Effectiveness of Keyword and Semantic Relation Extraction for Knowledge Map Generation

Virach Sornlertlamvanich[1(✉)] and Canasai Kruengkrai[2]

[1] Sirindhorn International Institute of Technology, Thammasat University, Bangkok, Thailand
virach@siit.tu.ac.th
[2] Graduate School of Information Sciences, Tohoku University, Sendai, Japan
canasai@ecei.tohoku.ac.jp

Abstract. We explore the named entity (NE) recognition and semantic relation extraction technique on the Thai cultural database. Within the limited domain and well-structured database, our proposed method can perform in an acceptable high accuracy to generate the tuples of semantic relation for expressing the essence of the record in terms of infobox and knowledge map. In this paper, we propose a semantic relation extraction approach based on simple relation templates that determine relation types and their arguments. We attempt to reduce semantic drift of the arguments by using named entity models as semantic constraints. Experimental results indicate that our approach is very promising. We successfully apply our approach to a cultural database and discover more than 18,000 relation instances with expected high accuracy.

Keywords: Named entity extraction · Semantic relation extraction · Cultural database · Infobox · Knowledge map

1 Introduction

Targeting on the user generated content (UGC) e.g. Thai Cultural Information Center website,[1] we are interested in relating the document units semantically to generate a network that can express in a knowledge map manner. In our approach, we focus on keyword and semantic relation extraction. Some language dependent problems have to be solved especially in handling the Thai language, which has no word delimiter or punctuation mark. We apply general tools for word segmentation and POS tagging, then extract the keyword according to the model trained from named entity (NE) tagged corpus.

The size of this cultural database has gradually increased to around 100,000 records (from November 2010 to December 2014). Each record contains a number of fields describing a specific cultural object. The content includes four main components: (1) cover image and thumbnails, (2) title, (3) description and (4) domain. We need to

[1] http://www.m-culture.in.th/.

Y. Murakami and D. Lin (Eds.): WLSI 2015, LNAI 9442, pp. 188–199, 2016.
DOI: 10.1007/978-3-319-31468-6_14

extract facts (hereafter referred to as relation instances) from the description. One can view relation instances as formal meaning representations of corresponding texts. These relation instances are useful for question answering and other applications i.e. summary as an infobox, or a network of information in knowledge map.

Recent research in semantic relation extraction has shown the possibility to automatically find such relation instances. Some approaches rely on high-quality syntactic parsers. For example, DIRT [10] and USP [12] discover relation instances based on the outputs from dependency parsers. Such parsers and annotated training corpora are difficult to obtain in non-English languages. Pattern-based approaches [1, 2, 11] seem to be more practical for languages with limited NLP resources. For example, TEXTRUNNER [2] can efficiently extract relation instances from a large-scale Web corpus with minimal supervision. It only requires a lightweight noun phrase chunker to identify relation arguments. More advanced approaches like SNE [7], RESOLVER [17] and SHERLOCK [13] exploit the outputs of TEXTRUNNER for learning.

Our cultural database allows us to make two assumptions:

(A1) Each record belongs to only one main cultural domain.
(A2) Each record has only one subject of relations.

The assumption (A1) seems to hold for most of records. We adopt the assumption (A2) from [6] that try to extract infobox-like relations from Wikipedia. Also, the assumption (A2) seems to hold for our data since the description provides the details about one cultural object whose name is expressed in the record title.

Based on the above two assumptions, we propose our strategy to semi-automatically extract relation instances from the cultural database. We focus on unary relation extraction similar to [4, 6]. We assume that the subject of the relation is the record title. Each relation remains only one argument to be extracted.

The Thai cultural database has been collected in the structure as described in Sect. 2. We describe our relation template in Sect. 3 and how to effectively find relation texts in a large database in Sect. 4. We use named entities to reduce semantic drift of the target arguments in Sect. 5. We examine the effect of the distances between the relation surfaces and the target arguments in Sect. 6.1 and provide preliminary results of our experiments in Sect. 6.2. The results indicate that our strategy of semantic relation extraction is very promising for real-world applications by applying to generate infobox and knowledge map of the Thai cultural database as described in Sect. 7.

2 Thai Cultural Database

The portal is hosted by Thai Ministry of Culture providing for cultural rural office to collect culture information online. The cultural information is structured to follow a template guideline mainly adopted from Dublin Core Metadata Element Set, Version 1.1.[2]

[2] http://dublincore.org/documents/2012/06/14/dces/.

There are 15 elements that have been introduced to annotate the record as elaborated in Table 1. The record is allowed to contain text, image, and video. In the period of November 2010 to December 2014, the number of uploaded records has already exceeded 100,000 records.

Table 1. The elements for annotating the content of the cultural information

Label	Definition
dc.title	Name of the culture resource
dc.subject	Set of tags or keywords representing the category of the resource
dc.description	Detail about the resource
dc.type	Type of attaching media i.e. image, video, sound, SWF
dc.relation	Reference identification to other resource
dc.coverage	Location of the resource
dc.creator	Person primarily responsible for making the resource
dc.publisher	Person responsible for making the resource available
dc.contributor	Person responsible for making contributions to the resource.
dc.rights	Information about rights held in and over the resource
dc.date	Point of time describing the last updating, creating, submitting, approving, contributing the detail of the resource
dc.identifier	Unambiguous reference to the resource within a given context
dc.language	Language of the resource
dc.source	Name of the attached media file
dc.format	File format, physical medium, or dimensions of the resource

As an example, Fig. 1 shows an excerpt of the front-end web page of the record number 35860 about the Phra Samut Chedi; (1) is the photo images of the record, (2) is the title of the record, (3) is the description of the record, (4) is the subject of the record.

The annotated information of title, description, and subject are the essential key fields that we use to identify the NE for keyword and semantic relation extraction. Subject is used to filter for the records of attraction (location), person, and artifact. These are the group of NE in which we are interested in this paper. Title is the target NE according to our assumption to identify the semantic relation to any occurrence of related NE in the description.

3 Relation Template

Table 2 shows the relation template. There are five main cultural domains in the database, and each main cultural domain has several sub-domains.

In our work [9], we focus on three cultural domains, including attraction, person and artifact, as shown in the first column. Based on these cultural domains, we expect that the subject of relations in each record (i.e., the record title) should be a place, a human or a man-made object, respectively. As a consequence, we can design a set of relations that correspond to the subject. For example, if the subject is a place, we may

Fig. 1. An excerpt of the front-end web page of the record number 35860 about the Phra Samut Chedi

Table 2. Relation template (LOC denotes location; PER denotes person; ORG denotes organization; DATE denotes date)

Domain	Relation	Surface	Argument
Cultural attraction	ISLOCATEDAT	ตั้งอยู่ที่	LOC
	ISBUILTIN	สร้าง(ขึ้น)*ใน สร้าง(ขึ้น)*เมื่อ ตั้ง(ขึ้น)*เมื่อ	DATE
	ISBUILTBY	สร้าง(ขึ้น)*โดย ตั้ง(ขึ้น)*โดย	PER, ORG
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	LOC, ORG
Cultural person	MARRIEDWITH	สมรสกับ	PER
	HASFATHERNAME	บิดาชื่อ	PER
	HASMOTHERNAME	มารดาชื่อ	PER
	HASOLDNAME	เดิมชื่อ ชื่อเดิม	PER
	HASBIRTHDATE	เกิด(เมื่อ)*	DATE
	BECOMEMONKIN	อุปสมบทเมื่อ	DATE
Cultural artifact	ISMADEBY	ผลิต(ขึ้น)*โดย ทำ(ขึ้น)*โดย ผลงานโดย	PER, ORG
	ISSOLDAT	จำหน่ายที่	LOC, ORG

need to know where it is, when it was built and who built it. We can formally write these expressions by ISLOCATEDAT, ISBUILTIN and ISBUILTBY. The second column shows our relations that are associated with the subject domains. The third column shows relation surfaces used for searching relation texts in which arguments

may co-occur. The word in parentheses with an asterisk indicates that it may or may not appear in the surface.

The answers to where, when and who questions are typically short and expressed in the form of noun phrases. Using noun phrases as relation arguments can lead to high recall but low precision. For example, the noun phrase occurring after the relation ISBUILTIN could be a place (is built in the area of…) or an expression of time (is built in the year of…). In our case, we expect the answer to be the expression of time, and hence returning the place is irrelevant. This issue can be thought of as semantic drift. Here, we attempt to reduce semantic drift of the target arguments by using named entities as semantic constraints. The forth column shows named entity types associated with the subject domains and their relations. Each relation can be expressed in more than one surface in the text. However, the surface list in Table 2 is not thoroughly expressed. Many other more can be extracted from the corpus.

4 Surface-Relation Mapping

Mapping text segments containing a given relation surface (e.g., "สร้างโดย" (is built by)) in a large database is not a trivial task. Here, we use Apache Solr[3] for indexing and searching the database. Apache Solr works well with English and also has extensions for handling non-English languages. To process Thai text, one just enables `ThaiWordFilterFactory` module in schema.xml. This module invokes the Java `BreakIterator` and specifies the locale to Thai (TH). The Java `BreakIterator` uses a simple dictionary-based method, which does not tolerate word boundary ambiguities and unknown words. For example, the words "สร้าง" (build) and "ก่อสร้าง" (construct) occur in the Java's system dictionary. Both convey the same meaning (to build). We can see that the first word is a part of the second word. However, these two words are indexed differently. This means if our query is "สร้าง" (build), we cannot retrieve the records containing "ก่อสร้าง" (construct). In other words, the dictionary-based search returns results with high precision but low recall.

In our work, we process Thai text in lower units called character clusters. A character cluster functions as an inseparable unit, which is larger than (or equal to) a character and smaller than (or equal to) a word. Once the character cluster is produced, it cannot be further divided into smaller units. For example, we can divide the word "ก่อสร้าง" (construct) into 5 character clusters like "ก่-อ-ส-ร้า-ง". As a result, if our query is "สร้าง" (build), we can retrieve the records containing "ก่อสร้าง" (construct). We refer to [16] for more details about character cluster based indexing. In our work, we implement our own `ThaiWordTokinizeFactory` module and plug it into Apache Solr by replacing the default `WhitespaceTokenizerFactory`. Our character cluster generator class is based on the spelling rules described in [8].

In Thai, sentence boundary markers (e.g., a full stop) are not explicitly written. The white spaces placing among text segments can function as word, phrase, clause or sentence boundaries (see the "รายละเอียด" (description) section in Fig. 1 for example).

[3] http://lucene.apache.org/solr/.

To obtain a relation text, which is not too short (one text segment) or too long (a whole paragraph), we proceed as follows. After finding the position of the target relation surface, we look up at most ±4 text segments to generate relation texts. This length should be enough for morphological analyzer and named entity recognizer.

5 Named Entity Recognition

We control semantic drift of the target arguments using named entities. We build our named entity (NE) recognizer from an annotated corpus developed by [15]. The original contents are from several news websites. The corpus consists of 7 NE types. We focus on 4 NE types according to our relation templates in Table 2. Once we obtained the NE corpus, we checked it and found several issues as follows:

1. Each NE tag contains nested NE tags. For example, the person name tag contains the forename and surname tags.
2. The corpus does not provide gold word boundaries and POS tags.
3. Each NE type is annotated separately.

For the first issue, we ignored the nested NE tags and trained our model with top NE tags (PER, ORG, LOC, DATE). For the second issue, we used a state-of-the-art Thai morphological analyzer [8] to obtain word boundaries and POS tags. In this work, we trained the morphological analyzer using ORCHID corpus [14] and TCL's lexicon [3]. We then converted the corpus format into the IOB tagging style for NE tags. Thus, the final form of our corpus contains three columns (word, POS tag, NE tag), where the first two columns are automatically generated and of course contain a number of errors. For the third issue, we trained the model separately for each NE type. We obtained 33231, 20398, 8585, 2783 samples for PER, ORG, LOC, DATE, respectively.

To ensure that our NE models work properly, we split samples into 90 %/10 % training/test sets and conducted some experiments. We trained our NE models using k-best MIRA (Margin Infused Relaxed Algorithm) [5]. We set k = 5 and the number of training iterations to 10. We denote 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. Table 3 summarizes all feature combinations used in our experiments. Our baseline features (I) include word unigrams/bigrams and NE tag bigrams. Since we obtained the word boundaries and POS tags automatically, we introduced them gradually to our features (II, III, IV) to observe their effects.

Figure 2 shows F1 results for the NE models. We used the `conlleval` script[4] for evaluation. We observe that PER is easy to identify, while ORG is difficult. Prefix/suffix features dramatically improve performance on ORG. Using all features (IV) gives best performance on PER (93.24 %), ORG (68.75 %) and LOC (83.78 %), while slightly drops performance on DATE (85.06 %). Thus, our final NE models used in relation extraction are based on all features (IV). Although these results are from the

[4] http://www.cnts.ua.ac.be/conll2000/chunking/conlleval.txt.

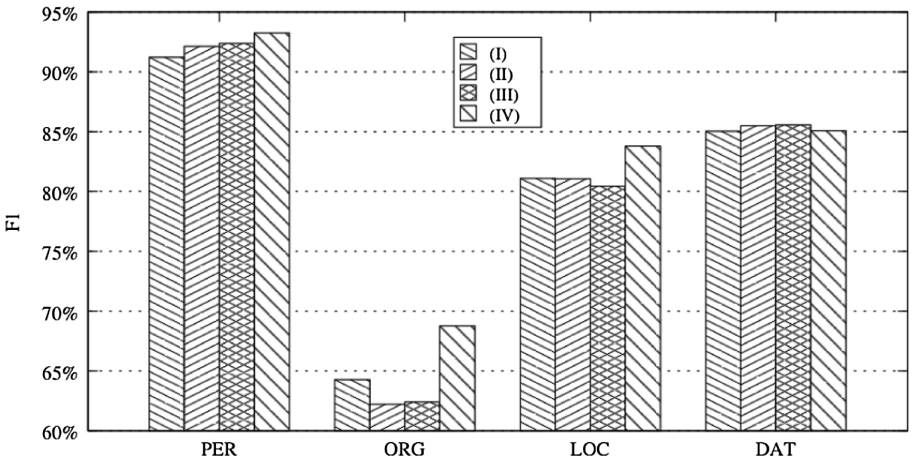


Fig. 2. F1 evaluation results of the NE models

news domain, we could expect similar performance when applying the NE models to our cultural domains.

We summarize our strategy as follows. After selecting the subject domain, we send its relation surfaces (shown in the 3rd column of Table 2) to Apache Solr. We then trim the resulting record descriptions to obtain the relation texts (described in Sect. 4). Next, we perform word segmentation and POS tagging simultaneously using our morphological analyzer and feed the results into our NE models (described in Sect. 5). We invoke the appropriate NE model based on our relation templates (described in Sect. 3). Finally, our system produces outputs in the form of RELATION(a, b), where a is a record title, and b is an argument specified by its NE type in the templates.

Table 3. NE 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** $\langle P_k(w_0) \rangle$, $k \in [2, 3] \times y_0$ $\langle S_k(w_0) \rangle$, $k \in [2, 3] \times y_0$ $\langle P_k(w_0), S_k(w_0) \rangle$, $k \in [2, 3] \times y_0$

6 Experiments

6.1 Effect of the Distances Between Relation Surfaces and Arguments

In this section, we examine the number of extracted instances for each relation (without considering its accuracy). Our assumption is that the target argument tends to be

relevant if it is adjacent (or close) to the relation surface. The relevance weakens with the distance. In our first example, the target argument "ตำบลปากน้ำ" (Tambon Paknam, a subdistrict name) is adjacent (distance = 0) to the relation surface "ตั้งอยู่ที่" (is located at). This target argument is relevant. Suppose there are intervening words (white space or punctuation mark) between them. The relevance tends to decrease. However, if we only select adjacent named entities to be the target arguments, the coverage may be limited. In our experiments, we varied the distances from 0 to 5 intervening words for observation.

Table 4 shows the numbers of relation instances when the distances are varied. For all relations, we observe that the numbers of relation instances do not significantly change after one word distance. For example, we cannot extract more relation instances for MARRIEDWITH + PER, even we increased the distance. This indicates that using named entities helps to bound the number of possible arguments.

Table 4. Numbers of relation instances when the distances are varied

Relation	Argument	Distance					
		0	1	2	3	4	5
Cultural attraction							
ISLOCATEDAT	LOC	356	574	591	624	678	757
ISBUILDIN	DATE	3825	11487	11538	11573	11633	11667
ISBUILDBY	PER, ORG	131	202	218	234	249	257
HASOLDNAME	LOC, ORG	0	9	21	26	27	29
Cultural person							
MARRIEDWITH	PER	132	177	177	177	177	177
HASFATHERNAME	PER	120	372	372	373	373	373
HASMOTHERNAME	PER	97	383	383	383	383	383
HASOLDNAME	PER	51	259	273	277	277	283
HASBIRTHDATE	DATE	4122	4745	4801	4947	4966	5075
BECOMEMONKIN	DATE	346	435	435	436	436	436
Cultural artifact							
ISMADEBY	PER, ORG	62	107	109	125	129	130
ISSOLDAT	LOC, ORG	31	31	56	59	62	64

6.2 Preliminary Results

To inspect the quality of relation instances extracted by our strategy, we randomly selected at most 50 instances of each relation for evaluation. Our evaluation procedure is as follows. Based on the assumptions (A1) and (A2), we expect that the subject (record title) of an instance should be relevant to its domain. We ignored instances whose subject is irrelevant. For example, the subject of the record no. 8026 is a person, but the volunteer assigned it to the cultural artifact domain. Note that this case rarely occurs, but exists. Next, a relation instance is considered to be correctly extracted if its argument exactly matches the fact. For example, if our system only extracts the first

name while the fact is the whole name, then we consider this instance to be incorrect. Finally, we set the maximum distance between the relation surface and its argument to 5. Table 5 shows the performance of our relation extraction. The overall results are surprisingly good, except those of HASOLDNAME and ISMADEBY. Table 6 shows some samples of relation instances produced by our system.

Table 5. Performance of the relation extraction

Relation	Argument	#Sample	#Correct	#Incorrect	Accuracy
Cultural attraction					
ISLOCATEDAT	LOC	50	49	1	98 %
ISBUILDIN	DATE	50	48	2	96 %
ISBUILDBY	PER, ORG	50	48	2	96 %
HASOLDNAME	LOC, ORG	27	23	4	85 %
Cultural person					
MARRIEDWITH	PER	50	49	1	98 %
HASFATHERNAME	PER	50	48	2	96 %
HASMOTHERNAME	PER	50	49	1	98 %
HASOLDNAME	PER	50	47	3	94 %
HASBIRTHDATE	DATE	50	48	2	96 %
BECOMEMONKIN	DATE	50	50	0	100 %
Cultural artifact					
ISMADEBY	PER, ORG	50	44	6	88 %
ISSOLDAT	LOC, ORG	50	49	1	98 %

7 Knowledge Map Generation

Relations between NE (or keyword) are successfully extracted as shown in the result in Table 6. The accuracy is acceptably high, ranging from 85 % to 100 % corresponding to the type of the relation. The tuples of relation are stored attaching to the record they belong to. Though the tuple of semantic relation is extracted from a part of the description, it determines the semantic modification to the title of the record. From the set of tuples of each record, the infobox of the record is generated to express the essence of the title we are looking for. NE's are used to modify the title which is also included in the set of NE. By mapping the NE found in the database, we can extensively trace the semantic modification of any target NE. Finally, the knowledge map, which is a network of the NE can be express to understand the relation among all NE's in the database.

Figure 3 shows the tuples of semantic relation extracted from the record of Phra Samut Chedi i.e.

ISBUILDIN(พระเจดีย์กลางน้ำ, พ.ศ. 2403)
Lit. ISBUILDIN(Phra Samut Chedi, BE 2403), and
ISLOCATEDAT(พระเจดีย์กลางน้ำ, ตำบลปากน้ำ).
Lit. ISLOCATEDIN(Phra Samut Chedi, Tambon Paknam).

Table 6. Relation instances produced by the system

Record no.	Relation instance
Cultural attraction	
38481	ISLOCATEDAT(วัดโพธิ์ศรี, บ้านโพธิ์ศรี ต.อินทร์บุรี)
114585	ISBUILDIN(วัดเบาวงกฏ, ประมาณปี พ.ศ.2471-2573)
114333	ISBUILDBY(วัดปีตุลาธิราชรังสฤษฎ์, กรมหลวงรักษ์รณเรศร์)
61446	HASOLDNAME(วัดหนองกันเกรา, วัดหนองตะเกรา)
Cultural person	
14125	MARRIEDWITH(นายเนาวรัตน์ พงษ์ไพบูลย์, นางประคองกูล อิศรางกูร ณ อยุธยา)
32530	HASFATHERNAME(พระครูประยุตนวการ, นายเหยม เดชมาก)
45389	HASMOTHERNAME(หลวงพ่อลิ้ง สุทสฺสโน, นางพริ้ง แก้วแดง)
144574	HASOLDNAME(พระครูมงคลวรวัฒน์, สวัสดีบพุศิริ)
145771	HASBIRTHDATE(อาจารย์ธนิสร์ ศรีกลีนดี, วันจันทร์ที่ 23 มกราคม 2494)
123678	BECOMEMONKIN(พระครูพิจิตรสิทธิคุณ, วันที่ ๑๖ เมษายน พ.ศ. ๒๕๒๘)
Cultural artifact	
160974	ISMADEBY(หนังสือประวัติคลองดำเนินสะดวก, พระครูสิริวรรณวิวัฒน์)
94286	ISSOLDAT(ข้าวเกรียบปากหม้อ, ตลาดเทศบาลพรานกระต่าย)

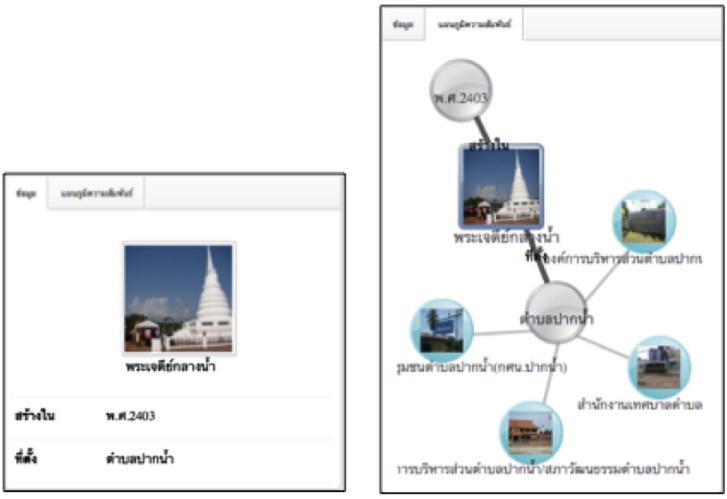
ISBUILDIN(พระเจดีย์กลางน้ำ, พ.ศ. 2403)

ISLOCATEDAT(พระเจดีย์กลางน้ำ, ตำบลปากน้ำ)

Fig. 3. Tuples of semantic relation extracted from the record of Phra Samut Chedi

In the infobox as shown in Fig. 4(1), it notifies when and where the Phra Samut Chedi was constructed. The summary information about the record in the form of infobox can help the audience to grasp the information about the record in quick. By knowing that the pagoda (Chedi) was founded in Tambon Paknam, we can trace further for what else are related to the NE of Tambon Paknam. As a result, we can find that

many other attractions are located in this Tambon Paknam. These records can then be attached to the location name of Tambon Paknam. The example of the knowledge map expression is shown in Fig. 4(2). The audience can traverse for other related information about the focus topic and understand the relation among the records. Further level of relation can be expended as far as they are connected with the extracted tuples of semantic relation.



(1) Infobox (2) Knowledge map

Fig. 4. Infobox and knowledge map extracted from the cultural database for the record of Phra Samut Chedi

8 Conclusion

We successfully applied our approach to a cultural database and could discover more than 18,000 relation instances with expected high accuracy. The outputs of our NE and relation extraction can be useful for other applications such as question answering or suggesting related topics based on semantic relations. For an example, attaching the tuples of semantic relation to the corresponding record, we can express the essence of the record in terms of infobox. In addition, by mapping among the NE's, a network of NE can be generated to form a knowledge map for better understanding the content of the cultural database.

In future work, many more other semantic relations are interested, especially in the cultural artifact domain. As an example, the relations like ISMADEOF, which requires the NE type like materials, can help in understanding the raw materials from what the artifacts are made. However, this NE type is not available in the current NE corpus. We will explore other techniques to constrain the noun phrases to prevent the semantic drift problem.

Acknowledgement. The experiments in this paper are conducted on the Thai Cultural Database of the Ministry of Culture, developed under the central information project since November 2010.

References

1. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Proceedings of ICDL, pp. 85–94 (2000)
2. Banko, M., Cafarella, M.J., Soderl, S., Broadhead, M., Etzioni, O.: Open information extraction from the web. In: Proceedings of IJCAI, pp. 2670–2676 (2007)
3. Charoenporn, T., Kruengkrai, C., Sornlertlamvanich, V., Isahara, H.: Acquiring semantic information in the TCL's computational lexicon. In: Proceedings of the Fourth Workshop on Asia Language Resources (2004)
4. Chen, H., Benson, E., Naseem, T., Barzilay, R.: In-domain relation discovery with meta-constraints via posterior regularization. In: Proceedings of ACL-HLT, pp. 530–540 (2011)
5. Crammer, K., McDonald, R., Pereira, F.: Scalable large-margin online learning for structured classification. In: Proceedings of NIPS Workshop on Learning with Structured Outputs (2005)
6. Hoffmann, R., Zhang, C., Weld, D.S.: Learning 5000 relational extractors. In: Proceedings of ACL (2010)
7. Kok, S., Domingos, P.: Extracting semantic networks from text via relational clustering. In: Daelemans, W., Goethals, B., Morik, K. (eds.) ECML PKDD 2008, Part I. LNCS (LNAI), vol. 5211, pp. 624–639. Springer, Heidelberg (2008)
8. Kruengkrai, C., Uchimoto, K., Kazama, J., Torisawa, K., Isahara, H., Jaruskulchai, C.: A word and character-cluster hybrid model for Thai word segmentation. In: Proceedings of InterBEST: Thai Word Segmentation Workshop (2009)
9. Kruengkrai, C., Sornlertlamvanich, V., Buranasing, W., Charoenporn, T.: Semantic relation extraction from a cultural database. In: Proceedings of The 3rd Workshop on South and Southeast Asian NLP (2012)
10. Lin, D., Pantel, P.: Dirt-discovery of inference rules from text. In: Proceedings of KDD, pp. 323–328 (2001)
11. Pantel, P., Pennacchiotti, M.: Espresso: leveraging generic patterns for automatically harvesting semantic relations. In: Proceedings of ACL, pp. 113–120 (2006)
12. Poon, H., Domingos, P.: Unsupervised semantic parsing. In: Proceedings of EMNLP, pp. 1–10 (2009)
13. Schoenmackers, S., Etzioni, O., Weld, D.S., Davis, J.: Learning first-order horn clauses from web text. In: Proceedings of EMNLP, pp. 1088–1098 (2010)
14. Sornlertlamvanich, V., Charoenporn, T., Isahara, H.: ORCHID: Thai part-of-speech tagged corpus. Technical report TR-NECTEC-1997-001, NECTEC (1997)
15. Theeramunkong, T., Boriboon, M., Haruechaiyasak, C., Kittiphattanabawon, N., Kosawat, K., Onsuwan, C., Siriwat, I., Suwanapong, T., Tongtep, N.: THAI-NEST: a framework for Thai named entity tagging specification and tools. In: Proceedings of CILC (2010)
16. Theeramunkong, T., Sornlertlamvanich, V., Tanhermhong, T., Chinnan, W.: Character cluster based Thai information retrieval. In Proceedings of IRAL, pp. 75–80 (2000)
17. Yates, A., Etzioni, O.: Unsupervised methods for determining object and relation synonyms on the web. J. Artif. Intell. Res. **34**, 255–296 (2009)