A SYNTACTIC ANALYSIS MODEL FOR VIETNAMESE QUESTIONS IN V-DLG~TABL SYSTEM

An Hoai Vo and Dang Tuan Nguyen

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

ABSTRACT

This paper introduces a syntactic analysis model that we propose to parse and process the Vietnamese questions about tablets in V-DLG~TABL system, which is a Vietnamese Question – Answering system working based on automatic dialog mechanism. The V-DLG~TABL system is built to support clients using Vietnamese questions for searching tablets based on interaction between the clients and the system. We apply the "Phrase Structure Grammar" of Noam Chomsky to develop a syntactic analysis model that is specific and suitable for the V-DLG~TABL system. This syntactic analysis model is used to implement the "V-DLG~TABL Syntactic Parsing and Processing" component of the system.

KEYWORDS

Syntax, Parsing, Question Answering, Automatic Dialog Mechanism, Vietnamese Language Processing.

1. INTRODUCTION

In this paper, we present a syntactic analysis model that we propose to parse and process Vietnamese questions about tablets in our V-DLG~TABL, which is an advanced Question – Answering system working with a dialog mechanism based on scenarios. We hope to build this system to help clients who want to buy tablets find the information about the ones they are interested in, based on their interaction with the system by using Vietnamese language.

To build a system with such functions, we design the architecture of V-DLG~TABL system based on major components as follows:

- The component "V-DLG~TABL Syntactic Parsing and Processing": the Vietnamese question about tablets that clients enter to system is automatically analyzed based on "V-DLG~TABL_PSG" grammar which has been defined in the system. Then, based on the syntactic structure of the question, the system determines the syntactic elements which contain the principal information of the question corresponding to the "information structure model" proposed in [1].
- The component "V-DLG~TABL Semantic Analyzing": the important elements of the syntactic structure of the question, which correspond to the components of the "information structure model" proposed in [1], will be retained and transformed into predicates in FOL (First-Order Logic) based on the implementation techniques used in [1] and the programming methods proposed in [2].
- The component "V-DLG~TABL Facts Database Querying": searching data in the database of facts, based on the methods and techniques proposed in [1].

- The component "V-DLG~TABL Answer Creating": creating Vietnamese answer by using the analyzed syntactic structure of the question, based on the method proposed in [1].
- The component "V-DLG~TABL Dialog": Operating the interactions between client and system, and making suggestions relating to the information that the client is interested in conversation process.

In this V-DLG~TABL system, structurally analyzing and processing the syntactic structure of Vietnamese question is a fundamental task. This paper is limited to present the syntactic analysis model that we propose to perform this task. In fact, this syntactic analysis model is used to build the component "V-DLG~TABL Syntactic Parsing and Processing" of V-DLG~TABL system. However, this research does not discuss in details on any implementation of the system.

Related works: Nguyễn Thành and Phạm Minh Tiến [1] built a basic Question – Answering system that allows clients to use some simple forms of Vietnamese questions to query information about tablets. In [1], the building of system has been based on Definite Clause Grammar (DCG) [2], and the methods of computational semantics [3]. Based on [2] and [3], there were some other Question – Answering systems built for Vietnamese language such as: [4], [5], [6], [7], [8].

2. SYNTACTIC ANALYSIS MODEL OF SYSTEM

In this research, we reuse the classification of Vietnamese questions about the tablets which has been proposed in [1]. According to [1], the questions about the tablets which are distinguished into three fundamental types:

- Type 1: Questions about the features of tablets, or components of tablets.
- Type 2: Yes/no questions about the components of tablets.
- Type 3: The others for finding the tablets based on the query's information.

The syntactic analysis model for processing Vietnamese questions in V-DLG~TABL system relates to the following aspects:

- The theoretic model of syntax that is applied for parsing Vietnamese questions.
- The representation of information structure of Vietnamese questions.
- The method for determining the information structure of Vietnamese question based on its syntactic structure.

2.1. Defining the syntactic elements

In order to define "V-DLG~TABL_PSG" grammar of the system in Definite Clause Grammar (DCG) [2], at first, we define the basic syntactic elements. We apply the way to name the syntactic elements of Phan Thị Thể [10], [11] for nouns and noun phrases: a name begins with a character ('n' for nouns, 'np' for noun phrase), and follows by an underscore symbol '_', and other description words.

- The np_tablet: this syntactic element is a noun phrase. It represents the name of tablet.
- The n_component_<name>: this syntactic element is a noun. It represents the name of component of the tablet. The name of component consists of symbol "n_component_" and name of component. The notation is: n_component_<name>.

  For example: n_component_screen, n_component_bluetooth.
- The n_property_<name>: this syntactic element is a noun. It describes in details the tablet or the component of tablet. Its notation is n_property_<name>.

For example: n_property_price, n_property_color.

2.2. Building the grammar of system

The syntactic rules of the system are defined in Definite Clause Grammar [2] through the following steps:

1) **Step 1**: Define the basic syntactic elements.

- Nouns and noun phrases:
  - np_tablet, n_tablet, pn_tablet_name: the elements describe the name of tablet.
  - n_component_<name>: the element describes the components of tablet.
  - n_property_<name>: the element describes the features of components or tablet.
- Verbs and other syntactic elements:
  - 'verb': the verbs.
  - 'interrog': the interrogative words.
  - 'wh_tablet': the questions about the tablets.

The definitions of basic syntactic elements are listed in Table 1.

Table 1: Basic syntactic elements of "V-DLG~TABL_PSG" grammar

No.	Syntactic elements	Description
1	np_tablet	Noun phrase for naming the tablets
2	n_tablet	Words describe the tablets
3	pn_tablet_name	Names of tablets
4	n_component_screen	Noun describes the screen
5	n_component_sim	Noun describes the SIM card
6	n_component_front_camera	Noun describes the front camera
7	n_component_back_camera	Noun describes the back camera
	….	
8	n_property_size	Noun describes the size
9	n_property_weight	Noun describes the weight
10	n_property_color	Noun describes the color
11	n_property_price	Noun describes the price
	….	
12	verb	Verbs used in the grammar
13	interrog	Words describe queried parts of questions
14	wh_tablet	Words ask about the tablets

2) **Step 2**: Define the syntactic rules for phrases and question types.

The definitions of the syntactic rules based on the syntactic structures of Vietnamese question types are described in the section 3.

2.3. Representing information structure of questions about tablets

After analyzing the syntax of Vietnamese question, the problem is how to determine the syntactic elements representing important information to ask. In order to solve this issue, the syntactic

structure of the question has to be mapped on a predefined representation model of the information structure of the question.

In this research, we reuse the "information structure model" proposed by Nguyễn Thành and Phạm Minh Tiến [1] for representing the content of Vietnamese questions about tablets. According to [1], the information structure of questions is analyzed by the following components (cf. [1]):

- The information about tablet.
- The components of tablet.
- The features about tablet or components.
- The description about the features of tablet or components.

However, the way that we approach to analyze and transform the syntactic structure of Vietnamese questions into "information structure model" has some differences from [1] as follows:

- In [1], Nguyễn Thành and Phạm Minh Tiến defined a Definite Clause Grammar (DCG) for analyzing the predefined question types based on their own "syntactic structure model" having the following elements [1]: *"product"*, *"functionWord"*, *"properties"* and *"value"*. These "syntactic elements" exactly correspond to the components of the "information structure model" of questions proposed in [1].
- In our research, we analyze the syntactic structure of Vietnamese questions by using the "Phrase Structure Grammar" theory of N. Chomsky [9]. Basing on the constituent structure of questions, we propose an algorithm for determining the syntactic elements corresponding to the components of the "information structure model" proposed in [1].

Briefly, Nguyễn Thành and Phạm Minh Tiến [1] analyzed Vietnamese questions about the tablets by using their own "syntactic structure model" that is not a real syntactic model, and it directly corresponds to the "information structure model" of the questions proposed in [1]. In [1], they did not use any grammar theory to analyze Vietnamese questions.

Otherwise, we analyze the syntax of Vietnamese questions based on the "Phrase Structure Grammar" theory of N. Chomsky [9] and then transform the syntactic structures into the "information structure model" which has been proposed in [1].

3. SYNTACTIC STRUCTURES OF QUESTION TYPES USED IN SYSTEM

In this section, we present the syntactic structures of the question types. These syntactic structures are viewed from the syntactic elements listed in Table 1, based on the "Phrase Structure Grammar" of N. Chomsky [9].

3.1. Questions about the features of the tablets or components

According to [1], the questions about the features of the tablets or components are type 1. We present the syntactic structures of this question type in Table 2.

Table 2: The syntactic structure of question type 1

Syntactic structure	Examples
<np_tablet> verb <n_component> verb <n_property> interrog	*"Máy tính bảng Nexus 7 có màn hình là loại gì?"* Syntactic tree: s(np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_component_screen('màn hình'), verb(là), n_property_type(loại)), interrog(gì)).
<np_tablet> verb <n_property> [preposition] <n_component> interrog	*"Máy tính bảng Nexus 7 có kích thước [của] màn hình bao nhiêu?"* Syntactic tree: s(np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_property_size('kích thước'), n_component_screen('màn hình')), interrog('bao nhiêu'))
<n_component> [preposition] <np_tablet> verb <n_property> interrog	*"Màn hình [của] máy tính bảng Nexus 7 là loại gì?"* Syntactic tree: s(np(n_component_screen('màn hình'), np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7'))), vp(verb(là), n_property_type(loại)), interrog(gì))
<n_property> [preposition] <n_component> <np_tablet> interrog	*"Loại [của] màn hình máy tính bảng Nexus 7 là gì?"* Syntactic tree: s(np(n_property_type(loại), n_component_screen('màn hình'), np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7'))), vp(verb(là), interrog(gì))

3.2. Yes/no questions about the components

According to [1], yes/no questions about components of tablets are type 2. We present the syntactic structures of this question type in Table 3.

Table 3: The syntactic structure of question type 2

Syntactic structure	Examples
<np_tablet> verb <n_component> verb <n_property> <literal> interrog	*"Máy tính bảng Nexus 7 có màn hình là loại 'cảm ứng' phải không?"* Syntactic tree: s(np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_component_screen('màn hình'), verb(là), n_property_type(loại), literal('cảm ứng')), interrog('phải không'))
<np_tablet> verb <n_property> [preposition] <n_component> verb <literal> interrog	*"Máy tính bảng Nexus 7 có kích thước [của] màn hình là '7 inch' phải không?"* Syntactic tree: s(np_tablet(n_tablet('máy tính

	bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_property_size('kích thước'), n_component_screen('màn hình'), verb(là), literal('7 inch')), interrog('phải không'))
<n_component> [preposition] <np_tablet> verb <n_property> verb <literal> interrog	*"Màn hình [của] máy tính bảng Nexus 7 có loại là 'cảm ứng' phải không?"* Syntactic tree: s(np(n_component_screen('màn hình'), np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_property_type(loại)), verb(là), literal('cảm ứng')), interrog('phải không'))
<n_property> [preposition] <n_component> [preposition] <np_tablet> verb <literal> interrog	*"Loại [của] màn hình [của] máy tính bảng Nexus 7 là 'cảm ứng' phải không?"* Syntactic tree: s(np(n_property_type(loại), n_component_screen('màn hình'), np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7'))), vp(verb(là), literal('cảm ứng')), interrog('phải không'))

3.3. Questions for finding the tablets by using query's information

According to [1], questions for finding the tablets based on the query's information are type 3. We present the syntactic structures of this question type in Table 4.

Table 4: The syntactic structure of question type 3

Syntactic structure	Examples
<wh_tablet> verb <n_component> verb <n_property> verb <literal> interrog	*"Máy tính bảng nào có màn hình là loại 'cảm ứng'?"* Syntactic tree: s(wh_tablet('máy tính bảng nào'), vp(verb(có), n_component_screen('màn hình'), verb(là), n_property_type(loại), literal('cảm ứng')))
<wh_tablet> verb <n_property> [preposition] <n_component> verb <literal> interrog	*"Máy tính bảng nào có kích thước [của] màn hình là '7 inch'?"* Syntactic tree: s(wh_tablet('máy tính bảng nào'), vp(verb(có), n_property('kích thước'), n_component_screen('màn hình'), verb(là), literal('7 inch')))
<n_component> [preposition] <wh_tablet> verb <n_property> verb <literal> interrog	*"Màn hình [của] máy tính bảng nào có loại là 'cảm ứng'?"* Syntactic tree: s(n_component_screen('màn hình'), wh_tablet('máy tính bảng nào'), vp(verb(có), n_property_type(loại), verb(là), literal('cảm ứng')))
<n_property> [preposition] <n_component> [preposition] <wh_tablet> verb <literal>	*"Loại [của] màn hình [của] máy tính bảng nào là 'cảm ứng'?"*

interrog	Syntactic tree: s(n_property_type(loại), n_component_screen(‘màn hình’), wh_tablet('máy tính bảng nào'), vp(verb(là), literal('cảm ứng'))
<wh_tablet> verb <n_component> interrog	*“Máy tính bảng nào có camera trước?”* Syntactic tree: s(wh_tablet('máy tính bảng nào'), vp(verb(có), n_component_front_camera(‘camera trước’))

Basing on the structures of these types presented in Table 2, Table 3 and Table 4, we define the syntactic rules for the grammar of the system. We use Definite Clause Grammar (DCG) [2] to define the syntactic rules for these question types which are handled in V-DLG~TABL.

In example 1, we illustrate a grammar which is built for analyzing a given question.

Example 1: Give the question “Máy tính bảng Nexus 7 có màn hình rộng bao nhiêu?”
(English: “How wide does the tablet Nexus 7 screen has?”)

- The syntactic element for the word “màn hình” is represented by *n_component_screen*.
- The syntactic component for the word “máy tính bảng” is represented by *n_tablet*.

The Definite Clause Grammar (DCG) in Table 5 is defined for analyzing the sentence in example 1.

Table 5: A Definite Clause Grammar (DCG) defined for analyzing the question in example 1

```
sentence(s(NP, VP, INTERROG)) --> np_tablet_nexus(NP),  vp_have_screen(VP),
interrog_how_many(INTERROG).
vp_have_screen(vp(V, N)) --> v_have(V), n_screen(N).
v_have(verb('có')) --> ['có'].
n_screen(n_component_screen('màn hình')) --> [màn, hình].
np_tablet_nexus(np_tablet(N, PN)) --> n_tablet(N), pn_nexus_7(PN).
n_tablet(n_tablet('máy tính bảng')) --> [máy, tính, bảng].
pn_nexus_7(pn_tablet_name('Nexus 7')) --> ['Nexus 7'].
interrog_how_many(interrog('rộng bao nhiêu')) --> [rộng, bao, nhiêu].
```

With the Definite Clause Grammar (DCG) in Table 5, the system can returns the syntactic tree in Prolog when a client inputs the following question:

sentence(S, [máy, tính, bảng, 'Nexus 7', có, màn, hình, rộng, bao, nhiêu], []).

The syntactic tree of example 1 is returned by Prolog as follows:

S = s(np_tablet(n_tablet('máy tính bảng'), pn_tablet_name('Nexus 7')), vp(verb(có), n_component_screen('màn hình')), interrog('rộng bao nhiêu')).

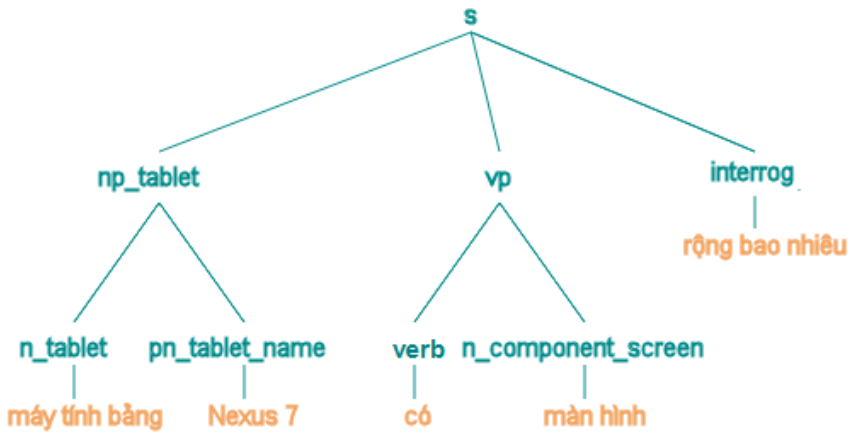

Figure 1: The syntactic tree of the question in example 1

4. CONCLUSIONS

Basing on the distinction of question types and the "information structure model" of Vietnamese questions about the tablets which are proposed in [1], we apply the "Phrase Structure Grammar" of N. Chomsky [9] to develop a syntactic analysis model that is specific and suitable for V-DLG~TABL system. This syntactic analysis model is used to implement "V-DLG~TABL Syntactic Parsing and Processing" component of system.

We have tested "V-DLG~TABL Syntactic Parsing and Processing" component of the system to evaluate the ability of answering Vietnamese questions. This component of system is able to answer exactly 141 of 150 tested Vietnamese questions about tablets.

ACKNOWLEDGEMENTS

This research is funded by University of Information Technology, Vietnam National University – Ho Chi Minh City (VNU-HCM), under grant number C2011CTTT-06.

REFERENCES

[1] Nguyễn Thành, Phạm Minh Tiến, "Xây dựng cơ chế hỏi đáp tiếng Việt cho hệ thống tìm kiếm sản phẩm máy tính bảng", B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2012.

[2] Fernando C. N. Pereira, Stuart M. Shieber, *Prolog and Natural-Language Analysis*, Digital Edition, Microtome Publishing, Brookline, Massachusetts, 2002.

[3] Patrick Blackburn, Johan Bos, *Representation and Inference for Natural Language: A First Course in Computational Semantics*, September 3, 1999.

[4] 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.

[5] Vũ Thế Nhân, Trần Thế Toàn, "Cơ chế phân tích nội dung câu hỏi dựa trên ngữ nghĩa hình thức cho hệ thống hỏi đáp tiếng Việt", B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2012.

[6] Lâm Thanh Cường, Huỳnh Ngọc Khuê, "Biểu diễn và xử lý ngữ nghĩa dựa trên FOL (First Order Logic) cho các dạng câu đơn tiếng Việt trong hệ thống hỏi đáp kiến thức xã hội", B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2012.

[7] Vương Đức Hiền, "Xây dựng công cụ truy vấn tiếng Việt về các phần mềm máy tính", B.Sc. Thesis in Computer Science, University of Information Technology, Vietnam National University – Ho Chi Minh City, 2013.

[8] 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. ISSN: 2278 - 1307 [Online]; 2319 - 4111 [Print].

[9] Noam Chomsky, *Syntactic Structures*, The Hague: Mouton & Co., 1957.

[10] Phan Thị Thể, "Cơ chế xử lý câu hỏi tiếng Việt cho hệ thống truy vấn thông tin đào tạo hệ tín chỉ", Master Thesis in Data Transmission and Computer Network, Posts and Telecommunications Institute of Technology, 2012.

[11] Phan Thị Thể, "Xây dựng cơ chế truy vấn dựa trên ngữ nghĩa của các cụm từ tiếng Việt cho hệ thống tìm kiếm việc làm", Master Thesis in Information Technology (Computer Science), University of Information Technology, Vietnam National University – Ho Chi Minh City, 2013.