A Large-Scale Semantic Structure for Chinese Sentences

Tang Li
Institutue for Infocomm Research
21 Heng Mui Keng Terrace
Singapore119613
Tangli@I2R.a-star.edu.sg

Ji Donghong, Yang Lingpeng
Institutue for Infocomm Research
21 Heng Mui Keng Terrace
Singapore119613
{dhji, lpyang}@I2R.a-star.edu.sg

Abstract

Motivated by a systematic analysis of Chinese semantic relationships, we constructed a Chinese semantic framework based on surface syntactic relationships, deep semantic relationships and feature structure to express dependencies between lexical meanings and conceptual structures, and relations that underlie those lexical meanings. Analyzing the semantic representations of 10000 Chinese sentences, we provide a model of semantically and syntactically annotated sentences from which reliable information on combinatorial possibilities of each semantic item targeted for analysis can be displayed. We also propose a semantic argument – head relation, 'basic conceptual structure' and the 'Head-Driven Principle'. Our results show that we can successfully disambiguate some troublesome sentences, and minimize the redundancy in language knowledge descriptions for natural language processing.

1 Introduction

To enable computer-based analysis of Chinese sentences in natural language texts we have developed a semantic framework, taking into account concepts used in the Berkeley FrameNet Project (Baker, Fillmore, & Lowe 1998; Fillmore & Baker 2001) and the Penn Chinese Tree Bank (Nianwen Xue; Fei Xia et al. 2000). The FrameNet Project, as a computational project, is creating a lexical resource for English, based on the principle of semantic frames. It has tried to concentrate on frames which help to explain the meanings of groups of words, rather than frames that cover just one word. The representation of the valences of its target words and descriptions of the semantic frames underlying the meanings of the words described are the mainly part of the database. The Penn Chinese Tree Bank analyzed the syntactic structure of a phrase or sentence for selected text, based on the current research in Chinese syntax and the linguistic expertise of those involved in this project. Different from Pan's syntactic structures and FrameNet's semantic frames, our object is to record exactly how the semantic features relates frames to those syntactic constituents. The key task is to determine the relationship between the two direct constituents in terms of the semantic relationship. The grammar functions are also considered for primarily identifying the relation. Here, we use methods developed for the analysis of semantic relationships to produce a framework based on the direct component link. Our framework is largely a semantic one, but it has adopted some crucial principles of syntactic analysis in the semantic structure analysis.

In this paper, we present our model of semantically and synactically annotated 10000 Chinese sentences. The focus is on the analysis of the semantic relationships between one word to another in a sentence. We also briefly discuss the annotation process.

2 Theoretical Framework and Case Study

The basic assumption of Frame Semantics (Fillmore 1976;1977; Fillmore & Atkins 1992; Petruck 1996) as it applies to the description of lexical meanings is that each word (in a given meaning) evokes a particular frame and possibly profiles some element or aspect of that frame. By being linked to frames, each word is directly connected with other words in its frame(s). where word dependence association are needed from surface syntactic structures which actually reflect the grammatical relationship to the deep semantics structure whereby semantic content are put into natural language. The meaning of a word, in most cases, is best demonstrated by reference to a semantic network. Referential meaning on its own is insufficient. Word meaning would include the other dimensions concerning the structure and function of words. Unlike English, in which there are two major types of evidence that help to determine the syntactic structure of a phrase or sentence: morphological information and distributional information (such as word order) , in Chinese the lack of conclusive morphological cues makes ambiguity analyses for one sentence more likely. Moreover, most Chinese sentences order are very flexible. Phrase omission, word movement,

ellipsis and binding also make it difficult to characterize their grammatical relation. So the semantic information provides important clues for Chinese sentence analyse. We have to rely on semantic knowledge to guide role assignment. Thus, we propose a method allowing a syntactic and semantic-based analysis of sequences and relationship of semantic items to obtain the common distribution of the relationship order.

3 Method

The analysis method that will be presented here is logically equivalent to the parsing of syntax and semantic dependency with feature constraints.

The key idea in our method is to avoid the complexity hierarchical tree sturcture. We are concerned with building structures that reflect basic relationships between one word and other in a single sentence. We use methods developed for the analysis of semantic relationships to produce a framework based on the order link. We started from an initial analysis based on the surface syntactics, then we analyzed deep semantic relationships, and attempted to improve it by removing the semantic order from the syntactic structure and reconnecting them in different places. Since many word phrase patterns are difficult for computers to recognize, trying to compromise between linguistic correctness and engineering convenience, we link the difference semantic roles on the flat level, while employing a few template rules. All semantic words are linked on the same level. They are non-hierarchical constructs. This flatted representation allows access to various levels of syntactic description tree simultaneously. In fact, the purpose of generalization is to get a regular expression from the original sentence.

We manually tagged two kind of relationship among our large-scale frameworks: 1. syntax-semantic relationship; 2. semantic feature relationship.

Our framework consists of a set of nodes and a set of arcs that join the nodes, with each word or concept corresponding to a node and links between any two nodes that are directly associated. The basic links in the framework are between one word item to another based on immediate semantic deperdency order. We summarized the immediate semantic relationship through a variety of semantic relation features such as agent, reason, result and so on. The feature of relationship between two nodes are labeled on the arc.

We developed the first fully instantiated semantic structure by manually labeling semantic representations in a machine-readable format. To make sure that our model can deal with various kinds of texts in real life situations, we have analysed 10000 sentences from large Web site corpora based on our formal model. Our aim is not to describe in detail any specific, but to capture at an abstract level the semantic relations between the direct components in a sentence. Our model's most important domain of application is to Chinese sentence analysis, but it may also be applicable to different languages. This semantic framwork constructs a model on the basis of a few rules.

The present paper indicates how situation types are represented, how these representations are composed from semantic representations of linguistic constituents, and how these type differences affect the expression of sentences.

3.1 Syntax-Semantics Relationship Labeling

This work flow includes linking and labeling of each relation between direct semantic items in single sentences, which reflects different semantic representation, and descriptions of the relations of each frame's basic conceptual structure in terms of semantic actions. A semantic representation is a feature that allows one word in the sentence to point at some other word to which it is related. A word in a sentence may have much direct representation, these are differentiated by the semantic action. By analyzing the direct se mantic representation, we can capture semantic relationships between words, reconstructing a framework for the order of Chinese sentences.

In most cases syntactic relationships are consistent with semantic relationships. The following framework shows show some important similarities between the structure of syntactic and semantic structure. For example, in

我在看电视.（'I am watching TV.'）

Syntactically, '我'（I）is subject, directly relating to the verbal predicate '看'（watch）, '电视'（TV）is object , also links to the verbal predicate directly. '在'（be doing）as a adverb is an adjoined predicate '看'（watch）, there is a direct relationship between the two nodes. Semantically, '我'（I）is the agent and '电视'（TV）is the recipients, both of them have a direct relationship with the activity '看'（watch）. So we link the different nodes as follows:

In cases where the relationship between syntax and semantics is inconsistent, by syntactic analysis, if there are multiple syntactic analyses among a sentence, we always choose the analysis

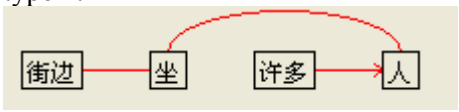
relationship that is consistent with the semantic relationship. For example, the Chinese sentence

街边坐着许多人。

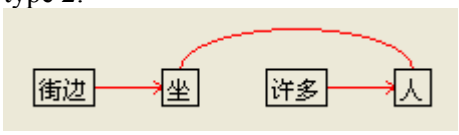
（many people sit beside the street.）

The above sentence can be analyzed either of the following two syntactic structures.

type 1:



type 2:



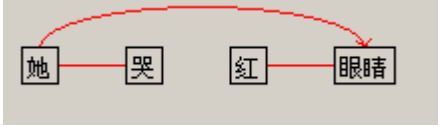
The two syntactic structures are analyzed with difference in the first node and the second node. In type 1, '街边'（beside of the street）is analyzed as subject, for type2, the linguist also analyzed it as adverb modifier, adjuncting to the predicate '坐'（seat）. But when this sentence is analyzed in terms of semantics, there is only one relationship structure similar as type 2. '人'（people）is analyzed as agent, '街边'（beside of the street）as localizer, attached to the activity '坐'（seat）. This semantic structure is consistent with the syntactic structure type 2. Only one structure can display both syntax and semantic relationship simultaneously. So we choose the second analysis.

If the syntactic relationship is different from the semantic relationship, we take no account of the syntactic order. In the Chinese sentence

她哭红了眼睛。

（she cry so much that her eyes become red.）

Within the surface syntactic structure, adjective '红'（red）will be analyzed as complementation and directly associated with main verb '哭'（cry）, which indicate result of predicate. Underlying the syntactic structure, '红'（red）actually point to '眼睛'（eyes）in semantic representation. There is no direct semantic relationship between '哭'（cry）and '红'（red）. The semantic network can be analyzed as: she cry + her eyes become red, the immediate relationship between 'he' as a possessor and 'belly' as a possession and that between 'belly' as entity and 'painful' as description. In this case we link the node '红'（red）to '眼睛'（eyes）directly based on semantic relationship.



3.2 'Head' Determination

The basic link is the direct link between two semantic units. In addition, a set of general rules for determining the directions has been identified.

1. That between Head and Its Modifier as a Case of Direct Relationship

The head (see below), and the modifiers that come before it, constitute a type of modification relationship, which is one of the typical cases of direct relationships, e.g,

A. Gao zige de ren
tall body DE person
the person with tall body

B. (to be compared with the above sentence)
ren de gezi gao
person DE body tall
'The person's body is tall.'

In the above sentence, *ren* 'person' and *gezi* 'body' hold a modification relationship, but *gao* 'tall' and *ren*'person' are related indirectly as the relationship between the two words is realized through that of *gezi* 'body'. Therefore, we say that the relationship that *ren* 'person' holds with *gezi* 'body' is a direct one, but that with *gao* is a rather indirect one.

2. That between An Action Verb and Its Patient as a Case of a Direct relationship

In case a head noun is an AGENT of an action verb within a modifying phrase, then the relationship between the Head none and the action verb is a direct one. The following sentences illustrate the point.

C. chi pingguo de nuhai.
Eat apples DE girl
'the girl who is eating apples.'

D. (to be compared with the above sentence)
nuhai chi pingguo
girl eat apples
'The girl is eating apples.'

In the above sentence, *nuhai* 'girl' is an AGENT of the action verb *chi* 'eat', the two words have a direct semantic relationship, therefore we link them directly and annotate 'girl' as a head. In contrary, the relationship between *nuhai* 'girl' and *pingguo* 'apples' is of an indirect type.

3. Other Cases of Direct Relationships

In case there is neither a modification nor an AGENT/PATIENT relationship, the whole phrase, which is still directly related to a following describing phrase, has to be embedded. E.g.,

E. ban shiqing yinggai guquan daju.
Handle problem should care-about overall situation
'People should care about the overall situation when they handle problems.'

F. chouyan hai shenti.
Smoke harm health

‘Smoking harms health.’

G. ta neng daying de shiqing wo ye neng daying.

He can accept DE event I also can accept

‘The event that he can accept are also acceptable to me.’

3.3 ‘Head’ Determination

Since Chinese lacks morphological cues, the grammatical markers (such as 的，把，被) and word order are comparatively important cues for the relationship determination. We have to rely on grammatical and semantic knowledge to guide role assignment.

In this study, we have proposed an approach that combines ‘basic conceptual structure’ and our ‘Head-Driven Principle’. According to the ‘Head-Driven Principle’, most structures are analyzed as having a ‘Head’ which is connected to various types of modifiers, such as Head-NP (adjective-noun, noun-adverbial pairs 我们都), Head-VP (adverbial-verb, verb-adverbial, adjective-verb…). In our framework, modification is represented by attaching tags with arrows to the core semantic item whereve the type of modification can be clearly identified. Since the SVO is the basic order in Chinese, there is no modifier relationship among the level of SVO. In our model, ‘Subject-Predicate Structures’ and ‘Verb-Object Structures’ are represented as non-head. In above example, the relation linking the ‘core’ noun and verb with their ‘adjunct’ is tagged with an arrow to indicate that it is a ‘head’. Both A and B label the ‘head’ as the core noun. E labels the ‘head’ as the core verb. Employing the ‘Head-Driven Principle’ for the construction of semantic models. Some ambiguous sentences can be clearly represented. The different meaning among sentence or phrase containing same words can also be described . Conside the following sentence and phrases:

学生喜欢老师。

（The students like the teachers.）

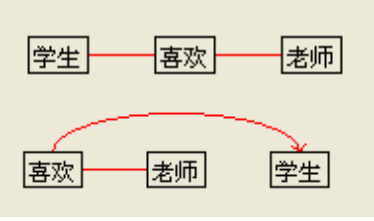
喜欢老师的学生

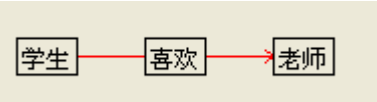
(the students who like the teachers)

学生喜欢的老师

(the teachers who the students like)

All of above examples containing same meaning words can have very different meaning, depending on the different word order and grammatical marker ‘的’（DE）. We use head tagging to construct different frameworks for these structures:





The above three head semantic structures clearly show us the different relationships among sentence and noun phrases with different meaning. The head words are connected to their modifier through arrow arcs. The first SVO relationship is also represented by non-head tagging.

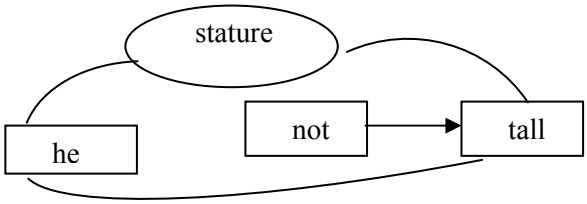
3.4 Feature Abstracting and Labeling

Based on the analysis of semantic relationships, we have been parsing feature structures to express dependencies between semantic features. In our analysis model, semantic feature means a variety of detailed semantic relationships. Most of the time, semantic features are not so easy to define. Some feature typologies have been provided, but there is still much discussions about the nature of a feature in a text. To avoid the confusion of feature classification, we proposed a method to abstract the semantic feature directly from sentences that contain the natural feature word. For those sentences without semantic features insert , we’ll labeling the semantic features refer to the categorys include in other sentences, attached on the relationship arcs. Thus we constructed a semantic framework based on multi-profile dimension. For example:

Ta gezi bu gao.

His stature isn’t tall.

He isn’t tall.



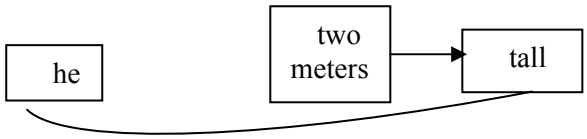
In traditional analysis, ‘stature’ is just a syntactic constituent in a sentence. However, the essential meaning of the sentence is ‘he is not tall’, ‘stature’ is semantic feature linking ‘he’ and ‘tall’ together, thus in our semantic analysis we link only ‘he’ and ‘tall’ semantically, ‘stature’ is taken as feature marking a semantic relationship, rather than an immediate constituent. This Chinese semantic structure, after feature abstraction, is very similar to its English counterpart. It facilitates the translation from one language into another.

In some sentences , there are not only semantic features but also their particular values included. Similarly we abstracted the values attached on the features. Thus we can expand the feature structures to express this level of detail. For example

Ta liang mi gao.

he two meters tall

‘He is two meters tall.’



In the above framework, ‘tall’ is the semantic feature describing staturs of the agent ‘he’ , and ‘two meters’ express the value of the feature. They provid different information at different level, constructing a feature sturcture.

4 The Advantages of Our Semantic Model

In developing our semantic frameworks, we also have articulated a framework of ‘Noun-Centrality’ as a supplement to the widely assumed ‘Verb-Centrality’ practice. We can successfully disambiguate some troublesome sentences, and minimize the redundancy in language knowledge description for natural language processing. We automatically learn a simpler, less redundant representation of the same information.

First, comparing syntactic order and semantic order, we used the reconstructed original order, giving some different order sentences similar results. Thus, variations of order in the same sentence can reveal the same relationships.

One semantic structure may correspond to more syntactic structures in Chinese, and this correspondence can be made specifically clear using our approach.

1.Ta da-le wo	2.Ta ba wo da-le	3.Wo BEI Ta da-le
She beat me	She BA me beat	I BEI she beat
‘She beat me.’	‘She beat me.’	‘I have been beaten by her.’

The above three sentences, their syntactic structures are clearly different from each other. That is, the direct object *wo* ‘me’ appears right after the main verb in (1) whereas the same logical object has moved to a pre-verbal position with the help of a special Chinese preposition *BA* in (2) and to a sentence-initial position with the help of *BEI* in (3). But underlying the difference syntactic structures, they share the same basic semantic structure, using semantic represented expression, the three sentences of above example can be described in below.

AGENT	*Ta* ‘she’
PATIENT	*Wo’me’*
ACTION	*Da* ‘beat’

Several different sentences which should be analyzed as having the same syntactic structure may have fundamentally different semantic structures. The following three sentences S1, S2and S3, for example, should be analyzed as having the syntactic structure, but their semantic structures are nevertheless represented as S1’, S2’ and S3’ respectively in our framework.

NP + V + Adj + NP

S1＝Ta xiao-tong le duzi

he laugh-painful ASP belly

‘He laughed so much that his belly was painful.’

S2＝Wo kan-tou le ni

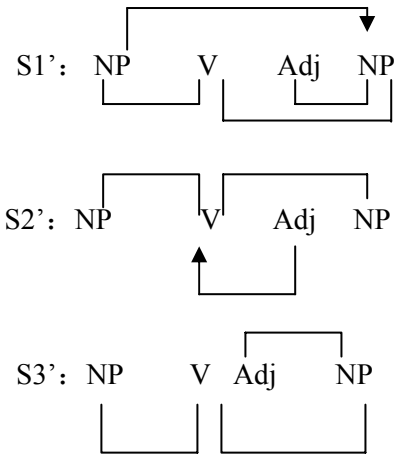
I see- through ASP you

‘I understand you thoroughly.’

S3＝Ta da po-le beizi

She broke up the cup

She broke up the cup.



On the other hand, many structural ambiguities in Chinese sentences are one of the major problems in Chinese syntactic analyses. One syntactic structure may correspond to two or more semantic structures, that is, various forms of structural ambiguity are widely observed in Chinese. Disregarding the semantic types will cause syntactic ambiguity. If this type of information is not available during parsing, important clues will be missing, and loss of accuracy will result. Consider the Chinese sentence

Ta de yifu zuo de piaoliang.

Her cloth do DE beautiful

Reading 1: ‘She has made the cloth beautifully

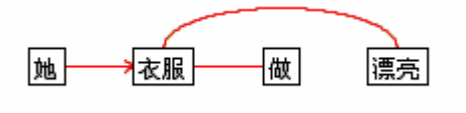
Reading 2: (Somebody) has made her cloth beautifully.'

Syntactically, the sentence, with either one of the above two semantic interpretations, should be analyzed as

But the two semantic structures have to be properly represented in a semantics-oriented framework. We do so as in type A and type B respectively.

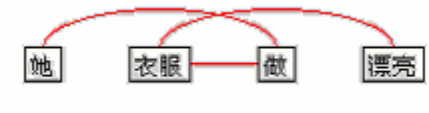
Type A: Ta de yifu zuo de piaoliang.

Her cloth do DE beautiful



Type B: Ta de yifu zuo de piaoliang.

Her cloth do DE beautiful



So under our proposal, the above two different types of semantic relations can be clearly represented..

5 Conclusion

In this paper we have demonstrated how our semantic model can be created to analyze and represent the semantic relationships of Chinese sentence structures. The semantic model project is producing a structured tree bank with a richer set of semantic and syntactic relationships of different words on the basis of the analysis of lexical meanings and conceptual structures that underlie those lexical meanings. We developed some methods for determining the relationship between direct semantic items based on the analysis of syntactic and semantic order. The key advantages of our semantic model are:

a) many ambiguous sentences can be clearly represented.

b) minimal redundancy in language knowledge description for natural language processing.

We hope to use the minimum analysis method to find the semantic order with equal relationship among new sentence. We then used the partition relationship as a training database to recognize new order as similar as these order structures.

We also have been creating feature sets parsing feature structures to expressing dependencies between semantic features. Furthermore, we abstracted the values attached to the features. Thus we can expand the feature structures to express this level of detail.

References

Baker C, Fillmore C, Lower J 1998 The Berkeley FrameNet Project, In *Proc. of ACL/COLING 1998*.

Daniel Gildea and Daniel Jurafsky 2002 Automatic Labeling of Semantic Roles. In *Proc. of ACL 2000*.

Nianwen Xue, Fei Xia 2000 The Bracketing Guidelines for the Pann Chinese Treebank, IRCS Report 00-08 University of Pennsylvania, Oct 2000

Dominique Dutoit, Thierry Poibeau 2002 Inferring Knowledge from a Large Semantic Network, In *Proc. of COLING 2000*

James Henderson, Paola Merlo, Ivan Petroff 2002 Using Syntactic Analysis to Increase Efficiency in Visualizing Text Collections, In *Proc. of ACL/COLING 2002*.