A Semi-Automatic Approach to Ontology Construction for Vietnamese High School Physics Subject

Binh Diep-Phuoc
Nguyen Thi Minh Khai Gifted High School
Soc Trang, Vietnam

An C. Tran
Can Tho University
Can Tho, Vietnam

ABSTRACT

Ontology is a knowledge representation formalism used in the Semantic Web to provide data understandable by both humans and computers. The success of the Semantic Web depends strongly on the development of ontologies so that it is considered the heart of the Semantic Web. With the evolution of the Semantic Web recently, ontology is becoming more and more important in the field of knowledge management and sharing. There is an actual demand for fast and easy ontology engineering to save time and effort in ontology construction to avoid the knowledge acquisition bottleneck. Therefore, this paper proposes a semi-automatic approach to ontology construction for Vietnamese high school physics subject including two steps. The first step is to manually build a "seeding" ontology based on the textbook glossary. Then, a pattern-based method is used to enrich the base ontology to save time and efforts. The evaluation result shows that the pattern-based method is suitable for enriching the ontology and provides a good trade-off between simplicity and enrichment result.

General Terms

Ontology construction, Ontology learning, Semantic web, .

Keywords

Ontology construction, semi-automatically, high school physics, pattern based

1. INTRODUCTION

Tim Berners-Lee, the inventor of the World Wide Web, defines the Semantic Web as "The Web of data with meaning in the sense that a computer program can learn enough about what the data means in order to process it" [1] . The Semantic Web aims primarily to provide a generic infrastructure for machine-processable Web content and is directly relevant to hypermedia research. It is a set of technologies that provide for the existence of knowledge on the Web in a format that can be understood and reasoned about by software applications [2].

Ontology is considered the heart of the Semantic Web and is crucial to its success and proliferation. It is a knowledge representation formalism used in the Semantic Web to provide data that can be understood by both humans and machines [3]. It models concepts and their relationships in a specific domain so that the domain knowledge can be shared and reused. In addition, ontology also brings inference capability to the Semantic Web based on the description logic. Currently, ontology is not only used in Semantic Web but also in many other fields such as artificial intelligence, information retrieval, etc. [4].

Since ontology serves as the knowledge base for other applications and its powerful capabilities in sharing, reasoning, etc., the demand for constructing ontologies in various application domains is very high. However, ontology development is a challenging, time-consuming, and labor-intensive task and requires the knowledge of domain experts. Therefore, many researchers have proposed new strategies for fast and efficient ontology development to solve this problem. One of the potential research directions attracting many researchers is ontology learning, which aims to build an ontology semi-automatically from data sources such as unstructured and structured texts.

This study proposes a method to semi-automatically construct a high school physics ontology, particularly the 12th grade Vietnamese physics, based on the pattern recognition technique. Firstly, a basic ontology (named BaseOntology) is constructed manually based on the grade twelfth physics textbook. That ontology includes some basic physics concepts extracted from the textbook, usually the glossary. Then,the pattern-based method was used to extract further concepts and instances to enrich the base ontology.

The paper is organized as follows. Section 2 presents the related work in ontology engineering, specifically ontology learning. Section 3 introduces the technical background related to the ontology and ontology development method. Section 4 describes our proposed methodology and provides a performance evaluation of the technique on the Vietnamese high school physics subject. Finally, section 5 draws a conclusion and future work.

2. RELATED WORKS

Ontology learning is the process of identifying terms, concepts, taxonomic relations, non-taxonomic relations, and axioms, each of which can be considered as a separate output. Based on the Semantic Web stack described in Figure 1 [5], Buitelaar P. et al. proposed a model for ontology learning referred to as "Ontology Learning Layer Cake" [2].

In the figure, all ontology learning tasks are organized in a layered cake of increasing complexity. At the lowest level, we have the terms, the basic building blocks in ontology learning. A term can be a single word or multiple words, which isthe linguistic realization of anything essential and relevant in a specific domain. For example, in the medical domain, terms can be disease, illness, and hospital. Synonyms are groups of words that are semantic variants of a term (e.g., illness and disease). Concepts consist of a labeling term, its synonyms, and sometimes also include instances of the concepts. The next level is the discovery of the relations between concepts. There are two types of relations, taxonomic relations and non-taxonomic relations. Taxonomic relations are relations (e.g., the doctor is a person) and are used to construct hierarchies.

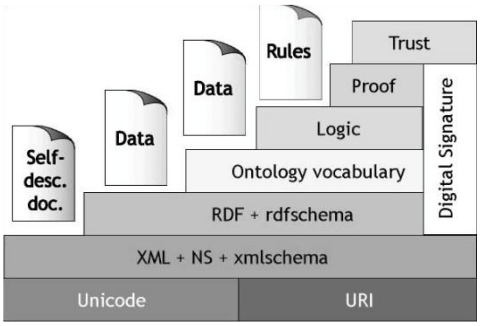


Fig. 1: A layered approach to the Semantic Web [5]

On the other hand, non-taxonomic relations such as a doctor curinga disease are less explicit and are often harder to identify and label. At the highest level in the layer cake are logical rules (or axioms) that can be defined over concepts and relations. They can be used to define constraints, verify the correctness of existing ontologies, and deduce new ontological elements.

Most systems use text copra of the target domain and utilized various techniques to construct the ontology. Techniques for ontology learning come from more established fields such as natural language processing, machine learning, information retrieval, and knowledge representation and reasoning. They can be divided into three main categories: statistics-based, linguistics-based, and logic-based and most systems use a combination of techniques from different categories. In Figure 2, Buitelaar et al. [2]show a nice overview of common techniques used in ontology learning.

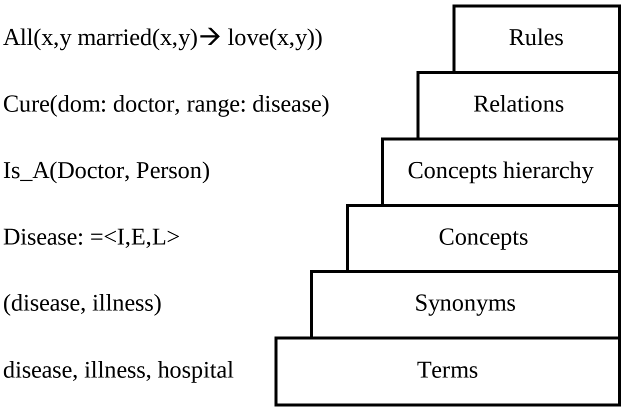


Figure 2. Ontology Learning Layer Cake [2]

Statistical-based techniques mainly come from information retrieval, machine learning, and data mining. These techniques are often based on the idea that the (co)-occurrence of lexical units usually means that they are related(e.g. "Bill Gates" and "Microsoft" tend to appear together). Clustering is one of the most commonly used methods where terms are assigned into groups based on a similarity measure [6]. In addition, a hierarchical relationship between two terms can also be discovered by using conditional probabilities of the occurrences of terms in documents as a measure [7]. A drawback of the statistics-based approach is that they tend to ignore the underlying semantics and other linguistics features. Thus, most system also incorporate linguistics-based techniques.

Linguistics-based techniques are probably the most widely-used techniques applicable to almost all ontology learning tasks (except for axiom discovery). They benefit from many state-of-the-art natural language processing tools already available. For instance, TreeTagger [8] and Link Grammar Parser [9] are a popular part-of-speech tagger and sentence parser. Another example is the Natural Language Toolkit [10],a comprehensive toolkit for natural language processing tasks. Syntactic and dependency analysis can help identify possible terms and relations. For example, nouns preceded by an article or an adjective can imply candidates for terms extraction, while verbs between two terms can be considered relation candidates. Lexical databases such as WordNet [11] are also beneficial for finding predefined concepts and relations including synonyms, hyponyms, and meronyms [12].

Methods for automatically building ontologies based on machine learning have also been proposed recently. For example, Tran A. C. et al. [13] proposed a method for learning axioms based on description logic learning techniques. In particular, this study aims to generate the hypothesis from the observed facts using the induction technique. Etzioni, O. et al. [14]and Zhang et al. [15] used the Conditional Random Fields (CRFs) to extract attributes and information. In the Fleischhacker and Völker’s study [16], the authors used inductive methods of enriching engineered ontologies with generating disjointness axioms. Recently, deep learning has also been applied in this area. Chen et al. [17] proposed an information extraction model to find and pair one of five types of relationships (Role, Part, At, Near, and Social) between each two named entities (Chinese entities) based on DBN. Wang et al. [18] used the CNN model to classify the text and then used the classified text and TF-IDF matrix to build the ontology for the shipping industry domain.

3. TECHNICAL BACKGROUND

3.1 Semantic web

The Semantic Web aims primarily at providing a generic infrastructure for machine-processable Web content and has direct relevance to hypermedia research. It is a set of technologies that envisions the existence of knowledge on the Web in a format that software applications can understand and reason about [5].Figure 1 shows the Semantic Web stack suggested by Tim Berners-Lee which describes the main layers of the Semantic Web design and vision.

At the bottom is the URI (Uniform Resource Identifier) which is a syntax to define the identifier for Semantic Web resources. The second layer is XML (eXtendable Markup Language), an independent and standardized file exchange format, particularly for sending the document across the Web. XML is used in Semantic Web for describing Semantic Web documents’ structure. The next layer is RDF, a triple data model for describing Semantic Web resource relationship. In this model, knowledge is represented in triples, called statements.

Ontology layer provides a more powerful modeling language for knowledge representation in Semantic Web. This is considered the highest-level of knowledge representation in Semantic Web. Ontology expands RDF Schema and allows more complex relationships and constrains and brings reasoning power to the Semantic Web. The Logic layer consists of rules that enable inferences while the Proof layer allows the explanation of given answers generated by the reasoners. Finally, the Trust layer emerges through the use of digital signatures to ensure the source is trustworthy.

The Semantic Web brings the following benefits [19]

- Information can be organized and searched based on its meaning rather than on the text.

- The way information is presented can be improved, e.g. information can be clustered by meaning or merged by its relevance, etc.
- The integration of information from heterogeneous source in different organizations through the use of semantic metadata.
- Semantic descriptions can be used to easily locate resources.

3.2 Ontology

Ontology is considered to be the heart of all Semantic Web applications. It is defined as a "formal, explicit specification of a shared conceptualization" [3]. This definition stresses two key points: 1) the conceptualization is formal and allows inference using computer, and 2) a practical ontology is designed for a specific domain of interest. An ontology can be visualized as a directed graph in which the nodes represent concepts, and the edges represent the relationships between concepts. The ontology is the most formal representation of knowledge. In other words, it is natural language independent and does not contain any lexical knowledge. Figure 3 shows the spectrum of knowledge formalization [20].

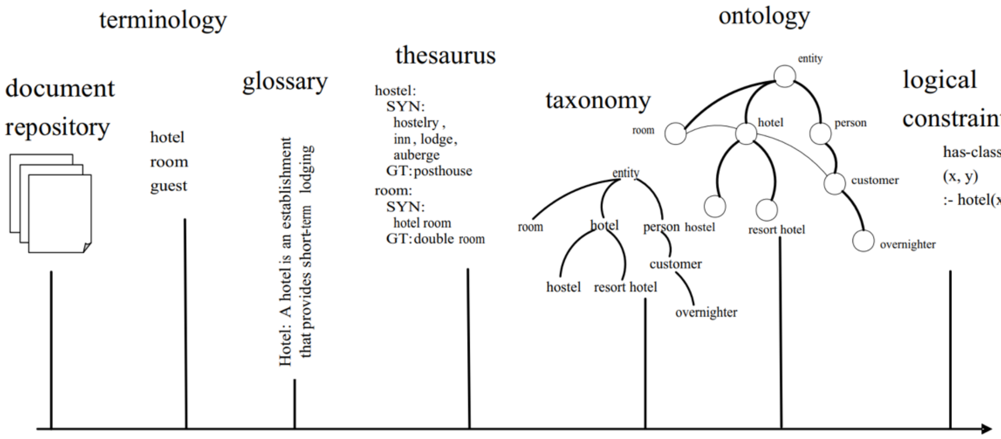


Fig. 3: Knowledge formalization spectrum [20]

At one end of the spectrum, we have text documents that have no structural requirements. At the other end of the spectrum, we have heavy-weight ontologies that extensively use logical rules (axioms) for specification. Terminology, glossary, and thesaurus can be called controlled vocabularies. A glossary consists of a list of terms that are enumerated, maintained, and regulated by independent authorities often for a specific domain. A glossary provides the definition for each term, while a thesaurus gives related terms as synonyms, antonyms, and so on. A taxonomy is a controlled vocabulary organized in a hierarchical (parent-child) structure. An ontology extends a taxonomy with non-taxonomic relations and possible logical constraints.

An ontology consists of concepts (also known as classes), relations (properties), instances and axioms. A more concise definition of an ontology is that of a 4-tuple <C, R, I, A>, where C is a set of concepts, R is a set of relations, I is a set of instances and A is a set of axioms [21]. A class is an abstract group, set or collection of objects that share common characteristics. Instances are the basic, fundamental components of the ontologies. A property is a particular aspect, characteristic, attribute, or relation used to describe a resource. There are two types of properties, namely object properties that relate instances to instances and datatype properties relate instances to datatype values. Finally, axioms provide information about classes and properties such as the equivalence of two classes, the range of properties, etc.

4. PATTERN-BASED HIGH SCHOOL PHYSICS ONTOLOGY BUILDING

This section describes our semi-automatic approach to constructing high school physics ontology, particularly for the twelfth grade. The construction process consists of two basic steps: 1) construct a basic ontology (BaseOntology) manually based on the physics textbook, particularly the textbook glossary, 2) enrich the basic ontology using the pattern-based method. Then the effectiveness of the proposed method is evaluated by checking the output ontology.

4.1 The proposed method

The proposed method for semi-automatic ontology building has two stages. In the first stage, a seeding ontology that contains basic concepts and relationships is constructed manually from a physics textbooks, as described in Figure 4.

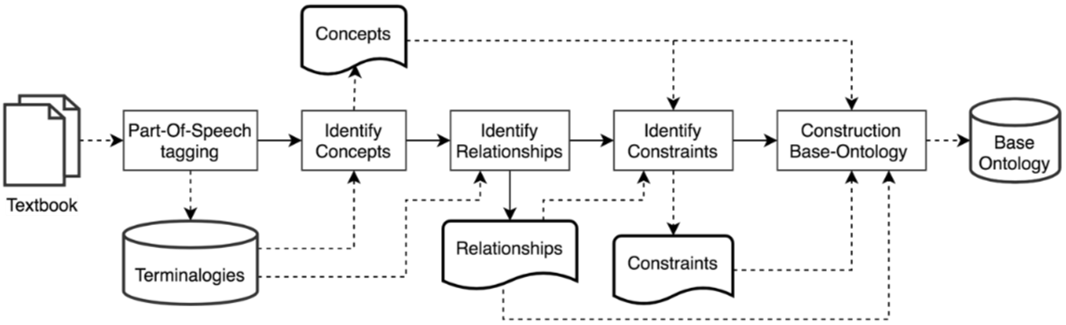


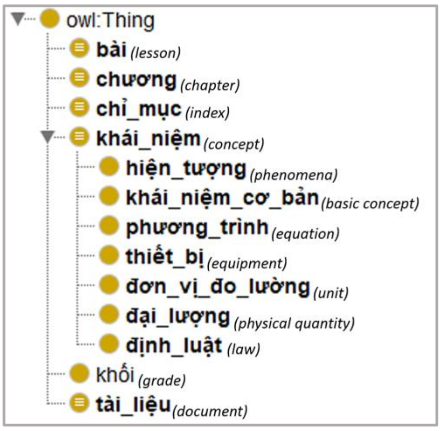
Fig. 4: Building the basic ontology (BaseOntology) manually from physics textbook

First, sentences in the digitalized physics textbook are POS tagged. This study uses VietSeg, a Vietnamese POS tagger provided at https://github.com/manhtai/vietseg. Tagged nouns and nouns phrases are extracted into a terminology database. There are totally of 1.528 terminologies extracted from this step.

Table 1. Several terms extracted from the textbook

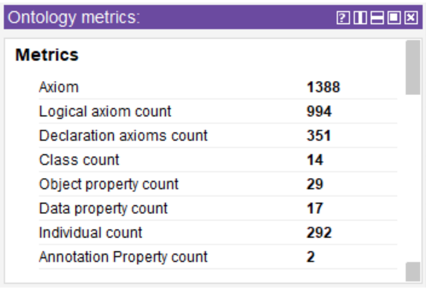
No	Terms
1	âm_sắc(timbre)
2	bảo_toàn_momen_động_lượng (conservation of angular momentum)
3	bức_xạ_hồng_ngoại(infrared radiation)
4	bức_xạ_sóng_điện_từ(electromagnetic wave radiation)
5	bức_xạ_tử_ngoại(ultraviolet radiation)
6	cảm_ứng_điện_từ(electromagnetic induction)
7	chu_kì_sóng(wave period)
8	con_lắc_dây(string pendulum)
9	con_lắc_lò_xo(spring pendulum)
10	công_suất(wattage)

Then the concepts, relationships, and constraints are manually identified by domain experts. They are divided into 14 classes, 29 object properties, 17 datatype properties, 292 individuals and 567 instances of properties. The BaseOntology class hierarchy and the metrics are taken from Protégé and they areshown in Figure 5 and Figure 6, respectively.



- owl:Thing
  - bài *(lesson)*
  - chương *(chapter)*
  - chỉ_mục *(index)*
  - khái_niệm *(concept)*
    - hiện_tượng *(phenomena)*
    - khái_niệm_cơ_bản *(basic concept)*
    - phương_trình *(equation)*
    - thiết_bị *(equipment)*
    - đơn_vị_đo_lường *(unit)*
    - đại_lượng *(physical quantity)*
    - định_luật *(law)*
  - khối *(grade)*
  - tài_liệu *(document)*

Fig. 5: The BaseOntology class hierarchy



Ontology metrics:

Metrics

Metrics	
Axiom	1388
Logical axiom count	994
Declaration axioms count	351
Class count	14
Object property count	29
Data property count	17
Individual count	292
Annotation Property count	2

Fig. 6: The BaseOntology's metrics

The second stage of the proposed method is described in Figure 7. This is called the enrichment stage, which extends the BaseOntology using the pattern-based method.

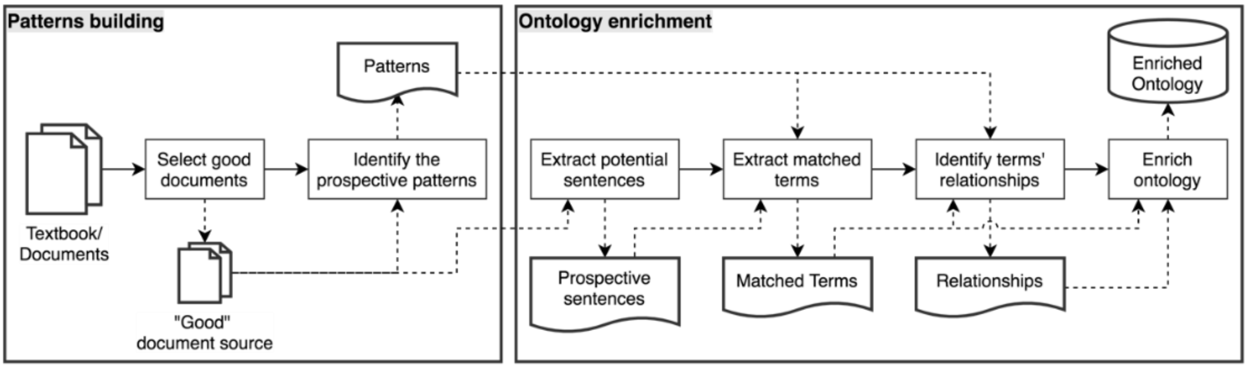


Fig. 7: Proposed pattern-based ontology enrichment process

To extend the BaseOntology, the potential documents were first selected from the possible sources and these documents were then used to analyze the prospective patterns that can be used to identify concepts and relationships from sentences. The document used to extend BaseOntology in this study is the High School Physics Dictionary published by the Vietnam Education Publishing House [22]. Prospective patterns in this study are listed in Table 2and are represented by the regular expression syntax.

Table 2. Proposed patterns for BaseOntology enrichment

No	Type	Pattern	Explanation
1	1	(công thức của)(.{3,100}?)(:)?($.+?$)	Identify formula for a concept
2		(.{10,40})(:)(.{100,})(\n)	Get the definition of a concept
3		(đo)(.{10,100}?)(bằng)(.+?)([\,\.\n])	Measurement equipment of an unit

4		(đơn vị của. ?)(.+?)(là)(.+?)(\, kí hiệu:)(.+)	
5		(định luật.{10,100}?)(:)(.+)(\n)	Definition of a concept
6	2	([^,\.;].{5,30})(còn gọi là)(.+?[^,\.])([();,\.\n])	
7		(.{10,100})(tuân theo)(.{10,100})	Identify concepts that follow a law
8		(.{10,100})(ứng dụng)(.{10,100})	Identify application of a concept
9	3	(khái niệm)(có liên quan đến khái niệm)(khái niệm)	Identify related concepts of a concept

To enrich the BaseOntology using the above proposed patterns, the sentences in the input document need to be extracted first. Then, the sentences are matched with the proposed patterns to take out the potential terms. Next, the extracted terms and patterns areused to identify the relationships between terms. Finally, the identified terms and relationships are added into the BaseOntology.

4.2 Experimental result

The High School Physics Dictionary [22] is utilized to evaluate the proposed method,. The evaluation result is shown in Table 3. For the first pattern, all the formulas of the concepts in the input document can be extracted. However, pattern 8 has the lowest accuracy as it can extract only 14/32 applications of the concepts in the dictionary. On average, 66.15% of concepts in the dictionary were extracted. Many terms could not be actually identified due to the complexity of the nature of language. Simultaneously, there are numerous ways to express the same idea and thus there are certain circumstances that the proposed patterns may not be recognized.

Table 3. The experimental result

No	Pattern	Accuracy (%)
1	(công thức của)(.{3,100}?)(:)?($.+?$)	100.00
2	(.{10,40})(:)(.{100,})(\n)	70.01
3	(đo)(.{10,100}?)(bằng)(.+?)([\,\.\n])	50.00
4	(đơn vị của. ?)(.+?)(là)(.+?)(\, kí hiệu:)(.+)	52.63
5	(định luật.{10,100}?)(:)(.+)(\n)	94.12
6	([^,\.;].{5,30})(còn gọi là)(.+?[^,\.])([();,\.\n])	38.57
7	(.{10,100})(tuân theo)(.{10,100})	90.91
8	(.{10,100})(ứng dụng)(.{10,100})	37.50
9	(khái niệm)(có liên quan đến khái niệm)(khái niệm)	61.61
Average		**66.15**

5. CONCLUSION AND FUTURE WORK

In the current study, a semi-automatic approach for high school physics ontology construction based on the pattern-based method is proposed. Using this method, a basic ontology, called BaseOntology, for high school physics subject has been constructed. This ontology includes only 292 individuals (physics concepts) and 567 instances of properties extracted manually from the 12th grade physics textbook. A high school physics dictionary is used to extend the BaseOntology based on the proposed patterns. There are9 patterns used to enhance the BaseOntology and they can identify 66.15% terms in the dictionary and expand the ontology.

To extract more terms in the input documents, more patterns may be identified to cover missed circumstances. In addition, it is possible that some methods be used based on deep learning techniques to increase the automatic level in ontology construction.

6. REFERENCES

[1] Berners-Lee, T. (1998). Semantic web road map.

[2] Buitelaar, P., Cimiano, P., & Magnini, B. (2005). Ontology learning from text: An overview. Ontology learning from text: Methods, evaluation and applications, 123.

[3] Antoniou, G., & Van Harmelen, F. (2004). A semantic web primer. MIT press.

[4] Studer, R., Grimm, S., & Abecker, A. (2007). Semantic web services. Springer.

[5] Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. Scientific American, 284(5), 34-43.

[6] Lindén, K., & Piitulainen, J. (2004). Discovering synonyms and other related words. In Proceedings of CompuTerm 2004: 3rd International Workshop on Computational Terminology (pp. 63-70).

[7] Fotzo, H. N., & Gallinari, P. (2004). Learning Generalization/ Specialization Relations between Concepts-Application for Automatically Building Thematic Document Hierarchies. In RIAO (pp. 143-155).

[8] Schmid, H. (2013). Probabilistic part-ofispeech tagging using decision trees. In New methods in language processing (p. 154).

[9] Sleator, D. D., & Temperley, D. (1995). Parsing English with a link grammar. arXiv preprint cmp-lg/9508004.

[10] Bird, S. (2008). Multidisciplinary instruction with the natural language toolkit. Association for Computational Linguistics.

[11] Fellbaum, C. (2010). WordNet. In Theory and applications of ontology: computer applications (pp. 231-

243). Springer, Dordrecht.

[12] Zhou, W., Liu, Z., Zhao, Y., Chen, G., Wu, Q., Huang, M. L., & Qiang, Y. (2006). A semi-automatic ontology learning based on WordNet and event-based natural language processing. In 2006 International Conference on Information and Automation (pp. 240-244). IEEE.

[13] Tran, A. C., Dietrich, J., Guesgen, H. W., & Marsland, S. (2017). Parallel symmetric class expression learning. The Journal of Machine Learning Research, 18(1), 2145-2178.

[14] Etzioni, O., Banko, M., Soderland, S., & Weld, D. S. (2008). Open information extraction from the web. Communications of the ACM, 51(12), 68-74.

[15] Zhang, J., Liu, J., & Wang, X. (2016). Simultaneous entities and relationship extraction from unstructured text. International Journal of Database Theory and Application, 9(6), 151-160.

[16] Fleischhacker, D., & Völker, J. (2011, October). Inductive learning of disjointness axioms. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems" (pp. 680-697). Springer, Berlin, Heidelberg.

[17] Chen, Y., Li, W., Liu, Y., Zheng, D., & Zhao, T. (2010). Exploring deep belief network for Chinese relation extraction. In CIPS-SIGHAN Joint Conference on Chinese Language Processing.

[18] Wang, J., Liu, J., & Kong, L. (2017). Ontology construction based on deep learning. In Advances in Computer Science and Ubiquitous Computing (pp. 505-510). Springer, Singapore.

[19] Davies, J., Studer, R., & Warren, P. (Eds.). (2006). Semantic Web technologies: trends and research in ontology-based systems. John Wiley & Sons.

[20] Navigli, R., & Velardi, P. (2008). From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions.

[21] Mädche, A., Staab, S., & Studer, R. (2004). Handbook on ontologies. Handbook on Ontologies, International Handbooks on Information Systems, 173-190.

[22] Hung V. V. & Khiet V. T. (2015), High School Physics Dictionary, The Vietnam Education Publishing House, (in Vietnamese).