A Sentence-level Semantic annotated Corpus Based on HNC Theory

Zhiying Liu[1,2], Yaohong Jin[1,2]
[1]Institute of Chinese Information Processing
[2]CPIC-BNU Joint Laboratory of Machine Translation
Beijing Normal University
Beijing, China
{liuzhy, jinyaohong}@bnu.edu.cn

Chuanjiang Miao
Department of Chinese and Bilingual Studies
Hong Kong Polytechnic University
Hong Kong, China
miaochj@hotmail.com

***Abstract*—The Sentence-level Semantic annotated Corpus (SSC) is directed by HNC (Hierarchical Network of Concepts) Theory in which the sentence is regarded as the basic unit and semantic and frame information are annotated. This paper introduces the building form of the semantic annotated corpus with XML, describes the content of semantic annotation on sentence level that the annotated contents can be classified three items: sentence category, which is the sentence semantic type; semantic chunks, which is the lower level semantic components in the sentence; sub-sentence (sentence ecdysis and chunk-extension-into-sentence) which is included in semantic chunks. SSC has been an important and valuable knowledge resource for language study and information processing.**

Keywords- HNC Theory; semantic annotation; semantic chunk; sentence category; XML

I. INTRODUCTION

With the continuous deepening of research and extension, linguists and computer experts realize that the semantics plays an important role in Chinese Information Processing. Meanwhile the focus of Chinese information processing is on the sentence processing instead of the word processing as far as language unit is concerned.

Nowadays, the resource building for sentence processing is beginning and short of the language knowledge, especially semantic knowledge. Building a semantic annotated corpus is important to understand and analyze the language. In our corpus, semantic information is annotated for continuous texts so that computer can understand the sentences by acquiring semantic information in texts.

II. BRIEFINGS OF SEMANTIC ANNOTATION BASED ON HNC THEORY

HNC(Hierarchical Network of Concepts) Theory (Huang Zengyang, 2004) argues that natural language understanding is a process of mapping from the space of natural language to the space of language concept, Each of which has its own symbolic system. The symbolic system in language space differs in thousands of ways, while the symbolic system in language concept space is unique in human society. The existence of the space of natural language depends on sound and character and the existence of the space of language concept depends on concept relevancy network [10].

HNC theory constructs an entire theoretical framework for Natural Language, gives the complete description of the deep structure of the sentence called sentence category on the basis of concept description system.

HNC designs a semantic network used to describe the concepts system of natural language wholly, on which builds the semantic descriptive patterns and constructs the sentence semantic structure expressions. These patterns and expressions formally describe the concept relevancy network on the sentence level. Sentences are infinite in number, but the number of concept categories of sentences, which are called sentence categories, is finite. infinite sentences can be described by finite sentence categories. Sentence category is the sentence classification by semantic categories from language deep layer.

A classic tagging example sentence is as follows:

```
创作取向体现了很强的时代特征。
 <s code="Y30" form="!0">
      <gbk type="YB">创作取向</gbk>
      <ek>体现了</ek>
      <gbk type="YC">很强的时代特征</gbk>.
</s>
```

Figure 1. Example of Tagging Form in HNC Semantic Corpus

So far, HNC semantic annotated corpus(SSC) has tagged 395 articles one million characters in detail which are selected from newspapers, mainly People Daily in recent years, The fields involve politics, economy, culture, sports, military, law, education, health, disaster, state, spirit etc. It takes HNC team 8 years to study and build. SSC has been an important and valuable knowledge resource for language study and information processing.

III. XML ANNOTATING FORM

This corpus adopts XML as annotating form. XML is short for eXtensible Markup Language which is a group of rules defining semantic marks. It contains, shapes, labels, structures, and protects information with symbols embedded in the text, called markup which enhances the meaning of information in certain ways, identifies each of parts and how they relate to each other. We can define these marks and grammar structure freely and annotate the information in texts by XML elements and attributes. In corpus, the semantic units are marked by special elements, and semantic knowledge is marked by attribute value. One element can be embedded in another element. Thus a tree

978-0-7695-4554-7/11 $26.00 © 2011 IEEE
DOI 10.1109/IALP.2011.11

structure is formed which shows the layer character of corpus.

For example, we define *s* element as sentence, *gbk* element as global object chunk, *ek* element as Eigen chunk in which the main verb is located, *fk* element as auxiliary chunk relative to the main chunk which *gbk* and *ek* belong to. Sentence category and sentence form are described in attribute values. By doing this, sentence semantic information is expressed distinctly.

```
It is the winter in the 6th year of the Republic.
  <s code="jD">
      <gbk type="1">It</gbk>
      <ek>is</ek>
      <gbk type="2">the winter</gbk>
      <fk type="Cn"> in the 6th year of the Republic</fk>.
  </s>
```

Figure 2. XML Tagging Form of SSC

This is a simple sentence which is expressed as *s* element. Sentence category is YesOrNo-judging sentence which is expressed as attribute value code="jD". The sentence contains two global object chunks which are expressed as *gbk* elements with the data content *It* and *the winter*. The order of *gbk* is expressed as attribute value type="1" and type="2".*fk* is auxiliary chunk which type is Cn that means time-condition.

IV. THE BASIC ANNOTATED CONTENTS

Since the natural language processing depends on semantic understanding, the SSC takes the meaning as the main annotated contents. The annotated mode is top-to-bottom. Larger language unit will be annotated first and smaller language one will be annotated later. That is, discourse and paragraphs are annotated first, sentences and semantic chunks are annotated later, and words last. Three aspects prove very important in tagging the meaning of sentences [1]:

- Sentence category, which categorizes the meaning of sentences.
- Semantic chunk, which constitutes the semantic components of the lower level.
- Sub-clause, which is imbedded in semantic chunk as a part of it.

A. Sentence Category

57 kinds of basic sentence categories are classified by deductive method in HNC theory. These sentence categories are the basic types of sentence meaning, which can be used to describe any sentence semantic types. In natural language, the semantic category of a sentence can be a kind of basic sentence category, or a combination of two or more basic sentence categories which is called compound sentence. For example,

Students are cleaning the classroom. (Basic action, X)

The corrupt officials finally pleaded guilty. (Reaction and state, X20S*10)

The first sentence belongs to a basic sentence category which sentence code is X. The second sentence belongs to a compound sentence category which is combined by reaction sentence category and state sentence category with the sentence code X20S*11.

Sentence code is regarded as the attribute of *s* element. It is annotated like this:

```
<s code="X">Students are cleaning the classroom. </s>
<s code="X20S*10">The corrupt officials finally pleaded guilty. </s>
```

Figure 3. code tagging

In SSC, sentences have all been annotated with sentence category codes which are classified as basic sentence category and compound sentence category.

TABLE I. ANNOTATION DATA OF SENTENCE CATEGORY

Type of Sentence Category	Sentences	Percentage
Basic sentence category	36099	69%
Compound sentence category	16224	31%

It shows that the numbers of sentences which belong to the basic sentence category are about twice as large as the numbers of sentences which belong to the compound sentence category in table 1.

In HNC Theory, sentences are described from two profiles: global action and global result. Action and result are causal relationship generally. In global sentence, the subject of sentence is Agent. In result sentence, the subject of sentence can be object, content but not Agent [2].

As to the sentence category, it is classified as global action sentence and global result sentence semantically. Global action sentence includes four kinds of types which are X, T, R, D. Global result sentence includes four kinds of types which are Y, P, S, jD. Global action sentence concludes action sentence(X), transfer sentence(T), relation sentence(R) and intellection sentence(D). Global result sentence concludes result sentence(Y), process sentence(P), state sentence(S) and basic logic definition sentence(jD).[10] The eight kinds of sentence category constitute the frame of basic sentence category. And it expands into 57 kinds of basic sentence categories.The Numerical data are as follows:

TABLE II. ANNOTATION DATA OF BASIC SENTENCE CATEGORY TYPES

Types of Basic sentence category		Sentences	Ratio
global action	X(action)	9988	27.67%
	T(transform)	4892	13.55%
	R(relation)	1198	3.32%
	D(determination)	2237	6.20%
global	Y(result)	4173	11.56%

Types of Basic sentence category		Sentences	Ratio
result	P(process)	843	2.34%
	S(state)	5072	14.05%
	jD(basic logic definition)	7696	21.32%

B. Semantic Chunks

Semantic chunks are lower semantic components than the sentence, which can be a word, a phrase, or a sub-sentence (sentence Ecdysis or chunk-extension-into-sentence), even can be the combination of the three mentioned above.

The borders of chunks need to mark when tagging. It is easy to mark them with container elements in XML. Elements are the building blocks of XML, dividing a document into a hierarchy of regions, each serving a specific purpose. It begins with a start tag and closes with an end tag. The chunks are contained in both tags as elements.

Different sentence category needs different semantic chunks. For example, action sentence category needs three main chunks: action, actor, object; and information transfer sentence category needs four main chunks: information transfer, transferor, receiver and transferred information.

As to the first sentence above, students are actor, are cleaning is action, the classroom is object.

It is not necessary to tag the meanings of main chunks since definite sentence category contains definite main chunks. When sentence category is known, the meanings of main chunks are doubtless. Therefore, we tag the type of gbk chunks with numbers which mark the order of gbk chunks instead of the meanings. Eigen chunk need not to tag the position information since it always appears on the second position among the chunks.

For the sentence above, it is annotated like this:

```
<s code="X">
    <gbk type="1"> Students </gbk>
    <ek> are cleaning </ek>
    <gbk type="2"> the classroom </gbk>.
</s>
```

Figure 4. Semantic chunks tagging

Besides, auxiliary chunks such as condition chunks which are made of time or space phrase are not the integrant components in sentence building. So the changes of auxiliary chunks have nothing to do with sentence category.

Sub-sentence is the sentence which is included in the semantic chunks. Sentence ecdysis and chunk-extension-into-sentence are the items adopted in HNC theory. Sentence ecdysis means that sentence changes into semantic chunk or the part of it. Chunk-extension-into-sentence means that a semantic chunk extends into a sentence. In language understanding, we should take them as sentence, describing their sentence category and semantic chunks. For example,

C. Sub-sentence: Sentence Ecdysis and Chunk-extension-into-sentence

Sub-sentence is the sentence which is included in the semantic chunks. Sentence ecdysis and chunk-extension-into-sentence are the items adopted in HNC theory. Sentence ecdysis means that sentence changes into semantic chunk or the part of it. Chunk-extension-into-sentence means that a semantic chunk extends into a sentence. In language understanding, we should take them as sentence, describing their sentence category and semantic chunks. For example,

```
The pain that economic crisis caused eased.
<ss code="P21">
    <gbk type="2">The pain</gbk>
    <gbk type="1">that the economic crisis</gbk>
    <ek>caused</ek>
</ss>
```

Figure 5. Sentence ecdysis tagging

The annotated part above is a sentence Ecdysis which sentence category code is P21(cause and effect sentence). Apparently, the pain is the headword and the attributive clause crisis causes is modifier. In fact, they can be reverted to a complete sentence form economic crisis causes the pain.

```
The schoolmaster hopes that they start a new study life.
<ss code="X20">
    <gbk type="1">they</gbk>
    <ek>start</ek>
    <gbk type="2">a new study life</gbk>
</ss>
```

Figure 6. Chunk-extension-into-sentence tagging

The annotated part is a chunk-extension-into-sentence which sentence category code is X20(basic reaction sentence).

Both chunk-extension-into-sentence and sentence Ecdysis are all semantic chunk or part of it. The difference of them is set according to the need of project. It is regulated that chunk-extension-into-sentence appears in definite 8 kinds of sentence categories. The chunk-extension-into-sentence is a chunk that must be a sentence form, but sentence Ecdysis is a chunk that can be either chunk or sentence form.

D. Separation of Semantic Chunk

In a sentence, some parts of a semantic chunk can be separated to different syntax positions, but they are one chunk on semantic layer. We use the *sep* element to describe this kind of semantic component. Separate chunk can appear on several positions, which can be both sub-element of sentence or sub-sentence and semantic chunk.

When tagging, we need to give the separate attribute information. We use the attribute from to describe that the position where it is separated. For example,

```
Beside candies, there are color pencils, a knife and some toys in the speaking trumpet.
<s code="jD">
    <sep from="gbk2">beside candies</sep>,
     <gbk type="1">there</gbk>
    <ek>are</ek>
    <gbk type="2"> color pencils, a knife and some toys </gbk>
    <fk type="Cn"> in the speaking trumpet </fk>.
</s>
```

Figure 7. Separation of semantic chunk tagging in English

The annotated part shows that the content of *sep* element is beside candies, the prepositional phrase here is a part of main semantic chunk, is separated from the 2nd gbk with the content color pencils, a knife and some toys.

It is interesting that separate phenomena are more common in Chinese than in English. In Chinese, both gbk chunks and Eigen chunk can separate into a few parts. For example,

```
他们被一阵敲门声打断了谈话。
<s code="X">
    <gbk type="2">他们</gbk>
    <gbk type="1">被一阵敲门声</gbk>
    <ek>打断了</ek>
    <sep from="gbk2">谈话</sep>。
</s>
```

Figure 8. Separation of semantic chunk tagging in Chinese

In Figure 8, The word 谈话 is the separate part which original position is in the second gbk. The deep structure of the sentence is 他们（的）谈话被一阵敲门声打断了. Such separation often occurs in passive voice.

```
我又把那份材料翻阅了一遍。
<s code="T19">
    <gbk type="1">我</gbk>
    <sep from="ek">又</sep>
    <gbk type="2">把那份材料</gbk>
    <ek>翻阅了一遍</ek>。
</s>
```

Figure 9. Separation of semantic chunk tagging in Chinese

In Figure 9, The word 又 is separated from eigen chunk 翻阅了一遍. The deep structure of the sentence is 我把那份材料又翻阅了一遍 which sentence form belongs to BA-sentence.

V. CONCLUSIONS

The building of sentence level semantic annotated corpus fills up the blank in the resource construction of Chinese Information Processing, which is important to the studies on both HNC theory and the fulfillment of HNC system of the sentence category analysis. Besides, it can also be helpful to many language researchers. It has great significance not only in Chinese information processing but also in language teaching and research.

In the future, we will extend our annotation to analyze the internal relationship of semantic chunks. Also we need to research the sentence group annotation which focuses on the study of Discourse and Discourse Situations. Besides, a corpus management system is to be developed with a friendly user interface.

ACKNOWLEDGMENT

The research is supported by “the Fundamental Research Funds for the Center Universities”.

REFERENCES

[1] Chuanjiang Miao Zhiying Liu, Sentence Level Semantic Tagging Based on Modern Chinese, Language Computing and Text Processing Based on Contents. Tsinghua University Press, 2005.

[2] Chuanjiang Miao, HNC (Hierarchical Network of Concepts) Theory Introduction, Tsinghua University Press, 2005.

[3] Chuanjiang Miao, Semantic Category of Sentence in Modern Chinese, Language Planning, pp. 56-58, 2006.

[4] Chuanjiang Miao, Semantic Research Based on the Sentence Category System of HNC Theory, Applied Linguistics, pp.126-133, Feb 2006.

[5] Elliott Rusty Harold, XML Bible(2nd edition). Electron Industry Press, 2002.

[6] Mark Birbeck, XML Advanced Programme(2nd edition), Engineer Industry Press, 2002.

[7] Yaohong Jin, Natural Language Understanding Based on the Theory of HNC (Hierarchical Network of Concepts), Sciences Press, pp.70-79, 2006.

[8] Yuhuan Chi, The Analysis and Processing of The Chinese Verbs’ Morphological Dilemma, 2005.

[9] Zengyang Huang, HNC (Hierarchical Network of Concepts) Theory, Tsinghua University Press, 1998.

[10] Zengyang Huang, The Basic Theorem and Mathematical and Physical Expressions of the Language Concept Space. Ocean Press, 2004.