Building a Chinese Lexical Taxonomy

Xiaopeng Bai, Nianwen Xue
Department of Computer Science, Brandeis University, Waltham, MA, USA 02453
{xpbai, xuen}@brandeis.edu

Abstract

In this paper, we present a Chinese lexical taxonomy, a hierarchically organization of Chinese lexical classes of nouns, verbs and adjectives. We first describe the structure of this taxonomy and then present the methods we used to build it. The distinctive characteristics of this lexical taxonomy are: 1) we use *definition frame* to describe each lexical class, as well as its members, 2) the lexical classes for nouns, verbs and adjectives are inter-connected. We also compare this taxonomy with the Chinese Proposition Bank, to look for possible ways to link these two independently developed language resources.

1 Introduction

A lexical semantic taxonomy is a hierarchical organization of lexical semantic classes. Such a taxonomy is a useful resource for Natural Language Processing, because it groups word senses into lexical semantic classes by their shared lexical meaning, and produces a finite set of lexical semantic classes. Since the lexical classes capture the shared meaning of individual senses, they can be used as a tagset to annotate words in a natural language corpus, which can then be used to train automatic lexical semantic classifiers. Compared with words sense disambiguation, where senses have to be defined for each word, classifying words based on their lexical classes is a more general task. The advantage is that there is no need to train classifiers for each individual word, as is typically the case for word sense disambiguation systems.

Building lexical semantic resources and systems has attracted much interest in the NLP and lexical semantics communities. (Picca et al., 2007, Ciaramita & Johnson, 2003) described a corpus annotated with the upper level synsets of WordNet (Fellbaum, 1998). (Gao et al, 2005) used lexical classes from Tongyici Cilin (Mei et al., 1983) for Chinese document retrieval, and (Tian et al, 2010) used the same resource to compute Chinese word similarity. One main drawback of these two lexical classification systems is that because the criteria for the lexical classification is not explicitly spelled out, when there is an out-of-vocabulary (OOV) sense, it is hard to determine its appropriate membership without going back to their original developers. Without explicit criteria, it is hard to ensure consistency when a new lexical taxonomy is established or an old one is extended. One desideratum in lexical taxonomy creation is consistency. Ideally, when a new word sense is put in taxonomy, different lexicographers/annotators should come up with the same class. This is also the biggest challenge in taxonomy/ontology development, and the key is to address this is to come up with concrete and explicit criteria that different lexicographers/annotators can follow so that there is no need to go back to the original creators every time a new word sense needs to be added to the taxonomy.

The rest of the article is organized as follows. In Section 2, we provide a brief review of related work. In Section 3, we present the structure and size of the current CLT as well as the corpus that is annotated with the lexical classes of the CLT. In Section 4, we show syntactic performances, semantic roles and selectional constraints are used to create the definition frame of each class. Comparison of CLT and Chinese Propbank (CPB) is performed in Section 5, and possible ways to link CLT to CPB are discussed in section 6.

2 Related Work

There have been several past efforts to produce (Chinese) lexical taxonomies aimed to provide lexical knowledge for NLP tasks (Chen, 1998; Chen, 2001; Wang etc., 2003). (Wang et al, 2003)

Proceedings of the Second CIPS-SIGHAN Joint Conference on Chinese Language Processing, pages 18–26,
Tianjin, China, 20-21 DEC. 2012

used lexical classes to describe word sense in SKCC (Semantic Knowledge Base of Contemporary Chinese), along with syntactic and argument structure features.

WordNet (Fellbaum, 1998). Gather senses with similar lexical meaning according to one or more dictionaries, and the lexical classes (synset in WN) are generated based on the judgment of word sense similarity. The judgment of similarity between word senses is depend on either the sense definition in dictionary or the intuition of developer. Such method is easy to use, but could be suffered with inconsistency among sense definitions (from different dictionaries) and different developers/annotators. It doesn't cost much at the initial stage of building taxonomy, but causes significant high cost to maintaining and expanding.

HowNet (Dong & Dong, 2006). HowNet uses "meaning primitives" (sememe in HN) as tagset to describe word senses, the computing of sense similarity and the generating of lexical classes can be automatically done. There is inconsistency problem encountered when adapting this method in such aspect: creating "meaning primitives" and expanding them in the future; selecting proper "meaning primitives" for defining word senses in consistent way.

As we argued in Section 1, a concrete definition for each class in a lexical taxonomy is required to ensure consistency. However, current Chinese lexical taxonomies generally do not provide such definitions. People have to create and extend their taxonomies by using dictionaries or the taxonomy made by other researchers, or by relying on their own intuition. Our work differs from others in that we use concrete linguistic features to define lexical classes. These class definitions can be used to extend the taxonomy by other researchers when new word senses need to be added to the taxonomy.

3 Status of CLT

In this section, we describe the structure and scale of the CLT taxonomy, as well as the corpus annotated with the lexical classes of this taxonomy.

3.1 Structure of CLT

CLT is a hierarchical structure formed by lexical classes, and each lexical class is a set of word senses that have shared lexical meaning and linguistic features. Currently we have three sub-taxonomies for nouns, verbs and adjectives respectively. Each sub-taxonomy has one root class, which dominates any number of terminal and non-terminal lexical classes. A given class can have one parent, one or more sisters and one or more children. Terminal classes do not have children. Table 1 shows part of the verb taxonomy in CLT.

1 自主变化 (self changing)
1.1 过程 (process)
——1.1.1 存现 (exist): 出土, 出现
——1.1.2 位移 (move): 流入, 上升
——1.1.3 变化 (transform): 消融, 变化
1.2 状态 (status)
——1.2.1 境遇 (situation)
————1.2.1.1 情绪 (emotion): 费心, 感恩
————1.2.1.2 生理状态 (physical situation): 打鼾, 咳
————1.2.1.3 其他 (other): 见鬼, 失礼
——1.2.2 自然现象 (natural phenomenon): 结冰, 降温
——1.2.3 一般状态 (circumstance): 无力, 作罢
——1.2.4 运动 (motion): 摆动, 翻卷
1.3 经历 (experience)
——1.3.1 经历 (experience): 处身, 拘泥
——1.3.2 感知意向 (attitude): 向往, 对得起
——1.3.3 所有 (possess): 装有, 有着
——1.3.4 影响 (influence): 震撼, 照耀
——1.3.5 产生 (generate): 组成, 泛起

Table 1: part of verb taxonomy

In table 1, node "1 自主变化 (self changing)" is a non-terminal class that has three children: "1.1 过程 (process)", "1.2 状态 (status)" and "1.3 经历 (experience)". These three classes are also non-terminal classes. They are sisters that inherit all the features of their parent "1 自主变化 (self changing)", and they also have some unique features of their own that distinguish themselves from one another. Classes "1.1.1 存现 (exist)", "1.1.2 位移 (move)" and "1.1.3 变化 (transform)" are terminal classes, because they have no child, and they are sisters. "1.2.2 自然现象 (natural phenomenon)" is a terminal class, while its brother, "1.2.1 境遇 (situation)" is a non-terminal class, since it has three children. The depth of taxonomy is not even, and among sister classes, some classes might be terminal nodes while others might be

non-terminal classes. Only terminal node classes contain word senses, while non-terminal classes have only the definition of the class, which we will discuss in detail in Section 4.

3.2 Scale of CLT

The members of each terminal class are word senses. The sense entries from Xiandai Hanyu Cidian (XH, 5th edition, Commercial Press, China) are our starting point. Different word senses of a polysemous word may be grouped together into the same lexical class or put into different lexical classes. For example, verb 落 has two senses in the XH Dictionary. One is the action of things dropping as a result of gravity, as in 树叶落下 ("The leaves dropped on the ground"). Another denotes the action of descending, as in 飞机落地 ("The aircraft landed"). These two senses are grouped into the same lexical class "1.1.2 位移 (move)". 1357 word types in corpus are polysemous and have more than one sense and are classified into different lexical classes.

There are 33480 word types and 46934 sense entries in the CLT that belong to 153 terminal classes.

Noun taxonomy. 25801 noun senses are grouped into 97 terminal classes. The maximum depth of the noun taxonomy is 5. Table 2 is part of noun taxonomy.

1 具体物 (concrete)
——1.1 生物 (living creature)
————1.1.1 人 (human)
————1.1.1.1 身份 (identification): 学生, 冠军
————1.1.1.2 关系 (relative): 司令, 科长
————1.1.1.3 超人 (superman): 观音, 上帝
————1.1.1.4 其他 (other): 汉人, 小伙子
————1.1.2 动物 (animal)
————1.1.2.1 兽 (beast): 狗, 老虎
————1.1.2.2 鸟 (bird): 麻雀, 大雁
————1.1.2.3 鱼 (fish): 鲤鱼, 青蛙
————1.1.2.4 虫 (insect): 蜈蚣, 苍蝇
————1.1.2.5 微生物 (micro living): 结核菌, 酵母
————1.1.3 植物 (botany)
————1.1.3.1 草木 (plant): 常青藤, 报春花
————1.1.3.2 果实 (fruit): 银杏果, 鸭梨
————1.1.4 群体 (group)
————1.1.4.1 机构 (institute): 总统府, 医学院
————1.1.4.2 团体 (organization): 训练团, 媒体
————1.1.4.3 其他 (other): 猪群, 人类
————1.1.5 生物部分 part
————1.1.5.1 肢体 (body): 触手, 右腿
————1.1.5.2 器官 (organ): 小肠, 五脏
————1.1.5.3 其他 (other): 落叶, 鹅毛
——1.2 非生物 (non-living creature)

Table 2: part of noun taxonomy

Verb taxonomy. 15920 verb senses are grouped into 37 terminal classes. The maximum depth is 4. Table 1 shows part of verb taxonomy.

Adjective taxonomy. The adjective senses taxonomy is the smallest. There are 5213 adjective senses in 19 terminal classes. Table 3 is part of adjective taxonomy.

1 生物属性值 (attribute value of living creature)
——1.1 生理 (physiological): 年轻, 疲劳
——1.2 心理 (mental): 困, 反感
——1.3 品性 (ethic): 酸, 清高
——1.4 状况 (situation): 背运, 没出息
2 其他属性值 (other attribute value)
——2.1 物理 (physical)
————2.1.1 可度量值 (measurable): 深, 粗
————2.1.2 不可度量值 (unmeasurable): 黏, 松
——2.2 内容值 (content): 深, 粗犷
——2.3 状态值 (situation): 顺, 袅袅
——2.4 其他 (other): 毒, 经济
3 方式事件值 (attribute of behavior and event): 正面, 自动
4 时空值 (attribute of spatio-temporal)
——4.1 时间值 (temporal): 原先, 悠久
——4.2 空间值 (spatio): 浩渺, 闹哄哄

Table 3: part of adjective taxonomy

3.3 Corpus Annotation

We also used the CLT to annotate a Chinese text corpus. The corpus we annotated is called the Chinese Sense Corpus, which consists of texts of Chinese textbooks. The corpus has 2,008 texts, 51,343 word types, 1,475,913 word tokens, and 2,186,853 character instances. The corpus is developed by National University of Singapore (Singapore), Commercial Press (China) and Peking University (China). We also used this corpus to extract the linguistic features to help create the sense classes.

4 Definition Frame for CLT

According to (B. Levin, 1993), the syntactic behaviors of word are determined by the meaning of the word. Therefore, we assume that senses with similar syntactic behaviors or other linguistics features (e.g. argument structure), can be considered as in one lexical class. Table 4 shows the definition frame of verb lexical classes "1.1.1 存现 (exist)" and "1.1.2 位移 (move)" and table 5 is the definition frame of noun class "2.1.3 生理属性值 (physiological attribute)".

1.1.1 存现 (exist) (v.) Syntactic performance: + subject, + object Argument structure: subject: Theme, Location; object: Theme, Location Selectional restriction: N.A 1.1.2 位移 (move) (v.) Syntactic performance: + subject, + object Argument structure: subject: Theme; object: Location Selectional restriction: N.A

Table 4: verb classes "1.1.1 存现 (exist)" and "1.1.2 位移 (move)"

2.1.3 生理属性值 (physiological attribute) (n.) Syntactic performance: *modifier Semantic role: subject: Theme; object: Content, Experiencer Selectional restriction: in modifier-head structure, the modifier can only be nouns of Living Creature

Table 5: definition frame of noun class "2.1.3 生理属性值 (physiological attribute)"

4.1 Linguistic Features in Definition Frame

There are three components in the definition frame, and each one presents a type of linguistic features of word sense:

Syntactic performance. Each sense is eligible to occupy certain syntactic positions in sentence. Senses in the same lexical classes have similar syntactic performances. We have syntactic frames to test the syntactic performances of word senses. For example, "verb (object)" frame is used to test whether a verb sense takes object. "verb (head)" is used to test whether a verb sense occupies adverbial position. "noun (head)" tests whether a noun sense occupies modifier position. "(head) adjective" tests whether a adjective occupies complement position. In table 4, operator "+" means "takes", both "1.1.1 存现 (exist)" and "1.1.2 位移 (move)" take subject and object. Operator "*" means "cannot occupy", senses of "2.1.3 生理属性值 (physiological attribute)" class cannot occupy the modifier position in "noun (head)" frame.

Argument structure/ semantic role. For verb senses, those in the same lexical class may share same argument structure: same number of arguments and same semantic roles. For noun senses, it concerns what specific semantic roles a noun sense acts. We have a scheme to identify the number of arguments that verb sense governs, and a semantic roles list noun acts.

The identification of arguments of a word sense is based on its syntactic frame. If a particular noun sense can be in the subject or object position, we identify the semantic roles of the noun sense in the positions. Notice that it is possible for a syntactic position to have more than one type of arguments. In table 4, since both "1.1.1 存现 (exist)" and "1.1.2 位移 (move)" take a subject and an object, the semantic roles of their arguments are identified in these positions. That is why we specify the syntactic positions before the semantic role labels. These two verb classes have similar syntactic behaviors and selectional restrictions, but they are distinguished from each other by their argument structure.

We have 10 semantic roles for arguments: Agent, Theme, Patient, Experiencer, Participant, Result, Content, Instrument, Time, and Location.

Selectional restrictions. Also known as semantic preferences, selectional restriction denotes semantic constrains between word senses within a syntactic constructions.

The definition frame is set of linguistic features for creating lexical classes and identifying which class a particular word sense should be assigned to. There are three components in each definition frame, and they are used sequentially. If the syntactic features can be used to create sub-classes, or assign a particular word sense to a proper lexical class, we will not use argument structure and selectional restriction features. In other words, syntactic structures are given precedence over the other two types of features.

Some of the selectional restriction features are lexical classes in the CLT. For the "2.1.3 生理属性值 (physiological attribute)" class, it takes noun class "1.1 生物 (living creature)" as a selectional restriction. From a particular lexical class, we can trace other lexical classes via the lexical class tags in definition frame of that class. This makes the lexical classes inter-connected, a point we will discuss in greater detail in Section 4.3.

4.2 How Definition Frame Works

In this subsection we present three examples to show how a definition frame works. Example 1 shows how to use definition frames to distinguish different senses. Example 2 shows how the senses of a polysemous word are determined to belong to one lexical class. Sample 3 shows how senses of a polysemous word are determined to belong to different lexical classes.

Example 1: distinguishing word senses. Sample members from verb class "1.1.1 存现 (exist)" and "1.1.2 位移 (move)" to show how senses belong together, and how they are separated to different classes. Table 6 gives some member senses of these two classes:

1.1.1 存现 (exist) (v.)
出土 (to be excavated), 充满 (fulfill), 出现 (appear), 发生(happen)
1.1.2 位移 (move) (v.)
通过(1, pass), 上升(1, raise), 后退 (fall back), 落入 (fall into)

Table 6: sample senses (the number inside the parentheses indicates the sense number from XH)

For these 8 verb senses, they all take both subject and object:

1) [这件 文物]/subject 出土 [于 龙门石窟]/object
the antique excavate Yu Longmen Shiku.
The antique is excavated in Longmen Shiku.
2) [难闻的味道]/subject 充满 了 [房间]/object
smelly De scent fulfill Le room
The room is fulfilled with smelly scent.
3) [太阳]/subject 出现 [在 东方]/object
sun appear at east
The sun appeared from the east.
4) [事故]/subject 发生 [在 南京路]/object
accident happen at Nanjing Road
The accident is happened at Nanjing Road.
5) [火车]/subject 通过 [隧道]/object
train pass tunnel
The train passed the tunnel.
6) [飞机]/subject 上升 [到 高空]/object
aircraft raise to high altitude
The aircraft has raised to high altitude.
7) [洪水]/subject 后退 [到 警戒线 以外]/object
flood fall back to alarm line behind
The flood has fallen back behind the alarm line.
8) [树叶]/subject 落入 [水中]/object
leaf fall into water inside
The leaf is falling into the water.

In examples 1) to 8), the semantic role of the argument in the subject position is Theme, and the semantic role of the argument in the object position is Location. That's why the 8 senses are in verb class "1.1 过程 (process)". For 1) to 4), the semantic role of the argument in the subject position can be Location, and Theme for the argument in the object position (see example 1a) to 4a)), while this is illegal for 5) to 8) (see 5a) to 8a)):

1a) [龙门石窟]/subject 出土 了 [这件 文物]/object
Longmen Shiku excavate Le the antique
The antique is excavated in Longmen Shiku
2a) [房间]/subject 充满 了 [难闻 的 味道]/object
room fulfill Le smelly De scent
The room is fulfilled with smelly scent.
3a) [东方]/subject 出现 了 [太阳]/object
east appear Le sun
The sun appeared from the east.
4a) [南京路]/subject 发生 了 [事故]/object
Nanjing Road happen Le accident
The accident is happened at Nanjing Road.
5a) *[隧道]/subject 通过 [火车]/object
tunnel pass train
6a) *[高空]/subject 升上 [飞机]/aircraft
high altitude raise to aircraft
7a) *[警戒线]/subject 以外 后退 [洪水]/object
alarm line behind fall back flood
8a) *[水中]/subject 落入 [树叶]/object
water fall into leaf

Since the position of arguments of 通过, 上升, 后退 and 落入 cannot exchange (as which is legal to 出土, 充满, 出现 and 发生), they are put in class "1.1.2 位移 (move)", while 出土, 充满, 出现 and 发生 are classified into "1.1.1 存现 (exist)".

Example 2: senses of a polysemous word go to one lexical class. Chinese noun 阿姨 has three senses according to XH:

阿姨 (n.) 1. 母亲的姐妹 (sisters of mother, aunt) 2. 和母亲年龄差不多大的女性 (ladies at mother's age) 3. 保姆 (babysitter or maid)

Table 7: sense definitions of 阿姨 from XH

The three senses of 阿姨 denote human being, so they go to noun class "1.1.1 人 (human)", and we should choose each sense a lexical class from the children of "1.1.1 人 (human)". The candidates are "1.1.1.1 身份 (identification)", "1.1.1.2 关系 (relative)", "1.1.1.3 超人 (superman)" and "1.1.1.4 其他 (other)". We first exclude "1.1.1.3 超人 (superman)", which denotes fictional human, like 上帝 (God), 菩萨 (Buddha). If the senses cannot fit definition frame of either "1.1.1.1 身份 (identification)" or "1.1.1.2 关系 (relative)", then they will be put into "1.1.1.4 其他 (other)". Therefore, we need to test the senses only in the definition frames of "1.1.1.1 身份 (identification)" and "1.1.1.2 关系 (relative)". Table 8 and 9 are definition frames of "1.1.1.1 身份 (identification)", "1.1.1.2 关系 (relative)":

1.1.1.1 身份 (identification) (n.) Syntactic performance: subject, object, modifier, head Semantic roles: Agent, Theme, Experiencer, Patient, Participant Selectional restrictions: if occupy head position of "modifier-head" structure, the modifier can be nouns of country, city, organization.

Table 8: definition frame of "1.1.1.1 身份 (identification)"

1.1.1.2 关系 (relative) (n.) Syntactic performance: subject, object, modifier, head, parenthesis Semantic roles: Agent, Theme, Experiencer, Patient, Participant Selectional restrictions: if occupy head position of "modifier-head" structure, the modifier can be people's name

Table 9: definition frame of "1.1.1.2 关系 (relative)"

The three senses of 阿姨 can be used as "title for people" in a sentence, for people to call other people. And if they occur in the head position of a "modifier-head" structure, the modifier can be people's names, but not names of countries, cities or organizations:

9) 张 阿姨
zhang aunt/lady/maid
Mrs. Zhang/ Aunt Zhang

9a) *中国 阿姨 / 北京 阿姨 / 大学 阿姨
China aunt/ Beijing aunt/ university aunt

According to definition frame of "1.1.1.2 关系 (relative)", three senses of 阿姨 should be put into this class.

Example 3: senses of a polysemous word go to different lexical classes. In XH, Chinese verb 爆发 has two senses:

爆发 (v.) 1. 火山的岩浆冲破地壳，向四外迸出 (volcanic eruption) 2. 突然发生 (suddenly happen)

Table 10: sense definitions of 爆发

For the argument of the subject of either of the senses, the semantic roles are Theme, thus both of them are fallen into class "1 自主变化 (self changing)". Syntactically, sense 1 of 爆发 is intransitive, i.e. it cannot take object:

10) 火山 爆发 了
volcano erupt LE
The volcano is erupting.

10a) *爆发 [火山]/object 了
erupt volcano LE

While sense 2 is transitive:

11) [多个 城市]/subject 爆发 [抗议 活动]/object
several city suddenly happened protest event
Protests are suddenly happened in several cities.

According to the definition frame of sub-classes of "1 自主变化 (self changing)", "1.2 状态 (status)" is for intransitive verb senses, "1.1 过程 (process)" and "1.3 经历 (experience)" are for transitive senses. Therefore, sense 1 of 爆发 falls into either "1.1 过程 (process)" or "1.3 经历 (experience)", and sense 2 falls into "1.2 状态 (status)".

The subject of sense 2 is specific to volcano, which is a kind of geographic entity. According to

the selectional restrictions of sub-classes of "1.2 状态 (status)", only "1.2.2 自然现象 (natural phenomenon)" requires geographic entity for the subject, so the lexical class for sense 2 of 爆发 is "1.2.2 自然现象 (natural phenomenon)".

For sense 1, the semantic roles of arguments of subject and object are Theme and Location, and it barely takes other roles. Semantic roles required by "1.3 经历 (experience)" are Theme, Patient, Content, Result and Experiencer, thus sense 1 of 爆发 is not belong to "1.3 经历 (experience)". Additionally, the positions of the arguments of sense 1 are exchangeable, which matches the definition frame of "1.1.1 存现 (exist)", so sense 1 of 爆发 is grouped into class "1.1.1 存现 (exist)".

4.3 Inter-Connectivity of Classes

The classes in sub-taxonomies are inter-connected, via the selectional restriction part of the definition frame of lexical classes. For example, the selectional restriction part of definition frame of "1.1.1 人 (human)":

1.1.1 人 (human) (n.) Syntactic performance: …… Semantic roles: …… Selectional restrictions: When occupying subject position in "subject-predicate" structure, requires predicates denoting: verb senses of social act, intended mental act; When occupying head position in "modifier-head" structure, requires modifiers denoting: noun senses of institute or organization, or adjective senses of human physiological, mental or social features.

Table 11: the selectional restriction part of definition frame of "1.1.1 人 (human)"

According to the selectional restrictions, senses of "1.1.1 人 (human)" collocate with verb senses of social act or intended mental act, noun senses of institute or organization, adjective senses of human physiological, mental or social features. Most of these senses can match classes in the taxonomy. There are verb classes "3.1.2 社会行为 (social behavior)", "3.3 社会活动 (social act)" denoting the meaning of social act, "3.4 心理活动 (mental act)" denoting intended meaning of intended mental act. We have noun classes with institution and organization meanings: "1.1.4.1 机构 (institute)" and "1.1.4.2 团体 (organization)". And there are adjective classes "1.2 心理 (mental)" denoting human mental features, and "1.3 品性 (ethic)" denoting human social features. So, noun class "1.1.1 人 (human)" is connected with verb classes "3.1.2 社会行为 (social behavior)", "3.3 社会活动 (social act)" and "3.4 心理活动 (mental act)", and with adjective classes "1.2 心理 (mental)" and "1.3 品性 (ethic)".

4.4 Complications

The motivation we use definition frame in building lexical taxonomy is to ensure the consistency for identifying lexical classes for word senses. The definition frame is a schema we follow when trying to assign a particular word sense to a proper lexical class and we want it to play an essential role in building and extending lexical taxonomy, but there are complications as a result of the morphological processes in Chinese.

The morphology structure of a word can mirror the syntactic structure of a phrase at the syntactic level, and this creates difficulties when classifying the words. For example, according to the definition frame of noun class "2.1 属性 (attribute)", senses belonging to this class denote a kind of attribute of entities and cannot be the subject by itself in a "subject-predicate" structure. For example, 颜色 (color) belongs to this class, the sentence 颜色 很好看 (color is beautiful) cannot be understood unless we add "host word" to form "modifier-head" structure to specify "whose/what thing's color is beautiful". So, 衣服 的 颜色 很 好看 (color of the cloth is beautiful) is interpretable, because the "host word" 衣服 is added forming "modifier-head" structure 衣服 的 颜色 (color of the cloth). In some cases, such the "host word" is a morpheme of a word. For example, in 月色 ("color of the moon"), the morpheme 月("moon") is the "host word" of 色 ("color"), so for the sense 月色, it breaks the syntactic performance rule in definition frame, therefore we cannot treat 月色 as member of "2.1 属性 (attribute)". But lexical semantically, 月色 denotes a particular attribute of moon, it doesn't make any sense if we do not put 月色 in "2.1 属性 (attribute)". Such cases also happen for verb senses, and some verb senses have

an object morpheme, like 拜师 (to become a student to a mentor), 播音 (broadcast).

5 Linkability of CLT and CPB

Propbank is a corpus that annotates predicates with argument labels. It is based on Treebank, where the syntactic trees present the syntactic relations between a predicate and its arguments. Verb senses in Propbank are called "framesets", which are defined based on the argument structure of a predicate. Annotation of the arguments of a verb sense follows the framesets of the sense. Chinese Propbank (CPB) (Xue and Palmer, 2009) is based on the Chinese Treebank (Xue et al, 2005).

As one type of features for formally describing the lexical semantic meaning of a word sense, argument structure plays essential role in the CLT as well. CLT uses semantic roles of arguments globally, which is a major difference between CLT and CPB. Table 12 presents a sample of frameset of the verb "爱".

```
<id>爱</id>
<frameset cdef="" edef="" id="f1">
  <role argnum="0" argrole="love giver"/>
  <role argnum="1" argrole="thing, person loved"/>
  <frame>
      <mapping>
      <V/>
      <mapitem src="sbj" trg="arg0"/>
      <mapitem src="npobj" trg="arg1"/>
    <comment/>
</mapping>
```

Table 12: sample of frameset of "爱"

The "argrole" field is the semantic role of argument, which in CPB is individually for each frameset. There is not a global list of semantic roles for the CPB, as shown in table 12. Verb sense is described by selectional restrictions that are similar to noun lexical classes in the CLT. For "爱" in Table 12 , ARG0 is "love giver", which can be nouns denoting people; ARG1 is "thing/person loved", which can be entities or person. The lacking of global semantic role list makes the verb senses in CPB are isolated from each other and are not connected.

Although CLT and CPB are independently developed language resources, lexical meanings of verb in both are represented by argument structure. Therefore, we believe CLT and CPB are linkable by replacing CPB's semantic roles with CLT's.

6 Conclusion and Future Work

In this paper, we presented the Chinese Lexical Taxonomy, and the Chinese Sense Corpus annotated with the lexical classes in the taxonomy. Each lexical class in CLT is described via a definition frame, which is collection of linguistic features. We show the definition frame reduces the possible inconsistency that may happen in taxonomy creation. Compared to WordNet and HowNet style, CLT is being unique on the way we create it. The methodology creating CLT enables its predictivity for the possible lexical classes of an OOV word sense. It also maintains the inter-consistency among different annotators. The definition frame is the key to our goal, which is constituted of steps can be followed both in making corpus annotation and taxonomy expanding.

We also compare the CLT with the CPB. The absence of a global semantic role list in the CPB makes verb senses disconnected from each other. Since there is not a global list of semantic roles in the CPB, we will use the semantic roles of the CLT to annotate arguments in CPB. We will also add new semantic roles if the current semantic roles are insufficient for the CPB. We will also acquire a list of syntactic frames and alternations to create a more fine-grained definition frames for the CLT.

Acknowledgement:
This work is supported partial by DARPA via Contract HR0011-11-C-0145. All views expressed in this paper are those of the authors and do not necessarily represent the view of DARPA.

References

Bai, Xiaopeng. 2012. Building Word Sense Taxonomy and Automatic Annotation for Mandarin Chinese. PhD Thesis, National University of Singapore.

Bai, Xiaopeng. 2008. The Word Sense Category based on Semantic Features of Argument. *Proceeding of Chinese Lexical Semantics Workshop*, Singapore.

Chen, Qunxiu. 2001. Expanding of Machine Tractable Dictionary of Contemporary Chinese Predicate Verbs and Research on Relations of Slots Centering on Noun of Contemporary Chinese. *Applied Linguistics (Yuyan Wenzi Yingyong)*, No. 4, P98-04.

Chen Xiaohe. 1998. A Lexical Classification System for Language Engineering. *Applied Linguistics (Yuyan Wenzi Yingyong)*, No. 2, P71-76.

Ciaramita, Massimiliano & Johnson, Mark. 2003. Supersense tagging of unknown nouns in WordNet. *Proceedings of the 2003 conference on Empirical methods in natural language processing*, Stroudsburg, PA, USA.

Dong, Zhendong & Dong, Qiang. 2006. Hownet And the Computation of Meaning. World Scientific Publishing Company, Singapore.

Fellbaum, Christiane. 1998. WordNet: An Electronic Lexical Database. MIT Press, USA.

Gao, Liqi et al. 2005. Thesaurus-Based Semantic Smoothing in Language Modeling for Chinese Document Retrieval. *International Conference on Mutilingual Information Processing*.

Picca, David et al. 2007. Semantic Domains and Supersense Tagging for Domain-Specific Ontology Learning. *Conference RIAO2007*, Pittsburgh, PA, USA.

Levin, Beth. 1993. English Verb Classes and Alternations: A Preliminary Investigation. The University of Chicago Press, Chicago, US.

TIAN, Jiu-le et al. 2010. Words Similarity Algorithm Based on Tongyici Cilin in Semantic Web Adaptive Learning System. *Journal of Jilin University (Information Science Edition)*, 2010-06.

Wang, Hui et al. 2003. The Specification of The Semantic Knowledge Base of Contemporary Chinese. *Journal of Chinese Language and Computing*, 13(2).

Xue, Nianwen, et al. 2005. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2):207-238.

Xue, Nianwen. 2008. Labeling Chinese Predicates with Semantic Roles. *Computational Linguistics*, 34 (2): 225-255.

Xue, Nianewn and Palmer, Martha. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143-172.