AN APPROACH
TO PARSING VIETNAMESE NOUN COMPOUNDS

Dien Dinh   **Kiem Hoang**

Department of Information Technology
National University of HCM City
227 Nguyen Van Cu St., HCM city, Vietnam.
ddien@az.com.vn   hkiem@citd.edu.vn

Abstract

Noun Compound (NC for short) is a very popular phenomenon in many languages (especially in materials of Science & Technology) and it plays a very important role in understanding natural languages. So, this paper will present a semantic grammar approach to parsing Vietnamese Noun Compounds in the aspect of syntactic structures, semantic relation and automatic processing models. We also make use of the researches on the conceptual hierarchical systems WordNet and the case grammar of Fillmore. From the basis of this NC model, we can deduce many possible interpretations of Vietnamese NC.

1   Introduction

In studying English NCs, many sophisticated works have been carried out with various approaches, in which there were two following typical works: Mac Donald (refer [4]) made use of the semantic relational model based on the conceptual representation NETL and Fillmore's case grammar, and Mark Lauer (refer [5]) used the "Statistical Language Learning" model. In studying Vietnamese NCs, Nguyen Tai Can (refer [6]) mentioned to NCs while he wrote about the structure of Vietnamese Noun Phrases. He considered NCs as the short-form (head Noun and its modifiers) of the Noun Phrase.

In this paper, we apply the semantic grammar and Fillmore's case grammar which are similar to those of Mac Donald but we base ourselves on the American WordNet for English to build the NC-model for Vietnamese. The NC in this paper will be understood as a combination of one head Noun and its modifier(s) to form a compound which functions as a noun with its own meaning as a whole, such as: *hệ thống* máy tính (computer *system*). We won't examine NCs that have such idiomatic meanings as: *kỷ luật sắt* (iron discipline). The word order, syntactic structure, semantic relation of NCs in this paper will be the normal, unmarked forms.

2   Structure of Vietnamese Noun Compounds

In order to study the Vietnamese NCs, firstly, we must start from the organization of Vietnamese Noun Phrases (NP), and then we determine our boundary of Vietnamese NCs in NPs. In syntax aspect, the modifier always plays the modifying role to the head noun. In word-order aspect, the head noun is always in the first position, before the positions of modifiers. In semantics aspect, the modifier is used to name a thing

which is the feature of the thing identified by the head noun, or to point out a thing which is related to the thing identified by the head noun, this relation may be *possession, material, content, time, space attributes, reason, purpose*, etc. In order to recognize these implicit semantic relationships automatically, we need a semantically approach to extract ("parse") these possible relations.

3   Semantics of Vietnamese Noun Compounds

This paper will take advantage of the conceptual tree in WordNet to construct the Vietnamese NC. WordNet system, however, was built for English words, so, we must re-compile into Vietnamese conceptual tree with the same original ideas. The conceptual system here is the knowledge-based system constructed on the hierarchical model consisting of nodes and relations. The model consists of 2 main types: concept and relation.
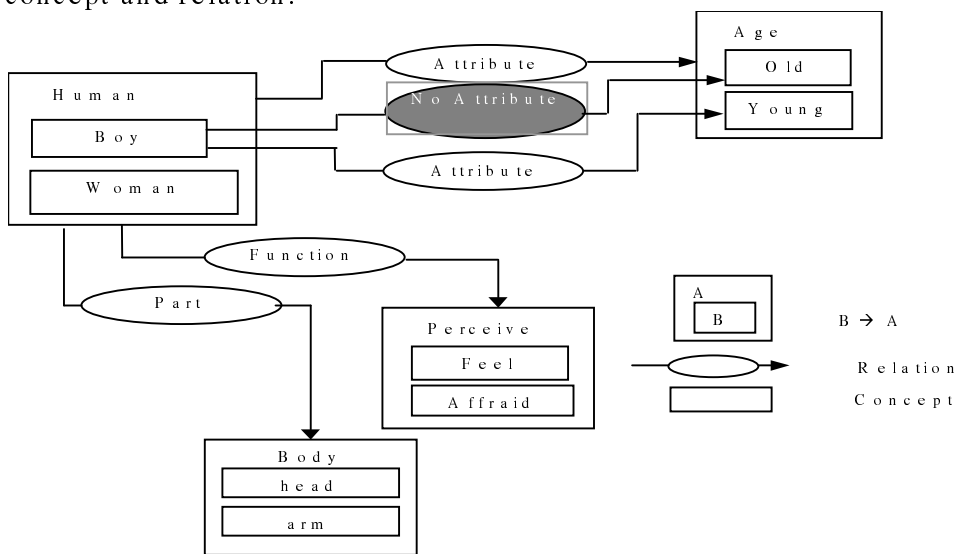

Figure 1: *Semantic network of typical concepts and relations.*

In figure 1, there are main relations as follows: *Has_Attribute* (point to Adjectives), *Has_Function* (point to Verbs), *Has_part* (point to Nouns), etc. We also see the concepts: *Human, Boy, Woman, Perceive, Feel, Afraid, Age, Old, Young,* and relationships between concepts: *Has_Attribute, Has_Part*, etc.

4   Vietnamese Noun Compound Model

In this model, we first organize the conceptual dictionary as follows: each word will be listed with all its possible parts-of-speech, in each of which will consist of distinguished meanings which will be linked (pointed) to an object – its corresponding semantic frame structure. Then, following the fields in each object, we determine possible semantic relationships between 2 objects and finally we choose the most reasonable relationship between them.

Firstly, we build objects with its semantic frame structure for each meaning of each word of each part-of-speech (N, V, Adj), then using case grammar (refer [3]) to parse Vietnamese NCs.

4.1 Parsing a typical Vietnamese NC

For example, considering the NC: "sân gạch" consisting of 2 words: "sân" and "gạch" . We will go through 3 following steps:

Step 1: to look up Vietnamese dictionary to get parts-of-speech and its meaning(s):

- Word "sân" in Vietnamese has only one part-of-speech with the meaning as: "*the land around the house or place for amusement, sports,...*" (pls.refer [3]).
- Word "gạch" has 2 parts-of-speech with 2 main meanings as a Noun: 1. *Material for constructing* 2. *Matter of (liver, egg) in crabs*; with 1 meaning as a Verb: to *make a line.*
- Basing ourselves on syntactical parsing of Vietnamese NC, we know that: the 1st word ("sân") must be a Noun and the head Noun (aspect of grammar as well as semantics), and the 2nd word ("gạch") may be Noun, Verb, Adjective. So, we only examine above-mentioned cases.

Step 2: to examine the semantic frame of words in NC:

- The semantic frame of Noun "sân" and "gạch" (meaning 1 and 2).
- The semantic frame of Verb "gạch".

Step 3: to examine the semantic relation of words in NC:

- To examine the semantic relation between the head noun "sân" (sân_dt) and the 1st meaning of the sub-noun "gạch" (gạch_dt1), then the 2nd meaning of the sub-noun "gạch" (gạch_dt2), and finally the meaning of verb "gạch".
- If the relationship between the head noun and its modifiers is closer or clearer, this relation will be evaluated with higher marks (higher potential). Finally, we choose the highest-mark relation. For example: in the above-mentioned case, the material relation "yard *made of* brick" with highest mark is chosen.

4.2 Semantic Relations between Objects

4.2.1 The relation between Noun and Noun:

- If the head noun matches the field "Meronym" or "Hyponym" of the sub-noun, we have the PART-OF relation. Ex: "chân bàn" (chân *của* cái bàn: leg *of* table).
- If the sub-noun is related to (e.g.: Objective, Reason, Purpose, … of Verbs) the field "Function" of the head noun, we have the PURPOSE relation. Ex: "ly nước" (water glass) => "ly (glass) *chứa* (to contain) nước (water)".
- If the head noun belongs to class "material object", its sub-noun modifies the material attributes, external forms of the head noun, e.g.: "sân gạch" (brick yard) => "sân"(yard) *làm bằng* (made of) "gạch" brick;...
- If the sub-noun is noun or pronoun of human, its relation is often POSSESSION relation, e.g.: "sách giáo viên" ("sách" *của* "giáo viên" : book *of* teacher).
- In addition, there are other relations between 2 nouns, e.g.: "hệ thống máy tính" (computer system) => "hệ thống (system) *bao gồm các* (consists of) máy tính" computer: the CONTENT relation.

4.2.2 The relation between Noun and Verb:

- If the verb matches the field "Function" of the head-noun, we have the PURPOSE relation. Ex: "sân chơi" (sân *để* chơi : yard *to* play).
- If the head N is the nominalization of V ("sự", "việc",..), the sub-N will be related to the head N through roles in Case Grammar of that nominalized V. For example: "việc sưu tầm tem" (stamp collection) => "tem" (stamp) is the *object* of "sưu tầm" (to collect").

4.2.3 The relation between Noun and Adjective:

- If the head noun belongs to class "material object", its adjective modifies the material attributes or the external forms of the head noun, e.g.: "cái bàn dài" (long table) => "cái bàn" (table) has size "dài" (long), "hoa vàng" (yellow flower) => "hoa" (flower) has color "vàng" (yellow),...
- If the head noun belongs to class "abstract thing", its adjective modifies the spiritual attributes or the internal content of the head noun, e.g.: "bộ phim hay" (interesting film) => "bộ phim" (film) has content "hay"(interesting),...
- If its adjective is Descriptive, it will describe the attributes of the noun which it modifies, e.g.: "sách đẹp" (pretty book); If its adjective is Relational, it will describe the relation between the head noun and the thing which that adjective indicates. e.g.: *"Người bạn cũ"* (old friend) => *the relation (beween I and my friend) is old*.

The semantic relationship between the head noun and its modifier is determined by the linkages between the semantic frame of each meaning of each word. In each semantic frame, there are pointers indicating to the related objects (e.g.: link to Hypernym, Hyponym, Meronym, Function, Agent, …). We will base ourselves on the measurement of distance of these linkages to determine whether its semantic relation is close or far in order to choose the most reasonable relation.

5   Conclusion

The above model is only a suggestion to parsing Vietnamese NC, which bases itself on the semantic relation model in English NC and semantic database of English words (WordNet). In order to complete this model, however, we must solve many difficult issues, such as: to re-build the semantic database for Vietnamese words, which will consume much time and effort. We also must make statistical calculation for all forms of Vietnamese NC and determine the semantic relation for each type and limit ourselves to types that can be solved by this model.

References

[1] Đỗ Hữu Châu. *Lexicography and Semantics of Vietnamese*. Publisher of National University of Hà Nội, VN (in Vietnamese), 1997.

[2] George A.Miller. *Introduction to WordNet*. (online lexical database) http://www.cogsci.princeton.edu/~wn/5Papers.pdf . Princeton, USA. 1996.

[3] Hoàng Phê. *Vietnamese Dictionary*. Center of Lexicography, Publisher of Đà Nẵng (in Vietnamese), VN, 1998.

[4] Mac Donald. *Understanding Noun Compounds*. Carnegie-Mellon Uni, Pittsburgh, PA, USA, 1982.

[5] Mark Lauer. *Designing SLL: experiments on Noun Compounds*. Macquarie Uni NSW, Australia (PhD thesis), 1995.

[6] Nguyễn Tài Cẩn. *Vietnamese Syntax*. Publisher of National University of Hà Nội, VN (in Vietnamese), 1998.