

Handling Dislocated and Discontinuous Constituents in Chinese Semantic Role Labeling

Nianwen Xue

Department of Computer and Information Science
University of Pennsylvania
Philadelphia, PA 19104

Abstract

This paper discusses two outstanding issues, dislocated and discontinuous arguments, in the construction of the Penn Chinese PropBank, a corpus that has already been treebanked. It describes the verb-specific approach in our semantic role annotation in which *arguments* and *adjuncts* are treated differently in the sense that arguments are assigned labels that are interpretable only within the scope of the verb while adjuncts receive labels that reflect a more global classification. The paper discusses the rationale behind this approach in comparison with other semantic role labeling projects.

1 Introduction

Recent years have seen an emergence of what can be generally called semantic parsing using statistical approaches (Gildea and Jurafsky, 2002; Gildea and Palmer, 2002) due to the availability of corpora annotated with semantic structures. Most of the semantic annotation or parsing so far deals with some form of semantic role labeling presumably due to its wider acceptance among researchers and the ease of annotation. The types of semantic role labeling vary from the use of very general role labels such as agent, theme, beneficiary (Chen et al., 2004), to labels that are meaningful to a specific situation (Baker et al., 1998), to verb-specific labels (Kingsbury and Palmer, 2002; Xue and Palmer, 2003). The

difference between the various approaches can be characterized in terms of levels of abstraction. The Propbank style of annotation can be considered to be the least abstract, as the role labels (*arg0*, *arg1*, etc.) in a propbank are meaningful only with regard to a specific verb. The FrameNet role labels are more abstract than a propbank label in the sense that they abstract away from any specific verb and instead apply to a class of related verbs (or nouns that have predicate-argument structures). Using labels that have global meanings abstracts from specific verbs or a class of verbs and instead applies to all verbs (or any other categories that denote a relation). Each of these approaches have its own advantages but an argument can be made that the more abstract it is, the more difficult it is for the computer to acquire that concept automatically, using the currently dominant statistical paradigm. By sticking to less abstract semantic roles, there might be a better chance of labeling them correctly by the computer. In addition, it can also be argued that it is difficult to generalize the situations that occur in natural language to a very limited set of semantic role labels that have global meanings. For a simple sentence like "A belongs to B", it is hard to say what "global" roles to assign to "A" and "B". We could call it "theme", but then the concept of "theme" would be so loose that no one would know what to do with it. Conversely, if we use role labels such as *arg0* and *arg1*, that are interpretable only within the scope of this verb, we might know what to do with the things that are so annotated when we translate them into another language. Thirdly, it is really important to proceed in replicable steps in order to measure progress. Even if we

later on decide that we can make verb-independent generalizations, verb-specific annotation would be a good first step in that direction.

On the other hand, using the less abstract role labels might lead to loss of generalizations. For example, in the Penn English Propbank (Kingsbury and Palmer, 2002), the arguments for “buy” and “sell” are defined as follows:

```
Frameset sell.01 "sell":
Roles: Arg0: Seller
       Arg1: Thing Sold
       Arg2: Buyer
       Arg3: Price Paid
       Arg4: Benefactive
```

```
Frameset buy.01 "purchase":
Roles: Arg0: buyer
       Arg1: thing bought
       Arg2: seller
       Arg3: price paid
       Arg4: benefactive
```

In contrast, the FrameNet annotation is independent of verbs and applies to a class of verbs or nouns. For example, “charge”, “lease”, “rent”, “retail”, “retailer”, “sale”, “sell”, “vend”, “buy”, “purchase”, “purchaser”, “rent” are all treated as realizations of the concept “Commerce_goods-transfer”, called ‘Frame’ in the FrameNet, in which there are four roles:

```
Commerce_goods-transfer
Buyer [Byr]
Goods [Gds]
Money [Mny]
Seller [Slr]
```

So while in the Propbank “Jess” would be labeled as *arg0* in (1a) and *arg2* in (1b), in the FrameNet “Jess” would be labeled as *Byr* in both occurrences. So there seems to be some loss of generalization in the Propbank annotation. However, once the propbank-style annotation is available, it is not hard to build more abstract annotations in a bottom-up manner. Based on these considerations, the Penn Chinese PropBank (Xue and Palmer, 2003) adopted a verb-specific approach in semantic role labeling for the arguments. Semantic adjuncts, on the other hand, are generally independent of verbs and thus

are assigned labels that denote more global concepts such as “temporal”, “location”, etc.

- (1) a. Jess BOUGHT a textbook.
- b. Lee SOLD a textbook to Jess.

The remainder of this paper is organized as follows. In Section 2, we will discuss the formal encoding devices of the Penn Chinese PropBank in greater detail. In Section 3, we will discuss the dislocated constituents and how they are represented in the Chinese PropBank. In Section 4, we will describe how we treat discontinuous arguments. Section 5 concludes this paper.

2 The Chinese Propbank annotation scheme

In this section, we will discuss the encoding scheme for the Chinese PropBank in greater detail. As we have briefly mentioned in the last section, the arguments and the adjuncts are annotated with different formal schemes based on the observation that arguments are generally more specific to a verb¹ while adjuncts are generally independent of verbs. We number the arguments of a verb sequentially, starting from 0. The argument number is prefixed with “arg”. The adjuncts, on the other hand, are always tagged “argM”, followed by a secondary tag that indicates the adjunct type. The actual predicate is marked as “REL”:

- (2) a. 中/China 美/the U.S. 交往/contact 的/DE
 大门/door 打开/open 了/ASP。
 “The door of contact between China and the
 U. S. opened.”
 arg1: 中美交往的大门
 REL: 打开
- b. 70年代/70s 初/beginning, 中/China
 美/the U.S. 两/two 国/country 领导人/leader
 果断/decisively 打开/open 了/ASP
 中/China 美/the U.S. 交往/contact
 的/DE 大门/door。
 “In the beginning of the 1970s, the leaders
 of China and the U.S. decisively opened the
 door of contact.”
 argM-TMP: 70年代初

¹The same also applies to nominal predicates. We limit our discussion to verbs in this paper.

arg0: 中美两国领导人
 argM-ADV: 果断
 REL: 打开
 arg1: 中美交往的大门

Note that even “中/China 美/the U.S. 交往/contact 的/DE 大门/door” occurs in different syntactic positions in (2a) and (2b). The role label an argument receives is independent of its syntactic realizations. The semantic roles or expected arguments can be realized in different ways syntactically. It should also be pointed out that the line drawn between arguments and adjuncts here is slightly different from what has been generally assumed in the theoretical linguistic literature, which is generally based on the obligatory/optional dichotomy. In a lot of cases, some constituents are clearly arguments but they are also clearly optional. For example, in the unaccusative (or pseudo-passive) construction, the agent is clearly optional syntactically and it is clearly an argument. In (2a), for example, the “door-opener” is optional but is clearly an argument.

In the 3500 or so verbs that we have done so far, the largest number of arguments a verb can have is 5. The four verbs that have five arguments are “缩短/shorten”, “提高/improve”, “扩大/expand”, “贬/devalue”. Interestingly, the semantic roles for these verbs are also similar:

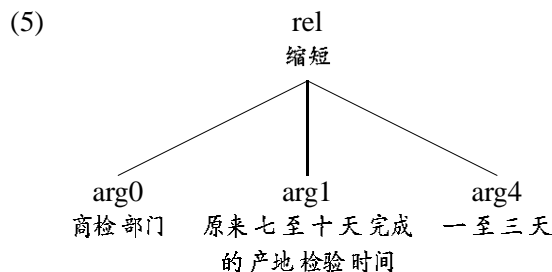
arg0: cause or agent
 arg1: theme
 arg2: range
 arg3: starting point
 arg4: end point

In any given occurrence, not all arguments have to be realized. In fact, it is rare that all five arguments of these verbs are realized.

(3) 商检/commercial inspection 部门/department 将/BA 原来/original 七/7 至/to 十/10 天/day 完成/finish 的/DE 产地/product origin 检验/check 时间/time 缩短/shorten 到/to 一/1 至/to 三/three 天/day。

“The commercial inspection departments shortened the time of product origin checking from the original 7 to 10 days to 1 to 3 days.”

arg0: 商检 部门
 arg1: 原来七至十天完成的产地检验时间
 arg4: 一至三天



Secondary tags for semantic adjuncts	
ADV	adverbial, default tag
BNF	beneficiary
CND	condition
DIR	direction
EXT	extent
FRQ	frequency
LOC	locative
MNR	manner
PRP	purpose or reason
TMP	temporal
TPC	topic
Secondary tags for discontinuous arguments	
CRD	coordinated arguments
PRD	predicate
PSR	possessor
PSE	possessee
Secondary tags for arguments to phrasal verbs	
AS	为, 是, 作, 做
AT	在, 于
INTO	成, 入, 进
ONTO	上
TO	到, 至
TOWARDS	向, 往

Table 1: Secondary tags

Some verbs can take on different sets of arguments that are realized in different sets of subcategorization frames. For example, the verb “存”, when it means “exist”, two arguments are expected: the thing that exists and the location or domain in which it exists. When it means “deposit”, three arguments are expected: the entity that makes the deposit, the sum of money deposited and the financial institution which the deposit is made. When it means “preserve”, we expect three different arguments: the agent that does the preservation, the thing preserved and the instrument used in the preservation. Since each of these three senses can be realized in different subcategorization frames, in the propbank annotation convention, these senses are called *frameset* meaning sets of subcategorization frames that realize a particular sense. The examples in (4) illustrates the three frameset of “存”.

(4) Framesets of “存”

Frameset 1: “deposit”

Semantic roles (or expected arguments):

arg0: entity making deposit

arg1: sum of money

arg2: financial institution

Examples:

(a) [**argM-TMP** 二十/20 年/year 前/ago] [**arg0** 每/each 人/person] [**argM-ADV** 平均/average] [**argM-ADV** 才/only] [**rel** 存/deposit] [**arg1** 二十/20 元/Yuan 钱/money]

“Twenty years ago, on average each person has only a deposit of 20 yuan.”

(b) [**arg1** 大批/large amount 资金/fund] [**rel** 存/deposit] [**arg2** 在/in 中/mid-sized 小/small 金融/financial 机构/institution]

“A large amount of fund is deposited in mid-sized or small institutions.”

Frameset 2: “exist”

arg0: location

arg1: thing that exists

Examples:

(c) [**argM-TMP** 现/now] [**arg0** 全球/the whole world] [**argM-ADV** 仅/only] [**rel** 存/exist][**arg1** 一千多/thousand 只/CL 大/giant 熊猫/panda]

“There exist only a little more than one thousand giant pandas in the whole world.”

Frameset 3: “preserve”

arg0: preserver

arg1: thing preserved

arg2: instrument

Examples:

(d) [**arg2** 方志/chronicle] 可以/can [**rel** 存/preserve] [**arg1** 史/history], 资/maintain 治/order, 教化/civilize
“Chronicles can be used to preserve history, maintain social order and teach civilized behavior.”

(e) [**arg0** 我们/we] [**arg2** 用/with 方志/chronicle] 存/preserve [**arg1** 史/history], 资/maintain 治/order, 教化/civilize

“We can use chronicles to preserver history, maintain social order and teach civilized behaviors.”

Given what we have described so far, each verb instance can be viewed as a semantic dependency tree with the verb predicate as the head.

3 Dislocated constituents

Ideally all the arguments and adjuncts of a predicate are all in a localized syntactic domain. However, it is well established that natural language allows long-distance dependencies and Chinese is no exception. For example, Chinese allows topicalization where the topicalized argument can be arbitrarily far away from its predicate. In (6), the topicalized argument “华人/Chinese 法官/judge 律师/lawyer” is non-adjacent from its predicate “培养/train”.

(6). A topicalized sentence

(IP (IP (NP-SBJ (QP (CD 三/three)) (ADJP (JJ 大/main)) (NP (NN 法典/law))) (VP (VV 需/need) (VP (VV 加快/accelerate) (NP-OBJ (NN 出台/promulgate) (NN 进程/process))))))

(PU ,)

(IP (NP-TPC-1 (NP (NN 华人/Chinese)) (NP (NN 法官/judge) (PU 、) (NN 律师/lawyer)))

(NP-SBJ (-NONE- *pro*))

(VP (ADVP (AD 也/also))

(VP (VV 需/need)

(VP (VV 加快/accelerate)

(IP-OBJ

(NP-SBJ (-NONE- *PRO*))

(VP (VV 培养)

(NP-OBJ

(-NONE-*T*-1))))))

“The promulgation of the three main laws should be sped up, and the training of Chinese judges and lawyers should be accelerated.”

Such a long-distance dependency can be “localized” by positing an empty category close to the predicate and linking the topicalized argument to this empty category. Since our annotation is performed on the parse trees in the Penn Chinese Treebank (Xue et al., 2004), which has already implemented these empty categories for such topicalized arguments, we can just annotate the empty categories and the user can just follow the indices to get to the right argument.

However, not all non-local dependencies are explicitly represented in the Chinese Treebank. In

general, the non-local dependencies represented in the treebank are the result of general syntactic processes such as topicalization, relativization, ba- and bei-constructions. Other non-local dependencies are less general or even lexical in the sense that they only hold for certain specific verbs or certain class of verbs. In this case, it is appropriate to capture these types of non-local dependencies in our propbank-style of annotation, which represents verb-specific lexical dependencies.

One type of such of dependencies is akin to the English PP-attachment problem (Hindle and Rooth, 1991; Abney et al., 1999; Pantel and Lin, 2000) where a prepositional phrase can either be dependent on a verb or the noun phrase that is its object (or both in the case of true ambiguity). This is illustrated in (7), where in (7a) the prepositional phrase preceding the verb should actually be interpreted as an argument of the object noun phrase "见解/opinion". We generally do not expect "发表/deliver" to take an argument that is the content of something, rather we expect it to take an argument that is an opinion, or a speech. On the other hand, we fully expect "见解/opinion" to have some sort of content. In contrast, the prepositional phrase in (7b) is in fact an argument of the verb "发表/deliver"

(7) a. 曾荫权/Zeng Yinquan [**pp** 就/on 建立/establish 国际/international 金融/financial 新/new 秩序/order][**v** 发表/express] [**np** 见解/view]。 "Zeng Yinquan expressed his own view on the establishment of a new international financial order."

b. 在/at 酒会/banquet 上/on, 蔡/Cai 大使/ambassador [**pp** 对/to 一向/always 关心/support 祖国/motherland 建设/development 的/DE 海外/overseas 父老兄弟/compatriot] [**v** 发表/deliver] 了/ASP [**np** 热情/enthusiasm 洋溢/overflow 的/DE 讲话/speech]。 "At the banquet, Ambassador Cai made an enthusiastic speech to the overseas compatriots."

To show that is not an idiosyncratic phenomenon, some more examples are given in (8). Note that the dependency between the NP and the PP is non-local with an intervening verb. While in English there is a straightforward way of representing this dependency structurally in the treebank by attaching them at different levels, it is not as straightforward to

represent this dependency structurally in Chinese. Using empty categories and indices seems less appealing in this case since this is not a general syntactic process. Given what we have described so far, in our propbank annotation, we do not treat the PP as an argument or adjunct to the verb. We defer the annotation of the PP until we annotate the relational nouns such as "兴趣/interest" that have their own predicate-argument structures.

(8) a. 美国/the U.S. 商界/business community [**pp** 对/towards 与/with 上海/Shanghai 建立/establish 良好/good 的/DE 关系/relation] 很/very [**v** 感/feel] [**np** 兴趣/interest]。

"The U. S. business community is also very interested in establishing good relations with Shanghai."

b. 近/recent 几/few 周/week 来/since, 叛军/rebels [**pp** 对/toward 马克尼/Makeni 等/etc. 地/place] [**v** 加强/intensify] [**np** 攻势/offensive]

"In the recent few weeks, the rebels intensified their offensive toward places like Makeni."

c. 美/the U.S. 英/Britain 也/also 声称/declare 将/will [**pp** 对/toward 伊/Iraq] [**v** 保持/maintain] [**np** 强大/strong 的/DE 军事/military 压力/pressure], 并/and 支持/support 反对派/opposition 推翻/overthrow 现/current 政权/regime。

"The United States and Great Britain also declared that they would maintain strong military pressure on Iraq and support the opposition's effort to overthrow the current regime."

d. 他/he 说/say, [**pp** 对/toward 法国/France 同/with 台湾/Taiwan 进行/maintain 纯属/purely 民间/unofficial 的/DE 经贸/economic and trading 往来/relation] 我们/we 不/not [**v** 持/hold] [**np** 异议/opposition]。

"He said: 'we do not oppose the unofficial economic and trading relations that France maintains with Taiwan'."

Another type of non-local dependency is between a PP and a nominalized verb, with an intervening light verb or semi-light verb. This is illustrated in (9). In this case, we treat the PP as an argument of the nominalized verb, not as an argument of the intervening light verb or semi-light verb.

(9) a. 他/he 希望/hope 澳门/Macao 政府/government [**pp** 就/on 涉及/concerning 澳门/Macao 平稳/smooth 过渡/transition 和/and 政权/sovereignty 顺利/smooth 交接/hand-over 的/DE 重大/important 问题/issue]

进一步/further [v 加强/strengthen] [np 同/with 中方/Chinese side 的/DE 磋商/discussion 与/and 合作/cooperation]。

“He hopes that the Macao government will step up its discussion and cooperation with the Chinese side on the important issues concerning Macao’s stable transition and the smooth handover of sovereignty.”

b. 双方/two sides [pp 对/over 两/two 军/military 30多/over 30 年/year 来/since 的/DE 友好/friendly 合作/cooperation] [v 表示/express] [np 满意/]。

“Both sides expressed their satisfaction over their friendly cooperation of over 30 years.”

c. 如果/if 100/100 条/piece 新闻/news 中/among 有/have 一/one 条/piece 是/be 假/fraudulent 的/， 读者/reader [pp 对/towards 另外/other 99/99 条/piece] 也/also 会/will [v 产生/arise] [np 怀疑/doubt]。

“If among the 100 pieces of news one piece is made up, the reader will also doubt the other 99 pieces.”

d. 国际/international 社会/community [pp 对/towards 此/this] 必须/must [v 保持/maintain] [np 高度/high degree 的/DE 警惕/vigilance]。

“The international community must maintain a high degree of vigilance on this.”

e. 许多/many 与会者/participants [pp 对/towards 香港/Hong Kong 的/DE 前途/future] [v 感到/feel] [np 放心/at ease]， 并/and 持/hold 积极/positive 的/DE 看法/outlook。

“Many participants of the meeting are at ease over Hong Kong’s future and hold a positive outlook.”

4 Discontinuous arguments

Discontinuous arguments are cases where an argument to a given predicate consists of parts that are not adjacent to one another. For example, in (10a), the event “西非/West Africa 经济/economy 增长/growth” is a lone argument to the predicate “恢复/resume”. In (10b), the constituent that denotes the event is split into the subject portion and the predicate portion. It is intuitively clear that the two sentences in (10) are semantically similar and should be treated similarly. It is the event in its entirety, not the subject portion “西非/est Africa 经济/economy” or the predicate portion “增长/growth” individually, that is the argument to the verb predicate “恢复/resume”, even when they are discontinuous as in (10b). In our

probank annotation, both portions receiving the label *arg0*, with the predicate receiving a secondary tag *-prd* when the subject and the predicate portion are split.

(10) a. [arg0 西非/West Africa 经济/economy 增长/growth] 已/already 明显/clearly [rel 恢复/resume]。

“West African economy has clearly resumed growing.”

b. [arg0 西非/African 经济/economy] 明显/clearly [rel 恢复/resume] [arg0-prd 增长/growth]。

“African economy has clearly resumed growing.”

Discontinuous arguments can also occur with predicates that take more than one arguments. (11a) shows that “估计” takes two arguments in (11a) and when *arg0* is not explicitly expressed, *arg1* is a split argument in (11b).

(11) a. [arg0 一些/some 专家/expert] [rel 估计/estimate]， [arg1 信息/information 高速/super 公路/highway 建成/construct-succeed 后/after， 知识/knowledge 对/towards 经济/economic 增长/growth 的/DE 贡献率/contribution 将/will 可能/likely 由/from 本/this 世纪/century 初/beginning 的/DE 百分之五/5 percent 至/to 百分之二十/20 percent 上升/increase 到/to 百分之九十/90 percent]。

“Some experts estimate that after the successful construction of the information superhighway, the contribution of knowledge to the economic growth will increase from the five to twenty percent in the beginning of this century to ninety percent.”

b. [arg1 非洲/African 经济/economy] 1997年/1997 [rel 估计/estimate] [arg1-prd 增长/increase 百分之三/3 percent]

“It is estimated that African economy grew 3 percent in 1997.”

Another construction that generally allows discontinuous arguments are the so-called “subject-predicate” verb compounds in which the first verb is semantically the subject of the second verb. The subject of the first verb is generally the subject of the verb compound as a whole. For example, in (12a), the argument of the verb “迅速/rapid” (second verb in the verb compound) is “外资/foreign investment 金融/financial 机构/institution 业务量/business volume 增长/increase”， where “增长/increase” is the predicate of this event. This argument is “split”

because the intervening “也/also” does not modify “增长/increase”. Instead, it is a modifier of “迅速/rapid”. Similarly, in (12b), “历史上/historically” does not modify “发展/develop”. Instead it is a modifier of “快” and “大”. Notice that the two events “大寨/Dazhai 发展/develop”, “大寨/Dazhai 变化/change” share the same subject “大寨/Dazhai” and they are arguments of “快” and “大” respectively.

(12) a. [arg0 外资/foreign investment 金融/financial 机构/institution 业务量/business volume] 也/also [arg0-prd 增长/increase] [rel 迅速/rapid]

“The business volume of foreign financial institutions has also increased rapidly.”

b. 搞/conduct 市场/market 经济/economy 以来/since 的/DE 这/these 几/few 年/year, 是/be [arg0 大寨/Dazhai] [历史/history 上/in] [arg0-prd 发展/develop] 最/most [rel 快/rapid]、[arg0-prd 变化/change] 最/most [rel 大/dramatic] 的/DE 几/few 年/year

“These few years after the introduction of the market economy are the periods of time when Dazhai’s economy grows the fastest and its change is the most dramatic.”

Analogous to the subject-predicate split is the possessor-possessee split. Here the possessor and possessee are abstract notions and does not necessarily indicate a strict possession relation between the possessor and the possessee. Generally this relation can be put to the DE-insertion test. It is possible to insert “的” between the possessor and the possessee and treat the whole thing as the argument to the predicate. For example, in (13a), the argument of “加快/accelerate” is “三/three 大/main 法案/law (的/DE) 出台/promulgation 进程/process”. The argument is labeled as *arg1* because it is possible for “加快/accelerate” to take another argument that denotes the initiator, which would be labeled as *arg0*. While the split subject-predicate argument is verbal or clausal, the possessor-possessee construction is nominal. However, a case can be made that both the predicate and possessee indicate relations. The possessor and the possessee can either occupy the subject and object positions as in (13a), (13b) and (13c), or the topic and subject positions as in (13d) and (13e), or the topic and object positions as in (13f).

(13) a. [arg1-psr 三/three 大/main 法案/law]

需/need [rel 加快/accelerate] [arg1-pse 出台/promulgation 进程/process]

“The promulgation process of the three main laws needs to be accelerated.”

b. [arg1-psr 中国/China 经济/economy 增长/growth] 也/also 将/will [rel 放慢/slow down] [arg1-pse 速度/speed]

“The speed of the Chinese economy will also slow down.”

c. 今天/today 子夜/midnight 新年/New Year 钟声/bell 响起/ring 时/when, [arg1-psr 太湖/Taihu Lake 的/DE 历史/history] 将/will [rel 翻开/turn] [arg1-pse 新/new 的//DE 一/one 页/page]。

“When the New Year bell rings at midnight tonight the history of the Taihu Lake will turn a new page.”

d. [arg1-psr 这些/these 国家/country] [arg1-pse 货币/currency] 已/already 大幅度/large-scale [rel 贬值/devalue]

“The currencies of these countries has undergone a large-scale devaluation.”

e. [arg0-psr 茅台酒/Maotai liquor] [arg0-pse 制作/brew 工艺/process] [rel 复杂/complicated], [arg0-pse 生产/production 周期/cycle] [rel 长/long] “Maotai liquor’s brewing process is complicated and its production cycle is long.”

f. [arg0-psr 不/not 能/can 如期/on time 扭亏为盈/turn around 的/DE 企业/enterprise] 我们/we 要/will 坚决/decisively [rel 撤换/change] [arg-pse 负责人/leader]

“For those enterprises that can not turn around on schedule, we should decisively change their leaders.”

Notice that in (13e), “茅台酒” is the possessor of both “制作工艺” and “生产周期”. “茅台酒的制作工艺” and “茅台酒的生产周期” are the arguments of “复杂” and “长” respectively. It should be pointed out that not all topics can be analyzed as the possessor of some possessee. For example, in (14), it is not immediately clear that any of the constituent is its possessee. In this case, we will analyze it as a semantic adjunct and assign to it the tag *argM-TPC*.

(14) Example of real topic construction [TOPIC 至于/as to 居住/reside 在/in 内地/inland 的/DE 儿童/children], 他们/they 必须/must 出示/show 居留权/residence 证明书/certificate, 证明其/their 居留权/residence。

“As for children, they must show their residence certificate to prove their residence.”

Another type of discontinuous arguments generally occurs with predicates that require multiple participants and it can be argued that these participants collectively play a role in relation to the predicate. In Chinese, generally one of these participants is realized as the subject and the other participants are realized as prepositional phrase, as in (15). There are two apparent options. One is assign the same label to all the participants indicating they all play the same role with regard to the predicate and the other is to assign different labels to each participant. The latter option is undesirable in two ways. One is that semantically, this way of annotation suggests that all of these participants play different semantic roles with regard to the predicate, which they are not. The other problem is there can be arbitrarily many participants. Assigning different roles to each participant would lead to an explosion of the possible argument structures for the predicate, which is an undesirable outcome. Based on these considerations, we assign the same label to all the participants.

(15) 其中/among them [**arg0-crd** 不少/quite a few 公司/company] [**arg0-crd** 与/with 中国/China 公司] /company 合作/cooperate.

“Among them, a lot of companies are cooperating with Chinese companies.”

5 Summary

In this paper we discussed the verb-specific approach in semantic role labeling in the construction of the Penn Chinese Propbank and argued that this approach has the potential of providing a replicable way of annotating semantic relations. Making an annotation effort replicable ensures that we are taking solid steps in advancing to the eventual goal of natural language understanding. We also argued that by staying close to the text and being less abstract, we increase the possibility that the semantic relations annotated can be automatically acquired by the computer. If it is determined that higher level of abstraction is desirable, this type of annotation also provides a solid basis for making that advance. We also described how we handled two important issues in the development of this resource, namely, dislocated and discontinuous arguments. When this resource is finished, which is expected in the near fu-

ture, we will have about 4800 verbs annotated with their semantic roles in the 250K-word Penn Chinese Treebank (Xue et al., 2004). This annotation will also be accompanied with a description of the semantic roles of each of the 4800 verbs with examples of their possible syntactic realizations.

References

- Donald Hindle and Mats Rooth. 1991. Structural ambiguity and lexical relations. In *The 29th Annual Meeting of the Association for Computational Linguistics*, University of California, Berkeley.
- C. Baker and C. Fillmore and J. Lowe. 1998. The Berkeley framenet project. In *Proceedings of COLING-ACL*, Singapore.
- S. Abney and R. Schapire and Y. Singer. 1999. Boosting applied to tagging and PP attachment. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999*.
- P. Pantel and D. Lin. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics, pages 101-108, Hong Kong, October 2000*.
- D. Gildea and D. Jurafsky. 2002. Automatic labeling for semantic roles. *Computational Linguistics*, 28(3):245–288.
- Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002)*, Las Palmas, Spain.
- Dan Gildea and Martha Palmer. 2002. The Necessity of Parsing for Predicate Argument Recognition. In *Proceedings of the 40th Meeting of the Association for Computational Linguistics*, Philadelphia, PA.
- Nianwen Xue and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.
- Keh-Jiann Chen and Chu-Ren Huang and Feng-Yi Chen and Chi-Ching Luo and Ming-Chung Chang and Chao-Jan Chen. 2004. Sinica Treebank: Design Criteria, Representational Issues and Implementation. In Anne Abeillé ed., *Building and Using Parsed Corpora*. Kluwer.
- Nianwen Xue and Fei Xia and Fu-dong Chiou and Martha Palmer. 2004. The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 10(4):1–30.