E.2473 พ.ศ. 2473
Sermon hall ศาลาการเปรียญ	IsbuitIn สร้างเมื่อ	B.E. 2520 พ.ศ. 2520

III Experimental Results

This method is implemented in PHP and used Apache SOLR for indexing and searching, running on OS X 10.8.2, with Intel Core i5 processor running at 2.5GHz and 4GB of RAM @1600MHz.

The performance of relation extraction is verified with a cultural data set. There are more than 100,000 records collected during November, 2010 to June, 2013 . For each document, it contains at least three components, that is title, description and category.

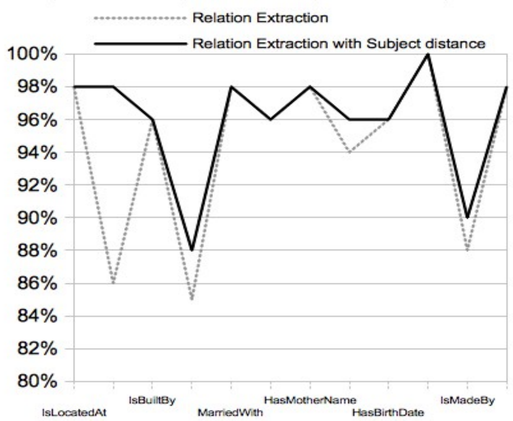
The test document set includes three domains, that is *place, person and artifact* with their relations. The highest accuracy of relation extraction is BeComeMonkIn, as shown in Table 6

Table 6. EXPERIMENTAL RESULTS

Relation	Argument	#Samples	#Correct		#Incor-rect		Accuracy	
			RE	RE + SD	RE	RE + SD	RE (%)	RE+ SD (%)
Place								
IsLocatedAt	LOC	60	59	59	1	1	98	98
IsBuiltIn	DATE	73	63	71	10	2	86	98
IsBuiltBy	PER, ORG	50	48	48	2	2	96	96
HasOldName	LOC, ORG	50	42	44	8	6	85	88
Person								
MarriedWith	PER	50	49	49	1	1	98	98
HasFatherName	PER	50	48	48	2	2	96	96
HasMotherName	PER	50	49	49	1	1	98	98
HasOldName	PER	50	47	48	3	2	94	96
HasBirthDate	DATE	50	48	48	2	2	96	96
BeComeMonkIn	DATE	50	50	50	0	0	100	100
Artifact								
IsMadeBy	PER, ORG	50	44	45	6	5	88	90
IsSoldAt	LOC, ORG	50	49	49	1	1	98	98

An accuracy comparison between two methods is shown in Fig.3

Figure 3 AN ACCURACY COMPARISON BETWEEN TWO METHODS



The accuracy of using relation template and relation extraction enhancement with subject distance is better than using only relation template.

The samples of relation instances produced by the approach is shown in table 7.

Table 7. THE RELATION INSTANCES .

Domain	Subject	Relation	Named Enity
Place	วัดชัยมังคลาราม	IsLocatedAt	ตำบลลำไทร
	วัดเกษตรสามัคคี	IsBuiltIn	24 มิถุนายน 2520
	วัดดอนตลุง	IsBuiltBy	สมเด็จพระศรีสุริยวงศ์
	วัดหนองกันเกรา	HasOldName	วัดหนองตะเกรา
Person	นายเนาวรัตน์ พงษ์ไพบูลย์	MarriedWith	นางประคองกูล อิศรางงกูร ณ อยุธยา
	พระครูสุตธรรมานุรักษ์	HasFatherName	นายเลื่อน
	นายมนตรี ตราโมท	HasMotherName	นางทองอยู่
	พระครูโสภณธรรมนาถ	HasOldName	ประดิษฐ์ นามสกุลคล้ประเสริฐ
	นายศิวกานท์ ปทุมสูติ	HasBirthDate	วันจันทร์ที่ ๑๗ สิงหาคม ๒๔๙๖
	พระครูปริยัติธรรมกิจ (มาโนชญ์)	BeComeMonkIn	วันที่ ๓๐ มิถุนายน พ.ศ. ๒๕๓๒
Artifact	หัตถกรรมจากเศษไม้ข้าวเกรียบปากหม้อ,	IsMadeBy IsSoldAt	นายสมบูรณ์ สมโพธิ์ ตลาดเทศบาลพรานกระต่าย

IV Conclusion and future work

The paper presents the relation extraction from a cultural database. The experimental results show the effectiveness of the proposed method especially for discovering more than 18,000 relation instances with expected high accuracy.

Possible future work will include an improved method to lower the false relation extraction rate, and at the same time, capability in extracting more relation instances. A repository of cultural domain documents containing a wider variety of contents is also in progress.

References

[1] Ang Sun, Ralph Grishman and Satoshi Sekine, “Semi-supervised Relation Extraction with Large-scale Word Clustering”, The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011.

[2] Dekang Lin and Patrick Pantel, “DIRT – Discovery of Inference Rules from Text ”,the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Page 323-328.

[3] Eugene Agichtein, Luis Gravano, Jeff Pavel, Viktoriya Sokolova and Aleksandr Voskoboynik,”Snowball: A Prototype System for Extracting Relations from Large Text Collections”, the 2001 ACM SIGMOD International Conference on Management of Data, Vol. 30, No. 2 ,June 2001.

[4] Pantel and Pennacchiotti , “Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relations”, the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 2006.

[5] Hoffmann, R., Zhang, C., and Weld, D. S. ,Learning 5000 Relational Extractors. In In ACL, 2010.

[6] Chen, H., Benson, E., Naseem, T., and Barzilay, R. In-domain Relation Discovery with Meta-constraints via Posterior Regularization. In Proceedings of ACL-HLT, page 530–540, 2011.

[7] Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong, T., and Chinnan, W. Character Cluster Based Thai Information Retrieval. In Proceedings of IRAL, pages 75–80, 2000.

[8] Canasai Kruengkrai, Virach Sornlertlamvanich, Watchira Buranasing, Thatsanee Charoenporn, “Semantic Relation Extraction from a Cultural Database”, In Proceedings of the 3rd Workshop on South and Southeast Asian Natural Language Processing, 2012.

[9] Kruengkrai, C., Uchimoto, K., Kazama, J., Torisawa, K., Isahara, H., and Jaruskulchai, C. “A Word and Character-Cluster Hybrid Model for Thai Word Segmentation.”, In Proceedings of InterBEST: Thai

Word Segmentation Workshop,2009.

[10] Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., and Tongtep, N. “Thai-nest: A Framework for Thai Named entity Tagging Specification and Tools”. In *Proceedings of CILC.* , 2010.

[11] Sornlertlamvanich, V., Charoenporn, T., and Isahara, H. ,” *ORCHID: Thai Part-Of-Speech Tagged Corpus”.* Technical Report TR-NECTEC-1997-001, 1997.

[12] Charoenporn, T., Kruengkrai, C., Sornlertlamvanich, V., and Isahara, H. Acquiring Semantic Information in the TCL’s Computational Lexicon. In Proceedings of the Fourth Workshop on Asia Language Resources, 2004.

[13] Crammer, K., McDonald, R., and Pereira, F. Scalable Large-Margin Online Learning for Structured Classification. In Proceedings of NIPS Workshop on Learning With Structured Outputs. , 2005.

[14] The conlleval Perl script, Conference on Computational Natural Language Learning , 2000.