Semantic Disambiguation of Chinese Homonyms in pinyin-hanzi Conversion

Li Sujian, Zhang Jian, Liu Qun

Institute of Computing Technology,

Chinese Academy of Sciences,

Beijing, China, 100080

Email: {lisujian, zhangjian, liuqun}@mtgroup.ict.ac.cn

Wan Jiancheng

Computer Science Department,

Shandong University of Technology,

Jinan, Shandong, China, 250016

Email: wanjch@dms.sdut.edu.cn

Abstract

Homonymous phenomenon is very common in Chinese. And it brings much difficulty with Chinese pinyin input method. To solve this problem, we propose Chinese thesaurus «Cilin» as the norm to generalize Chinese words into conceptual-level lexical information and form a hierarchical network of semantic classes. And we also present a method of semantic analysis based on hierarchical relations. The focus of semantic analysis is the computation and checking of semantic attributes which are triggered by syntactic relations. Hierarchical semantic information conducts restriction on syntactic analysis. Experimental result shows that the parser using conceptual-level lexical information as well as semantic analysis resolves homonymous ambiguity with 94.09% of accuracy.

1 Introduction

Today's computers were invented by Westerners, and keyboards are designed for languages with alphabets. Of all the conversions from alphabets to hanzi(Chinese characters), Chinese pinyin input method is the mainstream. But homonymous ambiguities always demand a second choice during pinyin-hanzi conversion, and it makes inputting Chinese much slower. To solve this problem, context information, syntactic and semantic knowledge are used to reduce the uncertainty of the first choice during conversion.

At present, there are two main branches: one is the statistical method, MSR(Microsoft Research) China has been engaging in this work[Lee, 1998]. It focuses on word-based n-gram language modeling as well as probabilistic context-free grammar (PCFG) and applies the model to pinyin-hanzi conversion.

The other branch is the basic syntactic and semantic method. This method must continue to strengthen the basic theoretical research about syntactic and semantic knowledge. At present, syntactic parsing is relatively mature, and semantic analysis becomes one highlight. So every breakthrough that represents and implements semantic knowledge will improve the method of NLP(Natural Language Processing). The current methods of

semantic analysis include Valent Grammar[Zhan and Liu, 1997], ATN[Winograd, 1983], Dependency Grammar[Feng, 1998], HNC theory[Wang, 1999] etc. These methods are useful to some extent. However, the definition of semantic knowledge doesn't come to an agreement, so there is still a long way to culminate the perfect representation and implementation of semantic knowledge. The methods mentioned above also can't solve the semantic problems thoroughly. We don't try to solve all the semantic problems. But from the practical point of view, based on Wittgenstein's theory "meaning is usage", we proposed one framework about the hierarchical relations of semantic classes and their hierarchical computation to solve the problem of Chinese homonyms.

The rest of the paper is organized as follows. First we give some necessary definitions about semantic classes; then we will introduce the computation of semantic attributes in detail; section 4 describes the basic relations implicit in our semantic mechanism; section 5 makes use of one example to show how to implement semantic analysis; and section 6 presents some experimental results; and finally we give our conclusions.

2 About Semantic Class

As Austine[Andrew, 1998] described, the individual words in language name objects — sentences are combinations of such names. For example, in Chinese, "Zhang San" is clearly the name of one person, while "Cat" is a certain kind of animals that bear the characteristics of the cat. The same to "Apple"，it is the name we give to a certain kind of fruits, and "Hit" is one kind of behavior, etc. Given some names, we can get what they mean no matter which order they are given by. For instance, when three Chinese words "张三(Zhang San)","吃(Eat)","苹果(Apple)" are given, the conception "Zhang San eats the apple" is implicit in these words, instead of "An apple eats Zhang San" or "Zhang San and an apple eat". This is because we have had the common sense that one person can eat one kind of fruit. Here Zhang San is a person and apple is one kind of fruit. As for these conception "ZhangSan", "Eat", "Apple", "Person", "Fruit", we can regard them as different classes. One meaningful sentence is constructed through their relations. This idea also consists with the ontology. Therefore, we give names to the entities, features and activities in the real world, and construct one net of semantic classes interacted through their relationship. These semantic classes are the fundamental part of one language and other words merely further elucidate these classes.

Combined with the characteristics of Chinese, We introduce several definitions:

1. Set of Chinese words $W = \{ w_1, w_2, \ldots, w_m \}$, $m > 0$. W is composed of all the Chinese words. $w_i (0 < i \leq m)$ represents the Chinese name of an entity, or an activity, or a feature, in the real world.
2. Set of attributes of words $A = \{ a_1, a_2, \ldots, a_n \}$, $n > 0$. A represents a set of the possible attributes of Chinese words. $a_i (0 < i \leq n)$ is one kind of attribute about some Chinese word. The value of a_i is assigned to every particular semantic class during the hierarchical computation.
3. Set of semantic classes $SC = \{ SC_1, SC_2, \ldots, SC_t \}$, $t > 0$. $SC_i (0 < i \leq t)$ represents one particular semantic class. From the view of ontology, every word in Chinese belongs to one class which has its own attributes and behavior. Thus the size of

this set is smaller than or equal to that of the set W. There are complicated relations among these semantic classes. For example, one class is perhaps the ancestor, brother or offspring of another one.

4. For each SC_i (0<i<t), we define it as follows:

$$
\begin{aligned}
&SC_i : SC_{m_1} : SC_{m_2} : \cdots \\
&\{ \; w_{i1}, w_{i2}, \cdots, w_{ip_i}; \\
&\quad ! \; SC_{ia}; \\
&\quad SC_i :< a_{i1} >; \\
&\quad SC_i :< a_{i2} >; \\
&\quad \cdots; \\
&\quad SC_i :< a_{iq_i} >; \\
&\} \\
&w_{ij} \in W(0 < j \le p_i), a_{ij} \in A(0 < j \le q_i)
\end{aligned}
$$

For any semantic class SC_i, its definition is given in the object-oriented form. Without specification, it only inherits from a certain semantic class by default according to the naming mechanism. SC_i can also have more than one parental class such as SC_{m1}, SC_{m2} and so on if possible. SC_i includes p_i Chinese words $w_{ij}(0<j\leqslant p_i)$ which can be seen as the value of one attribute which is not explicitly indicated in SC_i. We use '!' to indicate its antonymous semantic class SC_{ia} which is optional. SC_i includes q_i attributes whose values $SC_i : <a_{ij}>(1\leqslant j\leqslant q_i)$ can be preassigned or computed through inheritance.

From the definition of the semantic class above, we can see that a semantic class is extracted from one collection of Chinese words that have the same or similar usage and features. Attributes and their computation are imported into the semantic classes. Relations among semantic classes are constructed through computation, and then we can use them for the maintenance of words such as testing the validity of the words' usage. A semantic class has the characteristics of encapsulation and inheritance. And at the same time, it is the extraction from certain Chinese words and at least includes the conception of one word, then one class can also be regarded as an object without being instantiated. So one semantic class has dual functionality. As an abstract conception it can produce its subclasses that behave variously; as an object it can participate in semantic computation. Our semantic knowledge base is composed of such a network of all the semantic classes.

3 The Computation of Hierarchical Semantic Relations

Binary relations between two words or phrases play an important part in the composition of sentences. At the same time every phrase has one principal word, and thus the final form of binary relations can be thought of as relations between two principal words. Therefore, we extract the semantic constraints between two Chinese words and represent this kind of association through attributes. If we need to enumerate and check all the semantic constraints between two words, we call this method an extensive one. However, the number of Chinese words is very large, and it's ponderous to use this method to describe all the semantic knowledge even though it is possible. Since

semantic classes have been introduced, we can generalize constraints between two words to those between two semantic classes. Then we could obtain a highly summarized semantic knowledge base. This method is called an intensive method. The intensive method is fundamental to our semantic analysis. Thus, we generalize to the level of semantic class to demonstrate Chinese compositive laws.

Syntactic analysis, which is the basis of semantic analysis, is one important step in NLP[Yu, 1997]. For R_{SA} =x • y, x,y∈SA, SA represents syntactic attribute. This formula is the logical representation of syntactic analysis. From it we can get the principal-subordinate relationship of x and y and return the syntactic relation $R_{SA.}$. For semantic analysis, we define the formula R_{sc}=X • Y, here X, Y∈SC. X and Y correspond to x and y in the same position of the sentence respectively. Thus the two semantic classes X and Y also get their principal-subordinate relation by default. We can see that syntactic relations trigger semantic analysis and that semantic computation is conducted according to the principal semantic class. As the following formula,

$$X_1 \quad X_2 \ : \ C(t) \ ABr \qquad (1)$$

where X_1 and X_2 represent two semantic classes. A and B represent the corresponding syntactic attributes. The syntactic relation between X_1 and X_2 satisfies ABr. C(t) represents that X_t is the principal semantic class.

We specified fifteen kinds of Chinese syntactic relations with the form of AB_r that are used to trigger semantic computation. At the same time, for all Chinese notional words we classify the semantic classes into three kinds: semantic classes for entities(SCE), semantic classes for activities(SCA), and semantic classes for features(SCF). Then we will explain respectively how these three kinds are triggered to conduct semantic computation.

3.1 Semantic Class for Entities(SCE)

SCE is designed for entities which usually are static. For formula(1), the principal semantic class X_t belongs to SCE.

1. Unit-noun relation(UNr): According to formula(1), ABr is equal to the syntactic relation UNr, and the principal semantic class X_t is triggered by UNr and gets the value of its attribute—X_t : <unit attribute>. This attribute is special in Chinese, because when we refer to the quantitative entities, the entity or entities usually go after number word plus one unit word. The assignment of the values can be implemented through the inheritance of semantic classes. A subclass can inherit from its ancestor and get the corresponding value of the attribute. The value enters into the subordinate semantic class and conducts checking. If the checking returns a true value, then the meaning is right, otherwise it is wrong.
2. Modifier-noun relation(NPr): As in formula(1), X_t is triggered by NPr and gets the value of X_t:<modificatory attribute>. The value of this attribute is usually related with some SCF(semantic class for features) which describes the principal class.
3. Noun-noun relation(NNr): If the two parameters of this relation conform to CR relation discussed later, X_t will be stimulated to get the value of X_t:<constitutive attribute> and check if the value is consistent with the subordinate class.

4. Verb-noun relation(VNr): X_t is triggered by VNr to get the value of X_t:<action attribute>.
5. Noun-orientation relation(NOr): X_t is triggered by NOr to get the value of X_t:<orientation attribute>.
6. Preposition-object relation(POr): X_t is triggered by POr to get the value of its attribute—X_t:<preposition attribute>.
7. Numeral-unit relation(NUr): In Chinese we ascribe the unit words to SCE. This relation stimulates X_t to get the value of X_t:<numeral attribute>, which specify the validity of numerals.

3.2 Semantic Class for Activities(SCA)

SCA is designed for activities which are dynamic. In Chinese most verbs belong to SCA. Here we defined six attributes to conduct semantic computation. As the principal semantic class, X_t belongs to SCA, triggered by the following relations.

1. Verb-modifier relation(VPr): Here ABr represents VPr. X_t is triggered to compute the value of its attribute-X_t:<modifactory attribute>, which is designed to modify a certain SCA.
2. Subject-predicate relation(SPr): X_t is triggered by SPr to get the value of X_t:<agent attribute>, which represents the actor of some action.
3. Verb-complement relation(VCr): X_t is triggered by VCr to get the value of X_t:<complemental attribute>, which is used as the complement to an activity.
4. Verb-quanlifier relation(VQr): We compute and get the value of X_t's attribute—X_t:<quantifier attribute>, which is a special complement to SCA. Not all the SCAs have this attribute.
5. Verb-object relation(VOr): Here the two parameters triggered by this syntactic relation are similar to those triggered by VNr. The difference is that the principal semantic class here is SCA, not SCE. And we need to get the value of X_t's attribute—X_t:<object attribute>, which is the object of one action.
6. Verb-repeated relation(VVr): Sometimes, two actions appear and act as predicates in the same sentence in Chinese. Not all the combinations of two actions are legal. Here, VVr stimulates the semantic analysis and gets the value of X_t:<coordinative attribute>.

3.3 Semantic Class for Features(SCF)

SCF is used to describe one or another feature of an entity or activity. Here we design two attributes for SCF. As the principal semantic class, X_t belongs to SCF, triggered by the following relations.

1. Adjective-modifier relation(APr): APr triggers Xt to get the value of its attribute—Xt:<modificatory attribute>, which is used to modify a certain SCF.

2. Adjective-complement relation(ACr): ACr triggers X_t to get the value of its attribute—X_t:<complemental attribute>, which acts as the complement to some SCF.

4 Basic Semantic Relations

We have adopted the mechanism of Chinese thesaurus «Cilin» which conducts semantic classification for Chinese words. It comprises 12 major categories, 94 medium categories, and 1428 minor categories [Mei, 1983, Yuan et al, 1998]. Every category is regarded as one semantic class. We map these semantic classes to SCE, SCA and SCF respectively. We specify several attributes for every semantic class, maintaining «Cilin»'s net structure at the same time. The values of some semantic classes' attributes are assigned by hand in advance. Other semantic classes can obtain the values of their attributes through inheritance. This net structure of semantic classes contains several basic semantic relations of Chinese words. These relations are Synonymous Relation, Antonymous Relation, Constitutive Relation, Logical Multi-hierarchical Relation, and Non-monotonous Relation.

1. Synonymous Relation(SR): Synonym words are structured in synsets, underlying a linguistic concept. Every synset is connected with a semantic class, representing a textual definition that can be described in a logical form which is the building block of our knowledge base. This formulation that one class includes at least one word, can provide an elegant manner of localizing ambiguities[John and Jerry, 1988]. In fact Synonymous Relation is a relation of words and it can be seen as the basis of all the semantic relations.
2. Antonymous Relation(AR): $\exists X, Y \in SC, \quad X :! Y \text{ or } Y :! X$ represents X and Y are two semantic classes with antonymous conception. This kind of relation has the characteristic of symmetry. That is, if X is the antonymous semantic class of Y, Y must be the antonymous semantic class of X. When two semantic classes have the antonymous relation, the words they contain also have the corresponding antonymous relation.
3. Logical Multi-hierarchical Relation(LMHR): represents $\exists X, Y \in SC, \quad IS\_A(X, Y)$ that X is one offspring of Y. In «Cilin», X inherits and only inherits from Y, that is, if there exists IS_A(X,Y1)，IS_A(X,Y2)，we only get Y1=Y2. From the definition of Semantic Class above, we can see one semantic class inherits from one semantic class by default. At the same time, it can also inherit from other semantic classes. So we can get that Y1≠Y2 is right. Therefore, one semantic class can inherit more than one ancestor and we can get a network of semantic classes. This relation is the basis to assign values to attributes in some semantic classes.
4. Constitutive Relation(CR)：We assume two semantic classes X and Y satisfy this relation. They have three possible cases : X is the constitutive part of Y ; X is a member of Y ; X is the constitutive material of Y. No matter which case, is used $\exists X, Y \in SC, \quad PART\_OF(X, Y)$ to represent this relation. In this system, we describe this relation through the assignment of values to some attributes, e.g,

Constitutive attribute can embody such a relation. This relation is implicit in the assignment of attributes, however, it can demonstrate the idea of inheritance and the ability of inductive learning.

5. Nonmonotonous Reasoning Relation(NMRR): we define this kind of relation as follows.

$\exists X, Y \in SC, \exists a \in A,$

IS_A(X, Y) and (a in Y) and (a in X)

if $X:<a> \neq Y:<a>$

then NMRR(X, Y)

Here (a in Y) represents that a is one attribute of Y, and X: <a> represents the value of the attribute **a** in semantic class X. We assume that X is an offspring of Y and that both X and Y have the attribute **a**. By default, X:<a> inherits from Y. However, if X: <a> isn't equal to Y: <a>, but assigned a new value. Then we call the relation between X and Y as nonmonotonous reasoning relation. Although this relation isn't independent because it should satisfy IS_A() relation, it represents a general phenomenon in nature.

A network of semantic classes can represent those relations discussed above. On the other hand, those relations demonstrate inductive and reasoning ability, and bring convenience to semantic computation and analysis. Thus, we can abstain the system from repeated work and display an intelligent ability of learning.

5 Exemplification of Hierarchical Semantic Analysis

One concrete example of discriminating Chinese homonyms is used to demonstrate how to conduct semantic analysis. Semantic analysis is conducted to deal with the ambiguities that syntactic analysis can't solve. According to the Chinese characteristics[Wan and Yao, 1998], there exist binary relations between every two parts of one sentence. And further semantic constraints are made for these binary relations. Through the computation of these attributes, we check whether these relations make sense. Thus, we transfer the emphasis of semantic analysis to the computation of attributes. The following is one fragment of grammar rules about noun phrase and verb phrase.

1) $VP \rightarrow VP \quad NP \quad // \quad C(1) \quad VOr(\alpha_1, \alpha_2)$
2) $VP \rightarrow VP \quad Adj \quad // \quad C(1) \quad VCr(\alpha_1, \alpha_2)$
3) $VP \rightarrow Verb$
4) $NP \rightarrow VP \quad 'de' \quad NP \quad // \quad C(3) \quad VNr(\alpha_1, \alpha_3) \quad W(2, '的')$
5) $NP \rightarrow SHU \quad NP \quad // \quad C(2) \quad UNr(\alpha_1, \alpha_2)$
6) $NP \rightarrow Noun$
7) $SHU \rightarrow SHS \quad Unt \quad // \quad C(2) \quad NUr(\alpha_1, \alpha_2)$
8) $SHS \rightarrow Numb$

The functions after symbol "//" are used to conduct semantic analysis. Every time a reduction completes, the procedure applies these functions to the input sentence and the

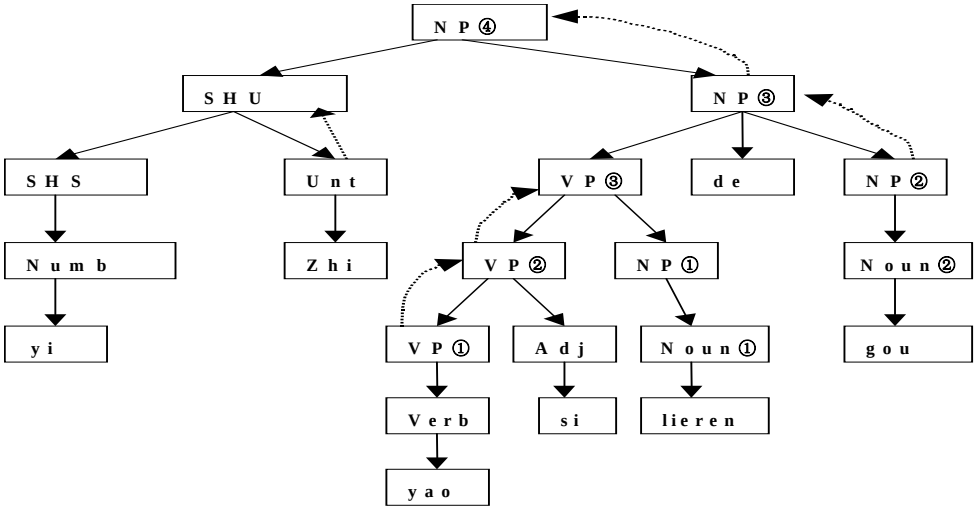
actions of the procedure are summarized in table 1. We use one example to explain in detail how to conduct semantic computation and checking.

Function	Description
W(n,'ch')	The nth part in the right will convert to the Chinese character 'ch'
C(n)	The semantic class mapped with the nth part in the right will be the principal class
XYr(α_i , α_j)	The ith and jth part in the right will be triggered by the relation Xyr

Table 1:procedures of semantic analysis

Suppose that the Chinese pinyin stream "yi.zhi.yao.si.lieren.de.gou" has been correctly segmented. Now we see how to get a stream of Chinese characters. After syntactic analysis we can get one dendriform structure which is also called a syntactic tree[fig.1].

In fig.1, the real lines represent the syntactic constitution of the tree and the dotted lines represent the transferring of the principal parts. The nodes which have no brothers will transfer themselves upwards. The terminal nodes are words in the pinyin stream. Every pinyin word corresponds to one syntactic attribute, but it still maps with several candidates of character words. We can see the syntactic result in Table 2. Then we make further semantic analysis on the ambiguous results. Table 3 summarizes the process of our semantic analysis. One row traces one possible procedure after the completion of one reduction when parsing the input sentence, corresponding to one multi-branch subtree in fig.1. After all the steps, one noun phrase is constructed and no ambiguity exists any more. Now we get the only result —"一.只.咬.死.猎人.的.狗".



F i g u r e 1 : A S y n t a c t i c T r e e f o r p i n y i n s t r e a m "y i . z h i . y a o . s i . l i e r e n . d e . g o u "

Pingyin word	Yi	Zhi	Yao	Si	lieren	de	Gou
Syntactic	Numb	Unt	Verb	Adj	Noun	de	Noun

attribute							
Chinese candidates	一，亿	只，支，指，枝	摇，邀，咬，吆	死，私	猎人	的	狗，沟，钩，诟

Table 2: Result of Syntactic Analysis

Formula number	Action	Syntactic relation	Principal semantic class and its computation	Result
(7)	SHS and Unt reduct	NUr	semantic classes of pinyin “zhi” get their values of numeral attribute	Chinese character “一(one)” not “亿(billion)” conforms to pinyin “yi”
(2)	VP① and Adj reduct	VCr	Semantic classes of pinyin “yao” get their values of complemental attribute	Chinese character “死(to death)” conforms to pinyin “si”, “摇(shake)”or”咬(bite)” maps with “yao”
(1)	VP② and NP① reduct	VOr	Semantic classes of “摇”and ”咬” get their object attribute	Still “摇(shake)”or”咬(bite)” maps with “yao”, “猎人(hunter)” is the result of “lieren”
(4)	VP③,”de” and NP② reduct	VNr	Semantic classes of pinyin”gou” get their values of action attribute	‘de’ converts to ‘的’. “狗(dog)” conforms to pinyin “gou”, and “咬(bite)” is correct character of pinyin “yao”
(5)	SHU and NP③ reduct	UNr	Semantic class of “狗(dog)” gets its value of unit attribute	“只” is correct character conforming to pinyin “zhi”

Table 3: An application of semantic analysis

6 Experimental Results

We have a dictionary of fifty thousand Chinese words, a knowledge base of about one hundred syntactic rules, and a preliminary prototype of semantic framework. Of course our work is not enough. However, the result is satisfying. We get the precision before analysis, after syntactic analysis and after analysis respectively as in table 4. Without any analysis we select the first candidate of one pinyin word and get the precision of 58.75%. With syntactic and semantic analysis the results are 79.15% and 94.09% respectively.

	Precision
Before analysis	58.75%
After syntactic analysis	79.15%
After semantic analysis	94.09%

Table 4: Experimental results

7 Conclusion

In this paper, a new theoretical method based on hierarchical computation has been presented for semantic analysis. We have adopted bottom-to-up syntactic analysis that triggers the computation of semantic attributes when making reduction. We can see that the system has been improved and gotten a satisfactory performance with semantic computation. However, what we have done is only part of the research on Chinese homonyms. Syntactic and semantic knowledge base still needs to be further developed.

References

[Lee, 1998], Lee Kai-Fu, The Direction of Our Research, http://www.research.microsoft.com/research/china/ directionsonresearch .asp, 1998.

[Zhan and Liu, 1997], Zhan Weidong & Liu Qun, Applications and Unsolved Problems of Semantic Classification on Words in a Chinese-English Machine Translation system, Language Engineering ,1997.

[Winograd 1983], Winograd, Terry: Language as a Congnitive Process. Volume I: Syntax. Reading, Mass. 1983.

[Feng, 1998], Feng Zhiwei, Some Formal Properties of Dependency Grammar, Proceedings 1998 International Conference on Chinese Information Processing. 1998.

[Wang, 1999], Wang Hou-feng, The ProcessingTactics based on HNC for Delimitation of Chinese sentences, Comm. COLIPS(Singapore), Vol.9, No.8, 1999.

[Wan and Yao, 1998], Wan Jiancheng, Yao Wenlin, Chinese syntactic and semantic analysis based on binary relations, Communications of COLIPS，Vol.8,No.1, 1998.

[Yu, 1997], Yu shiwen, Application of Syntactic Knowledge in Research of Linguistic Information Procession(in Chinese), Applied Linguistics, No.4, 1997.

[Andrew, 1998], Andrew Lilico, Wittgenstein & the Augustinian Picture of Language, http:// ourworld.compuserve.com/homepages/Andrew_Lilico/augustin.htm, 1998

[Mei, 1983], Mei Jiaju, Chinese thesaurus «Tongyici Cilin», Shanghai thesaurus Press, 1983.

[Yuan et al, 1998], Yuan Chunfa, Huang Changning, Xu Wei, Zhu Xiaodan, A study on the combinatorial regulation of Chinese semantic classes, Communications of COLIPS，Vol.8,No.2, 1998.

[John and Jerry, 1988], John Bear and Jerry R. Hobbs, Localizing Expressions of Ambiguity, In Proceedings of the Second Conference on Applied Natural Language Processing, Association for Computational Linguistics, 1988.