A Framework for Vietnamese Text Document Retrieval System Based on Phrasal Semantic Analysis

Tuyen Thi-Thanh Do
Faculty of Software Engineering
University of Information Technology, VNU-HCM
Ho Chi Minh city, Vietnam
tuyendtt@uit.edu.vn

Dang Tuan Nguyen
Faculty of Computer Science
University of Information Technology, VNU-HCM
Ho Chi Minh city, Vietnam
dangnt@uit.edu.vn

***Abstract* — Document retrieval system for Vietnamese language has been recently focused on the semantic aspect. In many research works, the common and general approach of semantic document retrieval systems is based on query expansion. The query expansion is good idea for utilizing existing thesaurus and searching methods to solve the problem of semantic information retrieval. However, this semantic retrieval strategy has to process many expanded queries of an original query. To overcome the problem, a framework of Vietnamese text document retrieval system based on phrasal semantic analysis has been investigated. In the framework, the word expansion task will be performed at indexing process instead of searching process. In addition, the object for indexing is also changed from the keywords to the semantic classes. To evaluation the framework, a text document retrieval system has been build according to the framework for experiment. The retrieval result has F-measure of 84.27% promising some further research for improvements.**

Keywords-Information retrieval framework; semantic information retrieval; semantic class; Vietnamese query analysis

I. INTRODUCTION

Recently, there are many research works for improving the searching performance of semantic information retrieval systems. These systems try to retrieve the information which has the meaning related to the query instead of the keywords appeared in the query. This feature enables people to save times on selecting the appropriate information in a large number of digital documents. Therefore, more and more solutions of semantic information retrieval system have been proposed with many different models. However, for the necessity of Vietnamese document retrieval systems, the research works in Vietnamese text document retrieval system enabling semantic analysis are still not very many. Therefore, an investigation in the framework for Vietnamese text document retrieval system based on phrasal semantic analysis is conducted to show the main problems should be solved to build a system.

In this paper, the framework is proposed after studying many solutions for the semantic information retrieval problems and the characteristic of Vietnamese language. The framework introduces the semantic class and the semantic relation as searching object instead of keywords to ensure the system works on the meanings of phrases. To accomplish the task, the framework contains four main components to process documents and query to identify the semantic classes and semantic relations, index the information and then search the information.

This paper presents the proposed framework in four sections. Section 2 presents the related works in building a semantic information retrieval system to identify the main problems of a semantic information retrieval system. Section 3 presents the characteristic of Vietnamese language and the method of semantic analysis for Vietnamese language. Then, the framework is proposed after considering other systems and the semantic analysis method for Vietnamese language in Section 4. In Section 5, a Vietnamese phrase retrieval system based on phrasal semantic analysis for experiment in implementing the framework is presented. Finally, the conclusions and future works are presented in Section 6.

II. RELATED WORKS

The main problem of building a semantic information retrieval system is to identify the objects of searching method. The objects, for example, are terms in a keyword based information retrieval system [1]. The objective of the keyword information retrieval is to retrieve documents containing many words which have some similar orders as in the query. Because of the characteristic of this approach, the keyword information retrieval systems can be applied to any domains and in any languages without any important modifications except the top-word list. However, these systems cannot retrieve documents which contain the same information as the query but their information is presented in different words or different syntactic structures from the query.

In contrast to keyword systems, semantic information retrieval systems try to retrieve documents which have the meaning related to the meaning of the query. Therefore, the objects of searching method are concepts instead of keywords. In the research works [4], [5] and [6], the concepts are organized in an ontology in which each concept contains many property values. To retrieve the documents for the user query, the systems try to extract the information from the query and identify the concepts which are likely to contain the extracted information as their property values. Then, the systems search these concepts in the document collection to identify the expected documents. The result of retrieval is a list of documents which do not contain words of the original query but have similar meaning to the query. In [5], Szymanski has built ontology according to the "concept of semantic memory" [7]. (Cf. [4], [5], [6], [7]).

The searching objects can be "conceptual structures" as in [8]. Conceptual structures are constructed by three steps [8]: in first step, sentences are parsed by using a syntactic parser; then, the concepts of a knowledge domain are extracted by comparing the parsed result to the domain model concepts which are manually built according to the domain specific knowledge; finally, the relations of these concepts are identified based on the dependencies which are predefined. (Cf. [8]).

In [9], the searching objects are "named entities": they are groups of words which are used to express "concepts" or "names" in the real world. In this system, the most important task is identifying the real world named entities by using specific domain ontology. By using named entities as the searching objects, the system can retrieve more appropriate documents than keyword systems because it recognizes the entities represented by words in the queries. (Cf. [9]).

In [10], Ofer Egozi uses "category labels of a document" as objects of searching. Category labels are titles of articles, such as Wikipedia article's titles. To identify the category labels of a document or a query, its text content will be analyzed by the Explicit Semantic Analysis component to identify which articles are similar to it. Then, the titles of the similar articles are the category labels of the document or the query. These category labels will be indexed as words. (Cf. [10])

III. Vietnamese Phrasal Semantic Analysis

A semantic information retrieval system works with the meaning of phrases or sentences as mentioned above. The meaning can be formalized as concept or category label [4], [5], [6], [7], [8], [9], [10]. A concept is not represented as a group of words in which words have grammatical relations among them and the meaning of the concept is not always made up from the meaning of each word used to present it. For example, "computer science" is considered as a concept because it is already defined while "security task" is not a concept although the grammatical structures of them are the same. In addition, the "computer science" does not have the meaning about the computing machine but has the meaning about solving problems by using a computer which is not known easily from two words "computer" and "science".

Although concepts can be presented by a phrase, such as "computer science", they are usually presented by a word in English. For example, "blackbird" is a word because there is no word separator between "black" and "bird"; however, it implies a bird which is black.

In Vietnamese language, the concepts are defined in a different way. According to H. X. Cao's research work [2], in which there are no compound words, each concept is always defined by a phrase [2]. Each word in a phrase is a concept and the next right word is the sub-category of the previous word. For example, "xe đạp", which means "bicycle", is a phrase containing "xe" (transportation means) and "đạp" (bicycle). In the phrase, "đạp" is a sub-category of "xe". This means there are many kinds of "xe", such as "hơi" (car), "đạp" (bicycle), "máy" (motorbike), etc. and "đạp" is one of them. The phenomenon of using a phrase to name an object has also appeared in English but it is not frequently as in Vietnamese because if a new concept is created from some words in English, it will be written continuously without spaces between words, such as "blackbird" or "catfish".

An important characteristic of Vietnamese language is that there are many synonyms in Vietnamese because people in different regions may have dialectal words. For example, the North people call a pig "lợn" while the South people call it "heo". There are not any semantic differences between these words. In addition, the synonyms can be used together to express its meaning as shown in [11]. For example, "tìm" and "kiếm" have the same meaning which means "search" and they can be used together as "tìm kiếm" or "kiếm tìm" to express the behavior of searching. This problem makes searching process more complex to find the relevant documents.

Beside the synonymy problems, each word usually has many meanings because Vietnamese is a monosyllabic language in which words do not have morphology. For example "hàng" (merchandise/line) can be a noun, or a verb, which means "surrender". Therefore, each concept is presented by a phrase of words in which the next right word complements the left word in order to eliminate any lexical polysemy in the phrase. For example, a sentence "họ giao hàng" will be understood as "they deliver goods" because "hàng" does not mean "line" or "surrender" in this context. To eliminate the polysemy of the word "hàng", Vietnamese people should say "họ giao hàng hóa" to express the idea "they deliver goods". The word "hóa" has eliminated the polysemy of the word "hàng".

A. Lexicon Ontology

According to the characteristics of Vietnamese language, it is very important to identify the meaning of each word explicitly. To identify the meanings of each word, each meaning of each word has to be explicitly defined and the semantic relations between it and the meanings of other words have to be established also. This information cannot be organized in a dictionary. They should be described in a lexical ontology. In addition, a lexical ontology contains all relations of a semantic class and the others. These relations are used to specify the head of a phrase without head rules in dependency parsing [14]. The dependency parsing results are used to translate a phrase into logic form. For example, the word "black" is an adjective and the word "horse" is a noun. In the phrase "black horse", the word "horse" is the head because of the rule in which the adjective modifies the noun. In the lexicon ontology, the word "black" has a relation to the word "horse" indicating the word "black" modifies the word "horse".

In the lexicon ontology, each individual should be used to represent a specific meaning of a word. The individual is called semantic class which is the implementation of the semantic memory concept [7]. If two words have the same meaning, this meaning should be represented by an individual. By using an individual to represent a meaning instead of a word, the problem of synonymy in retrieval process is solved efficiently.

Example 1: "phương pháp nuôi heo" and "cách nuôi lợn" are two phrases which have the same meaning which is "breeding pig method". In the two phrases, the three words "phương", "pháp" and "cách" have the same meaning which is "method". Because two synonyms can be used together to express their same idea, the phrase "phương pháp" has the same meaning to word "phương" or "pháp". If these words are organized in an lexicon ontology, there are individuals:

- The individual cls_cách is the meaning of three words "phương", "pháp" and "cách".
- The individual cls_nuôi is the meaning of the word "nuôi" which means breed.
- The individual cls_heo is the meaning of two words "heo" and "lợn".

By using the ontology, the phrase "phương pháp nuôi heo" can be represented as "cls_cách cls_cách cls_nuôi cls_heo" which can be reduced as "cls_cách cls_nuôi cls_heo" because there are two consecutive individuals cls_cách; the phrase "cách nuôi heo" can be represented as "cls_cách cls_nuôi cls_heo". The two phrases have the same representation in the meaning therefore they have the same meaning.

The lexicon ontology also contains information about relations of the meanings. There should be six kinds of relation as follow:

- Sub class relation: indicate that a meaning is a sub-category of an other meaning.
- Antonym relation: indicate that two meaning contradict each other.
- Modify relation: indicate that a meaning can be used to modify an other meaning. This relation is established between a meaning of an adjective and a meaning of a noun.
- Actor relation: indicate that a meaning can be an actor of an other meaning. This relation is established between a meaning of noun or pronoun and a meaning of a verb that the noun or pronoun is subject of the verb.
- Direct object relation: this relation is established between a meaning of noun or pronoun and a meaning of a verb that the noun or pronoun is the direct object of the verb.
- Indirect object relation: this relation is established between a meaning of noun or pronoun and a meaning of a verb that the noun or pronoun is the indirect object of the verb.

B. Semantic Analysis Method

Semantic analysis method should be applied to phrases instead of complete sentences because the user queries are usually in phrases. In addition, a complete sentence can have a complex syntactic structure which cannot be easy to identify its semantic. To analyze the semantic of a phrase, the lexicon ontology will be used in a process of three following steps:

- Step 1: Each word in a phrases is identified the semantic class which is represented by the the word. The semantic class is a unique element in ontology which is used to assign the semantic for a word. For example, "heo" and "lợn" is two words and they have the same semantic class which is an individual named "cls_heo". This means the two words "heo" and "lợn" is the labels of one object which is formalized as "cls_heo".
- Step 2: The phrase is reduced by removing the semantic class whose the next right semantic class is identical in the phrase. For example, the phrase "cls_cách cls_cách cls_nuôi cls_heo" is reduced as "cls_cách cls_nuôi cls_heo" because the semantic class cls_cách is identical to its next right one.
- Step 3: The phrase is analyzed to identify the head of phrase and the heads of sub-phrases by using the relation between semantic classes defined in the ontology. For example, "sản xuất pin mặt trời" (produce solar pin) after identifying semantic classes, the result is "cls_sản cls_xuất cls_pin cls_mặt cls_trời". In Vietnamese, "sản xuất" is recognized as a compound word in which the word "sản" (to produce) is the main meaning and the word "xuất" (to export) is the subsidary. These two words form a phrase "sản xuất" (to produce). Therefore, there is a relation "comp(cls_xuất, cls_sản)" to describe this compound word. Similarly, there is a relation "subcls(cls_trời, cls_mặt)" to describe the relation of the word "mặt" and the word "trời" to form a phrase "mặt trời" (solar). By using the ontology, the phrase "sản xuất pin mặt trời" can be translated into a structure shown in Figure 1.

The result shows that the word "sản" is the head of "sản xuất pin mặt trời" and of "sản xuất"; "pin" is the head of "pin mặt trời" and "mặt" is the head of "trời" (in English, "produce" is the head of "produce solar pin" and "pin" is the head of "solar pin"). This structure can be reproduced in logical form as "comp(cls_xuất, cls_sản), comp(cls_pin, cls_sản), subcls(cls_mặt, cls_pin), subcls(cls_trời, cls_mặt)".

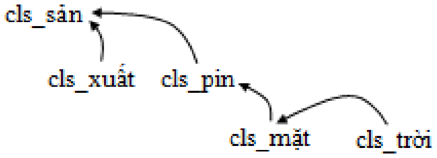

Figure 1. The structure of the phrase "sản xuất pin mặt trời".

The method of analyzing the semantic of a phrase as proposed above exposes the two important information of a phrase which are the meanings described by words and the relations between these meanings. By using the meanings, which are semantic classes, the two phrases "nuôi heo" and "nuôi lợn" (breed the pig) are easily identified the same because their semantic classes are the same "cls_nuôi cls_heo". By using the relations of the semantic classes, the two phrases "nuôi heo" (breed the pig) and "heo nuôi" (the pig breeds..) are easily identified the difference because the relations of "nuôi heo" are {dobj(cls_heo, cls_nuôi)} (cls_heo is the direct object of the behaviour cls_nuôi) while the relations of "heo nuôi" are {actor(cls_heo, cls_nuôi)} (cls_heo is the actor of the behaviour cls_nuôi). This information is very important to identify if two phrases are similar in semantic.

IV. FRAMEWORK FOR VIETNAMESE TEXT DOCUMENT RETRIEVAL SYSTEM BASED ON PHRASAL SEMANTIC ANALYSIS

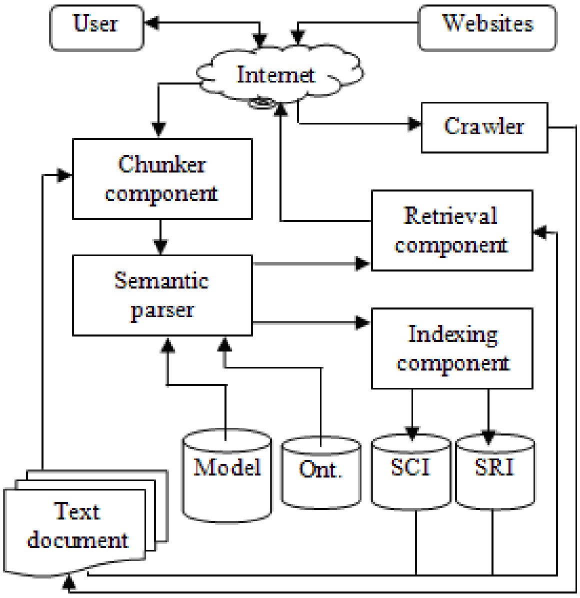

Figure 2. The framework for Vietnamese text document retrieval system based on phrasal semantic analysis.

The semantic document retrieval system for Vietnamese will work with the concept, as presented in [3], to enable the semantic feature in retrieval process. The semantic feature contains two aspects which are the meaning of word, called the semantic class, and the relation of any two semantic classes as mentioned above. Therefore, a framework for Vietnamese text document retrieval system based on phrasal semantic analysis is proposed as Figure 2.

A Vietnamese text document retrieval based on phrasal semantic analysis should have four main components.

A. *Chunker Component*

The chunker component splits a document or a user query into phrases. A phrase can be a clause, a verb phrase or a noun phrase. If a complex sentence is split into phrases, the clauses are tried to be split first, then the verb phrases and finally the noun phrases. The document has to be chunked into phrases because the relations of semantic classes are in a phrase. For example, the sentence "the black horse is in the stable and the small cat is in the garden" can be split into two clauses (1) "the black horse is in the stable" and (2) "the small cat is in the garden". In the clause (1), the word "black" can modify the word "horse" but cannot modify any word in the phrase (2). Similarly, each word in the phrase (1) cannot modify any word in the phrase (2). Therefore, it is not necessary to check whether a word in one phrase has a relation to a word in an other phrase. In reality, there are relations between the words of different sentences. This problem concerns to discourse analysis which is out of the phrasal semantic analysis.

B. *Semantic Parser*

The typical semantic parser of the proposed framework will identify semantic classes and relations between them, related to a word. This component should use a statistical model for identifying the semantic class of a word. To construct the model, many training phrases are manually analyzed to produce POS tag like representations in which each word is tagged with its semantic class. The results of semantic class tagging will be used to train a semantic class tagger. The semantic class tagger is used to identify the semantic class of every word in phrases being analyzing.

Example 2: To build a semantic class tagger for a set of three phrases (1) "phương pháp nuôi heo", (2) "cách nuôi lợn" and (3) "sản xuất pin mặt trời", the two step task is performed:

- Step 1: manually assign the semantic class to each word in each phrase to produce the POS tag like representation. The results are "phương/*cls_cách* pháp/*cls_cách* nuôi/*cls_nuôi* heo/*cls_heo*", "cách/*cls_cách* nuôi/*cls_nuôi* lợn/*cls_heo*" and "sản/*cls_sản* xuất/*cls_xuất* pin/*cls_pin* mặt/*cls_mặt* trời/*cls_trời*".
- Step 2: use the results in step 1 to train a semantic class tagger. The model of the semantic class tagger can be Hidden Markov Model [15], Maximum Entropy [16] or any statistical model. After training, the semantic class tagger can be used to identify the semantic classes of a phrase. For example, the phrase "phương pháp sản xuất pin mặt trời" (producing solar pin method) will be tagged as "phương/*cls_cách* pháp/*cls_cách* sản/*cls_sản* xuất/*cls_xuất* pin/*cls_pin* mặt/*cls_mặt* trời/*cls_trời*".

After identifying the semantic classes of a phrase, the parser applies the semantic analysis method to identify the relations between the semantic classes by using relations defined in the lexicon ontology. Then, the result is translated into logical form. For example, the tagged phrase "phương/*cls_cách* pháp/*cls_cách* sản/*cls_sản* xuất/*cls_xuất* pin/*cls_pin* mặt/*cls_mặt* trời/*cls_trời*" is represented as "*cls_cách cls_cách cls_sản cls_xuất cls_pin cls_mặt cls_trời*". Then, it is reduced to "*cls_cách cls_sản cls_xuất cls_pin cls_mặt cls_trời*". By using the lexicon ontology, the relations of these semantic classes are identified in order to specify the heads in the phrase as shown in Figure 3.

After identifying the relations between semantic classes in the phrase, these relations are translated into logic form as a set of logic functions: {comp(cls_sản, cls_cách), comp(cls_xuất, cls_sản), dobj(cls_pin, cls_sản), comp(cls_mặt, cls_pin), subcls(cls_trời, cls_mặt)}. These logic functions are the objects of indexing process like semantic classes.

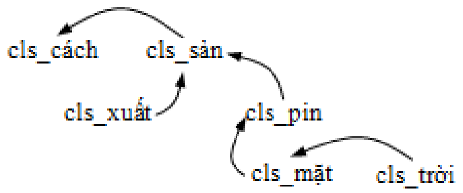

Figure 3. The structure of the phrase “phương pháp sản xuất pin mặt trời”.

C. Indexing Component

The indexing component indexes the semantic classes and semantic relations of every phrases of every document. The semantic classes are also expanded to their hypernyms for searching the hyponyms of semantic classes of the query. For example, assume there is a semantic class *cls_Fish* has hyponyms which are the semantic class *cls_Tuna* and *cls_Cod*; the phrase to be indexed contains semantic class *cls_Cod*. Therefore, the semantic class *cls_Cod* should be expanded to the semantic class *cls_Fish* for searching in the case a user query contains hypernym of the semantic classes. Then the list semantic classes being indexed also contains *cls_Fish*. If the query has semantic class *cls_Fish*, it is easy to retrieve a phrase containing the semantic class *cls_Cod* because that phrase has been expanded. When expanding the semantic classes, there is a problem of expanding level. The higher level is expanded, the larger index size is grown.

For working efficiently in the retrieval component, the index information should contain the identification number of a phrase and of the document containing the phrase. The indexing component works with phrases instead of documents because the system has to compute the semantic relateness of two phrases. Because the two types of semantic information are semantic class and semantic relation, the indexing component should create two indices which are Semantic Class Index (SCI) and Semantic Relation Index (SRI). The format of SCI and SRI are belong to the computation of semantic distance.

D. Retrieval Component

The retrieval component identifies the relatedness of the query and the phrases in the document collection by using the SCI and SRI. Before building this component, there should be a metric defined. This metric is used to identify the semantic distance which is to identify the relatedness of two phrases. To identify the expected documents, the system computes the semantic distance between the query and each phrase in the document collection by using two indices SCI and SRI. Then, the values calculated are used to identify the expected phrases. Based on the expected phrases, a list of documents containing some of these phrases is the retrieval result.

The semantic distance can be calculated according to the document similarity described in [13]. The document similarity calculation should be modified to work with semantic classes and semantic relations.

V. Experiment

A system of Vietnamese phrase retrieval system has been built according to the framework to test the capability of combination of the component. The phrase retrieval system, instead of the text document retrieval system, is built for evaluation because the list of documents identified expected documents totally depends on the list of expected phrases. In addition, it is easier to check the semantic relatedness of a query and a phrase than to check the semantic relatedness of a query and a whole document. The experimental system contain four main components as follow

- **Chunker component**. The chunker component is very simple. It uses separators which are conjunction words and sentence marks to identify the phrases. When a document is input, this component uses those separators to split it into phrases and specifies the identification number of the document and its phrases for indexing process.
- **Semantic parser**. The semantic parser uses the N-Gram model to identify the semantic classes, then it uses the Vietnamese Lexicon Ontology to identify the relations of the semantic classes of a phrases. The N-Gram model and Vietnamese Lexicon Ontology are created in [11].
- **Retrieval component**. This component computes the relatedness of two phrases based on semantic distance. The semantic distance of two Vietnamese phrases is based on the distance between two individuals on the lexicon ontology. The distance between two individuals is the number of edge to travel from one individual to the other. The semantic distance of two phrases is the sum of semantic class distance and semantic relation distance between them. The semantic class distance and the semantic relation distance are presented in detail in an other article.
- **Indexing component**. This component and the index format is developed from the research result in [12] to work efficiently with the retrieval component. The format of the indices and the semantic distance are presented in the same article.

After building the system, a test set contains 30 queries and 720 phrases for searching has been conducted manually in Vietnamese language. The test result is shown in Table I and indicates that the precision is 81.67%, the recall is 87.05% and F-measure is 84.27%.

VI. Conclusions and Future Works

This paper presents the framework for Vietnamese text document retrieval system based on phrasal semantic analysis. The framework is proposed after researching many solutions of semantic information retrieval and the characteristics of Vietnamese language in lexicon and grammar. The framework proposes using the semantic class and the semantic relatedness as searching object instead of word. In the framework, it is recommended to use a statistical model for semantic class identification and a lexical ontology for semantic relatedness identification according to the characteristic of Vietnamese language. To fully implement the framework, a metric of semantic distance should be defined to compute the relatedness of two phrases. Based on the metric, the format of the index is design to increase the efficiency of retrieval process.

According to the proposed framework, a Vietnamese phrase retrieval system has been built for experiment. The system shows the good capability of the combination of the components with the result of 81.67% in precision, 87.05% in recall and 84.27% in F-measure. The result shows that the proposed framework could be used to implement a semantic information retrieval system for Vietnamese.

In future, a large scale system should be built for experiment to evaluate the performance of the system. With the result of the large scale system, the framework should be modified to suit the practical conditions.

ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2013-26-06.

TABLE I: THE TEST RESULT OF THE EXPERIMENTAL SYSTEM

Query in Vietnamese	Query in English	Returned (n_1)	Correct (n_2)	Expected (n_0)	P (n_2/n_1)	R (n_2/n_0)
Nhà khoa học	Scientist	26	26	26	1.0	1.0
Giáo sư	Professor	4	4	7	1.0	0.571
Nhà máy điện	Electricity plant	64	63	77	0.984	0.818
Pin mặt trời	Solar battery	6	6	21	1.0	0.286
Năng lượng mặt trời	Solar energy	58	51	68	0.879	0.75
Máy bay nhẹ	Lightweight helicopter	25	24	24	0.96	1.0
Nhà máy điện hạt nhân	Nuclear electricity plant	77	68	80	0.883	0.85
Khảo sát địa điểm xây dựng nhà máy	To survey locations for building a plant	83	74	88	0.892	0.841
Xe tay ga	Scooter motocycle	41	41	41	1.0	1.0
Công ty thiết kế	Designer company	17	17	28	1.0	0.607
Khả năng bám cua	road pulling ability	43	19	19	0.442	1.0
Phương pháp cấp diện mới	Supplying electricity method	58	57	70	0.983	0.814
Nhà nghiên cứu	Reseacher	23	20	20	0.869	1.0
Phòng thí nghiệm	Laboratory	16	9	12	0.562	0.75
Phát triển công nghệ	To develop the technology	34	33	33	0.971	1.0
Tấm dán tường phát sáng	Lighting panel	17	16	23	0.941	0.696
Điều khiển xe lăn	Control the wheelchair	58	55	56	0.948	0.982
Chuyên gia	Professional	9	9	9	1.0	1.0
Giúp chuyên gia biết nguy hiểm	To help the professional to realize a danger	35	29	30	0.828	0.967
Sử dụng năng lượng mặt trời	To use the solar energy	73	68	81	0.932	0.839
Thiết kế nhỏ gọn	Small and tidy designer	22	22	23	1.0	0.957
Xe có hệ thống bánh lái	Car with helm system	113	96	129	0.849	0.744
Nguồn tạo khí carbon	Carbonic creating source	58	33	34	0.569	0.970
Dùng lưỡi điều khiển	To use tongue to control	40	20	21	0.5	0.952
Phương pháp giữ phản vật chất	Reserving antimatter method	31	25	26	0.806	0.962
Máy bay siêu nhẹ	Super lightweight plane	25	24	24	0.960	1.0
Xử lý chất thải	Waste subtance processing	22	3	3	0.136	1.0
Hội nghị về khí hậu	Conference on climate	221	3	3	0.136	1.0
Xây dựng hệ thống định hướng	To build a navigating system	57	38	41	0.667	0.927
Tàu chạy bằng năng lượng mặt trời	Ship using solar energy	91	84	101	0.923	0.832
Total					**0.8167**	**0.8705**

REFERENCES

[1] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. *Introduction to Information Retrieval*. NY, USA: Cambridge University Press, 2008, pp. 18-33.

[2] Cao Xuan Hao. *Tiếng Việt: mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa*. Vietnam: Giáo Dục publisher, 2007.

[3] Tuyen Thi-Thanh Do. “A Concept Identification Method for Vietnamese Concept-based Information Retrieval System,” in Proc. iiWAS, 2012, pp. 403 - 406.

[4] Miriam Fernández Sánchez. “Semantically enhanced Information Retrieval: An ontology-based approach.” Doctoral dissertation, Unitversidad de Autónoma, Spain, 2009.

[5] Julian Szymanski, Wlodzislaw Duch. (2012, April) “Information retrieval with semantic memory model”. Cognitive Systems Research. 14(1), pp. 84-100.

[6] Stein L. Tomassen and Darijus Strasunskas. (2010) "Measuring intrinsic quality of semantic search based on feature vectors." Int. J. Metadata, Semantics and Ontologies. 5(2), pp. 120-133.

[7] Rogers , T. T. , Lambon Ralph, M. A, Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., and Patterson, K. (2004) "The structure and deterioration of semantic memory: A neuropsychological and computational investigation". Psychological Review, 111(1), pp. 205-235.

[8] T. Rindflesch and A. Aronson. "Semantic processing in information retrieval," in Proc. SCAMC, 1993, pp. 611-615.

[9] Atanas Kiryakov, Borislav Popov, Ivan Terziev, Dimitar Manov, and Damyan Ognyanoff. (2004) "Semantic annotation, indexing, and retrieval." J. Web Semantics: Science, Services and Agents on the World Wide Web. 2(1), 2004, pp. 49–79.

[10] Ofer Egozi, Shaul Markovitch, Evgeniy Gabrilovich. (2011) "Concept-Based Information Retrieval Using Explicit Semantic Analysis." ACM Trans. Information Systems. 29(2), pp. 1-34.

[11] Tuyen Thi-Thanh Do. "Building a Vietnamese Lexicon Ontology for Syntactic parsing and Document Annotation," in Proc. of iiWAS, 2013, pp. 619-624.

[12] Tuyen Thi-Thanh Do. "Ontology-based Annotation and Indexing for Vietnamese Text Document," in Proc. of iiWAS, 2013, pp. 363-367.

[13] Ralf Steinberger, Bruno Pouliquen, Johan Hagman. "Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC," in Proc. of CICLing, 2002, LNCS vol. 2276, pp. 415-424.

[14] Kemal Oflazer. (2003) "Dependency parsing with an extended finite-state approach." Computational Linguistics. 29(4), pp 515-544.

[15] Michael Collins. "Discriminative training methods for Hidden Markov Models: Theory and experiments with perceptron algorithms," in Proc. of EMNLP, 2002, pp 1-8.

[16] Adwait Ratnaparkhi. "A maximum entropy model for part-of-speech tagging," in Proc. of EMNLP, 1996, pp. 133-142.