RESOLVING THE SEMANTICS OF VIETNAMESE QUESTIONS IN VNEWSQA/ICT SYSTEM

Son The Pham[1] and Dang Tuan Nguyen[2]

[1] Department of Information Science and Engineering, University of Information Technology, Vietnam National University – Ho Chi Minh City
[2] Faculty of Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City

ABSTRACT

Recently we have built a VNewsQA/ICT system which can read the titles of Vietnamese news in the domain of information and communication technology, then process and use them to answer the Vietnamese questions of users. The architecture of VNewsQA/ICT system has two main components: 1) the first component treats the simple Vietnamese sentences as its natural language textual data which is used to answer the user's questions; 2) the second component resolves the semantics of Vietnamese questions which query the system. This paper introduces a semantic representation model and a processing model to revolve the Vietnamese questions in VNewsQA/ICT system. These semantic representation and processing models are able to resolve the semantics of eight Vietnamese question classes which are used in our system.

KEYWORDS

Question Answering, Computational Semantics, Semantic Representation, Vietnamese Language Processing

1. INTRODUCTION

In order to answer Vietnamese questions about latest news of information and communication technology, we have built a VNewsQA/ICT system [1], which is based on a proposed model of Question – Answering system in [2], to allow users to query brief information in the Vietnamese news titles. In the first time, we tried to experiment the system with the news titles which are published at the website of online journal ICTNEWS [3].

The architecture of VNewsQA/ICT system [1] has two main components: 1) the first component treats the simple Vietnamese sentences as its natural language textual data which is used to answer the user's questions; 2) the second component resolves the semantics of Vietnamese questions which query the system. In [1], we focused on proposing the semantic models, and the semantic processing mechanism which was applied to determine the semantics of Vietnamese data sentences in VNewsQA/ICT system. In addition, we also revolved the following issues of semantic processing the Vietnamese data sentences: the time, the location, the manner, the negation, and the possession. However, we need other semantic representation and processing models to resolve the Vietnamese questions.

In this paper, we introduce a semantic representation model which is used in our semantic processing model to specially resolve the Vietnamese questions in VNewsQA/ICT system [1].

DOI : 10.5121/ijaia.2014.5202

Based on [2], [4], [5], [6], [7], [8], [9], [10], [11], these semantic representation and processing models are able to resolve eight Vietnamese question classes of our system.

2. RESOLVING VIETNAMESE QUESTIONS IN VNEWSQA/ICT SYSTEM

2.1. Semantic processing model for resolving Vietnamese questions

The semantic processing model for resolving Vietnamese questions in VNewsQA/ICT system is presented in Figure 1.

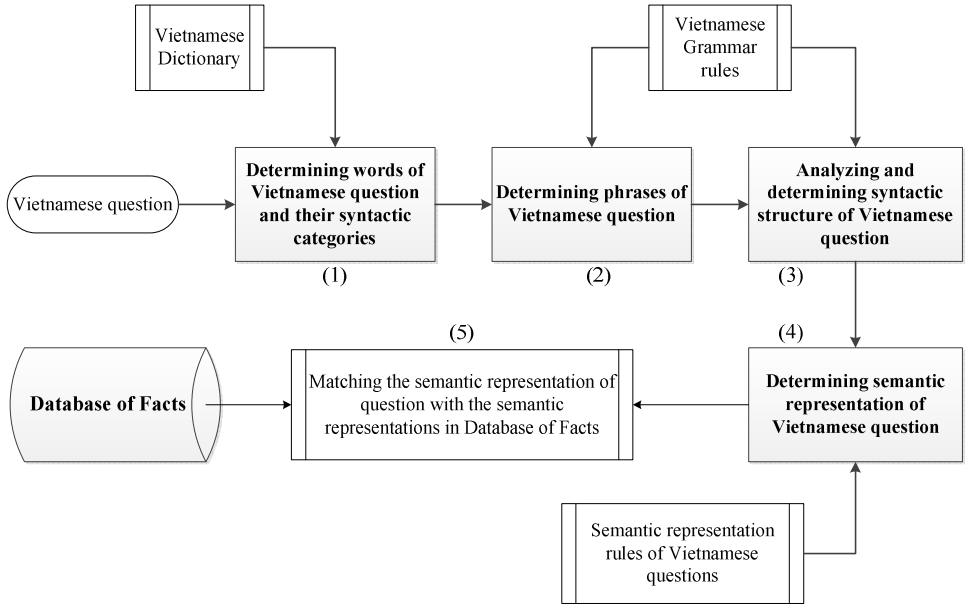

Figure 1: Semantic processing model for resolving Vietnamese Questions in VNewsQA/ICT system

The semantic processing model for resolving Vietnamese questions in the VNewsQA/ICT system includes five processing stages, corresponding with five stages (1) - (2) - (3) - (4) - (5) of the process in Figure 1.

- Stage 1: The system determines the words and their categories based on the "The Vietnamese Dictionary" and the "The Vietnamese grammar rules" of the system.
- Stage 2: Determine the phrases in the sentence. The system bases on the "The Vietnamese grammar rules" and the categories of the words to determine the phrases of question such as NP (noun phrase), QuaP (quantitative phrase), AdvP (adverb phrase), AdjP (adjective phrase), PreP (preposition phrase) and QueP (interrogative phrase).
  NP (noun phrase), QuaP (quantitative phrase), AdvP (adverb phrase), AdjP (adjective phrase), PreP (preposition phrase) will be arguments of predicates or functions. The QueP phrases also will be arguments of predicates or functions to contain the answer of question.

- Stage 3: This stage analyzes the syntactic structure of the question. After exactly determining the categories of the words in the question, the system bases on "The Vietnamese grammar rules" to determine the syntactic structure of the question.
- Stage 4: The system will determine the semantic representation of the question, based on "The semantic representation rules" and "GAP" technique proposed in [4], [5] and implemented in [2].
- Stage 5: Compare the semantic representation of the question with the semantic representation of the data sentences to search the answer of question.

2.2. Classification of questions in VNewsQA/ICT system

In VnewsQA/ICT system, Vietnamese questions are classified in eight classes, will present in section 3.

In the structure of Vietnamese questions [12], there are not only noun phrases, quantitative phrases, verb phrases, adjective phrases, preposition phrases, adverb phrases but also interrogative phrases. The interrogative phrases are classified into two types as follows:

- The interrogative phrases include a common noun (CN) combined with some interrogative words (IRG):

  CN + IRG

  e.g. "công ty nào" ("what company"), "công ty gì" ("what company").

- The interrogative phrases include some interrogative words (IRG) combined with a common noun (CN):

  IRG + CN

  e.g. "bao nhiêu công ty" ("how many company"), "mấy văn phòng" ("how many office"), v.v…

The noun phrases are processed follow two different types

- The noun phrase indicates person. (a)
- The noun phrase indicates things, facts, and phenomenon. (b)

In the Vietnamese grammar of system, we group (a) and (b) into the noun phrases. However, we have to define a function to determine that a noun phrase belong to type (a) or (b).

3. SEMANTIC REPRESENTATION MODEL FOR RESOLVING VIETNAMESE QUESTION CLASSES

In this section, we propose a method for presenting and calculating the semantic for eight types of question in VNewsQA/ICT system.

3.1. Questions about things, facts, phenomena [Class 1]

The interrogative words which are used in question class 1 include "gì", "nào" … (i.e. "what-questions" in English). The answers of the questions in class 1 are noun phrases describing things, facts, and phenomena.

The structure of interrogative phrases in class 1 which is used to query about things, facts, and phenomena is like as: CN + IRG. In which, CN is common noun indicating things, facts, phenomena and IRG is an interrogative word that includes: "gì", "nào".
According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection_rule> function(argument_3) verb(argument_1, argument_2)

Then, the semantic representations of the questions in class 1 have the structures as follows:

verb(What, argument_2) <connection_rule> function(argument_3) verb(What, argument_2)

Or

verb(argument_1, What) <connection_rule> function(argument_3) verb(argument_1, What)

The answers of the questions in class 1 are argument _1 or argument _2.

Example 1: *"Mobifone mở cái gì tại Myanmar?"* (ICTNEWS [3])
("What does Mobifone open in Myanmar?")

The semantic representation of this question is as follows:
mở(<MobiFone >, What) >-> Location(<tại Myanmar >)

3.2. Question about persons [Class 2]

The interrogative words which are used in class 2 include: "ai", "người nào", etc. (i.e. "who-questions" in English). The structure of question in question class 2 and the structure of question in class 1 are similar, but the query target of questions in class 2 is query about persons. The answers of questions in class 2 are noun phrases indicating persons.

According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection_rule> function(argument_3) verb(argument_1, argument_2)

Then, the semantic representations of the questions in class 2 have the structures as follows:

verb(Who, argument_2) <connection_rule> function(argument_3) verb(Who, argument_2)

Or

verb(argument_1, Whom) <connection_rule> function(argument_3) verb(argument_1, Whom)

The answers of the questions in class 2 are argument_1 or argument_2.

Example 2: *"Ai kiếm được 14.000 USD nhờ Flashback trong 3 tuần?"* (ICTNEWS [3])

("Who earns $ 14,000 in 3 weeks thanks to Flashback?")

The semantic representation of this question is as follows:
nhờ(kiếm(Who, <14.000 USD>), < Flashback >)>-->Time(<trong 3 tuần>)

3.3. Questions about time [Class 3]

The interrogative words which are used in question class 2 include: "khi nào", "lúc nào", "vào lúc nào", etc. (i.e. "when-questions" in English). The answers of the questions in class 3 are adverb phrases indicating time.

The structure of interrogative phrase in question class 3 is used to query about time is as follows:

(Pre) + CN + IRG

In which, CN is common noun indicating time, IRG is interrogative, Pre is preposition indicating time.

According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection_rule> time(argument_3) verb(argument_1, argument_2)

The semantic representation of the questions in class 3 has a structure as follows:

verb(argument_1, argument_2) <connection_rule> time(When)

The answers of the questions in class 3 are argument_3.

Example 3: *"FPT Polytechnic mở 2 chuyên ngành mới vào năm nào?"* (ICTNEWS [3])
("When did FPT Polytechnic open two new majors?")

The semantic representation of this question is as follows:
mở(<FPT Polytechnic>, <2 chuyên ngành mới>) >---> Time(When)

3.4. Questions about position, location [Class 4]

The interrogative words which are used in class 2 include: "đâu", "ở đâu", "nơi nào", "chỗ nào", "tại đâu", etc. (i.e. "where-questions" in English). The answers of the questions in class 4 are noun phrases indicating position, location.

According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection_rule> location(argument_3) verb(argument_1, argument_2)

Then, the semantic representations of the questions in class 4 have the structures as follows [1]:

verb(argument_1, argument_2) <connection_rule> location(Where) verb(argument_1, argument_2)

The answers of the questions in class 4 are argument_3.

Example 4: *"MobiFone mở văn phòng đại diện ở đâu?"* (ICTNEWS [3])
("Where does MobiFone open the representative office?")

The semantic representation of this question is as follows:
mở(<MobiFone>, <văn phòng đại diện>) >-> Location(Where)

3.5. Questions about the characteristics of objects, things [Class 5]

The interrogative words which are used in class 5 include: "như thế nào", "ra sao", "thế nào", etc. The answers of questions in class 5 are phrases describing object's size, shape, age, color, material, etc.

According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection> adjective(argument_3) verb(argument_1, argument_2)

The semantic representations of the questions in class 5 have the structures as follows:

verb(argument_1, argument_2) <connection> adjective(How) verb(argument_1, argument_2)

The answer of the questions in class 5 is argument_3.

Example 5: *"Công nghệ mới sạc smartphone và tablet như thế nào?"* (ICTNEWS [3])
("How does the new technology charge the smartphone and tablet?")

The semantic representation of this question is as follows:
sạc(<công nghệ mới>, <smartphone và tablet>)>-->Adjective(How)

3.6. Questions about the number of objects [Class 6]

The interrogative words which are used in class 6 include: "bao nhiêu", "mấy", "bấy nhiêu, etc. (i.e. "How much / How many - questions" in English). The answers of the questions in class 6 are quantitative phrases indicating the number of things, facts.

The structure of interrogative phrase in question class 6 which is used to query about quantity is as follows:

IRG + CN

In which, CN is common noun indicating person, things, facts, phenomena. IRG are some interrogative words.

According to [1], the semantic representations of the data sentences have the structures as follows:

verb(argument_1, argument_2) <connection_rule> function(argument_3) verb(argument_1, argument_2)

The semantic representations of the question in class 5 have forms as follows:

verb(How_many_much, argument_2) <connection_rule> function(argument_3)
verb(How_many_much, argument_2)

Or

verb(argument_1, How_many_much) <connection_rule> function(argument_3)
verb(argument_1, How_many_much)

The answers of the questions in class 6 are argument_1 or argument_2.

Example 6: *"Bao nhiêu doanh nghiệp Nhật tham dự Vietnam IT Day tại Nhật?"* (ICTNEWS [3])

("How many Japanese enterprises attended Vietnam IT Day in Japan?")

The semantic representation of this question is as follows:

tham_dự(How_many_much, <Vietnam IT Day>)->Location(<tại Nhật>)

3.7. General questions about objects [class 7]

The questions in class 7 are general questions. They are used to query information about objects. They are the questions which are structured from the questions belonging to the classes from 1 to 6. The questions in class (1) up to (6) are direct questions. The questions in class (7) are the questions for clarifying issues. To solve the questions in class 7, we base on the questions in class (1) up to (6) and build the mechanism for analyzing the query purpose of questions in class (7). The results of general questions about objects are the data sentences having concerned objects. The query object in class (7) includes two types:

- **The indefinite object:** the objects are composed common noun or the indefinite quantity phrase. These objects cannot be queried by the questions of class (7).
  For example, to query about a "telephone" object, users cannot put the questions: *"Có thông tin nào về điện thoại?"* (*"Which is information about the telephone?"*), or *"Những thông tin nào có liên quan đến điện thoại?"* (*"Which is information relate to the telephone?*)

- **The definite object:** the objects are identified by proper nouns, such as: "IPhone 8", "Ipad 8", etc. Therefore, the questions of class (7) are used to query information about definite object.

To resolve the questions in class (7), we will reuse the semantic representations mentioned above:

- The first semantic representation:

  verb_1(argument_1_1, argument_1_2) <connection_rule> function(argument_1_3)
  verb_1(argument_1_1, argument_1_2)

- The second semantic representation:

  verb_2(argument_2_1, argument_2_2) <connection_rule> function(argument_2_3)
  verb_2(argument_2_1, argument_2_2)

With these two semantic representations above, there are the cases that can happen:

- Two verbs (verb_1, verb_2) are different or similar.

- The argument_1_1 (the first representation) = the argument_2_2 (the second representation) = "the object that needs to be queried the information".

verb_1(argument_1_1, argument_1_2) <connection_rule> function(argument_1_3) verb_2(argument_2_1, argument_2_2) <connection_rule> function(argument_2_3)

- The argument_1_1 (the first representation) = the argument_2_1 (the second representation) = "the object that needs to be queried the information.

verb_1(argument_1_1, argument_1_2) <connection_rule> function(argument_1_3) verb_2(argument_2_1, argument_2_2) <connection_rule> function(argument_2_3)

- The argument_1_2 (the first representation) = the argument_2_1 (the second representation) = "the object that needs to be queried the information".

verb_1(argument_1_1, argument_1_2) <connection_rule> function(argument_1_3) verb_2(argument_2_1, argument_2_2) <connection_rule> function(argument_2_3)

- The argument_1_2 (the first representation) = the argument_2_2 (the second representation) = "the object that needs to be queried the information".

verb_1(argument_1_1, argument_1_2) <connection_rule> function(argument_1_3) verb_2(argument_2_1, argument_2_2) <connection_rule> function(argument_2_3)

Therefore, finding the answers of questions in class 7 will become the problem of finding the semantic representations of questions in the set of semantic representations of data sentences.

Example 7:

(a) "Viettel xây dựng phần mềm Quản lý quốc tịch cho Bộ Tư pháp." (ICTNEWS [3])
("Viettel builds citizenship management software for Department of Justice.")

(b) "FPT IS giúp Ngân hàng Nhà nước quản tiền bằng CNTT." (ICTNEWS [3])
("FPT IS helps State Bank manage cash in CNTT")

(c) "Học viện Công nghệ BCVT mở ngành An toàn thông tin vào năm 2013." (ICTNEWS [3])
("Posts & Telecommunications Institute of Technology opens Information Security industry in 2013.")

The semantic representation of the data sentence in example 7 has a structure as follows:

Sentence (a):
cho(xây_dựng(<Viettel>, <phần mềm Quản lý quốc tịch>), <Bộ Tư Pháp>)
Sentence (b):
quản(giúp(<FPT IS>, <Ngân hàng Nhà nước>), <tiền>) >--> Preposition(<bằng CNTT>)

Sentence (c):
mở(<Học viện Công nghệ BCVT>, <ngành An toàn thông tin>) >---> Time(<vào năm 2013>)

Using the questions in class 7 for querying, we have the questions and answers as follows:

- Question 1: *"Những tin tức liên quan đến Apple?"*
  ("Which news does relate to Apple?")
  Result: No information
- Question 2: *"Những thông tin liên quan đến Apple và iOS?"*
  ("Which information does relate to Apple and iOS?")
  Result: No information
- Question 3: *"Những tin tức nào liên quan đến FPT IS?"*
  ("Which news does relate to FPT IS?")
  Result: (b)
- Question 4: *"Những tin tức liên quan đến Viettel?"*
  ("Which information does relate to Viettel?")
  Result: (a)

3.8. Questions of "Yes – No" form [Class 8]

The structure of questions in class 8 includes two components: "information to be verified" (Tested-Inf) and "interrogative words" (IRG). Tested-Inf has the structure of the simple questions in class 1, class 2, class 3, class 4, class 5, and class 6. IRG includes: "có phải", "phải không", "đúng không", "không", "hay không", "có phải không", etc.
The structure of questions in class 8 is as follows:

IRG + Tested-Inf + IRG

Base on "Tested-Inf" and "IRG", we have many structural forms of questions as follows:

- "có phải" <Tested-Inf> "không"?
- <Tested-Inf> "không"?
- <Tested-Inf> "đúng không"?
- <Tested-Inf> "có phải không"?
- "có phải" < Tested-Inf> ?
- etc.

Example 8: *"Mỹ yêu cầu Trung Quốc dẹp loạn tin tặc."* (ICTNEWS [3])
("United States of American requires China to quell hacker")

Applying the question forms in class 8, we may have the questions as follows:
(a) *Có phải* Mỹ yêu cầu Trung Quốc dẹp loạn tin tặc *không*?
(b) Mỹ yêu cầu Trung Quốc dẹp loạn tin tặc *đúng không*?
(c) Mỹ yêu cầu Trung Quốc dẹp loạn tin tặc *có phải không*?

Base on interrogative phrase, if the system determines that the questions are class 8, the system will test "Tested-Inf". The processing stage of "Tested-Inf" is similar to the syntactic and semantic processing stage of Vietnamese questions. When having a semantic representation of the "Tested-Inf", the system will seeks in Database of Facts to find the fact that matches the semantic representation of the "Tested-Inf".

The semantic representation of the data sentence in example 8 has the structure as follows (C1):

dẹp_loạn(yêu_cầu(<Mỹ>, <Trung Quốc>), <tin tặc>)

The semantic representation of the questions (a), (b), and (c) has the structure as follows (C2):

dẹp_loạn(yêu_cầu(<Mỹ>, <Trung Quốc>), <tin tặc>)

Next, the system will compare the semantic representation of (C2) with the one of (C1) to answer the question. We note that the semantic representations of data sentence and question are identical for Yes – No questions in Vietnamese.

4. CONCLUSIONS

After building the VNewsQA/ICT system, we have used 400 news titles of ICTNEWS [3] as data sentences for operating the system. These data sentences have been processed by the semantic model and the semantic processing models proposed in [1]. The precision of syntactic and semantic processing for these news titles is 0.97.

The VNewsQA/ICT system is tested with 100 Vietnamese questions which are treated by the semantic representation and processing models introduced in this paper. The precision of answering is 0.96 for Vietnamese questions which are appropriate to the forms of eight question classes in this paper.

ACKNOWLEDGEMENTS

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

REFERENCES

[1] Son The Pham and Dang Tuan Nguyen, “Processing Vietnamese News Titles to Answer Relative Questions In VNewsQA/ICT System”, International Journal on Natural Language Computing (IJNLC), Vol. 2, No. 6, December 2013, pp. 39 - 51.

[2] 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 in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2012.

[3] ICTNEWS - Chuyên trang về CNTT của Báo điện tử Infonet. [Online]. Available at: http://www.ictnews.vn

[4] Fernando C. N. Pereira and Stuart M. Shieber, Prolog and Natural-Language Analysis, Microtome Publishing, 2005.

[5] Pierre M. Nugues, An Introduction to Language Processing with Perl & Prolog, Springer, 2006.

[6] Doug Arnold, LG519 Prolog and NLP Basics, Syntax and Semantics, Using Prolog, University of Essex, 2000.

[7] Sandiway Fong, LING 364: Introduction to Formal Semantics, Spring 2006. [Online]. Available at: http://dingo.sbs.arizona.edu/~sandiway/ling364/index.html.

[8] CSA4050: Advanced Topics in Natural Language Processing. [Online]. Available at: http://staff.um.edu.mt/mros1/csa4050/

[9] CSA5006: Logic, Representation and Inference. [Online]. Available at: http://staff.um.edu.mt/mros1/csa5006/

[10] CSM305: Introduction to Natural Language Processing. [Online]. Available at: http://staff.um.edu.mt/mros1/cs305/

[11] The Prolog Dictionary. [Online]. Available at: http://www.cse.unsw.edu.au/~billw/prologdict.html

[12] Cao Xuân Hạo (2001), Tiếng Việt mấy vấn đề ngữ âm, ngữ pháp, ngữ nghĩa, Nxb Giáo dục.

[13] Cao Xuân Hạo (2004), Tiếng Việt sơ thảo ngữ pháp chức năng, Nxb Giáo dục.