A COMPUTATIONAL APPROACH FOR ANALYZING INTER-SENTENTIAL ANAPHORIC PRONOUNS IN VIETNAMESE PARAGRAPHS

Trung Tran[1] and Dang Tuan Nguyen[2]

[1]Faculty of Computer Science, University of Information Technology, Vietnam National University - Ho Chi Minh City, Vietnam
ttrung@nlke-group.net
[2]Faculty of Computer Science, University of Information Technology, Vietnam National University - Ho Chi Minh City, Vietnam
dangnt@uit.edu.vn

ABSTRACT

This paper presents a strategy and a computational model for solving inter-sentential anaphoric pronouns in Vietnamese paragraphs composing simple sentences. The strategy is proposed based on grammatical features of nouns and the focus phenomenon when using pronouns in Vietnamese. In this research, we consider only nouns and pronouns which are human objects in the paragraph, and each anaphoric pronoun will appear one time in one sentence and can appear in adjacent sentences. The computational model is implemented in Prolog and based on applying and improving the models of Mark Johnson and Ewan Klein, had been improved by Covington and Schmitz, with theoretical background of Discourse Representation Theory.Analysis of test results shows that this approach which based on linguistic theories helps for well solving inter-sentential anaphoric pronouns in Vietnamese paragraphs.

KEYWORDS

Inter-sentential Anaphora, Anaphora Resolution, Discourse Representation

1. INTRODUCTION

Solving inter-sentential anaphoric pronouns in a Vietnamese paragraph is an important research topic in natural language processing, especially in text comprehension researches. Many authors have proposed different approaches with strategies and models for finding exact antecedents of anaphoric pronouns in paragraphs. In their researches, Mark Johnson and Ewan Klein [7], Covington and Schmitz [9], Blackburn and Bos [12] have proposed models based on Discourse Representation Theory [4] and constraints about number and gender of pronouns in English to find antecedents of anaphoric pronouns. A different approach using WordNet Ontology of Tyne Liang Dian-Song Wu [14] to identify the animate entity and the information about gender of the entity. The system also uses same characteristics about gender of the object and the distinction between animate, non-animate objects in English and proposes some heuristic rules to solve anaphoric pronouns. Some other researches, such as Michel Denber [11], proposed the solution based on characteristics about number and gender of objects in English, with additional constraints of animate, non-animate objects and the syntax of words in sentences. Another theory is also widely used as the basis of many researches for solving the anaphoric pronouns is Centering Theory, developed by Barbara J. Grosz, Aravind K. Joshi and Scott Weinstein [2] in the 1980s.

DOI : 10.5121/ijnlc.2013.2303

In this paper, we present a strategy and a computational model for solving inter-sentential anaphoric pronouns in Vietnamese paragraphs composing simple sentences. The strategy is proposed based on grammatical features of nouns and pronouns in Vietnamese and the focus phenomenon when using pronouns in Vietnamese. The model is designed accordant with the strategy based on Discourse Representation Theory [4] consists of four main components: the component for analysing the syntactic structure of the paragraph and sentences with the top-down method and describing by Unification-Based Grammar – UBG [8], [13], the component for describing lexical characteristics structures by Unification-Based Grammar [8], [13], the component for building Discourse Representation Structure, the component for finding antecedents with algorithms based on the strategy. To perform in Prolog, we apply the model of Mark Johnson and Ewan Klein [7], Covington and Schmitz [9] with some improvements accordant with the strategy of solving inter-sentential anaphoric pronouns in Vietnamese paragraphs as follows:

- Only resolve inter-sentential anaphoric pronouns, and the antecedent appears in the sentence preceding the sentence containing the pronoun.
- Do not analyse the paragraph into sentences using recursive method, instead determine the position of each sentence.
- Describe characteristics of lexical in Vietnamese grammar.
- The algorithm of finding the antecedent of inter-sentential anaphoric pronoun based on the strategy.

In this research, we limit the consideration of the following forms of paragraph:

Form 1: The paragraph having only one anaphoric pronoun:
Example 1: *"Nhân học môn vẽ. Anh dùng bút chì. Nghĩa hỏi anh."*
(English: "Nhân learns painting. He uses pencil. Nghĩa asks him.")
⇨ [anh = Nhân]
Form 2: The paragraph having two anaphoric pronouns which appear in different sentences:
Example 2: *"Lễ đọc sách. Anh thấy Chí. Anh ta đọc báo."*
(English: "Lễ reads book. He sees Chí. He reads newspaper.")
⇨ [anh = Lễ, anh ta = Chí]
Form 3: The paragraph having two anaphoric pronouns which appear in the same sentence:
Example 3: *"Lan học môn toán. Chị hỏi Mai. Chị ấy giúp chị."*
(English: "Lan learns maths. She asks Mai. She helps her.")
⇨ [chị = Lan, chị ấy = Mai]

2. BACKGROUND

2.1. Discourse Representation Theory

The Discourse Representation Theory model had been introduced in [4] with the basic idea: a natural language discourse will be presented in the context of representative structure, which is called Discourse Representation Structures – DRS. According to [4], a Discourse Representation Structures will include an order pair <U, Con>, where U is a list of discourse markers, or can be interpreted as objects of the discourse, and Con is a list of conditions, or can be interpreted as predicates or formulas that objects in U have to satisfy.
Example 4: Consider the paragraph having three sentences:

"Nhân học môn vẽ. Anh dùng bút chì. Nghĩa hỏi anh."

(English: “Nhân learns painting. He uses pencil. Nghĩa asks him.”)

This paragraph will have the following Discourse Representation Structure:

- Objects in set U: *X1 – Nhân, X2 – môn vẽ, X3 – bút chì , X4 – Nghĩa*.
- Conditions in set Con: *tên(X1, [Nhân]),môn_vẽ(X2), học(X1,X2), bút_chì(X3), dùng(X1, X3), tên(X4, [Nghĩa]), hỏi(X4, X1)*.

This structure is represented in table 1:

Table 1. The Discourse Representation Structure of the paragraph “Nhân học môn vẽ. Anh dùng bút chì. Nghĩa hỏi anh”.

X1, X2, X3, X4
tên(X1,[Nhân])
môn_vẽ(X2)
học(X1,X2)
bút_chì(X3)
dùng(X1,X3)
tên(X4,[Nghĩa])
hỏi(X4,X1)

2.2. Unification-Based Grammar

In [8], [13], the authors introduced theories Unification-based and Unification-based Grammar with the basic idea: Unification-Based Grammar is a formalism in which theories of grammar can be expressed, with the prominent role of unifying feature structures. In the analysis of the syntactic structure of sentences, in each constituent or lexical, can describe the additional characteristic structure of this constituent or lexical.

3. THE STRATEGY FOR SOLVING INTER-SENTENTIAL ANAPHORIC PRONOUNS IN VIETNAMESE PARAGRAPH

In this section, we will present the strategy for solving inter-sentential anaphoric pronouns in Vietnamese paragraphs composing simple sentences. This strategy is based on grammatical features of nouns and pronouns in Vietnamese as well as the focus phenomenon in the use of pronouns in Vietnamese paragraph.

In Vietnamese, nouns or pronouns only distinguish characteristic of human or animals, non-animate object, not distinguish gender. Although there are some pronouns such as “anh” or “cô” have the distinction of male and female, but also pronouns like “em”, “nó” do not specify the gender. Therefore, different from [7], [9], in this research, we do not use the gender characteristic, instead will be based on grammatical characteristics of nouns that distinguish human with animals or non-animate objects and role characteristic of nouns is the subject or object of verb in the sentence to find the antecedent for inter-sentential anaphoric pronoun. The main idea of using these two features is “depending on the anaphoric pronoun stand alone or stand with “ấy” / “ta” / “này”, will focus in nouns which indicate human and take the subject or object role of verb in preceding sentences”. These two constraints help to solve inter-sentential anaphoric pronouns for paragraphs composing more than two sentences in comparison with the model of [7], [9].The

paragraphs are considered in this research in the forms presented in Introduction section will have following characteristics:

- The number of sentences is not determined in the range from 3 to 5 sentences.
- There are only human anaphoric pronouns.
- There are only one or two human antecedents among several ones.
- The anaphoric pronouns and antecedents can appear in sentences that are not adjacent.
- Each anaphoric pronoun will appear one time in one sentence or can appear in adjacent sentences.

The strategy for solving inter-sentential anaphoric pronouns in Vietnamese paragraph is:

- Mark the order position of sentences in the paragraph.
- Find the exact antecedent at preceding sentences of the one containing the anaphoric pronoun.
- The anaphoric pronoun stands alone: Find the antecedent is focusing human object, takes the subject role of the verb in preceding sentences.
- The anaphoric pronoun + "ta" / "ấy" / "này": Find the antecedent is focusing human object, takes the object role of the verb in preceding sentences.

The finding strategy will be illustrated with the example in Introduction section as follows:

Example 5: *"Lễ đọc sách. Anh thấy Chí. Anh ta đọc báo."*
(English: "Lễ reads book. He sees Chí. He reads newspaper.")

- ⇨ Pronoun "anh" stand alone in the second sentence, will focus to proper noun "Lễ" indicating human object and taking the subject role of verb in the first sentence. Therefore, the antecedent of anaphoric pronoun "anh" is object "Lễ".
- ⇨ Pronoun "anh" plus "ta" in the third sentence, will focus to proper noun "Chí" indicating human object and taking the object role of verb in the second sentence. Therefore, the antecedent of anaphoric pronoun "anh ta" is object "Chí".

4. THE SYSTEM MODEL

In this section, we present the system model and algorithm to find antecedents of inter-sentential anaphoric pronouns that accordant with the strategy proposed above. The system model is designed based on Discourse Representation Theory [4], applied and improved the model of [7], [9], consists of four main components: the component for analysing the syntactic structure of the paragraph and sentences with the top-down method and describing by Unification-Based Grammar [8], [13], the component for describing lexical characteristics structures by Unification-Based Grammar [8], [13], the component for building Discourse Representation Structure, the component for finding the antecedent with the algorithm based on the strategy. The model is represented as follow:

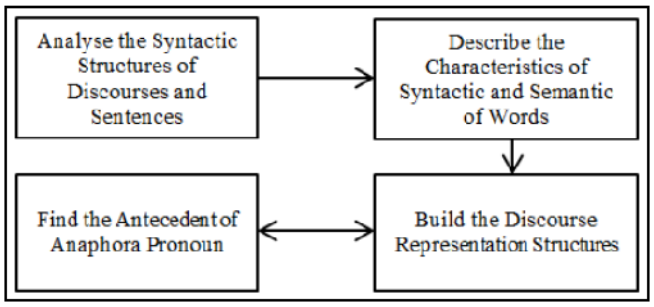

Figure 1. The general model for analyzing inter-sentential anaphoric pronoun.

Components of the system are demonstrated in more detail as follows:

4.1. Analysing the Syntactic Structures of Paragraphs and Sentences

This component with analyze the syntactic structure of the paragraph into sentences. Different from the model of [7], [9], in this component, we do not analyze using recursive method, instead clear separate into sentences and index the position for each sentence to distinguish the order of sentences in the paragraph. The analysis is performed by top-down rules as follows:

```
discourse --> statement_first, endpunct, statement_second,
endpunct, statement_third, endpunct.

discourse --> [].

endpunct --> ['.'].
```

Figure 2. Analyze the syntactic structure of paragraph into sentences using top-down method.

In this analysis, a paragraph is separated into three sentences. These sentences will be indexed the position order. In the description of analyzing syntactic structure of the paragraph into sentences by UBG in Prolog based on the model of [9], we define a flag `flag_position` to describe the position syntactic characteristic of each sentence in the paragraph. The characteristic `flag_position` of each sentence will take the value corresponding to the position of this sentence in the paragraph as follow:

```
discourse(D) --> {
                        S1 = syn~flag_position~[first],
                        D = sem~in~A,
                        S1 = sem~in~A,
                        S1 = sem~out~B,
                        S2 = syn~flag_position~[second],
                        S2 = sem~in~B,
                        S2 = sem~out~C,
```

```
            S3 = syn~flag_position~[third],
            S3 = sem~in~C,
            S3 = sem~out~E,
            D = sem~out~E
        },
        statement(S1),
        endpunct,
        statement(S2),
        endpunct,
        statement(S3),
        endpunct.
```

Figure 3. Analyze the syntactic structure of the paragraph into sentences in Prolog.

This component will analyze the constituent structure of each sentence into smaller constituents: noun phrases, verb phrases, lexical. In the process of analysis, based on the advantage of transferring data up and down between constituents of UBG, the position characteristic `flag_position` will be transferred to smaller constituents to determine the position of each constituent in the paragraph. In this research, we consider three types of simple sentence forming the paragraph will have the following constituent structures:

- Noun phrase + Verb phrase

Example 6: *"Lễ đọc sách."* (English: Lễ reads books.)

- Noun phrase + Adjective

Example 7: *"Lễ hạnh phúc."* (English: Lễ is happy.)

- Noun phrase + "là" (is) + Noun phrase

Example 8: *"Lễ là giám đốc."* (English: Lễ is manager.)

The analysis of the constituent structure of each sentence into smaller constituents will be performed by top-down rules as follow:

```
s --> np, vp.
s --> np, adj.
s --> np, [là], np.
np --> n(class:proper).
```

```
np --> [anh]; [anh ấy]; [cô]; [cô ấy]; [chị]; [chị ấy]; [ông];
[ông ấy]; [bà]; [bà ấy]; [em]; [em ấy]; [bạn]; [bạn ấy].

np --> n(class:common).

vp --> v, np.

vp --> v.
```

Figure 4. Analyze the constituent structure of each sentence into smaller constituents using top-down rules.

4.2. Describing the Syntactic and Semantic Characteristics of Words

After analyzing the constituent structure of each sentence into word level, this component will describe syntactic and semantic characteristics of each word depending on its category. Because of only considering three types of simple sentences described above, this component describes only words belong to three categories: noun, verb, adjective. To accordant with the strategy for solving the inter-sentential anaphoric pronoun proposed above, grammatical characteristics of noun are concentrated to be described: the unique index characteristic is defined exclusively for each noun, the position characteristic takes the value transferred from the position characteristic of the sentence, the characteristic indicates human or thing object (animal or non-animate object) is defined for each noun, the characteristic indicates the subject or object role of the noun with verb phrase will take the value transferred from the analysis of the structure of the sentence into noun phrase and verb phrase or the analysis of the structure of verb phrase into verb and noun phrase. The description of these characteristics helps to determine following points:

- Determine each object. Here we see proper nouns and common nouns are objects.
- Determine syntactic and semantic characteristics of each noun in the paragraph, become the premise for building the Discourse Representation Structure and determine which noun is the antecedent of the inter-sentential anaphoric pronoun.
- Determine actions and properties of each object in the paragraph.

In following table 2, we present syntactic and semantics characteristics of each category:

Table 2. Characteristics of Word Categories.

Word categories	**Syntactic characteristics**	**Semantic characteristics**
Noun	• The unique index i for each object. • The position index coincides with the position index of sentence, showing that the object appears in which sentence in paragraph. • The index indicates human or thing (animal or non-animate object). • The index indicates subject or object role of verb. • The index that distinguishes proper noun and common noun. Here we see proper nouns and common nouns are objects.	• Common meaning of the noun. • Describe the context of the DRS structure before and after considering the noun: adding the index of the object.
Verb	• The index that distinguishes transitive verb and intransitive verb. • The index denotes the arguments of the verb: o The first argument shows the subject of	• Common meaning of the verb. • Describe the context of the DRS structure before and

	the verb. o The second argument shows the object of the verb. o Intransitive verb has only first argument. Transitive verb has two arguments.	after considering the verb.
Adjective	• The index coincides with the index of the subject.	• Common meaning of the adjective. • Describe the context of the DRS structure before and after considering the adjective.

Characteristics of noun category will be illustrated through proper nouns and common nouns in the paragraph in Example 4 "*Nhân học môn vẽ. Anh dùng bút chì. Nghĩa hỏi anh.*" as follow:

Example 6: Consider proper noun "Nhân" at the first sentence.

- Syntactic characteristics:
  - The index `index` is generated uniquely for object "Nhân".
  - The index `flag_position` takes the value [first] transferred from the position index of the sentence in the analysis process of the structure of paragraph into consecutive sentences, indicates the position of the object is in the first sentence.
  - The index `flag_state` takes the value [subject] transferred from analysis process of the structure of sentence into noun phrase and verb phrase, indicates the role of noun "Nhân" is the subject of verb "học".
  - The index `flag_species` takes the value [human], indicates the object "Nhân" is human.
  - The index `class` takes the value [proper], indicates this is proper noun.
- Semantic characteristics:
  - This is the name of object "Nhân".
  - The context of the DRS structure after considering this proper noun, will add the index `index` of object "Nhân".

In Prolog, above characteristics of proper noun "Nhân" are described based on the model of [7], [9] as follows:

```
n(N) --> [nhân],
                {
                      unique_integer(I),
                      FSP = [human],
                      N = syn~   (index~I ..
                                 flag_position~FP ..
                                 flag_state~FST ..
                                 flag_species~FSP ..
                                 class~proper) ..
```

```
                    sem~ (in~  DRSList ..
                          out~ NewDRSList)
        }.
```

Figure 5. Describe characteristics of proper noun “Nhân” in Prolog

Example 7: Consider common noun “bút chì” (English: “pencil”) at the second sentence

- Syntactic characteristics:
  - The index `index` is generated uniquely for object “bút chì”.
  - The index `flag_position` takes the value [second] transferred from the position index of the sentence in the analysis process of the structure of paragraph into consecutive sentences, indicates the position of the object is in the second sentence.
  - The index `flag_state` takes the value [object] transferred from analysis process of the structure of verb phrase into verb and noun phrase, indicates the role of noun “bút chì” is the object of verb “dùng”.
  - The index `flag_species` takes the value [thing], indicates the object “bút chì” is the thing.
  - The index `class` takes the value [common], indicates this is common noun.
- Semantic characteristics:
  - The common meaning is to show that it is “bút chì”.
  - The context of the DRS structure after considering this common noun, will add index `index` of object “bút chì”.

In Prolog, above characteristics of common noun “bút chì” are described based on the model of [7], [9] as follows:

```
n(N) --> [bút,chì],
                {
                      unique_integer(I),
                      FSP = [thing],
                      N = syn~   (index~I ..
                                  flag_position~FP ..
                                  flag_state~FST ..
                                  flag_species~FSP ..
                                  class~common) ..
```

```
                    sem~ (in~ [drs(U,Con)|Super] ..

                          out~ [drs([I|U],NewCon)|Super])

}.
```

Figure 6. Describe characteristics of common noun "bút chì" in Prolog

4.3. Building the Discourse Representation Structure

After syntactic and semantic characteristics of each word had been described, this component will use these to build the Discourse Representation Structure – DRS of the paragraph by adding these characteristics into set U and set Con appropriately. Building the DRS structure helps to represent the meaning of the paragraph. The main idea here is determining each object and actions and properties of this object. This component will based on grammatical characteristics of each noun, adds the unique index to set U to determine each object, and then adds other characteristics of this noun to set Con to determine characteristics of this object. Syntactic and semantic characteristics of verb and adjective will be added to set Con to determine actions and properties of this object.

The building will be performed sequentially, right after described grammatical characteristics of each word, when perform the next sentence, will be based on the context of the DRS structure had been built from preceding sentences. Therefore, when specify the exact antecedent at one preceding sentence of the inter-sentential anaphoric pronoun, this component will add to set U and set Con actions and properties of this pronoun and associates with this antecedent.
In the following table 3, we present characteristics of predicates being added in the process of building the DRS structure.

Table 3. Build Set U and Set Con of the DRS Structure.

Word Categories	**Build Set U**	**Build Set Con**
Proper noun	• Add the index Index of the object.	• Add the characteristic: the object name condition. • Add the characteristic: the position condition of the object. • Add the characteristic: the subject or object role condition of the verb of the object. • Add the characteristic: indicate the object is human or thing.
Common noun	• Add the index Index of the object.	• Add the characteristic: the meaning condition of the object. • Add the characteristic: the position condition of the object. • Add the characteristic: the subject or object role condition of the verb of the object. • Add the characteristic: indicate the object is human or thing.
Transitive verb		• Add the characteristic: the meaning condition with two arguments: o The first argument shows the subject of the verb. o The second argument shows the object of the verb.
Intransitive verb		• Add the characteristic: the meaning condition with one argument: o This only one argument shows the subject of the verb.
Adjective		• Add the characteristic: the meaning condition with one argument:

		o This only argument shows the subject of the adjective.

The building DRS structure will be illustrated through adding syntactic and semantic characteristics of nouns described above as follow:

Consider proper noun “Nhân”:

- Build set U:
    - Add the `indexI` of object “Nhân”.
- Build set Con associated with `indexI`:
    - Add the characteristic: The object name condition – `named(I,[nhân])`.
    - Add the characteristic: The position condition of the object – `position(I,[first])`.
    - Add the characteristic: The subject role condition of the object – `state(I,[subject])`.
    - Add the characteristic: the human characteristic – `species(I,[human])`.

Consider common noun “bút chì” (English: “pencil”):

- Build set U:
    - Add the `indexI` of object “bút chì”.
- Build set Con associated with `indexI`:
    - Add the characteristic: The meaning condition of the object – `bút_chì(I)`.
    - Add the characteristic: The position condition of the object – `position(I,[second])`.
    - Add the characteristic: The object role condition of the object – `state(I,[object])`.
    - Add the characteristic: the thing characteristic – `species(I,[ thing])`.

4.4. Finding the Antecedent of the Inter-sentential Anaphoric Pronoun

This component is implemented together with the component building DRS structure, will find the antecedent of the inter-sentential anaphoric pronoun accordant with strategy proposed above, based on unique indexes and characteristic conditions of objects in set U and set Con of the DRS structure. The main idea here is to consider each object having unique index in set U and check characteristic conditions of this object in set Con agreement with conditions in strategy. Based on this idea, the algorithm finding the antecedent of inter-sentential anaphoric pronoun will be as follow:

- Consider the pronoun standing alone:

```
While (index I is in U)
    While (predicate associated with I is in Con)
        If ((position(I) < position(pronoun)) and
```

```
                        (state(I) is [subject]) and
                        (species(I) is [human]))
                Index of the antecedent = I
            End If
        End While
End While
```

Figure 7. The algorithm of finding antecedent for anaphoric pronoun standing alone

- Consider the pronoun standing with “ta” / “ấy” / “này” :

```
While (index I is in U)
    While (predicate associated with I is in Con)
        If ((position(I) < position(pronoun)) and
                        (state(I) is [object]) and
                        (species(I) is [human]))
                Index of the antecedent = I
        End If
    End While
End While
```

Figure 8. The algorithm of finding antecedent for anaphoric pronoun standing with “ta” / “ấy” / “này”

With two algorithms finding the antecedent for inter-sentential anaphoric pronoun described above and the component building DRS structure will perform continually after specify exact antecedent, the finding will have following prominent features:

- Limit considering objects to the time of considering the inter-sentential anaphoric pronoun, because only consider objects having unique index in set U.
- The found antecedent locates at the sentence which precedes the sentence containing the inter-sentential anaphoric pronoun. This is a different point from the model of [7], [9]: the antecedent can locate at the same sentence but precede the pronoun. The main idea here is indexing position for each sentence in the paragraph, so can determine the position of objects in the paragraph.
- The verb will take the found antecedent as the first or second argument, depending on the role of this antecedent is subject or object.
- The adjective will take the found antecedent as the argument.

The finding algorithm is applied for paragraph in Example 5 "*Lễ đọc sách. Anh thấy Chí. Anh ta đọc báo.*" (English: "Lễ reads book. He sees Chí. He reads newspaper.") as follows:

- Consider anaphoric pronoun "anh" stand alone at the second sentence:
    - Set U at the present time has 02 objects: X1 – Lễ, X2 – sách
    - Set Con at the present time has predicates: named(X1,[lễ]), position(X1,[first]), state(X1,[subject]), species(X1,[human]), sách(X2), position(X2,[first]), state(X2,[object]), species(X1,[thing])
    - Consider object X1:
        - The position of X1 is first < the position of "anh" is second.
        - The role of X1 is subject – subject of the verb.
        - X1 is human object.
    - At the result, the found antecedent of pronoun "anh" stand alone has the index X1.
    - Set Con will add predicate: đọc(X1,X2)
- Consider anaphoric pronoun "anh" stand with "ta" at the third sentence:
    - Set U at the present time has additional object: X3 – Chí
    - Set Con at the present time has additional predicates: named(X3,[chí]), position(X3,[second]), state(X3,[object]), species(X3,[human]), thấy(X1,X3)
    - Consider object X3:
        - The position of X3 is second < The position of "anh ta" is third.
        - The role of X3 is object – object of the verb.
        - X3 is human object.
    - At the result, the found antecedent of pronoun "and" stand with "ta" has the index X3.

The component building DRS structure will continue to perform and finally building the whole DRS structure of paragraph "*Lễ đọc sách. Anh thấy Chí. Anh ta đọc báo.*" as follow:

Table 4. The DRS structure of paragraph "*Lễ đọc sách. Anh thấy Chí. Anh ta đọc báo.*".

X1, X2, X3, X4
named(X1,[lễ])
position(X1,[first])
state(X1,[subject])
species(X1,[human])
sách(X2)
position(X2,[first])
state(X2,[object])
species(X1,[thing])
đọc(X1,X2)
named(X3,[chí])
position(X3,[second])
state(X3,[object])
species(X3,[human])
thấy(X1,X3)
báo(X4)
position(X4,[third])
state(X4,[object])
species(X4,[thing])
đọc(X3,X4)

5. DEVELOPMENT AND EVALUATION

We have tested 123 Vietnamese paragraphs satisfying characteristics described at Section 3. The system builds the DRS structure and determines the exact antecedent for anaphoric pronoun at 86 paragraphs. So, the successful rate is 70%. Analyze the result, we see that with the strategy and model proposed above, the paragraph that correctly identified will have following characteristics:

- Have only one human object that takes the subject role of the verb and appears at the sentence which precedes the sentence containing anaphoric pronoun standing alone.
- Have only one human object that takes the object role of the verb and appears at the sentence which precedes that sentence containing anaphoric pronoun standing with “ấy” / “ta” / “này”.
- Pronouns in paragraphs were defined in the system.

With Vietnamese paragraphs have not been successfully performed by the system, are divided into the following cases:

- Do not have any object appears before anaphoric pronoun. In these paragraphs, anaphoric pronoun can appear at the head of the first sentence, so there is no antecedent for these pronouns.
- There are more than one objects take the same subject or object role appear at the sentence precede the sentence containing the anaphoric pronoun. In this case, one pronoun standing alone or standing with “ấy” / “ta” / “nay”, there can be more than one candidate objects take the same subject or object role of verb at previous sentences. With current strategy, cannot determine exactly which candidate is the antecedent of anaphoric pronoun.
- There is pronoun “nó” (English: it) in the paragraph. In Vietnamese, pronoun “nó” can indicate the human object or thing (animal or non-animate object) depending on the context of the paragraph. In this research, we only consider nouns and pronouns indicate human object, so cannot solve pronoun “nó”.

The analysis show that the strategy and model proposed performed successfully for major of paragraphs.

6. CONCLUSION AND FUTURE WORK

We presented the strategy for solving inter-sentential anaphoric pronoun for Vietnamese paragraphs composing simple sentences. The strategy proposed based on syntactic and semantic characteristics of noun in Vietnamese and focus phenomenon into noun that takes the subject or object role of verb when considering the anaphoric pronoun standing alone or standing with “ấy” / “ta” / “này”. With this strategy, we built the system model with two algorithms implemented in Prolog, apply the model of [7], [9] with some differences and improvements accordant with strategy as follow:

- Only considering inter-sentential anaphoric pronouns, with the antecedent appears at the sentence precede the sentence containing pronoun.
- Do not analyse the paragraph into sentences using recursive method, instead determine the position of each sentence.
- There is no determiner in Vietnamese, so we remove the semantic role of determiner.
- Describe characteristics of words in Vietnamese grammar.

- The algorithms finding the antecedent for anaphoric pronoun based on the constraints of grammatical characteristics of noun: the position of noun in the sentence, the characteristic indicates that this is human object or thing (animal or non-animate object), the subject or object role of verb.
- Focusing to the antecedent is depending on the anaphoric pronoun standing alone or standing with "ấy" / "ta" / "này".

The strategy and model proposed performed successfully for major of Vietnamese paragraphs tested. The main reason is that we have based on grammar of words in Vietnamese, help to exactly analyze characteristics of paragraph, which propose appropriate treatment strategy. Analysing the experiment show that there are paragraphs that this strategy and model cannot resolve the inter-sentential anaphoric pronoun. This requires deeper understanding of Vietnamese linguistic theory so that can exactly analyse these cases. In future work, we continue to follow the current approach, further research on grammatical characteristics of lexical in Vietnamese paragraphs so that can resolve the inter-sentential anaphoric pronoun in following cases:

- Consider the thing (animal or non-animate object), not limit to human object. This requires the well research to resolve the pronoun "nó" in Vietnamese paragraphs.
- Consider the paragraphs that have more than one object take the same subject or object role of verb appear in the sentence precede the sentence containing anaphoric pronoun.

REFERENCES

[1] Aravind K. Joshi & Scott Weinstein, (1981) "Control of inference: Role of some aspects of discourse structure centering", *Proceedings of International Joint Conference on Artificial Intelligence (IJCAI'1981)*, pp385–387.

[2] Barbara J. Grosz, Aravind K. Joshi & Scott Weinstein, (1995) "Centering: a framework for modeling the local coherence of discourse", *Computational Linguistics*, Vol. 21, No. 2, pp203–225.

[3] Francis Cornish, (2009) "Inter-sentential anaphora and coherence relations in discourse: a perfect match", *Language Sciences*, Vol. 31, No. 5, pp572–592.

[4] Hans Kamp, "A theory of truth and semantic representation", (1981) *Formal methods in the study of language*, pp277–322, University of Amsterdam.

[5] Jaime G. Carbonell & Ralf D. Brown, (1988) "Anaphora Resolution: A Multi-Strategy Approach",*Proceedings of the 12th International Conference on Computational Linguistics*, pp96-101.

[6] José Abraços & José Gabriel Lopes, (1994) "Extending DRT with a focussing mechanism for pronominal anaphora and ellipsis resolution", *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pp1128-1132, Kyoto, Japan.

[7] Mark Johnson & Ewan Klein, (1986) "Discourse, anaphora and parsing", *Report No. CSLI-86-63*, Center for the Study of Language and Information, Stanford University, USA. Also in Proceedings of Coling86, pp669-675.

[8] Michael A. Covington, (2007) "GULP 4: An Extension of Prolog for Unification Based Grammar", *Research Report AI-1994-06*, Artificial Intelligence Center, The University of Georgia, USA.

[9] Michael A. Covington & Nora Schmitz, (1989) "An Implementation of Discourse Representation Theory", Advanced Computational Methods Center, The University of Georgia, USA.

[10] Michael A. Covington, Donald Nute, Nora Schmitz & David Goodman, (1988) "From English to Prolog via Discourse Representation Theory", *ACMC Research Report 01-0024*, The University of Georgia, USA.

[11] Michel Denber, (1998) "Automatic Resolution of Anaphora in English", Technical report, Eastman Kodak Co.

[12] Patrick Blackburn & Johan Bos, (1999) "Representation and Inference for Natural Language - Volume II: Working with Discourse Representation Structures", Department of Computational Linguistics, University of Saarland, Germany.

[13] Stuart M. Shieber, (2003) *An introduction to unification-based approaches to grammar*, Microtome Publishing Brookline, Massachusetts.

[14] Tyne Liang & Dian-Song Wu, (2004) "Automatic Pronominal Anaphora Resolution in English Texts", *Computational Linguistics and Chinese Language Processing*, Vol. 9, No.1, pp21-40.