Journal of Theoretical and Applied Information Technology
31st July 2020. Vol.98. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ONTOLOGY-BASED SEMANTIC EXTRACTION APPROACH FROM TEXT CORPUS

[1]KITTIPHONG SENGLOILUEAN, [2]NGAMNIJ ARCH-INT, [3]SOMJIT ARCH-INT

[1]Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand

[2]Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand

[3]Department of Computer Science, Faculty of Science, Khon Kaen University, Thailand

E-mail: [1]kittiphong.s@psu.ac.th, [2]ngamnij@kku.ac.th, [3]somjit@kku.ac.th

ABSTRACT

The Semantic Web is the salient technology of knowledge management, consisting of data extraction and annotation processes, which requires semantic representation to express data in an ontological format. The ontological extraction of unstructured data to enable the automatic generation of concepts and relations has led us to the presentation of our unique approach of automatic ontology extraction. However, domain experts are still required to modify the structure of ontological results, which makes the process very time-consuming and costly. Yet, there still exists the need for an ontology-based semantic extraction approach from text corpus to discover concepts, instances, and semantic relations between concepts or instances. This paper presents an approach of an ontology-based semantic extraction and the accompanying semantic extraction rules, as applied to tourism domain. The proposed semantic extraction rules are defined as extension rules working with GATE API. As a result, the efficiency of the proposed ontological extraction approach is validated through the Precision, Recall and F-measure scores, with average values of 91.48%, 89.12%, and 90.23%, respectively.

Keywords: *Ontology extraction, semantic extraction rules, unstructured data*

1. INTRODUCTION

Ontology is the fundamental concept of establishing Semantic Web standards [1, 9], generally designed and constructed by domain experts, in order to represent the common explicit concepts and relations garnered from their knowledge. Yet, the ontological engineering processes required to extract ontology from an unstructured text may be very costly and time-consuming. This has motivated several studies [2-4] involving the ontology extraction from unstructured texts, enabling semi-automatic mechanical ontological learning. However, limitations within the previous studies include the lack of flexibility and correctness in identifying the terms of common concepts and relations in the ontology extraction process. Other studies [5-7] clustered types of terms, such as *noun, verb, etc*., which were unable to find the relation of terms needed to link common ontological concepts.

This paper presents an ontology-based semantic extraction approach designed to enable the automatic derivation of concepts and relations from unstructured texts, as applied in the tourism domain. This research employs the GATE API [8] for the initial extraction of basic concepts. The semantic extraction rules have been extended to the GATE API in order to discover concept and concept relationship instances; such as concept hierarchy, and concept and instance properties, as well as to resolve the semantic identification problems, which occur in uncertain context expressions. The ontology extraction experiments were conducted in the domain of Tourism and were improved and validated through correctness and completeness evaluation metrics, namely Precision, Recall, and F-measure.

The remainder of this paper is organized as follows: Section 2 describes theoretical background and related works, Section 3 describes the proposed architecture of our ontology extraction, Section 4 presents the experiments and evaluation, and Section 5 presents our conclusion and plans for future work.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2. THEORETICAL BACKGROUND AND RELATED WORKS

The Semantic Web [1, 9] is a concept of enhanced technologies of existing Web standards (markup languages), which provides the means to understand information published on the Web, to humans, computers, or agent software. The information can be further processed, reused, and re-analyzed efficiently, through the use of semantic representation languages, such as Resource Description Framework (RDF) [10, 11], and Web Ontology Language (OWL) [12]; which have proved appropriate for ontological expression in various domains [13, 14]. The ontology consisting of concepts (or classes), relationships (or properties), individuals (or instances), restrictions, and rules; and are typically used to describe data with which to form explicit and non-ambiguous information.

Several studies have attempted to resolve the problems presented in ontology extraction. Jung [15] presented the Natural Language Processing-based (NLP) triple extractor, which employed language analysis techniques for generating the ontological triples (or statements). In the area of the lexicon-based approach, Anantharangachar, et al. [3] proposed a technique for extracting terminologies of text using the WordNet lexicon. The studies of Abedini, et al. and Hoffart, et al. [16, 17] proposed an approach for extracting semantic entities from text by matching words with entities of YAGO ontology, derived from the WordNet lexicon and various ontologies, such as DBpedia and SUMO. Focusing on the approach based on multi-strategy, the GATE was introduced as a framework for text extraction and annotation, which provides integrated methodologies of ontology extraction such as, language, pattern analysis, and lexicon-based matching [8]. This tool also provides a finite state transduction language, Java Annotation Patterns Engine (JAPE) [18], which is a flexible rules-expression mechanism for string pattern matching. Witte, et al. [19] presented the ontology extraction tool Owl Explorer which adopts JAPE to extract and generate text ontology through the proposed rules. Although the studies presented above provide the means to automatically generate text ontology, the flexibility and correctness of ontology extraction still remain limited. The automated extraction process and XML-to-Ontology process of our proposed approach has the flexibility to find specify extracted terms within more explicit concepts.

3. PROPOSED ARCHITECTURE

This paper presents an approach of an ontology-based semantic extraction applied to unstructured data, such as in the web page, illustrated below. Details of the ontology extraction process are described in the following paragraphs.

The extraction process is presented with an ontology learning approach through the extraction rule, working with GATE API, to extract knowledge from unstructured text to form an initial ontology or local ontology. An overview of the procedure of our proposed approach is shown in Figure1.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

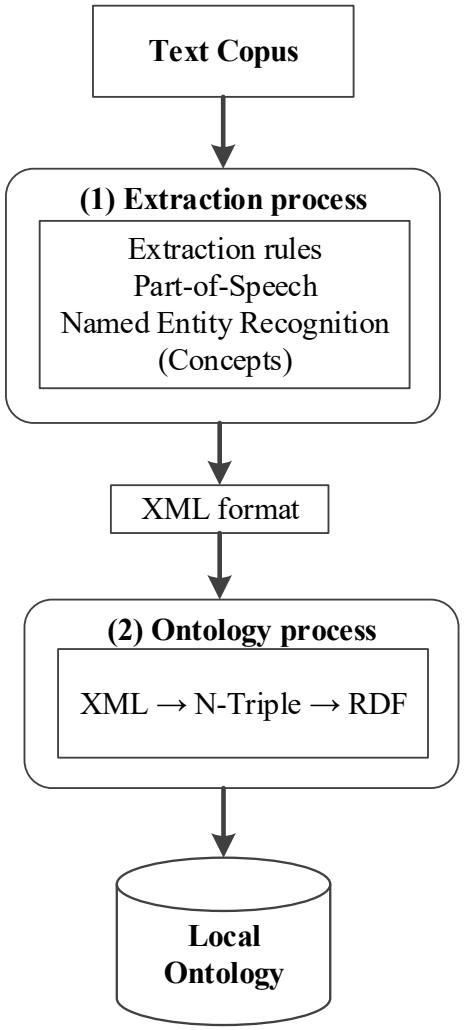

Figure 1: The procedure of an ontology-based semantic extraction approach.

The ontology extraction approach consists of two main sub-processes: (1) Extraction process, and (2) Ontology process, described as follows:

(1) Extraction process uses GATE API for extracting instances, and annotating the relevant concepts. GATE API enables developers to annotate specific concepts according to their needs, through the associated JAPE rules. The specific concepts are provided in the vocabulary of the developer's ontology. This way, specific concepts, which are sub-concepts of general concepts (i.e., *location, person, organization,* etc.), can be proposed by the developers. Figure 2 shows an example of the JAPE rule, which uses keywords to annotate the specific types of an organization instance within the tourism domain (such as *Hotel, Museum, Restaurant,* etc.), depending on the appearance of these concepts, in the organization instance's value. For example, the organization instance's value of "*Pullman Hotel*" may be determined to be an instance of the Hotel concept.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
Phase: TitlePhase
Input: Lookup Organization Token
Options: control = appelt

Rule: FindOrganization1
(
    {Organization}
):temp
-->
{
    AnnotationSet anno = bindings.get("temp");
    if(anno != null && anno.size() > 0)
    {
        int beginOffset = anno.firstNode().getOffset().intValue();
        int endOffset = anno.lastNode().getOffset().intValue();
        String mydocContent = doc.getContent().toString();
        String matchedString = mydocContent.substring(beginOffset, endOffset);

        String[] organTypes = new String[]{
            "Hotel", "Museum", "Hospital", "Transportation", "Restaurant", "Company"
        };
        for(int i = 0; i < organTypes.length; i++)
        {
            String lowerType = organTypes[i].toLowerCase();
            if(matchedString.contains(organTypes[i]) || matchedString.contains(lowerType))
            {
                FeatureMap features = Factory.newFeatureMap();
                features.put("rule", "FindOrganization1");
                outputAS.add(anno.firstNode(), anno.lastNode(), organTypes[i], features);
            }
        }
    }
}
```

Figure 2: The extraction rule implemented through JAPE language.

The annotated results ae exported to XML format. An example of such is shown in Figure 3.

(2) Ontology process transforms annotated results into ontological structure. The transformation is implemented through an ontology construction algorithm, shown in Figure 4. This algorithm requires as inputs; *a set of pairs of concepts*, their contexts, and the subject of the web page. Essentially, the algorithm defines the relationship between input entities. The outputs are expressed in N-Triple format [20, 21], as shown in Table 1. Lastly, these N-Triple statements are stored in RDF format, which are further used in the processes of ontology integration and ontology mapping, which enable the interoperability between extracted ontology and other existing ontologies. Sample results of the *Ontology process* are shown in Figure 5.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

```
<Root>
  <Organization>
    <Hotel>Park International Hotel</Hotel>
  </Organization>
  <Location>
    <City>London</City>
    <Country>United Kingdom</Country>
  </Location>
  <Contact>
    <Address>129 Cromwell Road London SW7 4DS, United Kingdom</Address>
    <Phone>020 7370 5711</Phone>
    <Email>reservations@parkinternationalhotel.com</Email>
    <Website>www.parkinternationalhotel.com</Website>
  </Contact>
  <Price>$182</Price>
</Root>
```

Figure 3: An example of annotated text in XML format.

Input : E is a set of XML elements and their sub-element
s is a subject of a wiki page
Output: T is a set of triples in RDF

i = null, a subject instance
$D = \phi$, a set of pairs of datatype properties and their values
$O = \phi$, a set of pairs of object properties and their object instances
$T = \phi$
for $(e_i, e_j) \in E$
  if $e_i \cong s$ **then**
    v_j = getNodeValue(e_j)
    i = newInstance(e_j, v_j)
    $T = T \cup \{(e_j, subClassOf, e_i)\}$
    // Find subClassOf
  else if e_j is a set of sub-elements **then**
    N_j = getChildNodes(e_j)
    i_j = newInstance(e_i, N_j)
    op_i = newObjectProperty(e_i)
    $O = O \cup \{(op_j, i_j)\}$
    // Find object property
  else if e_j is a node value **then**
    $v_j \leftarrow e_j$
    d_j = newDatatypeProperty(e_i)
    $D = D \cup \{(d_j, v_j)\}$
    // Find datatype property
  end if
end for
for $(op_i, i_i) \in O$
  $T = T \cup \{(i, op_i, i_i)\}$
  // add object property
end for
for $(d_i, v_i) \in D$
  $T = T \cup \{(i, d_i, v_i)\}$
  // add datatype property
end for

return T

Figure 4: The XML-to-Ontology construction algorithm.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1 presents an example of annotated text in N-Triple format, which describes the relationship between two entities in the form of subject-predicate-object. In this example, the generated instance wk:H00001 is identified as an instance of a class Hotel, wk:L00001 is identified as an instance of a class Location, and wk:C00001 is identified as an instance of a class Contact.

Table 1. An example of annotated text in triples.

Subject	Predicate	Object
wk:Hotel	rdfs:subClassOf	wk:Organization
wk:H00001	rdf:type	wk:Hotel
wk:H00001	rdf:label	"Park International Hotel"
wk:H00001	wk:hasPrice	"$182"
wk:H00001	wk:hasLocation	wk:L00001
wk:L00001	rdf:type	wk:Location
wk:H00001	wk:hasContact	wk:C00001
wk:C00001	rdf:type	wk:Contact
wk:L00001	wk:hasCity	"London"
wk:L00001	wk:hasCountry	"United Kingdom"
wk:C00001	wk:hasPhone	"020 7370 5711"
wk:C00001	wk:hasAddress	"129 Cromwell Road London SW7 4DS, United Kingdom"
wk:C00001	wk:hasWebsite	"www.parkinternationalhotel.com"
wk:C00001	wk:hasEmail	"reservations@parkinternationalhotel.com"

Note:

wk = http://www.semanticweb.org/webpage#
rdf = http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs = http://www.w3.org/2000/01/rdf-schema#

The generated N-Triple statements are then transformed to the ontology in RDF format, which may be used in the processes of ontology integration and ontology mapping, in order to enable the interoperability between the extracted ontology and others existing ontologies, as shown in Figure 5.

Journal of Theoretical and Applied Information Technology
31st July 2020. Vol.98. No 14

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

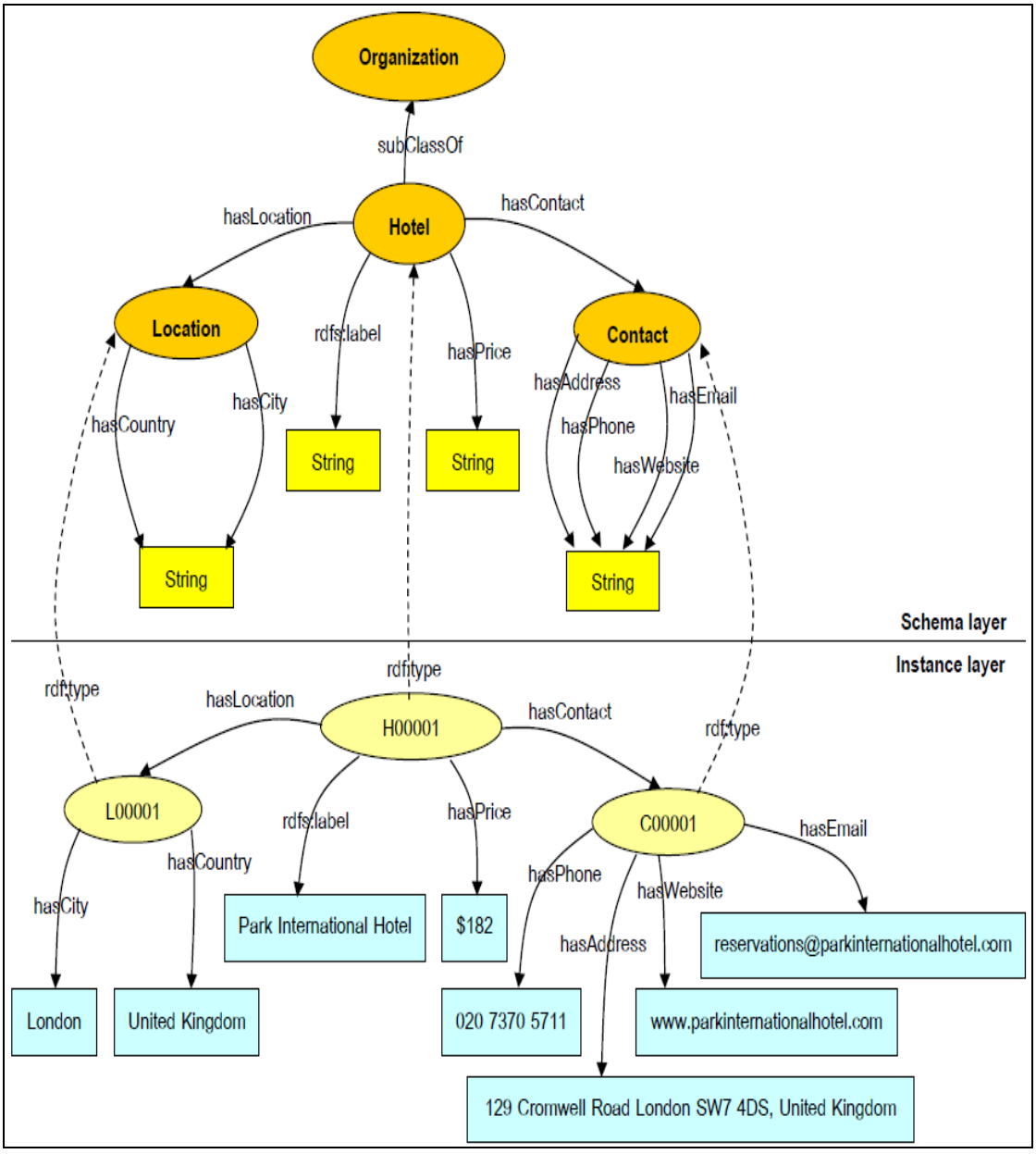

Figure 5: An example of annotated text in RDF graph format.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENTS AND EVALUATION

4.1 Experiments

In our experiments, we collected the dataset of unstructured texts in the Tourism domain from several popular Web pages. The dataset contained 1185 texts, derived from different providers. An example of a text from the dataset is shown in Figure 6.

Title: Park International Hotel

Park International Hotel offers comfortable accommodation and elegant Victorian architecture. Located in London, United Kingdom.
Address: 129 Cromwell Road London SW7 4DS, United Kingdom
Phone: 020 7370 5711
Email: reservations@parkinternationalhotel.com
Website: www.parkinternationalhotel.com
Price: $182 / night

Figure 6: An example of text in the dataset.

Each text was processed in the proposed ontology extraction approach, and the results were compared with reference answers in order to evaluate the efficiency and correctness of the approach. The reference answers in the ontology extraction of the text comprising a set of labeled terms were provided to perform our experiments. An example of reference answers of a text annotation is shown in Figure 7.

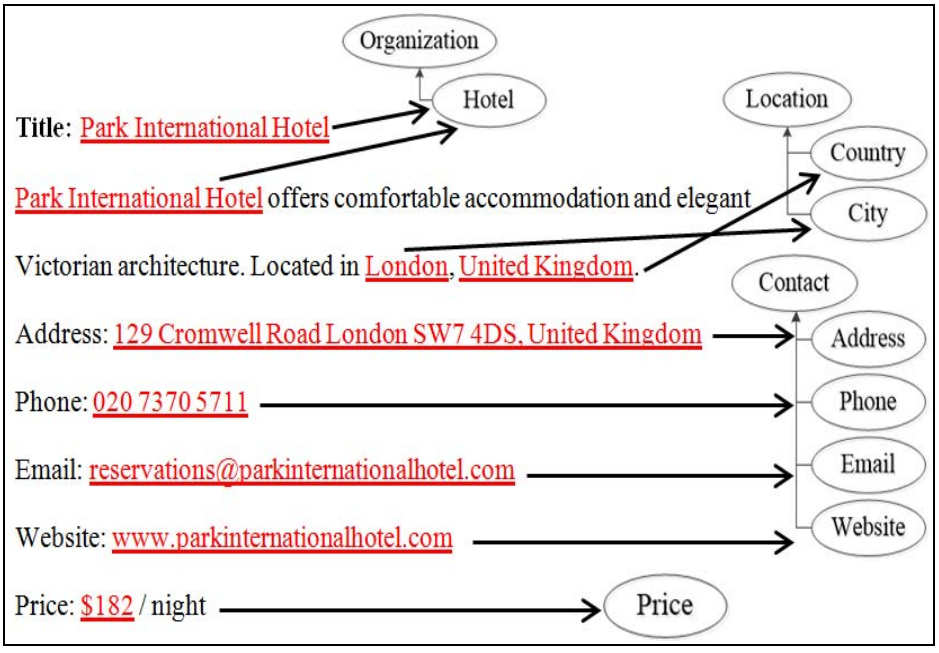

Figure 7: Answers of text annotation within common concepts.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4.2 Evaluation

In our evaluation of the proposed approach, we calculated the percentages of correctness (Precision), completeness (Recall), and average efficiency (F-measure), explained in Equations 1, 2, and 3.

$$Precision = \left(\frac{ce}{ce + te}\right) 100 \qquad (1)$$

While *ce is the number of entities correctly extracted, and te is the number of true entities not extracted.*

$$Recall = \left(\frac{ce}{ce + fe}\right) 100 \qquad (2)$$

While *ce is the number of entities correctly extracted, and fe is the number of entities extracted imprecisely.*

$$F - measure = 2\left(\frac{Precision * Recall}{Precision + Recall}\right) \qquad (3)$$

The evaluation indicated a high accuracy of ontology extraction with the total scores of Precision, Recall, and F-measure, at 91.48%, 89.12%, and 90.23%, respectively (Table 2). Our evaluation consisted of four groups: generated individuals, generated classes, generated datatype properties, and generated object property relations. The Precision, Recall, and F-measure scores of the group of individuals were 94.00%, 93.51%, and 93.75%; the group of classes were 90.24%, 85.91%, and 88.02%; the group of datatype property relations were 92.22%, 89.59%, and 90.89%; and the group of object properties were 89.44%, 87.46%, and 88.24%, respectively.

Table 2. Results of our method within various datasets.

List	Precision	Recall	F-measure
Individuals	94.00%	93.51%	98.75%
Classes	90.24%	85.91%	93.02%
Datatype Properties	92.22%	89.59%	95.89%
Object Properties	89.44%	87.46%	88.24%
Total	**91.48%**	**89.12%**	**90.23%**

4.3 Comparison with Other Approaches

This research study presents the differences between the proposed approach and other methods earlier mentioned in the theoretical background and related studies as follows.

R. Anantharangachar, S. Ramani, and S. Rajagopalan [3] suggest a solution for the information extraction from unstructured texts through Semantic Lexicon and Protégé to create and manage the ontology. In addition, Jena API is employed to create the ontology extraction. However, this study still has limitations on the process of creating and managing the ontology and that of the ontology extraction which is not automatically operated. Instead, it is administered by expert ontology engineers. Thus, it wastes time and costs a lot of money. Moreover, the rules used for ontology extraction lack of flexibility and accuracy, and there remains the problem of word similarities such as synonyms and homonyms.

M. Al-Yahya et al. [22] give a solution for the semantic annotation and information extraction by implementing the semantic annotation of the Arabic textual content and Protégé to create and manage the ontology. However, this study has limitations on its operation. That is, it is semi-automatically operated; it supports only Arabic domains; and it encounters the problem of word similarities. Furthermore, the lexical database has not yet covered Arabic words, and it may not be flexible when being applied to other languages.

Nevertheless, this present research study has different contributions from the above-mentioned research studies as detailed in the followings.

(1) It can extract Individuals from text corpus, and resolve similarities of Individuals.

(2) It can extract Classes from text corpus, and resolve similarities of Classes

(3) It can extract and create relationships of Datatype Properties.

(4) It can extract and create relationships of Object Properties.

(5) The results of measuring effectiveness of the proposed method are at a very high level.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

According to the results of the performance measurement in Table 2, it is evident that the proposed technique comes up with high accuracy and effectiveness. As for resolving the problem of word similarities, Wu and Palmer algorithm and WordNet are employed to find out word similarities.

5. CONCLUSION

This paper presents an ontology extraction approach, which extracts explicit information from unstructured texts. The rule-based extraction technique is a key component in our approach to solve the problem of identifying common concepts. Thus, the system generates the extracted terms into instances and classes automatically for constructing the ontology. To ensure the capability of the proposed approach, we conducted experiments within the Tourism domain; confirmed through Precision, Recall, and F-measure.

In the future, we will continue to develop this ontology extraction technique of Semantic web based ontology extraction and management system in the Tourism domain to further the process of ontology mapping and integration in order to populate more explicit information, and to create a knowledge base to enable specific semantic searches.

This present study on an ontology-based semantic extraction approach from text corpus still has the limitation on word similarities, particularly homonyms. For the further study, this kind of problem will be resolved to encourage the system to be more accurate and efficient.

6. ACKNOWLEDGEMENT

This study was granted by the Office of Higher Education Commission, Ministry of Education in Thailand.

REFRENCES:

[1] J. Ye, S. Dasiopoulou, G. Stevenson, G. Meditskos, E. Kontopoulos, I. Kompatsiaris, et al., "Semantic web technologies in pervasive computing: A survey and research roadmap," *Pervasive and Mobile Computing*, 2015.

[2] D. C. Wimalasuriya and D. Dou, "Ontology-based information extraction: An introduction and a survey of current approaches," *Journal of Information Science*, 2010.

[3] R. Anantharangachar, S. Ramani, and S. Rajagopalan, "Ontology Guided Information Extraction from Unstructured Text," *International Journal of Web & Semantic Technology*, 2013.

[4] A. Zouaq, M. Gagnon, and L. Jean-Louis, "An assessment of open relation extraction systems for the semantic web," *Information Systems*, vol. 71, pp. 228-239, 2017/11/01/ 2017.

[5] N. Sanchez-Pi, L. Martí, and A. C. Bicharra Garcia, "Improving ontology-based text classification: An occupational health and security application," *Journal of Applied Logic*, 2016.

[6] J. Park, W. Cho, and S. Rho, "Evaluating ontology extraction tools using a comprehensive evaluation framework," *Data & Knowledge Engineering*, vol. 69, pp. 1043-1061, 2010.

[7] Y.-B. Kang, P. Delir Haghighi, and F. Burstein, "CFinder: An intelligent key concept finder from text for ontology development," *Expert Systems with Applications*, vol. 41, pp. 4494-4504, 7// 2014.

[8] H. Cunningham, V. Tablan, A. Roberts, and K. Bontcheva, "Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics," *PLoS Comput Biol*, vol. 9, p. e1002854, 2013.

[9] Z. Wang, S. Xu, and L. Zhu, "Semantic relation extraction aware of N-gram features from unstructured biomedical text," *Journal of Biomedical Informatics*, vol. 86, pp. 59-70, 2018/10/01/ 2018.

[10] W3C. (2014). Resource Description Framework (RDF). Available: https://www.w3.org/RDF/

[11] L. Li, C. Zhou, J. He, J. Wang, X. Li, and X. Wu, "Collective semantic behavior extraction in social networks," *Journal of Computational Science*, vol. 28, pp. 236-244, 2018/09/01/ 2018.

[12] S. Bechhofer, F. v. Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, et al. (2004). OWL Web Ontology Language. Available: http://www.w3.org/TR/owl-ref/

[13] Y. Lu, Q. Li, Z. Zhou, and Y. Deng, "Ontology-based knowledge modeling for automated construction safety checking," *Safety Science*, vol. 79, pp. 11-18, 11// 2015.

[14] C. Ru, J. Tang, S. Li, S. Xie, and T. Wang, "Using semantic similarity to reduce wrong labels in distant supervision for relation extraction," *Information Processing & Management*, vol. 54, pp. 593-608, 2018/07/01/ 2018.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] J. J. Jung, "Semantic wiki-based knowledge management system by interleaving ontology mapping tool," *International Journal of Software Engineering and Knowledge Engineering*, vol. 23, pp. 51-63, 2013.

[16] F. Abedini, F. Mahmoudi, and A. H. Jadidinejad, "From Text to Knowledge: Semantic Entity Extraction using YAGO Ontology," *International Journal of Machine Learning and Computing*, 2011.

[17] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: a spatially and temporally enhanced knowledge base from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 28-61, 2013.

[18] M. Vanni and A. Neiderer, "General Architecture for Text Engineering (GATE) Developer for Entity Extraction: Overview for SYNCOIN," *Army Research Laboratory*, 2014.

[19] R. Witte, N. Khamis, and J. Rilling, "Flexible Ontology Population from Text: The OwlExporter," *in LREC*, 2010, pp. 3845-3850.

[20] W3C. (2014). N-Triples. Available: http://www.w3.org/TR/n-triples/

[21] S. Duari and V. Bhatnagar, "sCAKE: Semantic Connectivity Aware Keyword Extraction," *Information Sciences*, vol. 477, pp. 100-117, 2019/03/01/ 2019.

[22] M. Al-Yahya, M. Al-Shaman, N. Al-Otaiby, W. Al-Sultan, A. Al-Zahrani, and M. Al-Dalbahie, "Ontology-based semantic annotation of Arabic language text," *International Journal of Modern Education & Computer Science*, vol. 7, 2015.