Know-Ont Based Ontology Modeling Approach for Skill Knowledge Extraction

Khananat Jaroenchai and Churee Techawut

***Abstract*—Ontology is an important tool for organizing information into categorized data for a semantic web search engine in order to create an ontology which can support both the collection and the presentation of skill knowledge. This study used a Know-Ont based ontology modeling approach (KOOM) on a case study of a basketball shooting technique to create semi-automatic ontology engineering, which is expected to help facilitate better understanding and interpretation of the multidimensional table in a knowledge engineering process and reduce the need for expert support. The efficiency evaluation of skill knowledge extraction framework based on KOOM shows the following values: Precision =0.72, Recall =0.71, Accuracy = 0.94 and F-Measure =071., which are considered to be efficient.**

***Index Terms*—Ontology, ontology modeling, semantic web, knowledge engineering.**

I. Introduction

Ontology is an important tool for organizing information into categorized data for a semantic web search engine and knowledge distribution via the internet network. Skill is a very important knowledge which can be obtained from trial and error, and from the accumulated experience of an expert which then becomes knowledge and can be further applied in other related fields. At present, skill knowledge of agriculture [1] manufacturing industries [2] and sports science [3], [4] are all crucial. Most of the organizations want to collect and store knowledge to perform knowledge management in order to effectively explore and transfer practical skill knowledge suitable for each organization. Currently, there are 2 techniques for skill knowledge presentation; (1) figure and video, and (2) multidimensional table with a wide variety of statistical values. The multidimensional table consists of collected knowledge which can be analyzed and further studied to improve skill knowledge in other domains; therefore the multidimensional table is regarded as a reliable source of information and has been widely used in the ontology field.

The traditional way to create categorized knowledge from the multidimensional table is to directly extract words out of the table; then those extracted words are combined with text extracted from the whole documents. Knowledge engineering process is utilized to analyze the meaning of a group of words, assign a relationship between the word groups, and finally create knowledge ontology. However, there are some issues regarding this traditional method; (1) It is difficult to interpret relationship among multidimensional data because the knowledge engineer must have a deep understanding on knowledge of that particular domain in order to correctly assign meaning or relationship of values in the table, (2) Understanding and creating categorized ontology is a time consuming process, and (3) There is a chance that many words or values in the multidimensional table may be overlooked or excluded from the created ontology which will later prevent an effective semantic search.

There are not many references in ontology research field. Most of the references used traditional ontology modeling method to create the ontology. Know-Ont (Knowledge ontology) [5] is one of the examples which the knowledge engineer has to thoroughly study an organization's structure and manufacturing line with the help from specialists in many departments in order to create the ontology. There is also a semi-automatic method [6], [7] to create an ontology which can reduce the burden on the knowledge engineer by creating term extraction from text files in a form of unstructured text, and later organizing those terms into super class & subclass hierarchy.

This study used a Know-Ont based ontology modeling (KOOM) approach which (1) transforms a multidimensional table based on the relationship among the data in order to reduce the burden of trying to understand the dimensional data, (2) semi-automatically creates schema of knowledge ontology based on the structure of Know-Ont which support skill knowledge presentation by using natural language processing, ontology modeling, and machine learning, and (3) distinctly substitute a group of words in a multidimensional table with a concept or an instance on ontology in order to produce an effective semantic search; specifically in this process, the knowledge engineer can adjust the resulting ontology as needed.

This research report is organized as follows: Section II describes natural language processing, ontology modeling, and machine learning used in KOOM, Section III describes the details and components of KOOM by using a case study of a multidimensional table of knowledge regarding how to perform a jump shot in basketball, Section IV describes the experimental process of performance evaluation on KOOM's term mapping component, Section V is the discussion of the evaluation result which consists of precision, recall, accuracy, and F-Measure, and lastly, Section VI is conclusions and suggestions

II. Literature Review

KOOM was developed based on natural language processing, ontology modeling, and machine learning

Manuscript received October, 2018; revised November 12, 2018.

The authors are with the Department of Computer Science, Chiang Mai University, Chiang Mai, 50200 Thailand (e-mail: khananat09@gmail .com).

doi: 10.18178/ijke.2018.4.2.105

techniques as follows:

A. Natural Language Processing

Optical character recognition (OCR) is the automatic conversion of text image files into machine-encoded text. OCR is necessary when the imported data are in an image file format which the term cannot be extracted directly unless it is processed with OCR.

Semantic similarity measurement is a metric defined over a set of terms, where the idea of the distance between them is based on the likeness of their meaning. In this study, this process was used to compare and match similarities between (1) knowledge terms extracted from a multidimensional table, and (2) terms which are concepts of Know-Ont. Wu and Palmer semantic similarity [9], which is computed out as semantic similarity index by measuring the depth of concept between 2 terms which were categorized in Wordnet, was used in this study.

B. Ontology Modeling

UPON (Unified Process for Ontology building) [10] is an incremental methodology for ontology building. The process is based on the Software Development Unified Process used in the software engineering field. There are 5 workflows in the process: (1) Requirements (2) Analysis (3) Design (4) Implementation (5) Testing.

LEXON [6] is a basic building block of ontology representing a relationship between terms, mainly in a 5-tuple form but may also be presented in other forms depending on a usage condition. LEXON is a main component of DOGMA (Developing Ontology-Grounded Methods and Applications). There are 2 types of LEXON as follows:

LEXON used to present the relationship between terms in a form of <Subject, predicate, object> consists of <γ, Term1, Role, Role′, Term2>. γ (Gamma) is a context of Term, Term1, and Term2; they are linguistic terms which are important in creating ontology. Role and Role′ are lexicalizations of the pair roles of a binary conceptual relationship; the Role′ is the inverse of the Role, for example, if Role is "has part" then Role′ is "is part of".

LEXON upper form is used to describe a basic relationship of an object in each of the term, such as class, properties or instant. LEXON upper form consists of <γ, Term, ObjectType>. γ (Gamma) is a context of the term. The term is linguistic terms which are important in creating ontology. ObjectType is a property of term according to W3C standard which consists of Class, ObjectProperty, DataProperty or Instance.

Know-Ont [5] is knowledge ontology introduced in an industrial domain to improve the specialist development system. It stemmed from a case study of the manufacturing system of a particular industry organization by collecting information and creating a glossary of technical terms which are important in decision making in a specialist system. Know-Ont is a knowledge engineering process similar to skill knowledge which also requires some form of knowledge management. This is a reason that this study applied the concept of Know-Ont to the term extraction process in order to further develop categorized knowledge in other domains.

C. Machine learning

Reinforcement learning (RL) [11] to a kind of machine learning concerned with how an agent behaves in a variety of environment so as to maximize some notion of a cumulative reward under a policy which governs the result of the learning. Because identifying synset for terms will result in a significant improvement on the accuracy of categorization, the application of RL in this study should help identify the synset of extracted terms which belong to the same categories, series or data Section and correctly match the extracted terms with Know-Ont's term.

III. METHODOLOGY

The overview process of KOOM is shown in Fig. 1 The process is divided into 3 parts: pre-processing, extraction and conversion.

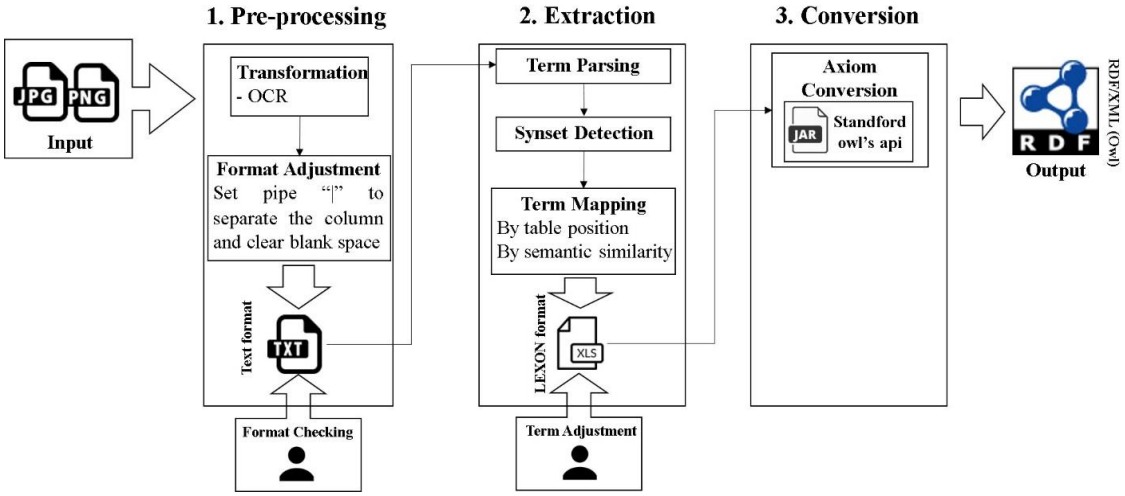

Fig. 1. Overview of the process.

Input is the multidimensional tables which contain knowledge. Most of the information used in this study was in the form of text files or picture files selected from published research articles in the domain of knowledge extraction of a jump shot technique in basketball.

Output is the Owl file which is ready to use and can be opened in the Protégé software for management and evaluation

A. Pre-processing

Pre-processing is a step where dimensional tables are prepared. If the information is already in the form of a text file, there is no need for transformation. On the other hand, if the information is in form of an image file, then it needs to be transformed into a text file before it can be used in the extraction step. Input information used in this study was a multidimensional table of a variation in ball throwing positions for a jump shot in basketball. There are 2 steps in pre-processing as follows:

1)Transformation

Transformation is the process of converting data from multidimensional tables in a picture file format into a structure text format using OCR. Terms in a multidimensional table (Fig. 2(b)) were identified and classified into a respective series, category, or data section (Fig. 2(a)).

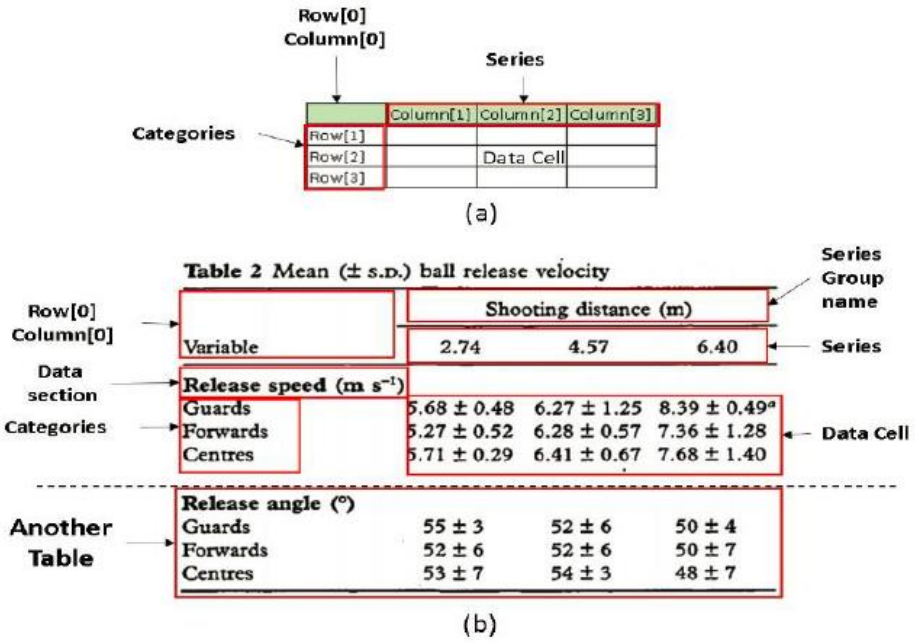

Fig. 2. (a) Components of multidimensional table, (b) Structure of example tables [2].

2)Format adjustment

Format adjustment is the process of arranging the composition of text data to make them suitable for term extraction and can be traced for their location in the multidimensional table. The result is in a text file format.

B. Extraction

Extraction is the process of retrieving terms from a text file to analyze and create an ontology in the form of LEXON. By using KOOM, more than 1 text file can be extracted from a single iteration. There are 3 main steps in the extraction process as follows:

1)Term parsing

Term parsing is the process of analyzing text file to categorize terms, identify their locations in the multidimensional tables, and describe other information such as row and column numbers.

2)Synset detection

Synset detection is the process of identifying a suitable synset for knowledge domains. For example, guard, forward and center are all in the same "player position" category in the table. These terms can be used to identify a synset and create a glossary. Synset detection in this study incorporated RL technique by assigning state a similarity between each pair of terms in the same synset as a rewarding score, and use the total reward score of each synset to select a suitable synset for each domain as shown in Fig. 3.

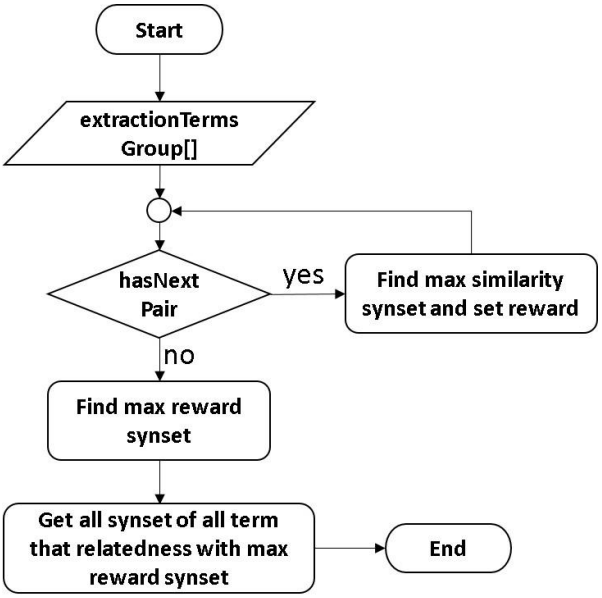

Fig. 2. Mapping synset by series and categories.

3)Term mapping

Term mapping consists of semantic mapping and table position mapping with details as follows:

- **Semantic mapping** is the process starting with creating the relatedness policy among terms in the table. Then the created policy is applied to the list of terms using the highest

similarity between extracted terms and Know-Ont terms to calculate the relatedness from Wu and Palmer's algorithm. If the comparing terms have a suitable synset derived from the synset detection step in Fig. 4, the suitable synset can then be used to calculate a similarity value. But if there is no suitable synset, the max similarity value will be derived by comparing all synsets from the extraction terms instead, as shown in Fig. 4.

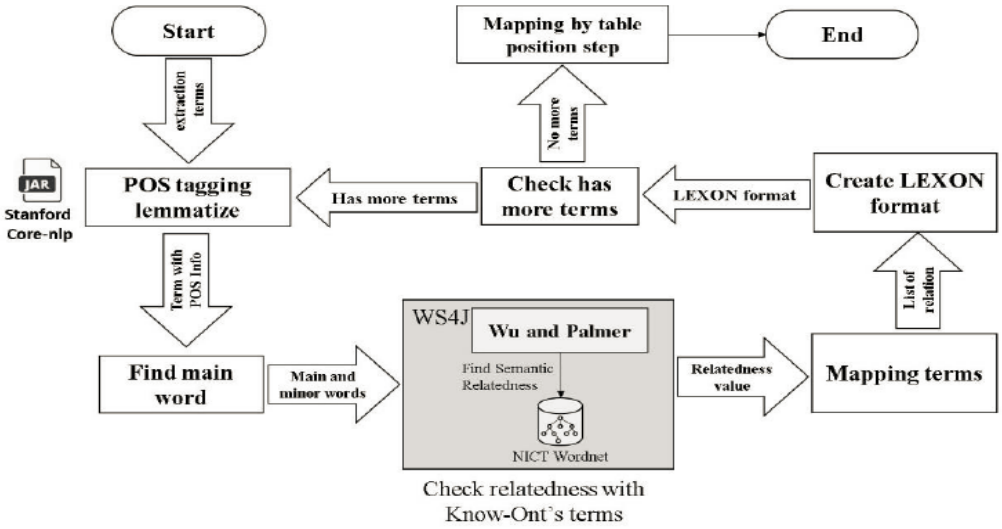

Fig. 3. Mapping term with know-ont term process.

- **Table position mapping** is the process of creating LEXON [6] from the relatedness value between terms derived from their positions in the multidimensional table, as shown in Fig. 5.

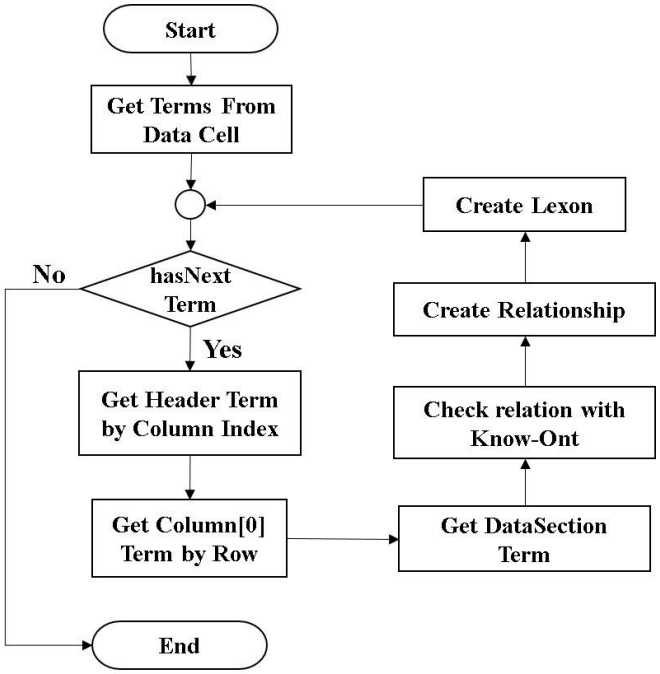

Fig. 4. Relation mapping by position in table process.

4)Term adjustment

Term adjustment is the process of verifying or editing a list of term relationship in LEXON. This process can also be used to identify the ObjectType property of each term in order to create an OWL file in the next step.

C. Axiom Conversion

Axiom conversion is the process of converting a LEXON format into an OWL format which can be used in a semantic web search. All terms in the LEXON format were used to create axiom in forms of subject, predicate, and object. Then Java Class Library was used to convert the axiom into an ontology structure and create an OWL file which can be opened in the Protégé software, as shown in Fig. 6.

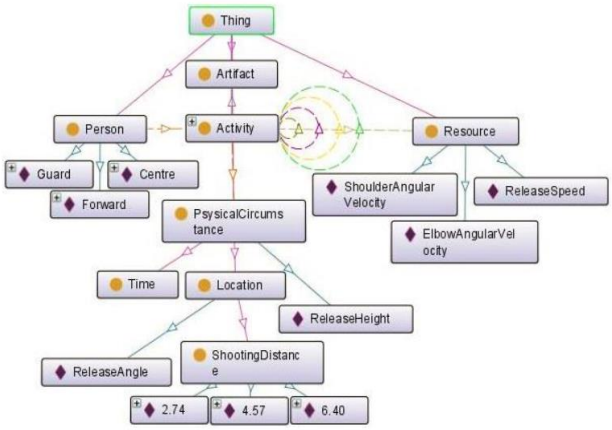

Fig. 5. Overview of resulting ontology.

IV. EVALUATION

A. Objectives

To calculate accuracy, precision, recall, and F-measure (effectiveness of model mapping) values in the term mapping process, both with and without a specified synset.

B. Variable

1)Dependent variable

Dependent Variable is calculated values which can be used to evaluate the effectiveness of the term mapping. The dependent variables range from 0-1, with value approaches 1

means good to very good efficiency.

-Accuracy is the accuracy of the term mapping

-Precision is the precision of the term mapping

-Recall the effectiveness of mapping the interested terms

-F-Measure is calculated from precision and recall. This value represents the overall efficiency of the term mapping.

2)Independent variable

Independent Variable is observations made from the term mapping results. In this study, they are the number of correctly or incorrectly mapped terms. There are 4 variables as follows:

True Positive (TP) a number of observation where the predicted $Term_p$ resulting in $Term_p$

False Positive (FP) a number of observation where the predicted $Term_p$ resulting in $Term_n$

False Negative (FN) a number of observation where the predicted $Term_n$ resulting in $Term_p$

True Negative (TN) a number of observation where the predicted $Term_n$ resulting in $Term_n$

$Term_p$ stands for Positive Term which is a major term for the calculation. $Term_n$ stands for Negative Term which is a minor term for the calculation. p and n values are a location of column and row, respectively, in the confusion matrix which contains all four independent variables.

C. Hypothesis

The term mapping has accuracy, precision, recall and F-measure values in a "good" range (> 0.7); this also means that KOOM process is efficient.

D. Evaluation Methods

Step 1 Prepare the term lists obtained from the extraction process.

Step 2 Prepare the term mapping answer keys.

Step 3 The extracted term list to the answer keys to validate the mapping process.

Step 4 Put the comparing results into confusion matrix and calculate the precision, recall, accuracy, and F-Measure values.

Repeat the step 4 for every predicted term (there are 8 mapping answer keys in this study) and calculate the average values of the precision, recall, accuracy, and F-Measure.

V. Result Analysis Discussions

The term extraction and categorization processes with specified synsets yielded the following values: precision = 0.72, recall =0.71, accuracy =0.94, and F-Measure =0.71. Since all values were greater than 0.7 it indicated that the KOOM model was accurate and efficient.

Additionally, when the terms were extracted and categorized without the specified synsets, the evaluation results were: precision =0.51, recall =0.58, accuracy =0.78, and F-Measure =0.54. Although the accuracy was higher than 0.7, the other three values were all lower than 0.7. The overall lower values are an indication of the low efficacy of KOOM model.

The evaluation results from the two different methods suggested that the semantic similarity calculation with the specified synsets for every term generated better pairing results than the calculation without the specified synsets. The unspecified synsets for terms prior to the semantic similarity calculation can lead to errors in the term mapping process because the selected synsets cannot match with the correct knowledge domain. Therefore, it can be comprehended that specifying synsets for all the terms prior to the semantic similarity calculation can greatly improve the efficacy of term mapping. The comparison of efficacy calculation between the two methods was shown in Fig. 7.

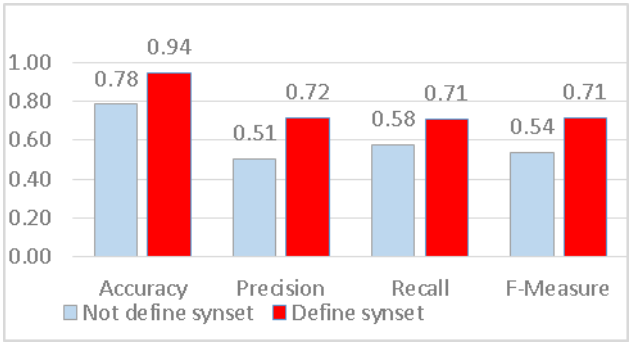

Fig. 6. Comparison of efficacy calculation of accuracy, precision, recall, and F-Measure between not define synset and define synset.

VI. Conclusions and Suggestions

The accuracy, precision, recall and F-measure values in this study all indicated a good efficacy of KOOM model with the accuracy at a very high level, and the precision, recall and F-measure at a high level. Therefore it can be concluded that the semi-automatic KOOM method used to create an ontology in this study is efficient and can be used to deliver an ontology which is suitable to the knowledge domain. This knowledge engineering approach can reduce a burden in the term categorization process, especially the work load of understanding the multidimensional tables. Also, this approach can semi-automatically and swiftly create a schema of knowledge ontology from multidimensional data based on the Know-Ont structure the structural ontology which support the presentation of a skill knowledge using the natural language processing, ontology modeling, and machine learning techniques resulting in a distinct ontology structure suitable for a valid semantic searching.

The advantage of KOOM is that this method can extract terms which are related to a particular domain without the need for the statistical analysis, unlike the traditional term extraction method which uses an unstructured text. Moreover, this approach can analyze many multidimensional tables in a single iteration. However, this does not mean that KOOM has no disadvantage. If there are too few multidimensional tables in each data source document, there might be a problem of the low number of the important terms related to a particular domain because they can only be sourced from the terms which appear in the multidimensional tables. As a result, important terms which do not appear in the tables will not be included in the ontology. Regardless, the knowledge engineer can still edit and add those terms into the ontology afterward.

Future studies in this topic should focus on (1) Creation of transforming and extraction processes for the semantic and relationship information from figures and videos in order to increase the number of available sources of skill knowledge data, (2) Improve the "clean and clear" process for the raw knowledge information from a table to obtain the data which

are suitable for the term categorization and extraction process in order to reduce the burden of data preparation for a knowledge engineer so the framework can be further developed. (3) Construct some further experiments to verify the efficiency of KOOM.

This study has integrated multi-disciplinary knowledge in order to solve the problem in skill knowledge extraction from the multidimensional table. The resulting framework and ontology can be applied to several other fields including (1) ontology engineering the framework can be used to import knowledge to the already existing ontology, (2) knowledge management the resulting ontology from this study can be used to improve the semantic web searching, and (3) natural language processing this study provide an approach for finding a suitable synset, as shown in Fig. 8.

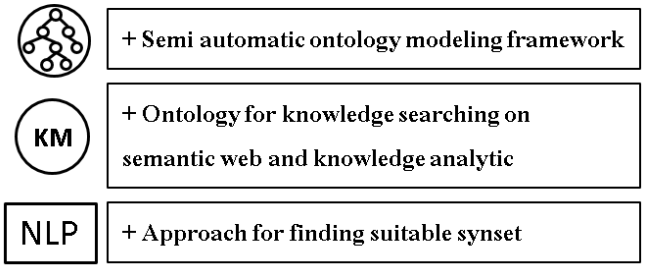

Fig. 7. Utilization for knowledge extraction.

REFERENCES

[1] C. Techawut, R. Sukhahuta, P. Manochai, J. Visithpanich, and Y. Khaosumain, *Easy Knowledge Engineering and Usability Evaluation of Longan Knowledge-Based System*, 2016.

[2] K. Chottikampon, S. Tang, S. Mathurosemontri, P. Sirisuwan, M. Inoda, H. Nishimoto, and H. Hamada, "Comparison knitting skills between experts and non-experts by measurement of the arm movement," *Digital Human Modeling. Applications in Health, Safety, Ergonomics and Risk Management: Human Modeling*, Springer International Publishing, pp. 3-13, 2015.

[3] S. Miller and R. Bartlett, "The relationship between basketball shooting kinematics, distance and playing position," *Journal of Sports Sciences*, vol. 14, no. 3, pp. 243-253, 1996.

[4] F. J. Rojas, M. Cepero, A. Oña, and M. Gutierrez, "Kinematic adjustments in the basketball jump shot against an opponent," *Ergonomics*, vol. 43, no. 10, pp. 1651-1660, 2000.

[5] H. Kumar and P. S. Park, "Know-ont: A knowledge ontology for an enterprise in an industrial domain," *International Journal Database Theory and Application*, vol. 3, no. 1, pp. 23-32, 2010.

[6] P. De Leenheer and R. Meersman, "Towards a formal foundation of DOGMA ontology: Part I," *Technical Report STAR-2005-06, VUB STARLab, Brussel*, 2005.

[7] M. Balakrishna and M. Srikanth, "Automatic ontology creation from text for national intelligence priorities framework (NIPF)," in *Proc. of 3rd International Ontology for the Intelligence Community (OIC) Conference*, 2018, pp. 8-12.

[8] A. D. P. Novelli and J. M. P. de Oliveira, "Simple method for ontology automatic extraction from documents," *Editorial Preface*, vol. 3, no. 12, 2012.

[9] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts. In Proceedings of the ACL workshop on empirical modeling of semantic equivalence and entailment," *Association for Computational Linguistics*, pp. 13-18, 2005.

[10] A. De Nicola, M. Missikoff, and R. Navigli, "A software engineering approach to ontology building," *Information Systems*, vol. 34, no. 2, pp. 258-275, 2009.

[11] L. Busoniu, R. Babuska, B. De Schutter, and D. Ernst, *Reinforcement Learning and Dynamic Programming Using Function Approximators*, CRC Press, 2010.

Khananat Jaroenchai is a master degree student at Department of Computer Science, Faculty of Science, Chiang Mai University. He graduated in B.S in information technology from Department of Information Technology, Faculty of Science, Meajo University, Thailand. His research interest are ontology and knowledge engineering.

Churee Techawut is an assistant professor at the Department of Computer Science, Faculty of Science, Chiang Mai University. She completed her Ph.D in computer science from the Asian Institute of Technology (AIT), Thailand. Her research interests are human computer interaction (HCI) and ontology engineering.