

PROCESSING VIETNAMESE NEWS TITLES TO ANSWER RELATIVE QUESTIONS IN VNEWSQA/ICT SYSTEM

Son The Pham and Dang Tuan Nguyen

Faculty of Computer Science, University of Information Technology, Vietnam National
University – Ho Chi Minh City

ABSTRACT

This paper introduces two important elements of our VNewsQA/ICT system: its semantic models of simple Vietnamese sentences and its semantic processing mechanism. The VNewsQA/ICT is a Vietnamese based Question Answering system which has the ability to gather information from some Vietnamese news title forms on the ICTnews websites (<http://www.ictnews.vn>), instead of using a database or a knowledge base, to answer the related Vietnamese questions in the domain of information and communications technology.

KEYWORDS

Question Answering, semantic representation, semantic processing, Vietnamese language processing.

1. INTRODUCTION

The VNewsQA/ICT is a Vietnamese based Question Answering system that we built to experiment and develop a system model which has the ability to answer simple Vietnamese questions on the domain of information and communications technology. To answer the related questions, the system retrieves the textual information which is gathered from several Vietnamese news titles on the ICTnews websites [1] to process and put them into system's database of facts.

In this research, we focus on introducing two main issues: 1) semantic models, and 2) semantic processing mechanism of VNewsQA/ICT system. These are two elements that help the system analyse the semantic of Vietnamese news titles which have the simple sentence structure. We also proposed a new method to represent the meaning about time, place, property, semantic relations ... in simple Vietnamese sentences based on the functions and "connection rules" that we defined for this system.

The information about the system model, architecture and features of the other components of the VNewsQA/ICT system won't be introduced in this paper.

2. THE SEMANTIC REPRESENTATION OF SIMPLE VIETNAMESE SENTENCES IN VNEWSQA/ICT SYSTEM

In this section, we propose a semantic representation method, based on the methods of computational semantics [2], [3], [4], [5], [6], [7], [8], [9], [10] and linguistic theories [11], for representing the content of simple Vietnamese sentences, which have one or two verbs, in VNewsQA/ICT system.

2.1. Definition of functions and connections

In VNewsQA/ICT system, we define some functions and “connection rules” to use in the semantic representation model.

2.1.1. Functions

The defined functions are used to represent the time, space ... of the sentence, or some relationships between two objects. The representation form of the functions as follows:

function_name(argument) function_name (argument _1, argument _2 ..., argument _n)
--

The arguments of the semantic functions can be preposition phrase, adjective phrase, or adverb phrase. The functions can take one or more arguments depending on the purpose of representation.

To represent the relationship between two objects, the representation form of the functions is as follows:

Function_name_1(Function_name_1(argument))
--

We also built some basic functions as follows:

- The function of time: Time(AdvP_time).
- The function of location: Location(PreP)
- The function of manner: Manner(AdvP).
- The functions representing the thing, matter, object: Object(NP) or Object(QuaP). These functions are used for the sentences which don't have verb.
- The function representing the possessive relation between two objects: Possessive(NP, PreP_Poss) or Possessive(QuaP, PreP_Poss).

In VNewsQA/ICT system, we distinguish the predicates from functions as follows:

- The predicates represent the meaning of a sentence: they will take the verb names to set the predicate names in the semantic representation model. The functions represent the functions of the phrase: the function names depend on the meaning of phrase. The predicates are a special case of the functions when the function names are coincided with the verb in the sentence.
- The arguments of the predicates have to be noun phrases or quantity phrases. Conversely, the arguments of the function are adverb phrases of time, preposition phrases of location, or defined symbols to represent the relations between the predicates and functions.

2.1.2. Connections

- The “connection rules” represent the relations between the functions and predicates, or between two phrases, or between phrases and predicates (this is the case of the sentences having two main verbs). We define and sign “connection rules” as follows:
- The connection rule $>->$ describes the modification of the locative preposition for the predicate in the semantic representation of the sentence.
- The connection rule $>-->$ describes the modification of the qualificative adverb for the predicate in the semantic representation of the sentence.
- The connection rule $>--->$ describes the modification of the adverb of time for the predicate in the semantic representation of the sentence.
- The connection rule $<->$ represents the relation between two verbs in the sentence having several verbs.

2.1.3. The priority of the connection rules

After define the connection rules, we define “the priority of the connection rules”. If certain connection rules have the same priority, the system will process them from left to right.

- The priorities of the connection rules $<->$, $>->$, $>-->$, $>--->$ are in descending order from left to right as follows:

$<->$, $>->$, $>-->$, $>--->$

The priority of the connection rules which we introduced above is defined under the form of “infix, prefix, and postfix operators” in Prolog [9].

2.1.4. Semantic representation of phrases

The phrases which are noun phrase (NP), preposition phrase (PreP), adjective phrase (AdjP), adverb phrase indicating the time (AdvP_time) are considered as “phrase constant”. They are used as the arguments of the predicates and functions in the semantic representation of the sentence.

The quantity phrase QuaP represents the quantity information about things, facts, and objects. They are split into two types:

- The definite quantity predicates: if a quantity predicate precedes the noun, it will quantify the noun. To represent the definite quantify predicates, we define functions as follows:

Definite(QuaP)

In which, QuaP is quantity phrase and Definite is function name.

- The indefinite quantity predicates: a quantity predicate precedes the noun and represents the indefinite number of the noun. To represent the indefinite quantify predicates, we define functions as follows:

Indefinite(QuaP)

In which, QuaP is quantity phrase and Indefinite is function name.

2.2. Semantic models of VNewsQA/ICT system

2.2.1. Sentences having one verbs

Vietnamese sentences having one verb are performed by VNewsQA/ICT system follow the model in Figure 1.

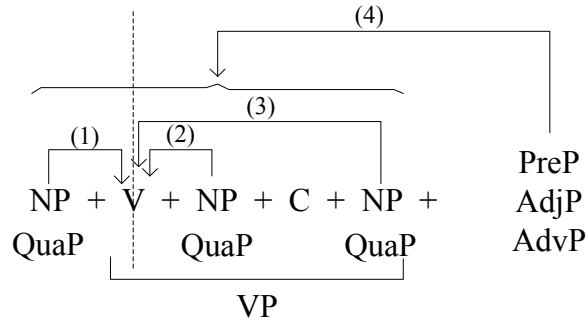


Figure 1: Sentences having one verb

The verb in Figure 1 is represented by a predicate as follows:

```
verb_predicate(argument 1, argument 2)
```

In which, the relation between “argument 1” and “argument 2” depends on the main verb of the sentence. The “argument 1” and “argument 2” takes the values of NPs or QuaPs.

The PreP, AdjP, AdvP_time phrases are circumstantial complements for the main verb by using “connection rule” (4) in Figure 1.

The representation of PreP, AdjP, and AdvP_time are defined as follows:

- PreP indicating the location is represented by the function Location(PreP) and “connection rule” >->.
- AdvP indicating the manner is represented by the function Manner(AdvP) and “connection rule” >->.
- AdvP indicating the time is represented by the function Time(AdvP) and “connection rule” >--->.

The priorities of the connection rules <->, >->, >->, >---> are in descending order from left to right.

Example 1: “*MobiFone mở văn phòng đại diện tại Myanmar.*” (ICTnews [1])

[English translation: “MobiFone opens the representative office in Myanmar.”]

- The noun phrases “MobiFone” and “văn phòng đại diện” are the argument of predicate “mở”.
- The location preposition “tại Myanmar” is the argument of function Location(PreP).
- The semantic representation of this sentence is as follows:
 $mở(<MobiFone >, <văn phòng đại diện >) >-> Location(< tại Myanmar >)$

2.2.2. Sentences having two verbs

In this sentence type, the VNewsQA/ICT system performs three cases:

- The sentence has two consecutive verbs;
- The sentence has two verbs which combine together by conjunction “và” (English translation: “and”);
- The sentence has two verbs and there is one phrase stands at the middle of them.

2.2.2.1. Sentence having two consecutive verbs

In this sentence type, the verb phrase of the sentence will be analyzed and represented as in Figure 2.

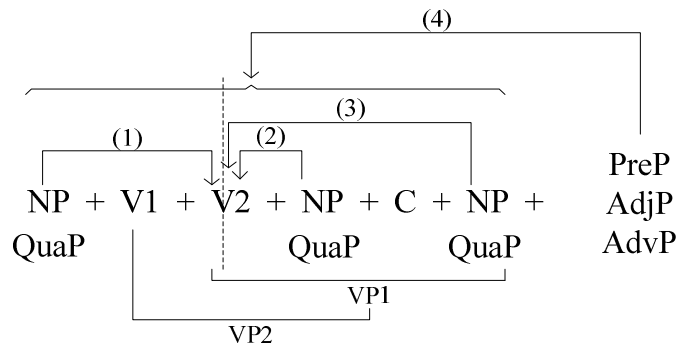


Figure 2: Sentence having two consecutive verbs

Two mandatory elements of this sentence type are VP1 and VP2. The structure of this sentence has the semantic representation form as follows:

verb_V1(argument_1, verb_V2(argument_1, argument_2))

In Figure 2, “argument_1” of “verb_V1” and “verb_V2” is a phrase preceding two verbs, but in the representation model we will set the value of the “argument_1” of the verb V2 is *same_Subject*, then the “argument_2” of the verb V2 is a phrase posterior to the verb V2.

Example 2: “Cyber Agent muốn xây dựng VinaGame thứ 2 tại Việt Nam.” (ICTnews [1])

[English translation: “Cyber Agent wants to build the second VinaGame in/at Viet Nam.”]

- The noun phrase “Cyber Agent” is the first argument of the verb “muốn” (English translation: “want”) and the verb “xây dựng” (English translation: “build”), and named “same_Subject”. The noun phrase “VinaGame thứ 2” is chosen as the second argument of the verb “xây dựng”.
- The location preposition “tại Việt Nam” (English translation: “in/at Vietnam”) is chosen as argument of the function Location(PreP).
- The connection rule >>> represents the relation between the function Location(PreP) and the predicate.
- The semantic representation of this sentence is as follows:

muốn(*<Cyber Agent >*, *xây_dựng* ((*<same_Subject >*, *<VinaGame thứ 2>*)) *>->* *Location*(*<tại Việt Nam>*)

2.2.2.2. Sentence has two verbs which combine together by conjunction “và”

In the semantic representation of this sentence type, the two verbs V1 and V2 are combined together by conjunction “và” (English translation: “and”) as described in Figure 3.

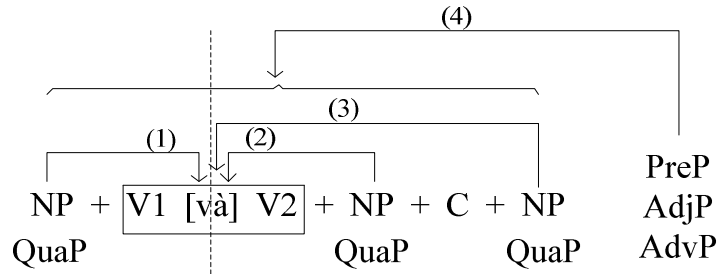


Figure 3: Sentence having two verbs which combine together by conjunction “và”

In this sentence structure in Figure 3, two mandatory elements of this sentence are V1 and V2. Both these verbs represent the semantic content of the sentence. We don't represent V1 and V2 as the form *name_verb_V1_and_name_verb_V2(argument_1, argument_2)*. Although both those verbs have the same arguments, the semantic content of each verb is different.

We propose a form to represent both verbs in the following model:

$$\boxed{\text{ver_V1}(\text{argument_1, argument_2}) \leftrightarrow \text{verb_V2}(\text{argument_1, argument_2})}$$

We use the connection rule \leftrightarrow to represent the combination of these two verbs.

Next, the processing of the phrases such as PreP, AdjP, and AdvP is similar to the process in Figure 1.

Example 3: “*Trẻ em dạo và ngắm công viên Angry Birds.*” (ICTnews [1])

[English translation: “The children walk and watch Angry Birds park.”]

- The noun phrase “the children” is the first argument of two predicates “dạo” and “ngắm”. Similar, the noun phrase “*công viên Angry Birds*” is the first argument of two predicates “dạo” and “ngắm”.
- The connection rule \leftrightarrow represents the relation between the verb “dạo” and the verb “ngắm”.
- The semantic representation of this sentence is as follows:

$$\text{dạo}(\text{trẻ em}, \text{ công viên Angry Birds }) \leftrightarrow \text{ngắm}(\text{trẻ em}, \text{ công viên Angry Birds })$$

2.2.2.3. Sentence having two verbs and one phrase standing in the middle of them

In this model, the two verbs V1, V2 stand far from each other as described in Figure 4.

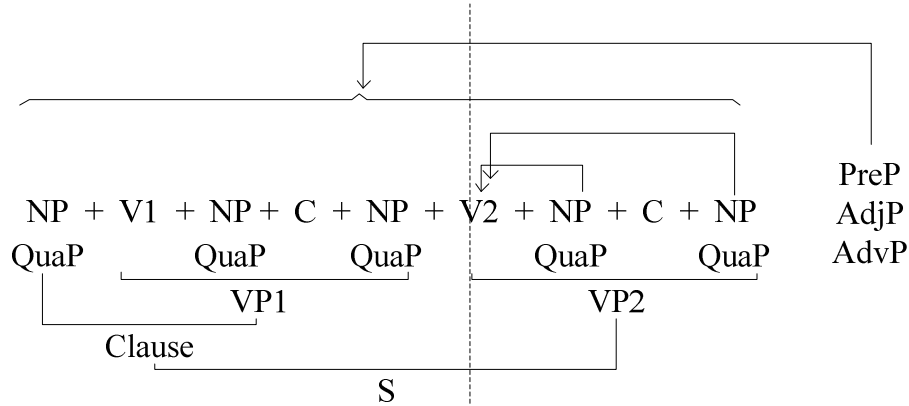


Figure 4: Sentence has two verbs standing far each other

In Figure 4, the composition of the “Clause” consists of noun phrase and verb phrase. Therefore, it is necessary to represent the semantic content of the clause by the verb V1:

verb_V1(argument_1, argument_2)

In which, “argument_1” and “argument_2” can be noun phrase (NP) or QuaP. With the verb V2, this verb also has two arguments as the following form:

verb_V2(argument_3, argument_4)

In which, “argument_3” is the semantic result of “Clause” (this is the semantic content of this verb V1). The “argument_4” will takes the value is NPs, or QuaPs. Follow the mentioned description, we have the representation form of the semantic content of this sentence type via the semantic form of the two verbs V1 and V2 as follows:

verb_V2(verb_V1(phrase_1, phrase_2), phrase_4)

The processing of the phrases such as PreP, AdjP, and AdvP is similar to the process in Figure 1.

Example 4: “*Viettel muốn đưa Việt Nam lên bản đồ công nghệ thế giới*” (ICTnews [1])

[English translation: “Viettel wants to put Viet Nam on the world map of technology.”]

- The phrase “Viettel muốn đưa Việt Nam” is a the clause, has the semantic representation form as follows:

muốn(<Viettel >, đưa(<Viettel >, < Việt Nam >))

This semantic representation of this clause is the first argument of the predicate “lên”.

- The clause “bản đồ công nghệ thế giới” is the second argument of the predicate “lên”.
- The semantic representation of this sentence is as follows:

lên(muốn(<Viettel >, đưa(<Viettel >, < Việt Nam >)), < bản đồ công nghệ thế giới >)

2.2.3. Sentence having only one verb phrase or verb phrase with preposition, adverb indicating time

This is a sentence type beginning with a verb phrase, it lacks the noun phrase preceding the verb. The noun phrase which is lack in the sentence is considered as a special argument of the predicate or function. We stipulate the symbol of the lacked phrase in the model as “_” or “no_subject”. The form of the semantic content of a sentence as follows:

$$\begin{matrix} \text{verb_V1}(\text{no_subject}, \text{argument}) \\ \text{verb_V2}(\text{verb_V1}(\text{no_subject}, \text{argument_1}), \text{argument_2}) \end{matrix}$$

Example 5: “Tra cứu mộ liệt sĩ nhờ mạng Internet” (ICTnews [1])

- This is the sentence type which has two verbs but lacks of one noun phrase or quantitative phrase preceding verb “tra cứu”. We consider the lacked phrase as the first argument “_” or “no_subject” of the predicate “tra cứu”.
- The phrase “mộ liệt sĩ” is noun phrase and will be chosen as the second argument of the predicate “tra cứu”. Therefore, we have the semantic form of the phrase “tra cứu mộ liệt sĩ” as follows: tra_cuu(<_>, <mộ liệt sĩ>) and this form is chosen as the first argument of the predicate “nhờ”.
- The phrase “mạng Internet” is noun phrase and will be chosen as the second argument of the predicate “nhờ”.
- The semantic representation of this sentence is as follows:

$$\text{nhờ}(\text{tra_cuu}(\text{<_>, <mộ liệt sĩ>}), \text{< mạng Internet >})$$

2.2.4. Sentence does not have verb

The structure of the sentence which does not have the verb is represented in Figure 5:

$$\begin{matrix} (1) & & (2) & & (3) & & (4) \\ \text{NP} & + & \text{PreP_loca} & + & \text{AdjP} & + & \text{AdvP} \\ \text{QP} & & \text{PreP_poss} & & & & \end{matrix}$$

Figure 5: Sentence does not have verb

The composition of this sentence includes the phrases NP, QuaP, PreP, and AdvP. The position of the phrase (1), (2), (3), (4) in Figure 6 can permute reciprocally. Because the sentence doesn’t have the verb, we will use the functions and the connection rules to represent the semantic of this sentence.

The type of the sentence which doesn’t have the verb is represented in Figure 6:

$$\begin{matrix} \text{Object}(\text{NP})^{(1)} & & \text{Location}(\text{PreP_loca})^{(2)} & & \text{Adjective}(\text{AdjP})^{(3)} & & \text{Time}(\text{AdvP})^{(4)} \\ \text{Object}(\text{QuaP}) \& \text{Possessive}(\text{NP/QuaP, PreP_poss}) \& \end{matrix}$$

Figure 6. The semantic representation of the sentence which does not have verb

In which,

- The function *Object(NP)*, *Object(QuaP)* represents the semantic of the NPs and QuaPs respectively.

- The function *Location(PreP_loca)*, *Possessive(NP/QuaP, PreP_poss)* represent the semantic of the location preposition and the possession preposition respectively.
- The function *Adjective(AdjP)* represents adjective phrase.
- The function *Time(AdvP)* represents the time adverb.
- The connection rule & combine the functions.

Example 6: “10 công dụng “khó đỡ” của Apple iPad.” (ICTnews [1])

[English translation: “10 usages “preposterous” of Apple iPad.”]

- The preposition phrase “của Apple iPad” indicates the possession and is used as the argument of function *Possessive(QuaP, PreP_Poss)*.
- The quantitative phrase “10 công dụng “khó đỡ”” is used as the argument of function *Possessive(QuaP, PreP_Poss)*.
- The semantic representation of this sentence as follows:

Possessive(<10 công dụng khó đỡ>, <của Apple iPad>)

3. THE SEMANTIC PROCESSING MODEL OF DATA SENTENCES IN VNEWSQA/ICT SYSTEM

In this section, we introduce the semantic processing model of data sentences in VNewsQA/ICT system which is built based on the semantic representation models of simple Vietnamese sentences in the section 2.

3.1. The Semantic Processing Model

The semantic processing model of simple Vietnamese sentences in the VNewsQA/ICT system is introduced in Figure 7.

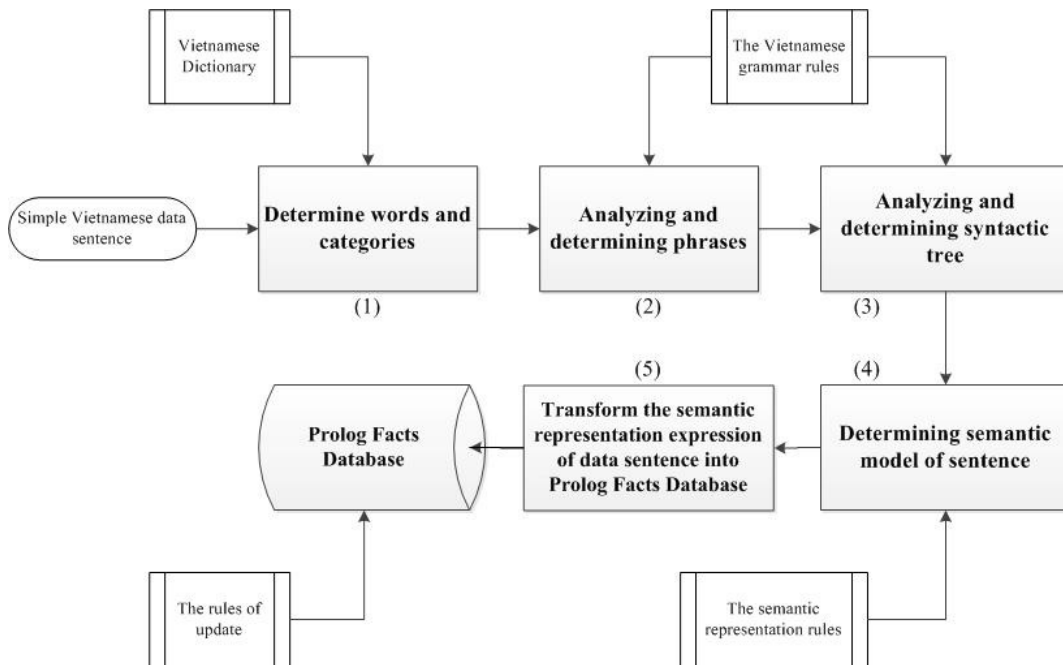


Figure 7: The semantic processing model of simple Vietnamese data sentences in VNewsQA/ICT system

The semantic processing model of simple Vietnamese sentences in VNewsQA/ICT system includes five processing stages, correspond with five stages (1) - (2) - (3) - (4) - (5) of the process in Figure 7.

- Stage 1: The system determines the words and the categories based on “The Vietnamese Dictionary” and the Vietnamese grammar rules of the system.
In this research, we accept the point of view about the simple word and compound word to convenient for the processing. For example, “văn phòng” (“office”) will be signed as “văn_phòng” and we consider as one word in the system.
The determining of words and categories in the sentence is based on “The Vietnamese Dictionary” which is built by us for the system. Every the word in the dictionary can have different categories labels.
- Stage 2: Determine the phrases in the sentence.
To determine the phrases have to base on “The Vietnamese grammar rules” and the categories of the words. The stage determining the phrases is important to determine exactly the arguments of the predicates or the functions in the semantic representation. Base on the phrases which are determined exactly, the system will determine the appropriate connection rule (if needed).
 - + If they are NP or QuaP, they will be the arguments of the predicate or the function *object(argument)*.
 - + If they are AdvP_time or phrase of the time, they will be the arguments of the function *time(argument)*.
 - + If they are PreP, they will be the arguments of the function *location(argument)*.
 - + If they are AdvP with the manner, they will be the arguments of the function *manner(argument)*.
- Stage 3: This stage analyzes the syntactic structure of the data sentence. After determining the categories of the words in the data sentence exactly, the system bases on “The Vietnamese grammar rules” to determine the syntactic structure of the sentence.
- Stage 4: After the stage 1, the stage 2, and the stage 3 are performed successfully, the system will determine the semantic representation model of the sentence. The process of choosing the semantic representation model of the data sentence is implemented with the support of “The semantic representation rules”.
The result of this stage is an expression representing the relation between the predicates and the functions. The arguments of the predicates or functions are the phrases determined in the stage 3.
- Stage 5: Transform the semantic representation expression of the data sentence into the Prolog Facts Database. Though the process of analyzing a semantic representation expression and combines with the rules of update, the system will transform the semantic representation expression of the sentence into the fact in Prolog Facts Database for user’s querying.

In this research, we define “The Vietnamese grammar rules” and “The semantic representation rules” in Definite Clause Grammar [2], [3], [4], [5], [6], [7], [8] and use the SWI-Prolog [12] to execute the rules.

Notice, with some semantic representation expressions which have the complete form as follows:

Predicate(argument 1, argument 2) <the connection rules> function(argument)

If in the semantic representation expression there are connection rules (>->, >-->, >--->) and *predicate(argument 1, argument 2)* (exclude the case of the sentences which don't have the verb), we have to create "a new fact" and put into "Prolog Facts Database". In detail, we use the expression *predicate(argument 1, argument 2)* to create a new fact into Prolog. The system creates and adds a new fact based on "The rules of update". This process uses the facts disaggregation mechanism.

Example 7: "Mobifone khai trương cửa hàng mới tại Vincom TP.HCM." (ICTnews [1])

[English translation: "Mobifone opens the new shopping in Vincom TP.HCM."]

Assume that we ask some questions about the content of this sentence:

- (a) Mobifone khai trương cái gì tại Vincom TP.HCM?
[English translation: "What does Mobifone open in Vincom TP.HCM?"]
- (b) Ai khai trương cửa hàng mới tại Vincom TP.HCM?
[English translation: "Who opens the new shopping in Vincom TP.HCM?"]
- (c) Tại TP.HCM Mobifone khai trương cái gì?
[English translation: "At TP.HCM, What does Mobifone open?"]
- (d) Tại TP.HCM ai khai trương cửa hàng mới?
[English translation: "Who opens the new shopping in Vincom TP.HCM?"]

If we use the semantic representation model as follows:

khai_truong(['Mobifone'], [cửa_hàng, mới]) >-> location([tại, 'Vincom', 'TP.HCM'])

With this semantic representation model, the system will find the answers for the above questions.

- Assume that we ask some other questions:
 - (e) Mobifone khai trương cái gì?
[English translation: "What does Mobifone open?"]
 - (f) Ai khai trương cửa hàng mới?
[English translation: "Who open the new shopping?"]

The system cannot find the answers for these questions. To overcome this issue, we create a new fact by taking the semantic representation expression of the predicate and use as a new fact. Then, the system has two facts:

khai_truong(['Mobifone'], [cửa_hàng, mới]) >-> location([tại, 'Vincom', 'TP.HCM'])
khai_truong(['Mobifone'], [cửa_hàng, mới])

When we add *khai_truong(['Mobifone'], [cửa_hàng, mới])* into the fact database, the system can find the results for the question (e) and (f) and also question (a), (b), (c), (d).

If the model lacks of the first or the second argument of the predicate, then do not need to create a new fact (similar to the case that the sentence does not have verb).

3.2. Application of Semantic Processing Model

The semantic processing model of simple Vietnamese sentences is applied to VnewsQA/ICT system to analyze the semantic of data sentences. In this section, we introduce one example to illustrate the stage processing one data sentence.

Example 8: “MobiFone ra mắt gói cước Opera Mini.” (ICTnews [1])

[English translation: “MobiFone launches Opera Mini package.”]

The processing of the sentence in example 8 is based on the semantic processing model in Figure 7 as follows:

- Stage 1: The system use “Vietnamese Dictionary” to determine the words and categories. The words in this sentence are determined and labeled as follows:
MobiFone|PN ra_mắt|V gói_cước|CN Opera_Mini|PN
- Stage 2: Base on the rules in “The grammar rules”, the system determines the following phrases:

MobiFone	ra_mắt	gói_cước	Opera_Mini
NP	V	NP	
NP	VP		

- Stage 3: The syntactic structure of this sentence is represented in two forms:

1) Syntactic form 1:

sentence(np(pn('Mobifone')), vp(v(ra_mắt), np(cn(gói_cước), pn('Opera_Mini'))))

2) Syntactic form 2:

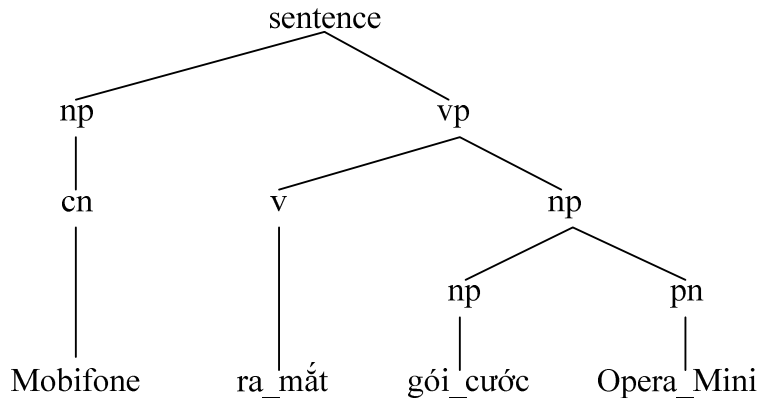


Figure 8: The syntactic structure of sentence “MobiFone ra mắt gói cước Opera Mini”

- Stage 4: Build the semantic representation expression of the data sentence. The semantic representation is represented as follows:

ra_mắt(['MobiFone'], [gói_cước, 'Opera_Mini'])

- Stage 5: Put the semantic representation expression of the data sentence into the fact database.

4. CONCLUSIONS

In this research, we present the semantic representation models and semantic processing mechanisms for simple Vietnamese sentences which are performed in VnewsQA/ICT system. We also introduce the use of the functions and connection rules in the semantic representation model. These models and mechanisms allow the system to analyse the Vietnamese news titles which are used as the data sentences of the system.

In the future papers, we will introduce more details about other information of VnewsQA/ICT system.

ACKNOWLEDGEMENTS

This research is funded by Vietnam National University – Ho Chi Minh City (VNU-HCM) under grant number B2012-26-05.

REFERENCES

- [1] ICTnews. [Online]. <http://www.ictnews.vn/Home/>
- [2] Fernando C. N. Pereira and Stuart M. Shieber, *Prolog and Natural-Language Analysis*, Microtome Publishing, 2005.
- [3] Pierre M. Nugues, *An Introduction to Language Processing with Perl & Prolog*, Springer, 2006.
- [4] Doug Arnold, *Prolog and NLP Basics, Syntax and Semantics, Using Prolog*, University of Essex, 2000.
- [5] Sandiway Fong, LING 364: Introduction to Formal Semantics, Spring 2006. [Online]. <http://dingo.sbs.arizona.edu/~sandiway/ling364/index.html>.
- [6] CSA4050: Advanced Topics in Natural Language Processing. [Online]. <http://staff.um.edu.mt/mros1/csa4050/>
- [7] CSA5006: Logic, Representation and Inference. [Online]. <http://staff.um.edu.mt/mros1/csa5006/>
- [8] CSM305: Introduction to Natural Language Processing. [Online]. <http://staff.um.edu.mt/mros1/cs305/>
- [9] The Prolog Dictionary. [Online]. <http://www.cse.unsw.edu.au/~billw/prologdict.html>
- [10] Phạm Thế Sơn, Hồ Quốc Thịnh, "Mô hình ngữ nghĩa cho câu trần thuật và câu hỏi tiếng Việt trong hệ thống vấn đáp kiến thức lịch sử Việt Nam", B.Sc. Thesis, Faculty of Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2012.
- [11] Noam Chomsky, *Syntactic Structures*, The Hague: Mouton & Co., 1957.
- [12] SWI-Prolog. [Online]. <http://www.swi-prolog.org/>