# Automatic Semantic Role Assignment for a Tree Structure

**Jia-Ming You**
Institute of Information Science
Academia Sinica
swimming@hp.iis.sinica.edu.tw

**Keh-Jiann Chen**
Institute of Information Science
Academia Sinica
Kchen@iis.sinica.edu.tw

**Abstract**

We present an automatic semantic roles labeling system for structured trees of Chinese sentences. It adopts dependency decision making and example-based approaches. The training data and extracted examples are from the Sinica Treebank, which is a Chinese Treebank with semantic role assigned for each constituent. It used 74 abstract semantic roles including thematic roles, such as 'agent'; 'theme', 'instrument', and secondary roles of 'location', 'time', 'manner' and roles for nominal modifiers. The design of role assignment algorithm is based on the different decision features, such as head-argument/modifier, case makers, sentence structures etc. It labels semantic roles of parsed sentences. Therefore the practical performance of the system depends on a good parser which labels the right structures of sentences. The system achieves 92.71% accuracy in labeling the semantic roles for pre-structure- bracketed texts which is considerably higher than the simple method using probabilistic model of head-modifier relations.

## 1. Introduction

For natural language understanding, the process of fine-grain semantic role assignment is one of the prominent steps, which provides semantic relations between constituents. The sense and sense relations between constituents are core meaning of a sentence.

Conventionally there are two kinds of methods for role assignments, one is using only statistical information (Gildea and Jurafsky, 2002) and the other is combining with grammar rules (Gildea and Hockenmaier, 2003). However using only grammar rules to assign semantic roles could lead to low coverage. On the other hand, performance of statistical methods relies on significant dependent features. Data driven is a suitable strategy for semantic roles assignments of general texts. We use the Sinica Treebank as information resource because of its various domains texts including politics, society, literature…etc and it is a Chinese Treebank with semantic role assigned for each constituent (Chen etc., 2003). It used 74 abstract semantic roles including thematic roles, such as 'agent'; 'theme', 'instrument', and secondary roles of 'location', 'time', 'manner' and modifiers of nouns, such as 'quantifier', 'predication', 'possessor', etc. The design of role assignment algorithm is based on the different decision features, such as head-argument/modifier, case makers, sentence structures etc. It labels semantic roles of parsed sentences by example-based probabilistic models.

### 1.1 Sinica Treebank

The Sinica Treebank has been developed and released to public since 2000 by Chinese Knowledge Information Processing (CKIP) group at Academia Sinica. The Sinica Treebank version 2.0 contains 38944 structural trees and 240,979 words in Chinese. Each structural tree is annotated with words, part-of-speech of words, syntactic structure brackets, and semantic roles. For conventional structural trees, only syntactic information was annotated. However, it is very important and yet difficult for Chinese to identify word relations with purely syntactic constraints (Xia et al., 2000). Thus, partial semantic information, i.e. semantic role for each constituent, was annotated in Chinese structural trees. The grammatical constraints are expressed in terms of linear order of semantic roles and their syntactic and semantic restrictions. Below is an example sentence of the Sinica Treebank.

Original sentence:
*他 'Ta'要 'yao' 張三 'ZhangSan'撿 'jian' 球 'qiu'。*
***He ask Zhang San to pick up the ball.***
Parsed tree:
*S(agent:NP(Head:Nhaa:他'He')|Head:VF2:要'ask'*
*|goal:NP(Head:Nba:張三'Zhang San') |theme:VP(Head:VC2:撿'pick'|goal:NP(Head:Nab:球'ball')))*

Figure 1: An example sentence of Sinica Treebank

In the Sinica Treebank, not only the semantic relations of a verbal predicate but also the modifier head relations were marked. There are 74 different semantic roles, i.e. the task of semantic role assignment has to establish the semantic relations among phrasal heads and their arguments/modifiers within 74 different choices. The set of semantic roles used in the Sinica Treebank is listed in the appendix.

## 2. Example-based Probabilistic Models for Assigning Semantic Roles

The idea of example-based approaches is that semantic roles are preserved for the same event frames. For a target sentence, if we can find same examples in the training corpus, we can assign the same semantic role for each constituent of the target sentence as the examples. However reoccurrence of exact same surface structures for a sentence is very rare, i.e. the probability of finding same example sentences in a corpus is very low. In fact, by observing structures of parsed trees, we find that most of semantic roles are uniquely determined by semantic relations between phrasal heads and their arguments/modifiers and semantic relations are determined by syntactic category, semantic class of related words. For example:

Original sentence:
*我們 'wo men' 都 'du' 喜歡 'xi huan' 蝴蝶 'hu die'。*
***We all like butterflies.***
Parsed tree:
***S(experiencer:NP(Head:Nhaa:我們 'we')|quantity:Dab:都 'all'|Head:VK1:喜歡 'like'***
***|goal:NP(Head:Nab:蝴蝶 'butterflies'))。***

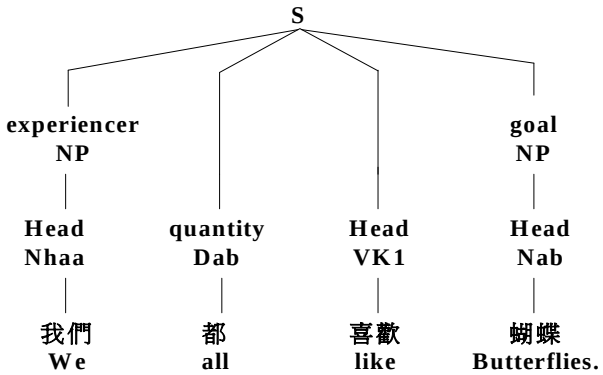

Figure 2: The illustration of the parsed tree.

In Figure2, 喜歡 'like' is the sentential head; 我們 'we' and 蝴蝶 'butterflies' are the arguments; 都 'all' is the modifier. As a result, the semantic role 'experiencer' of 我們 'we' is deduced from the relation between 我們 'we' and 喜歡 'like', since the event frame of 喜歡 'like' has the two arguments of experiencer and goal and the experiencer usually takes the subject position. The semantic roles of 蝴蝶 'butterflies' and 都 'all' are assigned by the same way. For the task of automatic role assignment, once phrase boundaries and phrasal head are known, the semantic relations will be resolved by looking for similar head-argument/modifier pairs in training data.

### 2.1 Example Exaction

To extract head-argument/modifier examples from the Sinica Treebank is trivial, since phrase boundaries and semantic roles, including phrasal head, are labeled. The extracted examples are pairs of head word and target word. The target word is represented by the head of the argument/modifier, since the semantic relations are established between the phrasal head and the head of argument/modifier phrase. An extracted word pair includes the following features.

Target word:
The head word of argument/modifier.
Target POS:
The part-of-speech of the target word.
Target semantic role:
Semantic role of the constituent contains the target word as phrasal head.
Head word:
The phrasal head.

Head POS:
The part-of-speech of the head word.
Phrase type:
The phrase which contains the head word and the constituent containing target word.
Position:
Shows whether target word appears before or after head word.

The examples we extracted from Figure 2 are listed below.

| Target thematic Role | experiencer | quantify | goal |
|---|---|---|---|
| Target word | 我們 'we' | 都 'all' | 蝴蝶 'butterflies' |
| Target POS | Nhaa | Dab | Nab |
| Head word | 喜歡 'like' | 喜歡 'like' | 喜歡 'like' |
| Head POS | VK1 | VK1 | VK1 |
| Phrase type | S | S | S |
| Position | before | before | after |

Table 1: The three head-argument/modifier pairs extracted from Figure 2.

### 2.2 Probabilistic Model for Semantic Role Assignment

It is possible that conflicting examples (or ambiguous role assignments) occur in the training data. We like to assign the most probable roles. The probability of each semantic role in a constituent with different features combinations are estimated from extract examples.

$$P(r \mid constituent)$$
$$= P(\,r \mid h, h\_pos, t, t\_pos, pt, position\,)^{1}$$
$$= \frac{\#(r, h, h\_pos, t, t\_pos, pt, position)}{\#(h, h\_pos, t, t\_pos, pt, position)}$$

Due to the sparseness of the training data, it's not possible to have example feature combinations matched all input cases. Therefore the similar examples will be matched. A back off process will be carried out to reduce feature constraints during the example matching. We will evaluate performances for various features combinations to see which features combinations are best suited for semantic roles assignments.

We choose four different feature combinations. Each has relatively high accuracy. The four classifiers will be back off in sequence. If none of the four classifiers is applicable, a baseline model of assigning the most common semantic role of target word is applied.

**if # of (h,h_pos,t,t_pos,pt,position) > threshold**
**P(r|constituent)=P(r|h,h_pos,t,t_pos,pt,position)**
**Else**
**if # of (h_pos,t,t_pos,pt,position) > threshold**
**P(r|constituent)=P(r|h_pos,t,t_pos,pt,position)**
**Else**
**if # of (h,h_pos,t_pos,pt,position) > threshold**
**P(r|constituent)=P(r|h,h_pos,t_pos,pt,position)**
**Else**
**if # of (h_pos,t_pos,pt,position) > threshold**
**P(r|constituent)=P(r|h_pos,t_pos,pt,position)**
**Else**
**Baseline model:**
P(r|constituent)=P(r| t, t_pos,pt)

## 3. Experiments

We adopt the Sinica Treebank as both training and testing data. It contains about 40,000 parsed sentences. We use 35,000 sentences as training data and the rest 5,000 as testing data. The table 2 shows the coverage of each classifier, their accuracies, and performance of each individual classifier without back off process. The table 3 shows combined performance of the four classifiers after back off processes in sequence. The baseline algorithm is the simple unigram approach to assign the most common role for the target word.

[1] r: semantic role; h: the head word; h_pos: part-of-speech of head word; t: the target word; t_pos: part-of-speech of target word; pt: the phrase type.

Because the accuracy of the four classifiers is considerably high, instead of using linear probability combinations we will rather use the most reliable classifier for each different features combination

| Classifiers | Coverage | Accuracy | Performance |
|---|---|---|---|
| P(rl h, h_pos, t, t_pos, pt, position) | 23.02% | 96.60% | 22.24% |
| P(rl h_pos, t, t_pos, pt, position) | 56.11% | 95.53% | 53.60% |
| P(rl h, h_pos, t_pos, pt, position) | 52.98% | 91.45% | 48.45% |
| P(rl h_pos, t_pos, pt position) | 97.38% | 89.87% | 87.52% |

Table 2: Coverage and accuracy of different features combinations

| Method | Accuracy |
|---|---|
| Backoff | 90.29% |
| Baseline: | 68.68% |

Table 3: The accuracy of our backoff method and the base line (the most common semantic roles)

### 3.1 Error Analyses

Although the accuracy of back off model is relatively high to the baseline model, it still has quite a room for improvement. After analyzed the errors, we draw following conclusions.

a) Semantic head vs. syntactic head

A semantic role for a prepositional phrase (PP) is mainly determined by the syntactic head of PP, i.e. preposition, and the semantic head of PP, i.e. the head word of the DUMMY-argument of PP. For example, in Figure 3, the two sentences are almost the same, only the contents of PP are different. Obviously, the semantic roles of PP (在 'in' 印尼 'Indonesia') is location, and the semantic role of PP (在 'in' 今年 'this year') is time. Therefore the semantic roles of the two PPs should be determined only within the scope of PP and not relevant to matrix verb.

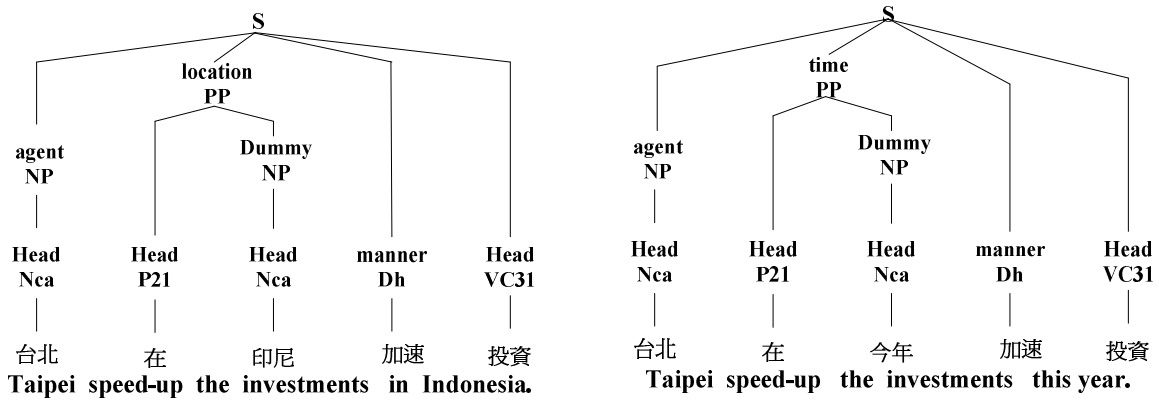


Figure 3: Parsed trees of "台北在印尼加速投資" and "台北在今年加速投資"

b) Structure-dependent semantic roles assignments

Complex structures are always the hardest part of semantic roles assignments. For example, the sentences with passive voice are the typical complex structures. In Figure 4, the semantic role of 蝴蝶 'Butterflies' is not solely determined by the head verb 吸引 'attracted' and itself. Instead we should inspect the existence of passive voice and then reverse the roles of subject and object.

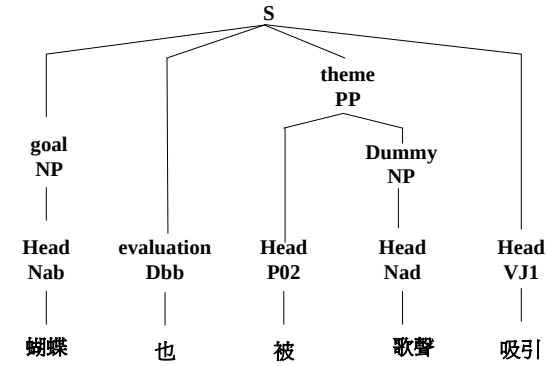

Figure 4: A parsed tree of passive sentence "蝴蝶也被歌聲吸引"

## 4 Refined Models

Chen & Huang (1996) had studied the task of semantic assignment during Chinese sentence parsing. They concluded that semantic roles are determined by the following 4 parameters.

1. Syntactic and semantic categories of the target word,
2. Case markers, i.e. prepositions and postpositions
3. Phrasal head, and
4. Sub-categorization frame and its syntactic patterns.

Therefore head-modifier/argument examples only resolve most of semantic role assignments. Some of complex cases need other parameters to determine their semantic roles. For instance, the argument roles of Bei sentences (passive sentence) should be determined by all four parameters.

The refined model contain two parts, one is the refinements of features data which provide more precisely information and the other is the improvements of back off process to deal with special semantic roles assignments.

### 4.1 Refinement of Features Extractions

The refinements of features extractions focus on two different cases, one is the features extractions of case-marked structures, such as PP and GP (postpositional phrases), and the other is the general semantic class identifications of synonyms.

The features of PP/GP include two different feature types: the internal and the external features. The internal features of phrases compose of phrasal head and Dummy-head; the external features are heads (main verbs) of the target phrases.

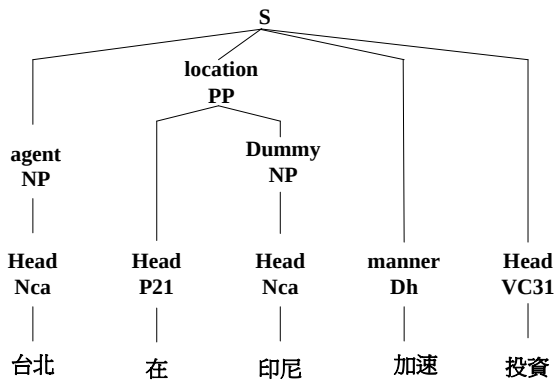

Figure 5: A parsed tree for demonstrating features extractions of PP

| | Internal relation | External relation |
|---|---|---|
| Target thematic role | Dummy | location |
| Target word | 印尼 'Indonesia' | 在..印尼 'in..Indonesia' |
| Target POS | Nca | P21 |
| Head word | 在 'in' | 投資 'invest' |
| Head POS | P21 | VC31 |
| Phrase type | PP | S |
| Position | after | before |

Table 4: The internal/external relations of Figure 5.

The semantic class identifications of synonyms are crucial for solving data sparseness problems. Some type of words are very productive, such as numbers, DM (determinative measurement), proper names. They need to be classified into different semantic classes. We use some tricks to classify them into specific word classes. For example we label 1 公斤 'one kilogram', 2 公斤 'two kilograms' as their canonical form 某公斤 'n kilograms'; 第一天 'the first day', 第二天 'the second day ' as 第某天 'the nth days'; 張三 'Zhang San', 李四 'Li Si' as a personal name…etc. With this method, we can increase the number of matched examples and resolve the problem of occurrences of unknown words in a large scale.

### 4.2 Dependency Decisions and Refined Back off Processes

The refined back off model aimed to solve semantic roles assignments for certain special structures. Using only head-modifier features could result into decision making with insufficient information. As illustrated before, the semantic role of 蝴蝶 'butterflies' in Figure 4 is 'agent' observed from the head-argument feature. But in fact the feature of passive voice 被 'passive' tells us that the subject role of 蝴蝶 'butterflies' should be the semantic role 'goal' instead of the usual role of 'agent'.

Therefore we enhanced our back off process by adding some dependency decisions. The dependency conditions include special grammar usage like passive form, quotation, topical sentences… etc. In the refined back off process, first we have to detect which dependency condition is happened and resolved it by using dependency features. For example, if the feature word 被 'passive' occurs in a sentence, we realize that the subjective priority of semantic roles should be reversed. For instance, 'goal' will take subject position instead of 'agent' ('goal' appears before 'agent').

### 4.3 Experiment Results

The experiments were carried out for the refined back off model with the same set of training data and testing data as in the previous experiments. Table 5 shows that the refined back off model gains 2.4 % accuracy rate than the original back off model. However most of the improvement is due to the refinements of features extractions and canonical representation for certain classes of words. A few improvements were contributed to the decision making on the cases of structure dependency.

| Method | Accuracy |
|---|---|
| Refined Backoff | 92.71% |
| Backoff | 90.29% |
| Baseline | 68.68% |

Table 5: Role assignment accuracies of refined backoff, backoff, and baseline models.

## 5 Conclusion and Future Works

Semantic roles are determined by the following 4 parameters.

1. Syntactic and semantic categories of the target word,
2. Case markers, i.e. prepositions and postpositions,
3. Phrasal head, and
4. Sub-categorization frame and its syntactic patterns.

We present an automatic semantic roles labeling system. It adopts dependency decision making and example-based approaches, which makes decision on the amount of parameters by observing the occurrence of dependency features and to utilize the minimal amount of feature combinations to assign semantic roles. It labels semantic roles of parsed sentences. Therefore the practical performance of the system depends on a good parser which labels the right structures of sentences. The system achieves 92.71% accuracy in labeling the semantic roles for pre-structure- bracketed texts which is considerably higher than the simple method using probabilistic model of head-modifier relations.

In the future, we will consider fine-grain semantic role assignment problems. The current semantic roles assignment is focus on one sentence. However, the occurrences of frame elements are not limited to a single sentence. For instance, "John bought the books from Mary". The semantic roles of 'John' and 'Mary' are agent and theme respectively. According to Fillmore's FrameNet, the frame element assignment for the above sentence should be 'John' the buyer, 'Mary' the seller, 'the books' the goods. The precondition of buy-frame says that the seller should be the owner of the goods. Therefore after the sentence parsing and logical reasoning, the following semantic relations should be established.

Event frame: Commerce-buy
Buyer: John
Seller: Mary
Goods: books
Additional frame: Own

Before the buy event
Owner: Mary
Possession: books
After the buy event
Owner: John
Possession: books

The semantic roles assignment is a process of crossing phrasal and sentential boundaries. Some semantic roles of an event might occur at left or right context. Therefore we have to analyze the relation between two consecutive events. The relations include causal relation, temporal relation, resultant relation, etc. How to resolve the above problems will be our future studies.

## References


Chen, Keh-Jiann, Chu-Ren Huang. 1996. *Information-based Case Grammar: A Unification-based Formalism for Parsing Chinese*. *Journal of Chinese Linguistics Monograph* Series No. 9.

Chen, Keh-Jiann, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo,Ming-Chung Chang, Chao-Jan Chen, and Zhao-Ming Gao, 2003. *Sinica Treebank: Design Criteria, Representational Issues and Implementation*. In Anne Abeille (Ed.) *Treebanks Building and Using Parsed Corpora*. Language and Speech series. Dordrecht:Kluwer, pp231-248.

Chu-Ren Huang, Keh-Jiann Chen, and Benjamin K. T'sou Eds. *Readings in Chinese Natural Language Processing*. 23-45. Berkeley: JCL.

Daniel Gildea and Daniel Jurafsky. 2002. *Automatic Labeling of Semantic Roles*. *Computational Linguistics*, 28(3):245-288

Daniel Gildea and Julia Hockenmaier. 2003. *Identifying Semantic Roles Using Combinatory Categorial Grammar.* Conference on Empirical Methods in Natural Language Processing (EMNLP).

Xia, Fei, 2000, *The Part-of-Speech Tagging Guidelines for the Penn Chinese Treebank (3.0)*. IRCS Report 00-07. Philadelphia, PA: University of Pennsylvania.

Appendix:

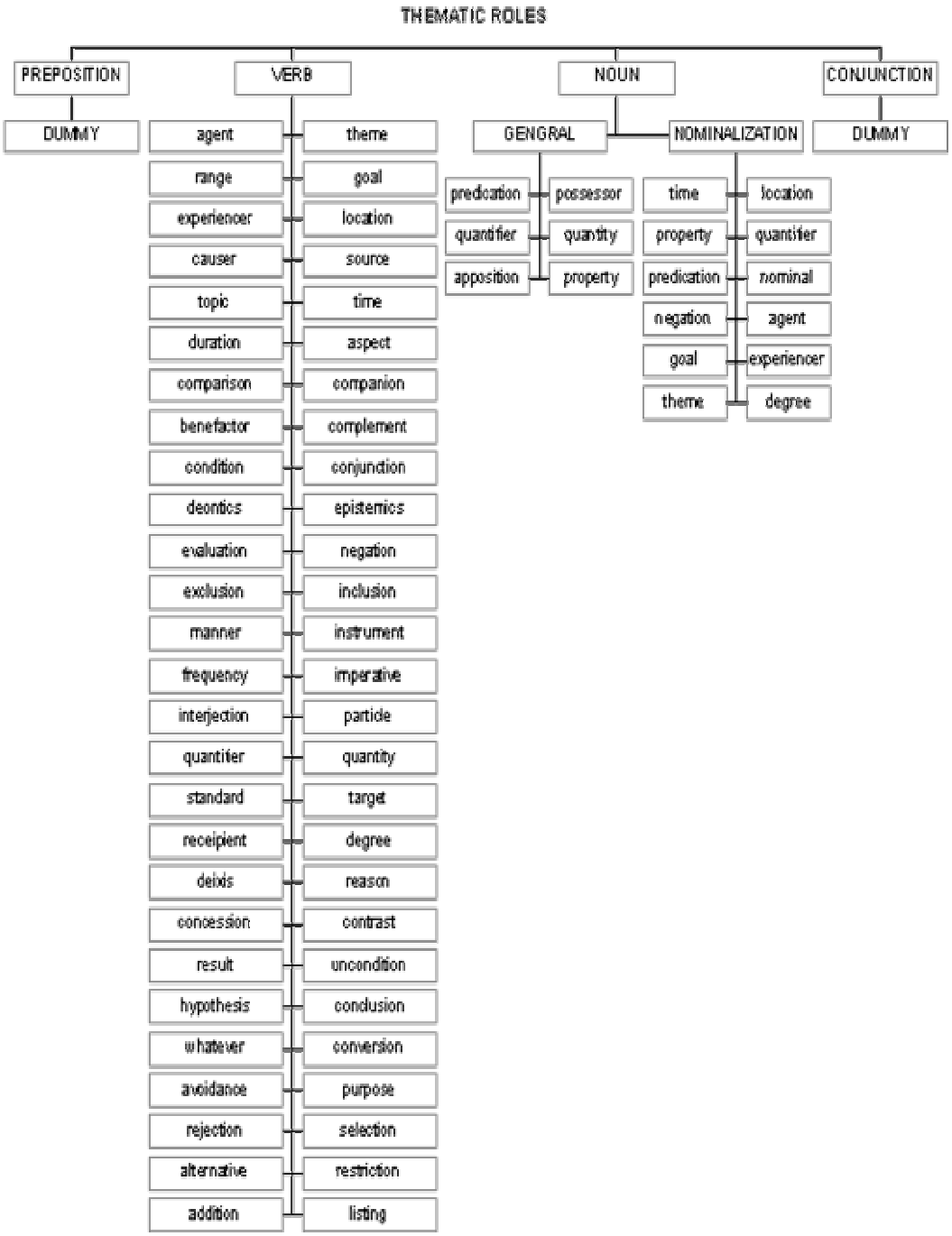


Figure 6: The detail classification of semantic roles in the Sinica Treebank