Journal of Computer Science and Cybernetics, V.31, N.3 (2015), 185– 202
DOI: 10.15625/1813-9663/31/3/5923

PHRASAL SEMANTIC DISTANCE FOR VIETNAMESE TEXTUAL DOCUMENT RETRIEVAL

DO THI THANH TUYEN[†] AND NGUYEN TUAN DANG[‡]

University of Information Technology, VNU-HCM;
[†]*tuyendtt@uit.edu.vn;* [‡]*dangnt@uit.edu.vn*

Abstract. In this paper, a computational semantic method is proposed to estimate the phrasal semantic distance used in our model of Vietnamese document retrieval system. The semantic distances between phrases are defined in terms of semantic classes and semantic relations to ensure that it can reflect how different two certain phrases are. To estimate the semantic distance, the semantic classes of a phase are identified by using the n-gram model. After identification of the semantic classes, their semantic relations are also identified by using a Vietnamese Lexicon Ontology. This handcrafted ontology contains defined semantic classes and their potential relations in Vietnamese language explicitly. For the evaluation purpose, a phrasal semantic retrieval system has been built to test with a data set of 720 phrases and 30 queries. The evaluation shows the precision of 96.6% and the recall of 78.4% on experiment results.
Keywords. Lexicon ontology, phrasal semantic analysis, semantic class, semantic distance, semantic information retrieval.

1. INTRODUCTION

Actually, most approaches of modern information retrieval systems are aimed at exploiting semantic features of phrases in both documents and queries to identify which documents are relevant to the user's needs. In fact, the systems conceived by such approaches are called "semantic information retrieval systems", which are distinguished from the other information retrieval systems working with documents of semantic web standard as in [1, 2].

In an information retrieval system, the key problem is how to estimate the "semantic similarity" between a keywords based query and each text document. To solve this problem, the searching unit which is used to calculate the "semantic similarity" has to be defined firstly. Then, a metric will be defined in terms of searching unit for calculating the semantic distance of a query and a document. In keyword information retrieval system [3], the searching unit is term and the metric is defined as a function which returns the weight of a term identified by its occurrence in the document collection. The weight of a term is calculated by using tf and idf values of the term. To calculate the similarity of a query and a document, they are represented as two multi-dimensional vectors according to the Vector Space Model [3]. By using the terms as the searching unit, the retrieval process only tries to find the document containing exact words which appear in the user query. It cannot find the documents which are written by the synonyms of words of user's query. This characteristic is a disadvantage of keyword information retrieval systems. In semantic information retrieval systems, the searching unit is not directly the term because it has to represent the meaning of the term.

In this paper, the concept and the concept relation are used as searching units. It means the search process works with the concepts and the relations existing in documents and in user's queries. This approach includes two issues. The first issue is how to identify the searching units, which are concepts and their relations from a phrase, and the second issue is how the semantic distances between pairs of phrase are calculated. The first issue will be solved by using n-gram model created upon training data in which each word is manually tagged with its concept, called semantic class. This n-gram model will be used to identify the concepts of phrases. After concept identification process, the distances between phrases are calculated with the semantic distance formulas defined to solve the second problem.

The paper is organized as follows: Section 2 reviews some related works about semantic information retrieval, Section 3 describes our proposed approach to estimate the semantic distance between Vietnamese phrases, Section 4 presents the experimental system built to evaluate the performance of the system when phrasal semantic distance is used, and Section 5 recaps our contributions and concludes the paper.

2. RELATED WORKS

The most crucial issue of semantic information retrieval systems is to find appropriate documents whose textual contents are relevant to the queries of user in natural language form. This challenge cannot be solved directly by invoking computer processing because the computer does not understand the natural language as human does now. Therefore, a universal approach in information retrieval domain for resolving this problem is to reduce it into an easier problem in which the retrieved documents must contain words which are related to words of the queries. The relations between words are synonymy, hypernym, hyponymy, holonymy and meronymy. According to this approach, many previous works tried to apply calculation methods used in keyword information retrieval method to calculate the semantic distance between the semantic representations of the queries and the searching documents. These methods can be divided into two classes: "query enrichment" (or "query expansion") and "semantic annotation".

2.1. Approaches of query enrichment

In most query enrichment methods, the query is represented as a set of derived queries which are considered as equivalent with the original query. The semantic distance between the original query and a document is defined as the semantic distance between the set of derived queries and that document. In this approach, the Vector Space Model [3] is applied to represent semantic vectors of queries and documents, and used to calculate semantic distances between them.

In [4–6], the set of queries is created by following a process of two steps. Firstly, the terms of the original query are extracted. An information extraction tool will be used to identify the named entities in this step. Then, the related terms of these terms are used to form new queries by using a thesaurus or ontology of a specific domain. These queries are used to retrieve documents. The retrieved documents may contain information related to the original query without containing any words of that query.

In a different approach of semantic information retrieval, Szymanski proposed a method to find documents containing the homonyms of the user queries in [5]. To do that, the terms of the query is extracted firstly. Then, these terms are used to identify the concepts which contains these terms by using ontology. These concepts are used to form new queries to search documents. The ontology

in [5] was built according to the concept of "semantic memory" [7]. For example, a user may input query "four wheels transportation". The query is processed to get the terms which are "four" "wheel" and "transportation". Assuming that there is a concept named "car" in the ontology which contains "four", "wheel" and "transportation" in it properties, then the word "car" is selected to enrich the query. As the result, many documents containing word "car" are returned to the user.

2.2. Approaches of semantic annotation

The semantic distance between a document and a query is calculated upon their own semantic representations called annotation. There are three types of annotation as follows:

The first type of semantic annotation uses the predicate argument structure. The predicate argument structure contains a "functor", which is usually a verb or a preposition following its nominal arguments. In Rindflesch's work [8], a text document is firstly split into sentences. Each sentence is then parsed to determine noun phrases, verb phrases, and preposition phrases. Next step, each of these phrases is replaced by the terminology of the application domain which has the same meaning to the original phrase by using the Metathesaurus. Finally, the new phrases are mapped into predicate argument structure by using concept models which are manually predefined. These predicate argument structures are used to calculate the semantic distances between queries and documents by matching. According to this approach, a set of logic formulas is used to represent a document which can be annotated exactly based on well-defned concept models and mapping operations.

The second type of semantic annotation uses ontological concept. According to [4,9], a document or a query is analyzed to extract the named entities such as proper name, address, etc. These extracted named entities are then referred to ontology to identify which concepts contain them as their property values. These concepts will be used as searching units. Similar to the ontological concepts, the searching units can be the semantic categories which are the results of text classifying process. In [10, 11], the semantic categories are the article's titles. A document or a query is classified as a category which is the title of an article if the content of the document or the query is similar to the content of the article. According to this approach, a document or a query is annotated by a list of article's titles. These article's titles are the searching units which are used to calculate the semantic distance in the same way as keyword information retrieval method.

The third type of semantic annotation uses propositional description logic. The annotation of a sentence is a logic expression which presents the meaning of the sentence in propositional description logic. The annotation is created after doing a two-step process [12]. In the first step, all concepts of the sentence are extracted by a syntactic parser. In the second step, these concepts are checked for the ability of forming complex concepts by using a dictionary. These complex concepts are used as propositions to form a logic expression. According to this type, a document is annotated as a set of logic expression which is also the searching unit. The dictionary used in annotation process is manually built. This dictionary contains all complex concepts in all documents which will be searched.

After translating the documents or queries into the appropriate semantic representation, the searching units can be used to calculate the semantic distances according to the document similarity calculation method described in [13] because the searching units can be used as descriptors of an inter-lingual.

3. PHRASAL SEMANTIC DISTANCE

3.1. Concept naming, synonymy and polysemy resolution

In [14], the phrasal semantic distance is defined according to linguistic characteristics of Vietnamese language to reflect the "semantic similarity" of two phrases in Vietnamese. There are three important characteristics addressed that are the concept naming method, the synonymy and the polysemy in Vietnamese.

According to [15], phrases are usually used to name concepts in Vietnamese language. Because there is no morphological and syntactic rules to explicitly identify if a word is an adjective, a verb or a noun in Vietnamese, it is not easy to identify whether a phrase is the name of a concept or not. For example, "*xe tải*" ("*lorry*") is a phrase according to [15] (it is a complex word in which "*xe*" is the main word and "*tải*" is the complementary word according to many Vietnamese linguists). This phrase contains two words: "*xe*" ("*vehicle*") and "*tải*" ("*lorry*", the sub-category of the word "*xe*"). The word "*tải*" also has a meaning "*transport*" which is a verb. Therefore, the phrase "*xe tải*" is ambiguous because it can be a name of a concept or a phrase containing a noun and a verb like "*xe chở*" ("*the vehicle transports*") without any differences in morphology or syntax. The lack of morphological and syntactic rules to explicitly identify the word category causes the ambiguous in keyword information retrieval method. The example in [16] shows that when using phrase 1) "*máy tính khoa học*" ("*scientific calculator*") to search by Google[1], the result includes many documents containing phrase 2) "*khoa học máy tính*" ("*computer science*") in high ranks (the highest rank is 2nd) because the phrase "*khoa học*" ("*scientific*"), which can be used as complex word, in phrase 1) does not have any differences in morphology from the phrase "*khoa học*" ("*science*") in phrase 2).

The problem of synonymy is also important because there are many synonyms in Vietnamese. These synonyms appear because of two reasons. The first reason is that people in different regions may have dialectal words. For example, the North people call a pig "*lợn*" while the South people call it "*heo*". The second reason is that the Vietnamese vocabulary is composed of pure Vietnamese words and the Hanji-Vietnamse words. For example, "*hiền*", which is a Hanji-Vietnamese word, and "*lành*", which is a pure Vietnamese word, have the same meaning ("*gentle*") while they are in the phrase "*hiền lành*". There are not any semantic differences between these synonyms. In addition, the synonyms can be used together to express their same meaning. According to [17], "*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 search process more complex to find the relevant documents. Beside the synonymy problem, the problem of polysemy also makes Vietnamese more difficult to identify the meaning of a word. For example, the phrase "*chim én*" ("*swallow*") and the phrase "*gà ác*" ("*black chicken*") have the same structure. In the phrase "*chim én*", the word "*chim*" ("*bird*") is complemented by its sub-category word "*én*" (it means "*én*" is a kind of "*chim*"). In the phrase "*gà ác*", the word "*gà*" ("*chicken*") is also complemented by its sub-category word "*ác*" (it means "*ác*" is a kind of "*gà*"). However, it is possible to write "*én*" instead of "*chim én*" while it is impossible to write "*ác*" instead of "*gà ác*" because "*én*" does not have any popular homonyms while "*ác*" has a popular homonym which means merciless. Therefore, Vietnamese people always use phrases in which the next right word is the sub-category of the left word to express a concept because the polysemy of words is eliminated by using this way.

[1]http://www.google.com.vn/, accessed on 20th Feb, 2015.

3.2. Conception of semantic class

According to the above characteristics of Vietnamese language, the conception of semantic class is proposed to solve the problems of synonymy and polysemy. In [14, 17, 18], a semantic class is defined as a unique sign representing a specific meaning of a word in a specific context. The definition of semantic class, which is originated from the concept of "*semantic memory*" [7], indicates that the synonyms have the same semantic class and all meanings of a word in a specific context are explicitly identified. Therefore, the semantic class is the foundation of the semantic distance calculation which reflects more accurately the similarity between phrases in meaning.

The semantic class has the same function as the part of speech of a word in Vietnamese language. Because there is no morphology in Vietnamese language, a word cannot be classifier into a certain word category with its morpheme. Vietnamese people can only determine the word category of a word if they know the meaning of that word in a sentence. For example, "*bàn*" can be a noun ("*table*") or a verb ("*discuss*"). Its word category is identified if the meaning of the sentence containing it is known. For that reason, the semantic class has been used to build a Vietnamese Lexicon Ontology (VLO) for syntactic parsing and semantic annotation in [17]. The Vietnamese Lexicon Ontology contains every meaning of a word in a specific context. For example, the word "*bàn*" has two meanings that are "*discuss*" and "*table*". Therefore, there are two concepts (or semantic classes) for the word "*bàn*" that are, supposedly, "*bàn_ discuss*" and "*bàn_ table*". In addition, the word "*thảo*" has two meanings "*discuss*" and "*grass*". Thus, the word "*bàn*" and "*thảo*" are labels of the semantic class "*bàn_ discuss*".

By using semantic classes, the question of searching word containing the synonyms of the query is solved efficiently. Because words of documents and of the queries are replaced by their semantic classes, all synonyms of a word are represented by a unique string. Therefore, the amount of comparing operators will be reduced.

When using semantic classes for analyzing phrases, an important problem is how to identify the semantic classes of a phrase. In [17], the problem of identifying the semantic classes is solved by applying POS tagging method. By this way, the POS tags are replaced by the appropriate semantic classes of the words in creating the training data step. Then, the Hidden Markov model [19] or Maximum Entropy model [20] can be used to train a semantic class tagger. The semantic identification process is done as the same way as the POS tagging does. In [17], the n-gram model is used with the accuracy of 74.05%.

3.3. Conception of semantic relation

The semantic relation is the dependency relation of two semantic classes. In this paper, the dependency relations are defined in VLO [17]. There are six important types of relations in VLO. The relations between semantic classes are important for dependency parser [21] to identify the dependencies of semantic classes of a phrase. The six types of relation are:

- Sub class relation (subcls): indicate that a meaning is a sub-category of a certain meaning.
- Antonym relation (ant): indicate that two meanings contrast to each other.
- Modify relation (comp): indicate that a meaning can be used to modify a certain meaning. This relation is established between a meaning of an adjective and a meaning of a noun.

- Actor relation (actor): indicate that a meaning can be an actor of a certain 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 (dobj): 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 (idobj): 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.

The semantic relations are used in two purposes. The first purpose is to identify the importance of a semantic class to the other. In a dependency relation, there are a head and its dependant. If there are two semantic classes combining a phrase, the head is important than the dependent when calculating the semantic distance. The dependency relations are also used to distinguish two sentences containing the same semantic classes. For example, two sentences "chim n c" ("the bird eats the fish") and "c n chim" ("the fish eats the bird") contain the same semantic classes which are represented by the words "chim", "n" and "c" but they do not have the same meaning. The two sentences are represented in relations as "dobj(c, n), actor(chim, n)" and "dobj(chim, n), actor(c, n)" which are different.

To identify the dependencies of semantic classes of a phrase, each word of the phrase is identified for the semantic class first. Then, a process of two steps is applied as follows [14]:

- Step 1: splitting the semantic classes into groups. The semantic classes are grouped by Hanji-Vietnamese or pure Vietnamese, then the semantic class categories. The semantic class categories are noun, verb, adverb, adjective, pronoun, number and conjunction.

- Step 2: identifying the dependency relation of the semantic classes. With each group, every two consecutive semantic classes are checked the relation of them by using the lexicon ontology. If there is a relation in these semantic classes, they combine a semantic class group which has all attributes of the head of two semantic classes. This step is repeated in every group until there are no relations among semantic classes or semantic class groups. Then, this step is applied with the groups because each group has the same function as a semantic class.

3.4. Phrasal semantic distance

The phrasal semantic distance of two phrases should be calculated with semantic classes and semantic relations of these phrases to indicate how different in meaning they are. The phrasal semantic distance is defined as follows:

Definition 1 The semantic distance of two semantic classes C_1 and C_2, denoted as dc, is the number of edges to travel from one semantic class to the other in the lexicon ontology. This is the edge-based distance described in [22, 23].

Example 1 Assuming that there is a part of a lexicon ontology sample, as shown in Figure 1, in which semantic class "*cls_ con*" is the direct hypernym of semantic class "*cls_ gà*" and semantic class "*cls_ sói*". The semantic distance of "*cls_ gà*" and "*cls_ sói*" is *dc(cls_ gà, cls_ sói)*=2 because there are two edges between node "*cls_ gà*" and node "*cls_ sói*".

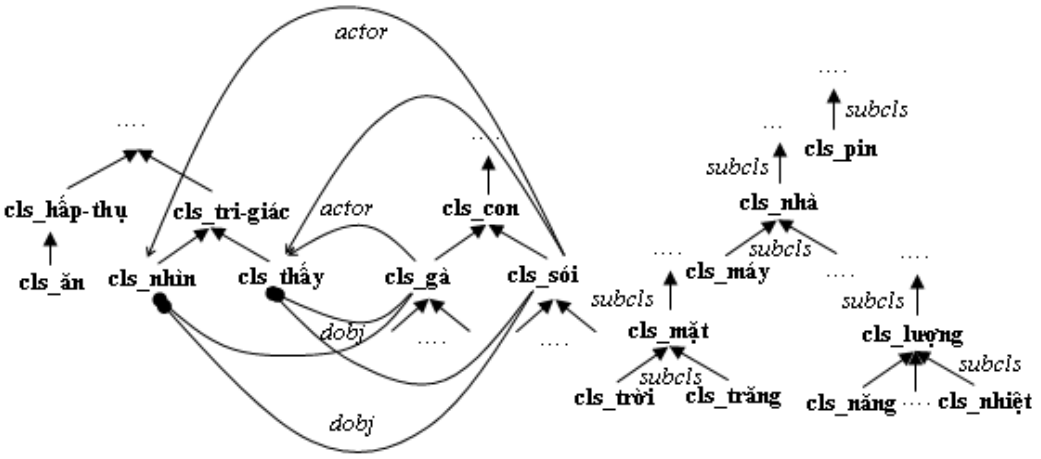

Figure 1: A part of a lexicon ontology sample

Definition 2 The semantic distance of two relations $R_1(C_1, C_2)$ and $R_2(C_3, C_4)$, denoted as d_r, is calculated according to the following formular

$$d_r = \begin{cases} d_c(C_1, C_3) + d_c(C_2, C_4), R_1 = R_2 \\ 2\omega + d_c(C_1, C_3) + d_c(C_2, C_4), R_1 \neq R_2 \end{cases}$$

ω is an integer constant which is larger than the maximum semantic distance of two arbitrary semantic classes in the lexicon ontology.

Example 2 Assuming that there is ontology as Example 1. The semantic distance of the two relations "*dobj(cls_gà, cls_thấy)*" and "*dobj(cls_sói, cls_thấy)*" is:

$$d_r = d_c(cls\_gà, cls\_sói) + d_c(cls\_thấy, cls\_thấy) = 2 + 0 = 2$$

and the semantic distance of the two relations "*dobj(cls_gà, cls_thấy)*" and "*dobj(cls_sói, cls_thấy)*" is

$$d_r = 2\omega + d_c(cls\_gà, cls\_sói) + d_c(cls\_thấy, cls\_thấy) = 2\omega + 2 + 0 = 2\omega$$

because relations are "*obj()*" and "*sub()*" which are not the same.

Definition 3 The semantic distance of a semantic class $\boldsymbol{c}$ and a set of semantic classes $\boldsymbol{C}$, denoted as $d_{cC}(c, C)$, is the minimum semantic distance between $\boldsymbol{c}$ and every semantic class in $\boldsymbol{C}$.

$$d_{cC}(c, C) = \min(d_c(c, i)), i \in C$$

Example 3 Assuming there is a part of lexicon ontology as example 1. According to the definition 3, the semantic distance between a semantic class "*cls_gà*" and the set of semantic class *–cls_sói, cls_thấy"* is d_{cC}=d_c(*cls_gà, cls_sói*)=2 and the semantic distance between a semantic class "*cls_gà*" and the set of semantic classes *–cls_sói, cls_thấy, cls_gà"* is $d_{cC} = d_c$(*cls_gà, cls_gà*) = 0.

Definition 4 The semantic distance of a semantic relation $\boldsymbol{r}$ and a set of semantic relations $\boldsymbol{R}$, denoted as $d_{rR}(r, R)$, is the minimum semantic distance between $\boldsymbol{r}$ and every semantic relation in $\boldsymbol{R}$.

$$d_{r,R}(r, R) = \min\left(d_r, (r, i)\right),\ i \in R$$

Example 4 Assuming there is ontology as example 1. The semantic distance between a semantic relation "*actor(cls_ gà, cls_ thấy)*" and the set of semantic relations –*dobj(cls_ gà, cls_ thấy), actor(cls_ sói, cls_ nhìn)*" is

$$d_{cR} = d_r(actor(cls\_\ gà,\ cls\_\ thấy),\ actor(cls\_\ sói,\ cls\_\ nhìn)) = 2 + 2 = 4.$$

because $d_r(actor(cls\_\ gà,\ cls\_\ nhìn), dobj(cls\_\ gà,\ cls\_\ nhìn)) = 2\omega > 4$ according to the Definition 2.

Definition 5 The semantic distance of two sets of semantic classes $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted as $d_{CC}(A, B)$, is identified as the following formula:

$$d_{cc} = (A, B) = \max\left(\sum_{i \in A} d_{cC}(i, B), \sum_{j \in B} d_{cC}(j, A)\right).$$

Example 5 Assuming there is ontology as Example 1. The semantic distance between a set of semantic class $\boldsymbol{A}$, which is –*cls_ gà, cls_ thấy*", and a set of semantic class $\boldsymbol{B}$, which is –*cls_ sói, cls_ nhìn, cls_ gà*", is

$$\begin{aligned} d_{CC}(A, B) =& \max(d_{cC}(cls\_\ gà, B) + d_{cC}(cls\_\ thấy, B), \\ & d_{cC}(cls\_\ sói, A) + d_{cC}(cls\_\ nhìn, A) + d_{cC}(cls\_\ gà, A)) \\ =& \max(0 + 2, 2 + 2 + 0) = 4 \end{aligned}$$

Definition 6 The semantic distance of two sets of semantic relations $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted as $d_{RR}(A, B)$, is identified as the following formula:

$$d_{RR}(A, B) = \max \sum_{i \in A} d_{r,R}(i, B), \sum_{j \in B} d_{r,R}(j, A)\cdot$$

Example 6 Assuming there is ontology as example 1. The semantic distance between a set of semantic relations A, which is –*actor(cls_ gà, cls_ thấy)*", and a set of semantic class B, which is –*dobj(cls_ sói, cls_ nhìn), actor(cls_ gà, cls_ nhìn)*", is

$$\begin{aligned} d_{RR}(A, B) =& \max(d_{rR}(actor(cls\_\ gà,\ cls\_\ thấy), B), \\ & d_{rR}(dobj(cls\_\ sói,\ cls\_\ nhìn), A) + d_{rR}(actor(cls\_\ gà,\ cls\_\ nhìn), A)) \\ =& \max(2, (2\omega + 4) + 2) = 2\omega + 6 \end{aligned}$$

Definition 7 Given two phrases P_1 and P_2, C_1 and C_2 are their sets of semantic classes and R_1 and R_2 are their sets of relations respectively. The semantic distance of the two phrases, denoted as $d_{sem}(P_1, P_2)$, is identified as the following formula:

$$d_{sem}(P_1, P_2) = d_{CC}(C_1, C_2) + d_{RR}(R_1, R_2)$$

Example 7 Assuming there is ontology as Example 1. Given two phrases P_1 "*gà thấy sói*" and P_2 "*sói nhìn gà*" the sets of semantic classes identified from P_1 and P_2 are respectively $C_1 = \{cls\_gà, cls\_thấy, cls\_sói\}$ and $C_2 = \{cls\_sói, cls\_nhìn, cls\_gà\}$. The sets of semantic relations identified from P_1 and P_2 are respectively R_1=–*actor(cls_ gà, cls_ thấy), dobj(cls_ sói, cls_ thấy)*" and R_2=–*actor(cls_ sói, cls_ nhìn), dobj(cls_ gà, cls_ nhìn)*". The semantic distance between P_1 and P_2 is calculated as follows:

$$\begin{aligned}
d_{CC}(C_1, C_2) =& \max(d_{cC}(cls_gà, C_2) + d_{cC}(cls_thấy, C_2) + d_{cC}(cls_sói, C_2),\\
& d_{cC}(cls_sói, C_1) + d_{cC}(cls_nhìn, C_1) + d_{cC}(cls_gà, C_1))\\
=& \max(0+2+0, 0+2+0) = 2.\\
d_{RR}(R_1, R_2) =& max(d_{rR}(actor(cls_gà,\ cls_thấy), R_2) + d_{rR}(dobj(cls_sói,\ cls_thấy), R_2),\\
& d_{rR}(actor(cls_sói,\ cls_nhìn), R_2) + d_{rR}(dobj(cls_gà,\ cls_nhìn), R_2))\\
=& \max(4+4, 4+4) = 8\\
d_{sem}(P_1, P_2) =& d_{CC}(C_1, C_2) + d_{RR}(R_1, R_2) = 10
\end{aligned}$$

Example 8 Assuming there is ontology as example 1. Given two phrases P_1 "*gà thấy sói*" and P_2 "*sói ăn gà*" the sets of semantic classes identified from P_1 and P_2 are respectively $C_1 = \{cls\_gà, cls\_thấy, cls\_sói\}$ and $C_2 = \{cls\_sói, cls\_ăn, cls\_gà\}$. The sets of semantic relations identified from P_1 and P_2 are respectively $R_1 = \{actor(cls\_gà,\ cls\_thấy), dobj(cls\_sói,\ cls\_thấy)\}$ and $R_2 = \{actor(cls\_sói,\ cls\_ăn), dobj(cls\_gà,\ cls\_ăn)\}$. The semantic distance between P_1 and P_2 is calculated as follows:

$$\begin{aligned}
d_{CC}(C_1, C_2) =& \max(d_{cC}(cls_gà, C_2) + d_{cC}(cls_thấy, C_2) + d_{cC}(cls_sói, C_2),\\
& d_{cC}(cls_sói, C_1) + d_{cC}(cls_ăn, C_1) + d_{cC}(cls_gà, C_1))\\
=& \max(0+4+0, 0+4+0) = 8.\\
d_{RR}(R_1, R_2) =& \max(d_{rR}(actor(cls_gà,\ cls_thấy), R_2) + d_{rR}(dobj(cls_sói,\ cls_thấy), R_2),\\
& d_{rR}(actor(cls_sói,\ cls_ăn), R_2) + d_{rR}(dobj(cls_gà,\ cls_ăn), R_2))\\
=& \max(6+6, 6+6) = 12\\
d_{sem}(P_1, P_2) =& d_{CC}(C_1, C_2) + d_{RR}(R_1, R_2) = 20.
\end{aligned}$$

The two Examples 7 and 8 show that the phrase "*gà thấy sói*" is more similar to the phrase "*sói nhìn gà*" than to the phrase "*sói ăn gà*" according to the lexicon ontology.

According to the above definitions, the semantic distance between two phrases is calculated as following:

- Step 1: Identify the set of semantic classes of each phrase by using semantic class tagger described in Section 3.2.
- Step 2: Identify the set of semantic relations of each semantic class set identified at step 1 by using the method described in section 3.3.
- Step 3: Calculate the semantic distance between two sets of semantic classes d_{CC}.
- Step 4: Calculate the semantic distance between two sets of semantic relations d_{RR}.
- Step 5: The sum of d_{CC} and d_{RR} is the semantic distance d_{sem} between two phrases.

4. EXPERIMENT

4.1. Setting up the experimental system

For the purpose of experiment, a semantic phrase retrieval system has been set up according to [24] with the model shown in Figure 2. The experimental system is composed of four components described as following:

- **Semantic class identifier**. This component receives a phrase from user as query or from the collection of testing phrases. Then, the component uses the semantic tagger model to identify the semantic class of each word in the phrase like a semantic class tagger.
- **Semantic relation identifier**. This component receives a set of semantic classes corresponding to a phrase, and uses Vietnamese Lexicon Ontology (VLO) [17] to identify the relations among semantic classes. This component returns a set of semantic classes and a set of relations of semantic classes. In case the user's query contains only one word, the set of relations is null. The output of this component can be used to index if the phrase is in the collection of phrases or to search if it is the user's query.
- **Index component**. As in [14], this component indexes the representation of the phrase in semantic class and semantic relation to the Semantic Class Index (SCI) and Semantic Relation Index (SRI) respective. The component is developed from [18] by adding SCI and replace the head-dependent relation by six types of relation defined in [17]. This component is used to improve the performance of the search process. Therefore, before indexing, the semantic classes are expanded up a number n of level in SCI and SRI for searching hyponyms. In the experiment, n is set to 1.

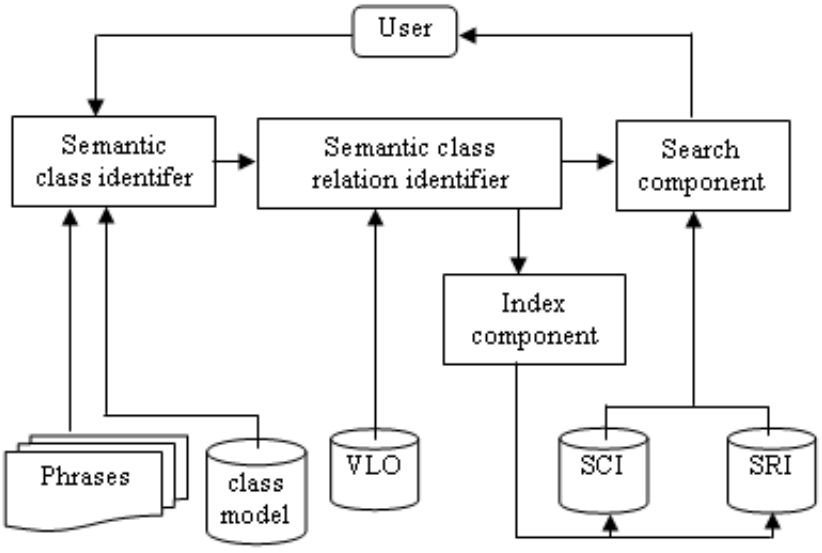

Figure 2: The model of semantic phrase based retrieval system

- **Search component**. This component receives the user query in the representation of semantic class and of semantic relation. The representation is used to identify which phrases

contain semantic classes and semantic relations of the user query by using two indices SCI and SRI. Then it calculates the semantic distances of these phrases and the query based on the definition of phrasal semantic distance. Finally, it sorts the phrases by descending semantic distances and returns the list of phrases to the user.

4.2. Evaluation

The testing data is composed of a collection of phrases C, a set of queries Q and sets of expected phrases R, which should be returned according to each query in Q. C consists of 720 phrases or sentences randomly chosen in many news web pages. Some phrases in C could be meaningless to check if the system retrieves phrases according to their semantic. After choosing C, the set of queries is made by selecting 30 queries according to these criteria:

- The query should contain verb frame to check if the semantic roles of the subject and the object of the verb are ensured.
- The query should only contain noun phrases to check if the heads of phrases are identified correctly.
- The query should contain words which can be either verbs or nouns, such as "*thiết kế*" ("*to design*" or "*design*").
- The query should contain polysemy words. For example, the word "*nhà*" can be used to refer the person, such as "*nhà khoa học*" ("*the scientist*"), or it can be used to refer the constructions, such as "*nhà máy điện*".
- The query should contain Hanji-Vietnamese words to check if the system can retrieves the similar meanings according to the structure of these words. For example, the word "*giáo sư*", which is consider as a phrase, is composed of two words "*giáo*" ("*profess*") and "*sư*" (suffix "*-or*"). Therefore, the similar meaning of the word "*giáo sư*" might be "*vũ sư*" ("*dancing teacher*"), or "*kỹ sư*" ("*engineer*").
- The query should contain some more words but its meaning is similar to a certain query in Q to check if the system can retrieve similar results by the similar queries.

With each query in Q, 720 phrases or sentences in C are checked to find the appropriate ones. The appropriate one should contain words which are synonyms or have the similar meaning to the words appeared in the query.

The experiment E_1 has been conducted with the above data set. The results [14], shown as in the Table 1, show the precision of 81.5%, the recall of 88.6% and the F-measure of 84.9%.

In order to evaluate the proposed method, the two experiments are also conducted. There are one experiment E_2 for keyword based information retrieval method without word segmentation and one experiment E_3 for the method with word segmentation.

In the experiment E_2 on keyword based information retrieval without word segmentation, the results, shown as in the Table 2, show the precision of 57.9%, the recall of 89.6% and the F-measure of 70.4%. In this method, the precision is low because there is no semantic feature analyzed. This method cannot retrieve the synonyms, distinguish the different meanings of polysemy words and recognize the compound words which can be used as concepts. However, the recall of this method is quite high because it retrieves all phrases containing keywords in each query.

Query in Vietnamese	**Query in English**	**Returned** ($\mathbf{n_1}$)	**Correct** ($\mathbf{n_2}$)	**Expected** ($\mathbf{n_0}$)	**P** ($\mathbf{n_2/n_1}$)	**R** ($\mathbf{n_2/n_0}$)
Nhà khoa học	Scientist	26	24	24	0.923	1.000
Giáo sư	Professor	4	4	6	1.000	0.667
Nhà máy điện	Electricity plant	64	63	77	0.984	0.818
Pin mặt trời	Solar battery	6	6	21	1.000	0.286
Năng lượng mặt trời	Solar energy	58	50	67	0.862	0.746
Máy bay nhẹ	Lightweight helicopter	25	24	24	0.960	1.000
Nhà máy điện hạt nhân	Nuclear electricity plant	77	68	80	0.883	0.850
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	1.000	1.000
Công ty thiết kế	Designer company	17	17	28	1.000	0.607
Khả năng bám cua	Road pulling ability	43	19	19	0.442	1.000
Phương pháp cấp điện mới	Supplying electricity method	58	51	60	0.879	0.850
Nhà nghiên cứu	Researcher	23	20	20	0.870	1.000
Phòng thí nghiệm	Laboratory	16	9	12	0.562	0.750
Phát triển công nghệ	To develop the technology	34	33	33	0.971	1.000
Tấm dán tường phát sáng	Lighting panel	17	15	22	0.882	0.682
Điều khiển xe lăn	Control the wheelchair	58	49	49	0.845	1.000
Chuyên gia	Professional	9	9	9	1.000	1.000
Giúp chuyên gia biết nguy hiểm	To help the professional to realize a danger	35	29	30	0.829	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	1.000	0.957
Xe có hệ thống bánh lái	Car with helm system	113	84	89	0.743	0.944
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	0.500	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	1.000
Xử lý chất thải	Waste substance processing	22	2	2	0.090	1.000
Hội nghị về khí hậu	Conference on climate	93	15	15	0.161	1.000
Xây dựng hệ thống định hướng	To build a navigating system	57	56	59	0.982	0.949
Tàu chạy bằng năng lượng mặt trời	Ship using solar energy	91	84	90	0.923	0.933
	Total				**0.815**	**0.886**

Table 1: E_1, the testing results of semantic phrase retrieval system using phrasal semantic distance (Source: [14])

Query in Vietnamese	**Query in English**	**Returned** ($\mathbf{n_1}$)	**Correct** ($\mathbf{n_2}$)	**Expected** ($\mathbf{n_0}$)	**P** ($\mathbf{n_2/n_1}$)	**R** ($\mathbf{n_2/n_0}$)
Nhà khoa học	Scientist	63	24	24	0.381	1.000
Giáo sư	Professor	8	6	6	0.750	1.000
Nhà máy điện	Electricity plant	94	61	77	0.649	0.792
Pin mặt trời	Solar battery	33	21	21	0.636	1.000
Năng lượng mặt trời	Solar energy	73	53	67	0.726	0.791
Máy bay nhẹ	Lightweight helicopter	63	24	24	0.381	1.000
Nhà máy điện hạt nhân	Nuclear electricity plant	104	65	80	0.625	0.813
Khảo sát địa điểm xây dựng nhà máy	To survey locations for building a plant	95	58	88	0.611	0.659
Xe tay ga	Scooter motocycle	35	28	41	0.800	0.683
Công ty thiết kế	Designer company	73	28	28	0.384	1.000
Khả năng bám cua	Road pulling ability	53	19	19	0.358	1.000
Phương pháp cấp điện mới	Supplying electricity method	42	30	60	0.714	0.500
Nhà nghiên cứu	Researcher	58	20	20	0.345	1.000
Phòng thí nghiệm	Laboratory	15	12	12	0.800	1.000
Phát triển công nghệ	To develop the technology	75	31	33	0.413	0.939
Tấm dán tường phát sáng	Lighting panel	47	22	22	0.468	1.000
Điều khiển xe lăn	Control the wheelchair	43	38	49	0.884	0.776
Chuyên gia	Professional	24	9	9	0.375	1.000
Giúp chuyên gia biết nguy hiểm	To help the professional to realize a danger	46	28	30	0.609	0.933
Sử dụng năng lượng mặt trời	To use the solar energy	93	67	81	0.720	0.827
Thiết kế nhỏ gọn	Small and tidy designer	31	22	23	0.710	0.957
Xe có hệ thống bánh lái	Car with helm system	123	77	89	0.626	0.865
Nguồn tạo khí carbon	Carbonic creating source	40	33	34	0.825	0.971
Dùng lưỡi điều khiển	To use tongue to control	29	20	21	0.690	0.952
Phương pháp giữ phản vật chất	Reserving antimatter method	41	20	26	0.488	0.769
Máy bay siêu nhẹ	Super lightweight plane	66	24	24	0.364	1.000
Xử lý chất thải	Waste substance processing	32	2	2	0.063	1.000
Hội nghị về khí hậu	Conference on climate	30	15	15	0.500	1.000
Xây dựng hệ hệ thống định hướng	To build a navigating system	73	59	59	0.808	1.000
Tàu chạy bằng năng lượng mặt trời	Ship using solar energy	93	63	90	0.677	0.700
	Total				**0.579**	**0.896**

Table 2: E_2, The testing results of on keyword based information retrieval without word segmentation

Query in Vietnamese	**Query in English**	**Returned** ($\mathbf{n_1}$)	**Correct** ($\mathbf{n_2}$)	**Expected** ($\mathbf{n_0}$)	**P** ($\mathbf{n_2/n_1}$)	**R** ($\mathbf{n_2/n_0}$)
Nhà khoa học	Scientist	9	9	24	1.000	0.375
Giáo sư	Professor	2	2	6	1.000	0.333
Nhà máy điện	Electricity plant	12	11	77	0.917	0.143
Pin mặt trời	Solar battery	25	21	21	0.840	1.000
Năng lượng mặt trời	Solar energy	46	45	67	0.978	0.672
Máy bay nhẹ	Lightweight helicopter	10	10	24	1.000	0.417
Nhà máy điện hạt nhân	Nuclear electricity plant	15	14	80	0.933	0.175
Khảo sát địa điểm xây dựng nhà máy	To survey locations for building a plant	46	46	88	1.000	0.523
Xe tay ga	Scooter motocycle	1	1	41	1.000	0.024
Công ty thiết kế	Designer company	28	28	28	1.000	1.000
Khả năng bám cua	road pulling ability	18	18	19	1.000	0.947
Phương pháp cấp điện mới	Supplying electricity method	17	12	60	0.706	0.200
Nhà nghiên cứu	Researcher	5	5	20	1.000	0.250
Phòng thí nghiệm	Laboratory	1	1	12	1.000	0.083
Phát triển công nghệ	To develop the technology	30	30	33	1.000	0.909
Tấm dán tường phát sáng	Lighting panel	25	18	22	0.720	0.818
Điều khiển xe lăn	Control the wheelchair	11	11	49	1.000	0.224
Chuyên gia	Professional	9	9	9	1.000	1.000
Giúp chuyên gia biết nguy hiểm	To help the professional to realize a danger	23	23	30	1.000	0.767
Sử dụng năng lượng mặt trời	To use the solar energy	60	56	81	0.933	0.691
Thiết kế nhỏ gọn	Small and tidy designer	22	22	23	1.000	0.957
Xe có hệ thống bánh lái	Car with helm system	73	61	89	0.836	0.685
Nguồn tạo khí carbon	Carbonic creating source	21	20	34	0.952	0.588
Dùng lưỡi điều khiển	To use tongue to control	21	19	21	0.905	0.905
Phương pháp giữ phản vật chất	Reserving antimatter method	21	16	26	0.762	0.615
Máy bay siêu nhẹ	Super lightweight plane	10	10	24	1.000	0.417
Xử lý chất thải	Waste substance processing	1	1	2	1.000	0.500
Hội nghị về khí hậu	Conference on climate	15	13	15	0.867	0.867
Xây dựng hệ thống định hướng	To build a navigating system	46	45	59	0.978	0.763
Tàu chạy bằng năng lượng mặt trời	Ship using solar energy	63	55	90	0.873	0.611
	Total				**0.940**	**0.582**

Table 3: E_3, The testing results of on keyword based information retrieval with word segmentation

In the experiment E_3 on keyword based information retrieval with word segmentation, the word segmentation tool used in this experiment is vnTokenizer [25]. This tool can identify compound words, person names and locations. The results of the experiment, shown as in the Table 3, show the precision of 94.0%, the recall of 58.2% and the F-measure of 71.9%. In this method, the precision is very high because the process of word segmentation in Vietnamese can be used as the concept identifying process. Therefore, the retrieval process works with concepts instead of morphemes. However, the recall is low because the synonyms, the hyponyms and the hypernyms are not analyzed in retrieval process. Therefore, the F-measure of E_3 is approximately the same as of E_2 (71.9% compares to 70.4%).

The results of the two experiments show that the proposed method is better than the keyword based information retrieval method with or without word segmentation.

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	18	18	24	1.000	0.750
Giáo sư	Professor	4	4	6	1.000	0.667
Nhà máy điện	Electricity plant	64	63	77	0.984	0.818
Pin mặt trời	Solar battery	6	6	21	1.000	0.286
Năng lượng mặt trời	Solar energy	50	44	67	0.880	0.657
Máy bay nhẹ	Lightweight helicopter	24	24	24	1.000	1.000
Nhà máy điện hạt nhân	Nuclear electricity plant	57	57	80	1.000	0.713
Khảo sát địa điểm xây dựng nhà máy	To survey locations for building a plant	52	52	88	1.000	0.591
Xe tay ga	Scooter motocycle	28	28	41	1.000	0.683
Công ty thiết kế	Designer company	17	17	28	1.000	0.607
Khả năng bám cua	road pulling ability	22	19	19	0.864	1.000
Phương pháp cấp điện mới	Supplying electricity method	28	27	60	0.964	0.450
Nhà nghiên cứu	Researcher	22	20	20	0.909	1.000
Phòng thí nghiệm	Laboratory	9	9	12	1.000	0.75
Phát triển công nghệ	To develop the technology	32	31	33	0.969	0.939
Tấm dán tường phát sáng	Lighting panel	16	16	22	0.938	0.682
Điều khiển xe lăn	Control the wheelchair	36	36	49	1.000	0.735
Chuyên gia	Professional	9	9	9	1.000	1.000
Giúp chuyên gia biết nguy hiểm	To help the professional to realize a danger	28	28	30	1.000	0.933
Sử dụng năng lượng mặt trời	To use the solar energy	65	62	81	0.954	0.765
Thiết kế nhỏ gọn	Small and tidy designer	22	22	23	1.000	0.957
Xe có hệ thống bánh lái	Car with helm system	85	72	89	0.847	0.809
Nguồn tạo khí carbon	Carbonic creating source	35	33	34	0.943	0.971
Dùng lưỡi điều khiển	To use tongue to control	20	20	21	1.000	0.952
Phương pháp giữ phản vật chất	Reserving antimatter method	23	19	26	0.826	0.731
Máy bay siêu nhẹ	Super lightweight plane	24	24	24	1.000	1.000
Xử lý chất thải	Waste substance processing	1	1	2	1.000	0.500
Hội nghị về khí hậu	Conference on climate	15	15	15	1.000	1.000
Xây dựng hệ thống định hướng	To build a navigating system	56	56	59	1.000	0.949
Tàu chạy bằng năng lượng mặt trời	Ship using solar energy Ship using solar energy	62	55	90	0.887	0.611
	Total				**0.966**	**0.784**

Table 4: The testing results of semantic phrase retrieval system using phrasal semantic distance after modifying VLO (Source: [14])

Method	**P(%)**	**R(%)**	**F-measure(%)**
Keyword-based information retrieval without word segmentation	57.9	89.6	70.4
Keyword-based information retrieval with word segmentation	94.0	58.2	71.9
Information retrieval using phrasal semantic distance	81.5	88.6	84.9
Information retrieval using phrasal semantic distance (after error analyzing)	96.6	78.4	86.5

Table 5: The experiment results of all methods

4.3. Error analysis

According to [14], although the precision and the recall of our method are high, there are some test cases in which the precision are very low. For example, the result of the query phrase *"hội nghị về khí hậu"* has the precision of 0.161, or the result of the query phase *"xử lý chất thải"* has the precision of 0.090, etc. In order to overcome these cases, the Vietnamese Lexicon Ontology (VLO) has been checked to find out the reasons which cause the low precision. There are two reasons [14]

- The relations of the words are sometimes very close. For example, the word "*chất thải*", which is a phrase in our point of view, contains the word "*chất*". In VLO, the word "*chất*" is the hypernym of many words which designate the material. Therefore, there are many results which are very weakly similar to the query.

- There are some lacks of meaning of polysemous words. For example, the word "*có*" has two meanings which correspond to ordinary verb, to have, and auxiliary verb, to do. In the sentence "*Nó có căn nhà*" ("*He has a house*"), the word "*có*" is an ordinary verb. In the sentence "*Nó có ăn bánh*" ("*He does eat the cake*"), the word "*có*" is an auxiliary verb.

In order to overcome the error, the relations of words have been corrected and the lacking meanings of polysemy words have been added in the VLO. After that, the experiment is conducted on the proposed method again. The results [14], shown in Table 4, show the precision of 96.6%, the recall of 78.4% and the F-measure of 86.5%. The results of all experiments are shown in Table 5 for comparison.

5. CONCLUSIONS AND FUTURE WORKS

This paper introduces a novel computational semantic method proposed to estimate the phrasal semantic distance between two Vietnamese phrases. The phrasal semantic distance is composed of two factors: the semantic class distance and the semantic relation distance. The semantic classes are defined to solve some problems of Vietnamese lexicons. Originated from "semantic memory concept" [7], a semantic class is a specific meaning of a word in a specific context. When using semantic classes instead of POS tag, the meaning of every word is explicitly identified. Therefore, the problem of polysemy can be solved. In addition, the semantic relations are also used to model the meaning of the combination of words in a phrase.

The semantic relations and semantic classes are organized in lexicon ontology. In this ontology, the relations of compound words are also included. Therefore, this ontology can be used instead of grammatical rules to identify the head of a phrase in dependency parsing.

After proposing the phrasal semantic distance, an experimental system has been set up and tested with testing data of 720 phrases and 30 queries. These testing data are manually created by identifying the meaning of each query and each phrase in synonym, hyponym and polysemy. The evaluation shows the performance with the precision of 96.6%, the recall of 78.4% and F-measure of 86.5%.

The application of the phrasal semantic distance in textual document retrieval system is simply to split documents into phrases and mark each phrase with the document containing it. Then, the results of phrase retrieval will be processed to get the document list.

However, in the future, the system should be evaluated with big test sets to consider possible disadvantages of our phrasal semantic distance and proposed method of estimation of phrasal similarity in order to build practical semantic document retrieval applications for Vietnamese language.

REFERENCES

[1] U. Shah, T. Finin, A. Joshi, R. S. Cost, and J. Matfield, "Information retrieval on the semantic web," in *Proceedings of the eleventh international conference on Information and knowledge management.* ACM, 2002, pp. 461–468.

[2] T. Finin, J. Mayfield, A. Joshi, R. S. Cost, and C. Fink, "Information retrieval and the semantic web," in *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on.* IEEE, 2005, p. 113a.

[3] C. D. Manning, P. Raghavan, H. Schütze *et al.*, *Introduction to information retrieval.* Cambridge university press Cambridge, 2008, vol. 1.

[4] M. Fernández Sánchez, "Semantically en enhanced information retrieval: an ontology-based aprroach," Ph.D. dissertation, Unitversidad de Autónoma, Madrid, 2009.

[5] J. Szymański and W. Duch, "Information retrieval with semantic memory model," *Cognitive Systems Research*, vol. 14, no. 1, pp. 84–100, 2012.

[6] S. L. Tomassen and D. Strasunskas, "Measuring intrinsic quality of semantic search based on feature vectors," *International Journal of Metadata, Semantics and Ontologies*, vol. 5, no. 2, pp. 120–133, 2010.

[7] T. T. Rogers, M. A. Lambon Ralph, P. Garrard, S. Bozeat, J. L. McClelland, J. R. Hodges, and K. Patterson, "Structure and deterioration of semantic memory: a neuropsychological and computational investigation." *Psychological review*, vol. 111, no. 1, pp. 205–235, 2004.

[8] T. C. Rindflesch and A. R. Aronson, "Semantic processing in information retrieval." in *Proceedings of the Annual Symposium on Computer Application in Medical Care.* American Medical Informatics Association, 1993, pp. 611–615.

[9] A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff, "Semantic annotation, indexing, and retrieval," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 2, no. 1, pp. 49–79, 2004.

[10] J. Gonzalo, I. Chugur, and F. Verdejo, "Sense clusters for information retrieval: evidence from semcor and the eurowordnet interlingual index," in *Proceedings of the ACL-2000 workshop on Word senses and multi-linguality.* Association for Computational Linguistics, 2000, pp. 10–18.

[11] O. Egozi, S. Markovitch, and E. Gabrilovich, "Concept-based information retrieval using explicit semantic analysis," *ACM Transactions on Information Systems (TOIS)*, vol. 29, no. 2, pp. 1–34, 2011.

[12] F. Giunchiglia, U. Kharkevich, and I. Zaihrayeu, "Concept search: Semantics enabled syntactic search," in *SemSearch 2008, CEUR Workshop Proceedings*, 2008, pp. 1–15.

[13] R. Steinberger, B. Pouliquen, and J. Hagman, "Cross-lingual document similarity calculation using the multilingual thesaurus eurovoc," in *Computational Linguistics and Intelligent Text Processing.* Springer, 2002, pp. 415–424.

[14] T. T. T. Do and D. T. Nguyen, "Mô hình tìm kiếm văn bản tiếng việt dựa trên ngữ nghĩa cụm từ truy vấn [Phrasal query based semantic information retrieval model for Vietnamese texts]," University of Information Technology, VNU-HCM, Scientific Report, VNU-HCM Research Project C2013-26-06, 8 April 2015.

[15] H. X. Cao, *Tiếng Việt mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa*, 3rd ed. Vietnam Education Publishing House, 2007 (in Vietnamese).

[16] T. T. T. Do, "A concept identification method for vietnamese concept-based information retrieval system," in *Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services*. ACM, 2012, pp. 403–406.

[17] ——, "Building a vietnamese lexicon ontology for syntactic parsing and document annotation," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. ACM, 2013, p. 619.

[18] ——, "Ontology-based annotation and indexing for vietnamese text document," in *Proceedings of International Conference on Information Integration and Web-based Applications & Services*. ACM, 2013, p. 363.

[19] M. Collins, "Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2002, pp. 1–8.

[20] A. Ratnaparkhi *et al.*, "A maximum entropy model for part-of-speech tagging," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, vol. 1. Philadelphia, USA, 1996, pp. 133–142.

[21] K. Oflazer, "Dependency parsing with an extended finite-state approach," *Computational Linguistics*, vol. 29, no. 4, pp. 515–544, 2003.

[22] G. Varelas, E. Voutsakis, P. Raftopoulou, E. G. Petrakis, and E. E. Milios, "Semantic similarity methods in wordnet and their application to information retrieval on the web," in *Proceedings of the 7th annual ACM international workshop on Web information and data management*. ACM, 2005, pp. 10–16.

[23] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," in *Proc. of ROCLING*, 1997, pp. 19–33.

[24] T. T. T. Do and D. T. Nguyen, "A framework for vietnamese text document retrieval system based on phrasal semantic analysis," *International Journal of Simulation Systems, Science & Technology*, vol. 15, no. 4, p. 52, 2014.

[25] P. H. Le, "vnTokenizer –Vietnamese word segmentation, (Accessed on 20th February, 2015)," Tech. Rep. [Online]. Available: http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer

Received on February 26 - 2015
Revised on August 30 - 2015