Natural Language Engineering **15** (1): 143–172. © 2008 Cambridge University Press
doi:10.1017/S1351324908004865 Printed in the United Kingdom

Adding semantic roles to the Chinese Treebank

NIANWEN XUE and MARTHA PALMER
Department of Linguistics and Center for Spoken Language Research, University of Colorado at Boulder, CO, U.S.A.
e-mail: Nianwen.Xue@Colorado.EDU

(*Received 9 September 2008; first published online 24 October 2008*)

Abstract

We report work on adding semantic role labels to the Chinese Treebank, a corpus already annotated with phrase structures. The work involves locating all verbs and their nominalizations in the corpus, and semi-automatically adding semantic role labels to their arguments, which are constituents in a parse tree. Although the same procedure is followed, different issues arise in the annotation of verbs and nominalized predicates. For verbs, identifying their arguments is generally straightforward given their syntactic structure in the Chinese Treebank as they tend to occupy well-defined syntactic positions. Our discussion focuses on the syntactic variations in the realization of the arguments as well as our approach to annotating dislocated and discontinuous arguments. In comparison, identifying the arguments for nominalized predicates is more challenging and we discuss criteria and procedures for distinguishing arguments from non-arguments. In particular we focus on the role of support verbs as well as the relevance of event/result distinctions in the annotation of the predicate-argument structure of nominalized predicates. We also present our approach to taking advantage of the syntactic structure in the Chinese Treebank to bootstrap the predicate-argument structure annotation of verbs. Finally, we discuss the creation of a lexical database of frame files and its role in guiding predicate-argument annotation. Procedures for ensuring annotation consistency and inter-annotator agreement evaluation results are also presented.

1 Introduction

In recent years, the development of linguistically interpreted corpora has seen a shift of emphasis from annotating syntactic structures to semantic structures. The predicate-argument structure is at the focal point of most of the semantic annotation projects. Representative efforts include FrameNet (Baker, Fillmore and Lowe 1998), the Proposition Bank (Palmer, Gildea and Kingsbury 2005), the tectogrammatical layer of the Prague Treebank (Hajič, Böhmová, Hajicová and Hladká 2003), the Salsa project (Burchardt, Erk, Frank, Kowalski, Pado and Pinkal 2006), and many others. The choice of the predicate-argument structure as the foundation for semantic annotation has a high level of consensus presumably because it is conceptually less controversial and practically easier to annotate. The predicate-argument structure annotation generally involves locating the predicates in a corpus and using them as anchors for the annotation of their arguments. The semantic labels assigned to

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

the arguments, also known as semantic role labels,[1] vary from project to project. They range from very general role labels such as agent, theme, beneficiary, e.g., the semantic components of the Sinica Treebank (Chen, Huang, Chen, Luo, Chang and Chen 2004), to labels that are meaningful to a specific situation, e.g., the FrameNet (Baker *et al.* 1998) and the Salsa Project for German (Burchardt *et al.* 2006), to predicate-specific labels, e.g., the English Proposition Bank (Palmer *et al.* 2005) and Nombank (Meyers, Reeves, Macleod, Szekely, Zielinska, Young and Grishman 2004). 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 it uses argument labels (*Arg0*, *Arg1*, etc.) that are meaningful only with regard to a specific predicate. Although there is certain degree of consistency in the way the arguments are labeled across predicates, for the most part, the numbered arguments can only be interpreted in conjunction with the specific predicate it is associated with. The FrameNet role labels have a higher degree of abstraction in the sense that they abstract away from any specific predicate and instead are applicable to a class of related verbs (or nouns that have predicate-argument structures). For example, the "Buyer" role in the "Commerce_goods-transfer" frame applies to a whole class of predicates including "charge," "lease," "retail," "retailer," "sale," "sell," "vend," "buy," "purchase," "purchaser," "rent." Using labels such as agent and theme that have global meanings abstracts away from specific predicates or classes of predicates and instead applies to all predicates (or any other category that denote a relation). Arguments can be made for using each type of argument label, but for practical purposes, it is possible to map less abstract (thus less general) argument labels to more abstract (thus more general) roles with some effort should that be desirable (Yi, Loper and Palmer 2007). One may view the Proposition Bank style of annotation as a stepping stone to more abstract semantic annotations. From an engineering point of view, it also makes sense to adopt a bottom-up approach that starts with less abstract annotation and the high annotation consistency reported in Palmer *et al.* (2005) attests the validity of this approach, which we adopt in this work. The semantic roles we add to the Chinese Treebank (CTB) (Xue and Palmer 2005) are essentially the Proposition Bank style labels adapted to Chinese, and this annotation framework will be examined in greater detail in Section 2.

The CTB is a syntactically annotated corpus, with sources from Xinhua newswire (mainland China), Hong Kong news, and Sinorama Magazine (Taiwan). More recently under DARPA GALE funding it has been expanded to include broadcast news, broadcast conversation, news groups, and web log data. It currently has over one million words and is fully segmented, POS-tagged and annotated with phrase structures according to a style book (Xue and Xia 2000) similar to that of the Penn English Treebank (Marcus, Santorini and Marcinkiewicz 1993). As part of the general trend to move from syntactic annotation to semantic annotation, in the past few years we have been adding Proposition Bank style semantic role labels to the CTB. The work reported here is on a 500,000 word subcorpus of the CTB

[1] We will use argument labels and semantic role labels interchangeably throughout the article.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

that has already been annotated with predicate-argument structure information. The annotated data consists of semantic role labels added to arguments of more than 80,000 verb instances for 11,000 verb types, as well as 20,000 noun instances for over 2,500 noun types. All verbs, except for modal verbs, have been annotated with predicate-argument structure information, and 10.5 per cent of the all noun instances are annotated as nominalized predicates. On average, a verb predicate has 2.92 arguments and 2.04 core arguments, while a nominal predicate has 1.97 arguments and 1.20 core arguments. In comparison, a verb predicate in the English Propbank averages 3.20 arguments and 2.5 core arguments. A verb in the Sinica Treebank has 1.97 arguments. The annotated data as well as the *frame files*, a predicate-argument structure lexicon we created to guide the annotation, has been released to the computational linguistics research community via the Linguistic Data Consortium.[2]

The semantic annotation we describe here is on top of the syntactic annotation in the sense that the semantic role labels are added to constituents in the syntactic parse tree. We start by describing the annotation of the predicate-argument structure of verbs as part of the Chinese Proposition Bank project (Xue and Palmer 2003; Xue 2004) and issues specific to verb annotation in Section 3. Section 4 describes our approach to the annotation of nominalized predicates as part of the Chinese Nombank project (Xue 2006b). Section 5 describes a method to extract subcategorization frames as a preprocessing step to speed up annotation for verbs. For any linguistic annotation projects, annotation consistency is of utmost importance for the resulting corpus to be useful. In Section 6, we discuss the procedures and the companion resources we have created to ensure annotation consistency. In particular, we briefly describe a valency lexicon, or frame files we have created to serve as lexical guidelines to support our annotation. We also present evaluation results using an inter-annotator agreement metric. Finally, we conclude in Section 7.

2 The Propbank annotation framework

As alluded to in the previous section, the Propbank annotation assigns predicate-specific argument labels to the constituents in a parse tree. The arguments of each predicate in the sentence, which are limited to verbs and their nominalizations in the work we report here, receive an argument label in the form of *ArgN*, where *N* is an integer between 0 and 5. These numbered arguments represent *core* arguments that are defined in relation to the predicate, which is labeled as *Rel*. Each core argument plays a unique role with regard to the predicate and generally the total number of core arguments for each predicate does not exceed six. The core arguments annotated for the verb 调查 (investigate) in Example (1)[3] are the NPs 警方 (the police) and

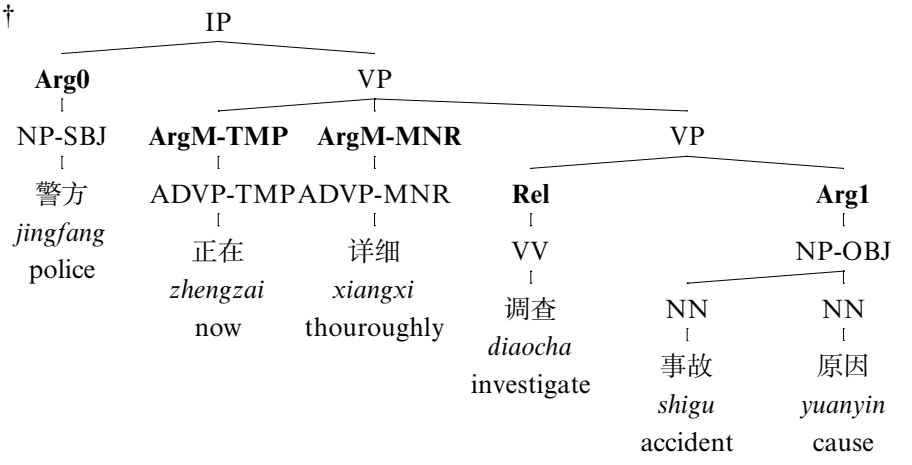
[2] http://www.ldc.upenn.edu.

[3] Most examples cited in this paper are directly from the annotated corpus this paper describes. There are also some constructed examples, mostly for purposes of contrast. These constructed examples are marked with † throughout the paper.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

事故 (accident) 原因 (cause), which are labeled as *Arg0* and *Arg1*, respectively. The semantic role labels added to the parse tree are in bold.

(1) †



"The police are thoroughly investigating the cause of the accident."

In addition to core arguments, we also annotate semantic adjuncts. Semantic adjuncts are not selected by the predicate and do not play unique roles with regard to a particular predicate. The flip side of the picture is that semantic adjuncts also do not tend to be selective as to which predicate they can occur with. They tend to represent general semantic notions such as temporality and location that are applicable to a heterogeneous set of predicates. It is not true, however, that temporal and locational elements are necessarily semantic adjuncts. For some predicates, temporal or locational elements may be required and are thus core arguments to the predicate. For example, 放, like its English translation "put," requires a locational element as one of its core arguments. Since semantic adjuncts are not required by a particular predicate, it makes little sense to define them in a predicate-specific manner. Therefore, the semantic role labels[4] assigned to semantic adjuncts are not predicate-specific and are represented in the form of *ArgM*, indicating adjunct, followed by a secondary tag representing the semantic classification of the adjunct. In (1), for example, the ADVP 正在 (right now), a temporal adjuct, is labeled *ArgM-TMP* and the ADVP 详细 (thoroughly) is marked *ArgM-MNR*. The complete list of semantic adjuncts is presented in Table 1. 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 based on the obligatory/optional dichotomy. In a lot of cases, some constituents are labeled as core arguments even though they are clearly optional.

[4] Semantic role or argument labels refer to both core arguments and semantic adjuncts unless noted otherwise.

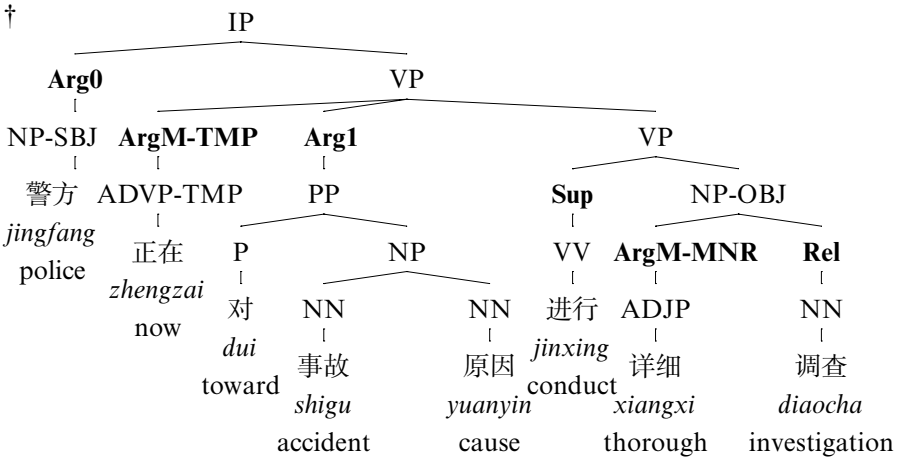
https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Table 1. *The complete list of functional tags*

ADV	Adverbial	FRQ	Frequency
BNF	Beneficiary	LOC	Locative
CND	Condition	MNR	Manner
DIR	Direction	PRP	Purpose or reason
DIS	Discourse marker	TMP	Temporal
DGR	Degree	TPC	Topic
EXT	Extent		

The predicate-specific argument labels of the Propbank annotation are designed to account for syntactic variations, the different ways in which the argument structure of a predicate can be realized. An argument is consistently labeled no matter how and where it is realized. In (2), a close paraphrase of (1), for example, the PP 对 (toward) 事故 (accident) 原因 (cause) still receives the semantic role label *Arg1* even though it undergoes a categorial change from NP to PP and is realized before the predicate 调查 (investigation) as a syntactic adjunct to the VP rather than as an object realized after the predicate. Notice also the categorial change of 调查 (investigation), a nominalized predicate in this sentence. In Chinese, nominalized predicates share the same form as their verb counterpart.

(2) †



"The police are conducting a thorough investigation of the cause of the accident."

Syntactic variations do not have to be accompanied by categorial changes of the predicate. In a book on English verbs, Levin (1993) discussed extensively how each class of verbs can be realized in sets of subcategorization frames that are related through diathesis alternations. Similar alternations can also be found in Chinese. One common type of alternation is the "subject of intransitive/object of transitive" alternation that can be found in verbs that can be used both transitively and intransitively. When used intransitively the only realized argument is in the subject position. The same argument will occur in the object position in a transitive frame. The Propbank annotation scheme captures this by assigning the same semantic role

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

label to this argument no matter where it occurs. 打开 (open) in (3) exemplifies a verb that can be used both transitively and intransitively. The noun phrase "中/China 美/the U.S. 交往/contact 的/DE 大门/door" occurs in the subject position in the intransitive frame[5] and in the object position when the verb is used transitively. In the semantic annotation discussed here, it is labeled *Arg1* in both frames, independent of its syntactic realization.

(3) †a. [**Arg1** 中 美 交往 的 大门] [**Rel** 打开] 了 。
zhong mei jiaowang de damen dakai de
China the U.S. contact DE door open ASP .
"The door of contact between China and the U.S. opened."

b. [**ArgM-TMP** 70年代 初] , [**Arg0** 中 美 两
qishiniandai chu zhong mei liang
70s beginning , China the U.S. two
国 领导人] [**ArgM-ADV** 果断] [**Rel** 打开] 了 [**Arg1**
guo lingdaoren guoduan dakai le
country leader decisively open ASP
中 美 交往 的 大门] 。
zhong mei jiaowang de damen
China the U.S. contact DE door .
"In the beginning of the 1970s, the leaders of China and the U.S. decisively opened the door of contact."

It is important to note that semantic roles are defined for each expected argument of a predicate, not just for arguments that are realized in a given sentence. In a given sentence, not all arguments have to be realized. The list of expected arguments and their semantic roles for each predicate are defined in frame files where there is a frame file for each predicate (Xue 2006a). The frame files are created by experienced annotators who examine a large number of sentences for each predicate so that semantic roles defined for that predicate have a solid empirical grounding. The frame files provide a conceptual framework in which the semantic roles are annotated, and they also serve as lexical guidelines that ensure consistent annotation (a point we will discuss further in Section 6). By annotating the realized arguments in a sentence with semantic roles with respect to the expected arguments in the frame files, the arguments that are missing in a given sentence can also be inferred. This is illustrated in (4), where 缩短 (shorten) has five expected arguments (4b). Only three of its arguments, *Arg0*, *Arg1*, and *Arg4* are realized and it can be inferred that *Arg2* and *Arg3* are missing.

[5] In theoretic Chinese linguistics literature, this is often analyzed as a case where the subject is dropped and the object is fronted in a general topicalization process. We believe this is more appropriately analyzed as a case of syntactic alternation, tied to certain class of verbs.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(4) a. [**Arg0** 商检 部门] 将 [**Arg1** 原来 七至十
shangjian bumen jiang yuanlai qi zhi shi
commercial inspection department BA original 7 to 10
天 完成 的 产地 检验 时间] [**Rel** 缩短 到]
tian wancheng de chandi jianyan shijian suoduan dao
day finish DE product origin inspection time shorten to
[**Arg4** 一至 三 天] 。
yi zhi san tian .
1 to 3 day .
"The commercial inspection departments shortened the time of product origin checking from the original 7–10 days to 1–3 days."

b. **Expected arguments for** 缩短:
Arg0: cause or agent
Arg1: theme
Arg2: range
Arg3: starting point
Arg4: end point

Some verbs can take on different sets of arguments that are realized in different sets of subcategorization frames. For example, the verb "存" expects two arguments when it means "exist:" 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 where 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 *framesets*, meaning sets of subcategorization frames that realize a particular sense. The examples in (5) illustrate the three framesets of "存."

(5) **Frameset 1**: "deposit"
Semantic roles:
Arg0: entity making deposit
Arg1: sum of money
Arg2: financial institution
a. [**ArgM-TMP** 二十 年 前] [**Arg0** 每 人] [**ArgM-ADV** 平均]
ershi nian qian mei ren pingjun
20 year ago each person average
[**ArgM-ADV** 才] [**Rel** 存] [**Arg1** 二十 元 钱] 。
cai cun ershi yuan qian .
only deposit 20 yuan money .
"Twenty years ago, on average each person has only a deposit of 20 yuan."

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

b. [**Arg1** 大批 资金] [**Rel** 存] [**Arg2** 在 中 小
dapi *zijin* *cun* *zai zhong* *xiao*
large amount fund deposit in mid-sized small
金融 机构]
jinrong *jigou*
financial institution
"A large amount of funds are deposited in mid-sized or small institutions."

Frameset 2: "exist"
Semantic roles:
Arg0: location
Arg1: thing that exists

c. [**ArgM-TMP** 现] [**Arg0** 全球] [**ArgM-ADV** 仅]
xian *quanqiu* *jin*
now the whole world only

[**Rel** 存] [**Arg1** 一千多 只 大熊猫] 。
cun *yiqianduo* *zhi* *daxongmao* .
exist over one thousand CL Giant Panda .

"There exist only a little more than 1,000 giant pandas in the whole world."

Frameset 3: "preserve"
semantic roles:
Arg0: preserver
Arg1: thing preserved
Arg2: instrument

d. [**Arg2** 方志] 可以[**Rel** 存] [**Arg1** 史] , 资 治 ,
fangzhi *keyi* *cun* *shi* , *zi* *zhi* ,
chronicle can preserve history , maintain order ,
教化 。
jiaohua .
civilize .
"Chronicles can be used to preserve history, maintain social order, and teach civilized behavior."

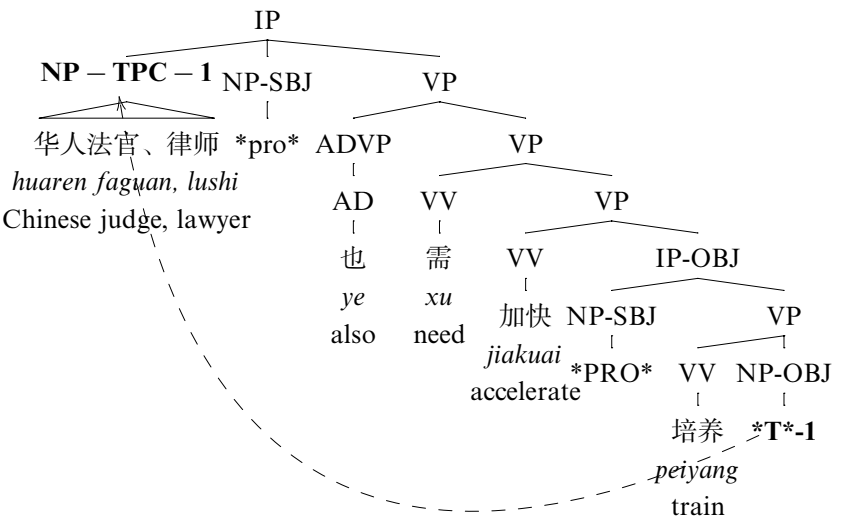
3 Verb annotation issues

As stated in the previous section, the primary goal in the predicate-argument structure annotation described here is to account for the variations in how the argument structure of a verb and its nominalization are realized. The syntactic variations of the argument structure of verbs can be a result of general syntactic processes such as topicalization, passivization, and BA-construction in Chinese. Through these general syntactic processes, arguments can be dislocated, typically fronted, from their canonical position, leading to long-distance dependencies between the verb and the dislocated argument. In the current semantic annotation framework,

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

the dislocated argument would still receive the same semantic role label as if it were in its canonical syntactic position. In the CTB, the long-distance dependencies as a consequence of general syntactic processes are "localized" by the postulation of an empty category adjacent to the verb and the linking of the dislocated argument to this empty category. For example, in (6), the topicalized NP "华人/Chinese 法官/judge 律师/lawyer" is linked to the empty catgegory *T*, which marks the canonical position of the NP, through co-indexation. Since the predicate-argument structure annotation is performed on the parse trees in the underlying CTB, we are able to simplify our annotation by just adding a semantic role label to this empty category. The real argument can be obtained by following the traces through co-indexation.

(6)



"The training of Chinese judges and laywers need to be accelerated."

In addition to the general syntactic processes, syntactic variations for the argument structure of verbs can also be a result of diathesis alternations in the sense of Levin (1993). These diathesis alternations are not accounted for in the CTB because they are more lexical in nature and they are only licensed for particular classes of verbs. We have shown in the previous section that verbs like "打开/open" in (3) expect two arguments and allow the "object of transitive/subject of intransitive alternation." "变成/turn into" in (7) exemplifies another group of verbs that typically expects four arguments, *Arg0*, the cause or agent that brings about the change, *Arg1*, the thing undergoing the change, *Arg2*, the starting state, and *Arg3*, the end state. As illustrated in (7), these verbs allow four-way alternations and not all arguments have to be realized in any given occurrence of the verb. Giving a complete survey of the possible alternations in Chinese is beyond the scope of this article, but interested readers are referred to Xue (2006a) for more details.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(7) a. [**Arg0** 该 公司] 把 [**Arg1** 电脑] [**Arg2** 由 文字 处理器]
gai gongsi *ba* *diannao* *you wenzi chuliqi*
this company BA computer from word processor
[**Rel** 变成] [**Arg3** 高 智能 的 电视 、
biancheng *gao zhineng de dianshi*
turn into high intelligence DE TV ,
录像 一体机] 。
luxiang *yitiji*
Video player combo .

"This company has turned the computer from a word processor into a highly intelligent TV-video player combo."

b. [**Arg0** 该 公司] 把 [**Arg1** 电脑] [**Rel** 变成] [**Arg3** 高
gai gongsi *ba* *dianno* *biancheng* *gao*
this company BA computer turn into high
智能 的 电视 、录像 一体机] 。
zhineng de dianshi luxiang yitiji
intelligence DE TV , Video player combo .

"This company has turned the computer into a highly intelligent TV-video player combo."

c. [**Arg1** 电脑] [**Arg2** 由 文字 处理器] [**Rel** 变成] [**Arg3** 高
diannao *you wenzi chuliqi* *biancheng* *gao*
computer from word processor turn into high
智能 的 电视 、录像 一体机] 。
zhineng de dianshi luxiang yitiji
intelligence DE TV , Video player combo .

"The computer has been turned from a word processor into a highly intelligent TV-video player combo."

d. [**Arg1** 电脑] [**Rel** 变成] [**Arg3** 高 智能 的 电视 、
diannao *biancheng* *gao zhineng de dianshi*
computer turn into high intelligence DE TV ,
录像 一体机] 。
luxiang *yitiji*
Video player combo .

"The computer has been turned into a highly intelligent TV–video player combo."

In the remainder of this section we discuss two complications that arise when annotating the argument-structure of verbs. The first case involves syntactic adjuncts of the verb, generally a PP, that are actually semantically related to postverbal NP object of the verb. The second case concerns discontinuous arguments.

3.1 "PP-attachment"

For English the PP-attachment problem (Hindle and Rooth 1991; Abney, Schapire and Singer 1999; Pantel and Lin 2000) involves a prepositional phrase that can either

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

be semantically dependent on a verb (e.g., "I ate a pizza with friends") or the noun phrase that is the object of the verb (e.g., "I ate a pizza with onions").[6] Since both the NP and the PP are postverbal in English, this ambiguity can be resolved syntactically by attaching the PP at either the VP level or the NP level in a parse tree, based on the semantic dependency. In Chinese, similar PPs almost always occur before the verb, even if they are semantically dependent on the NP object, which typically occurs after the verb. The intervening verb makes it impossible for the preverbal PP to form a contiguous constituent with the postverbal NP object even when this is semantically desirable. This is illustrated in (8), where in (8a) the PP preceding the verb should be interpreted as an argument of the object noun phrase 见解 (opinion). We generally do not expect the verb "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to have some sort of content. In light of this observation, in our propbank annotation, we do not treat the PP "就/on 建立/establish 国际/international 金融/financial 新/new 秩序/order" as an argument or adjunct to the verb. Instead, we consider the PP as semantically related to the nominal predicate 见解. In contrast, the prepositional phrase in (8b) is in fact an argument of the verb "发表/deliver." In short, we represent the different semantic dependencies with semantic roles instead of using different levels of attachment in the syntactic structure of the sentence, which is hard to do in Chinese.

(8) a. 曾荫权 [**PP** 就 建立 国际 金融 新 秩序]
zengyinquan *jiu jianli guoji jinrong xin zhixu*
Zeng Yinquan on establish international financial new order

[**V** 发表] [**NP** 见解] 。
fabiao *jianjie*
express view .

"Zeng Yinquan expressed his own opinion on the establishment of a new international financial order."

b. 在 酒会 上 ，蔡大使 [**PP** 对 一向 关心
zai jiuhui shang caidashi dui yixiang guanxin
at banquet , Ambassador Cai to always support
祖国 建设 的 海外 父老兄弟] [**V** 发表] 了
zuguo jianshe de haiwai fulaoxongdi fabiao le
motherland development DE overseas compatriot deliver ASP

[**NP** 热情 洋溢 的 讲话] 。
reqing yangyi de jianghua
enthusiam overflow DE speech .

"At the banquet, Ambassador Cai delivered an enthusiastic speech to the overseas compatriots who have always supported the development of their motherland."

[6] In rare cases, the PP can be semantically related to both the verb and its object NP.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Another type of nonlocal 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 predicate, not as an argument of the intervening verb. The annotation of the predicate-argument structure of nominalized predicates is discussed in Section 4.

(9) 双方 [**PP** 对 两 军 3 0多 年 来 的 友好
shuangfang *dui* *liang jun* *sanshiduo nian lai* *de* *youhao*
two side over two military over 30 year since DE friendly
合作] [**V** 表示] [**NP** 满意] 。
hezuo *biaoshi* *manyi* .
cooperation express satisfaction .
"Both sides expressed their satisfaction between the two militaries over their friendly cooperation of over 30 years."

3.2 Discontinuous arguments

What can be characterized as "discontinuous arguments" are cases where a constituent that occurs as one argument in one sentence can also be realized as multiple arguments (generally two) for the same predicate in another sentence, without causing changes in the meaning of the sentences. 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 propbank 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 经济 增长] 明显 [**Rel** 恢复] 。
xifei *jingji* *zengzhang* *mingxian* *huifu*
West Africa economy growth clearly resume .
"West African economy has clearly resumed growing."

†b. [**Arg0** 西非 经济] 明显 [**Rel** 恢复]
xifei *jingji* *mingxian* *huifu*
West Africa economy clearly resume

[**Arg0-PRD** 增长] 。
zengzhang
growth .

"West African economy has clearly resumed growing."

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

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

of the verb compound as a whole. For example, in (11), "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.

(11) 搞 市场 经济 以来 的 这 几 年 ，是 [**Arg0** 大寨]
gao shichang jingji yilai de zhe ji nian shi dajai
conduct market economy since DE these few year , be Dazhai
[历史 上] [**Arg0-PRD** 发展] 最 [**Rel** 快] 、
lishi shang fazhan zui kuai
history in develop most rapid ,
[**Arg0-PRD** 变化] 最 [**Rel** 大] 的 几 年 。
bianhua zui da de ji nian
change most dramatic DE few year .
"These few years after the introduction of the market economy are the periods of time when Dazahi'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 do not necessarily indicate a strict possession relation between the possesser and the possessee. Generally this relation can be put to the DE-insertion test. It is possible to insert "的" beween the possessor and the possessee and treat the whole thing as an argument to the predicate. For example, in (12), the argument of "加快/accelerate" is "三/three 大/main 法典/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 the possessee indicate relations.

(12) [**Arg1-PSR** 三 大 法典] 需 [**Rel** 加快]
san da fadian xu jiakuai
three main law need accelerate
[**Arg1-PSE** 出台 进程] 。
chutai jincheng
promulgation process .
"The promulgation process of the three main laws needs to be accelerated."

In (12), the possessor and the possessee are realized as the subject and the object in the syntactic structure, but they also have other syntactic manifestations. Chinese is often described as a topic-comment language (Li and Thompson 1976), where the notion of topic figures prominently in syntax. Topics are generally realized as NPs in a presubject position and they are labeled with the functional tag "-TPC" in the CTB (Xue, Xia, Chiou and Palmer 2005), but in the semantic role annotation described here, it is determined that the topic is generally not a semantic argument

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

to the verb predicate. Instead, in many cases a loosely defined possession relation holds either between the topic and subject, or between the topic and the object. (13) is an example where there is a possession relation between the topic and the subject.

(13) [**Arg0-PSR** 茅台酒] [**Arg0-PSE** 制作 工艺] [**Rel** 复杂] ,
maotaiju *zhizuo gongyi* *fuza*
Maotai Liquor brew process complicated ,
[**Arg0-PSE** 生产 周期] [**Rel** 长] 。
shengchan zhouqi *chang*
production cycle long .
"Maotai liquor's brewing process is complicated and its production cycle is long."

Notice that in (13), "茅台酒/Maotai Liquor" is the possessor of both "制作/brew 工艺/process" and " 生产/production 周期/cycle." "茅台酒/Maotai 的/DE 制作/brew 工艺/process " and "茅台酒/Maotai Liquor 的/DE 生产/production 周期/cycle" are the arguments of "复杂/complicated" and "长/long," 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 constituents is the possessee of the topic. In this case, we will analyze it as a semantic adjunct and assign to it the tag *ArgM-TPC*.

(14) [**ArgM-TPC** 至于 居住 在 内地 的 儿童] , 他们 必须 出示
zhiyu juzhu zai neidi de ertong *tamen bixu chushi*
as to reside in inland DE children , they must show
居留权 证明书 , 证明 其 居留权 。
juliuquan zhengmingshu *zhengming qi juliuquan*
residence certificate , prove their residence .
"As for children, they must show their residence certificate to prove their residence."

Another type of discontinuous argument 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 phrases, as in (15). There are two apparent options. One is to 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.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(15) 其中 [**Arg0-CRD** 不少 公司] [**Arg0-CRD** 与 中国
qizhong *bushao* *gongshi* *yu* *zhongguo*
among them quite a few company with Chinese
公司] 合作 。
gongsi *hezuo*
company cooperate .
"Among them, a lot of companies are cooperating with Chinese companies."

4 Annotating nominalizations

A subtask of the Chinese predicate-argument structure annotation is the annotation of nominalized predicates as part of the Chinese Nombank project. Unlike the English Nombank Project (Meyers *et al.* 2004), the focus of the Chinese Nombank annotation is limited to nominalizations of verbs, to the exclusion of relation nouns that are also considered to be predicates. Within this limited scope, the annotation of nominalized predicates closely parallels the annotation of verbs in the Chinese Proposition Bank project. The relationship between the two projects is particularly synergistic for Chinese given that Chinese verbs and their nominalizations share the same form and a frame file created for a verb can to a large extent be used to guide the annotation of their nominalizations. Verbs and their nominalizations often share the senses except that in some cases only a subset of the verb senses are realized in their nominalized forms. For example, among the many senses for the verb "发展," there are two prominent senses. When it means "evolve," five arguments are expected: the cause of the evolution, which is often not realized, the entity evolving, the starting point of the evolution, the end point of the evolution, and the range of the evolution. When it means "recruit," two arguments are expected: the recruiter and the entity recruited. However, only the former sense has a corresponding nominalized form (16), while the latter does not.

(16) 海峡 两 岸 今后 可 共同 规划
haixia *liang* *an* *jinhou* *ke* *gongtong* *guihua*
Taiwan Straits two side from now on can together plan
[**Arg1** 两岸 关系] 的 [**Rel** 发展] 。
liangan *guanxi* *de* *fazhan*
Cross-Strait relations DE progression .
"The two sides of the Taiwan Straits can plan the progression of the cross-Strait relations hereafter."

In the remainder of this section, we discuss a few issues that are specific to the predicate-argument annotation of nominalized predicates. In Section 4.1, we discuss the markability of nominalized predicates and their arguments. In Section 4.2, we discuss the role of support verbs, and finally in Section 4.3 we describe how the phenomenon of incorporation is addressed.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

4.1 Markable predicates and markable arguments

Unlike the annotation of verbs where almost each verb instance is a target of annotation, our annotation of nominalized predicates is selective in the sense that not all noun instances are targets of annotation. To identify nominalized predicates that are our target of annotation, we start by sifting through all nouns that have the same written form as some verb and annotate the subset that have similar argument structures as their verbal counterparts. Not all nouns that have a corresponding written verbal form are nominalizations. For example, 行政can be used as both a noun (executive authority) or a verb (exercise executive authority). While the nominal form and verbal forms are clearly related, they do not share the same arguments. The verbal form is a predicate that requires an agent, an executive that exercises the authority, while the nominal form does not need one and is not a predicate. Similar nouns are 旅游 ("travel" or "tourism"), 健身 ("exercise" or "fitness"), 纺织 ("weave" or "textile"), etc. For some nouns, some of their senses are nominalizations while others are not. For example, 教授can be used as a noun (professor) or a verb (to teach). When it is used as a title, e.g., 王教授 (Professor Wang), clearly it is not a predicate of any kind. On the other hand, in 英文教授 (English professor), it is a predicate that takes "英文/English" as its argument. In fact it is closely related to its verbal form, e.g., 他/he 教授/teach 英文/English (He teaches English). For the purpose of this project, we are only interested in nouns that are nominalizations of their verbal forms.[7]

Even when a noun is a true nominalized predicate, not all of its modifiers are legitimate arguments or adjuncts of this predicate. This is another aspect where the annotation of verbs differs from that of nominalized predicates. Some modifiers can only co-occur with the nominalized form and cannot co-occur with its corresponding verbal form. We are only interested in arguments and adjuncts that can co-occur with both the nominal and verbal forms of the predicate. When making this judgment our criteria are semantic rather than syntactic. For example, an adverb that is an adjunct to a verb is almost always realized as an adjective when it modifies the nominalized predicate, but it is still considered to be an adjunct to the nominalized predicate even though its syntactic category has changed. In Chinese due to the lack of morphological variation, the adjectival and adverbial modifiers often share the same form, just as the nominal and verbal forms of a predicate do. The most reliable difference is their syntactic distribution. For example,"充分/thorough" is an adjunct (**ARGM-MNR**) of the nominalized predicate "调查/survey" because it can also be used as an adverbial modifier of its verbal counterpart. This is illustrated in (17). Also notice that *Arg1* is a noun phrase when the predicate is a verb while it is a prepositional phrase when the predicate is nominalized.

[7] We agree with one reviewer's comment that not all nominalized predicates take arguments. Nominalized predicates that refer to the event as an individual do not take arguments. They only take arguments when they refer to the event as an event (Chierchia 1984). This issue is addressed in the current annotation framework because figuring out whether a nominalized predicate takes an argument in a particular context is an important part of the annotation task.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(17) †a. 联合国 工发 组织 近期 [**ArgM-MNR**
lianheguo gongfa zuzhi jinqi
UN industry development organization recently
充分] [**Rel** 调查] 了 [**Arg1** 图们江 地区
chongfen diaocha le tumenjiang diqu
thoroughly survey ASP Tumenjiang region
工业 项目 现状] .
gongye xiangmu xianzhuang
industrial project status
"The UN Industry Development Organization recently thoroughly surveyed the status of the industrial projects in the Tumenjiang region."

b. 联合国 工发 组织 近期 [**Arg1** 对
lianheguo gongfa zuxhi jinqi dui
UN industry development organization recently regarding
图们江 地区 工业 项目 现状] 进行 了
tumenjiang diqu gongye xiangmu xianzhuang jinxing le
Tumenjiang region industrial project status conduct ASP
[**ArgM-MNR** 充分] [**Rel** 调查]
chongfen diaocha
thorough survey
"The UN Industry Development Organization recently conducted a thorough survey regarding the status of the industrial projects in Tumenjiang Region."

Certain modifiers of nominalized predicates are typically noun-specific and should not be marked as an argument or adjunct of the nominalized predicates. With a few exceptions, determiner phrases (DP) and quantity phrases (QP) modifying the nominalized predicates are not annotated as arguments or adjuncts of the nominalized predicate because they are not possible arguments or adjuncts of their verbal form. This is illustrated in Example 18:

(18) 最近 的 一 项 [**Rel** 调查] 显示 . . .
zuijin de yi xiang diaocha xianshi
recent DE one CL survey show . . .
"A recent survey shows . . . "

Exceptions are made for determiner phrases and quantity phrases that can also be modifiers of verbs, e.g., "五/five 次/times," "三/three 天/days," etc. These modifiers are typically expressions of duration and frequency.

When the nominalized predicate is the head of a relative clause, the relative clause as a whole is generally not an argument or adjunct of this nominalized predicate. However, its arguments or adjuncts may be found inside the relative clause as in (19).

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(19) [**Arg0** 亚洲 金融 危机] [**Arg1** 给 世界 经济] [**Sup** 造成]
yazhou jinrong weiji *gei shijie jingji* *zaocheng*
Asian financial crisis on world economy create
的 [**Rel** 影响] 比 以前 估计 的 严重 。
de *yingxiang* *bi* *yiqian* *guji* *de* *yanzhong*
DE effect than previously estimate DE serious
"The effect that the Asian financial crisis created on the world economy is more serious than what is previously estimated."

One exception is when the relative clause involves the use of stative verbs that can be used as adverbial phrases if the predicate is in its verbal form. In this case, the entire relative clause is marked as ARGM of the nominalized predicate.

4.2 The role of support verbs

In many cases all arguments of a nominalized predicate can be located within the NP of which it is the head. Syntactically the arguments are realized as modifiers of the predicate. In Chinese, with rare exceptions, these modifiers are to the left of the head. Depending on their semantic relation to the predicate, they should either be tagged as an argument (ARGn) or an adjunct (ARGM). The argument/adjunct distinction is drawn along the same lines as the arguments and adjuncts of the verbs: the arguments are selected by the predicate and thus must fulfill the selectional restrictions of the predicate. The *Arg0* of "合作/cooperation," for example, must be of multiple parties and must be animated entities that are capable of cooperating. The adjuncts (ARGM), on the other hand, can modify a wide range of predicates. Nominalization is generally accompanied by adjective/adverb conversion: verbal predicates are modified by adverbs while nominalized verbs are modified by adjectives, even though these modifiers share the same semantic content. Therefore, we use the same functional tags to categorize the ARGMs of the verbal and nominal predicates, independently of their syntactic category.

(20) 这 一 地区 成为 [**Arg0** 海峡 两 岸]
zhe yi diqu chengwei *haixia liang an*
this one region become Straits two side
[**Arg1** 科技 、经贸] [**Rel** 合作]
keji *jingmao* *hezuo*
scientific technological , economic and trade cooperation
的 最佳 地带 。
de zuijia didai
DE best place .
"This region became the best place for scientific and technological, economic and trade cooperation."

As is often the case, the nominalized predicate occurs with a support verb. Some support verbs have little or no semantic content and are generally there to

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

fulfill a syntactic function. Other support verbs do add meaning to the nominalized predicate:

(21) [**ArgM-LOC** 在 国际 事务 中], [**Arg0** 欧盟
zai guoji shiwu zhong *oumeng*
at international affairs inside , European Union
同 中国] [**Sup** 进行] 了 [**ArgM-MNR** 很 好] 的
tong zhongguo *jinxing* *le* *hen hao de*
and China conduct LE very good DE
[**Rel** 合作] 。
hezuo
cooperation .
"In international affairs, China and European Union also have had very good cooperation."

By definition, a support verb is considered to be a support verb only if it at least shares an argument with the nominalized predicate. Whether the nominalized predicate also shares the adjunct modifiers of the support verb is a much more difficult question. This answer partly depends on the level of semantic content of the support verb and partly depends on the nature of the adjunct. In general, the less semantic content the support verb has, the more likely the adjunct is licensed by the nominalized predicate. This is another area where much of the ambiguity resolution needs to be done. Our policy is to mark the adjunct as belonging to the nominalized predicate when it is clearly licensed by the nominalization, not the light verb. In (22) the locative prepositional phrase "在/in 工业/industry 、农业/agriculture、经贸/trade 、文教/culture and education 、等/etc. 方面/aspect" is clearly licensed by the nominalized predicate "合作/cooperation" and therefore should be marked as such. This is also made possible by the fact that "进行/conduct" has little semantic content.

(22) [**Arg0** 中外] [**ArgM-LOC** 在 工业 、农业
zhongwai *zai gongye* *nongye*
China and foreign countries in industry , agriculture
、经贸 、文教 等 方面] [**Sup** 进行] 了
jingmao *wenjiao* *deng fangmian* *jinxing* *le*
, trade , culture and education etc. aspect conduct LE
{ 多 种 } [**Rel** 合作] 。
duo zhong *hezuo*
many kinds cooperation .
"China and foreign nations cooperated in many ways in industry, agriculture, trade, culture and education and other areas."

In (23), it is clear that the underlined adverbial is licensed by the support verb "扩大/expand," not by the nominalized predicate "合作/cooperation". It is also worth noting that this particular support verb has its own semantic content.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(23) 通过 进一步 [**Sup** 扩大] [**ArgM-DIR** 对 外]
tongguo jinyibu kuoda dui wai
through further expand toward outside
[**Rel** 合作] . . .
hezuo
cooperation . . .
"through further expanding foreign cooperation. . . "

It is possible that a verb is a support verb in some cases but not in others. A verb should not be marked as a support verb when it does not share an argument with the nominalized predicate, even though it may share an argument in other context.

4.3 Event/Result distinction and argument incorporation

The argument structure of a nominalized predicate is largely independent of whether it denotes an event or a result in the sense that the number and type of arguments the predicate takes do not change regardless of whether the predicate has an event or result interpretation. 调查 (investigation) has an event reading in (24a) indicating the act of the survey. In (24b), on the other hand, it has a result reading, indicating the result of the survey. However, in both cases it has two expected arguments, *Arg0* the agent that conducts the survey and *Arg1* the target of the survey, even though only *Arg1* is realized in (24b). The event/result distinction is clearly an important one but since the main concern of this project is the argument structure, it is set aside for future work.

(24) a. [**Arg0** 联合国 工发 组织] [**ArgM-TMP**
lianheguo gongfa zuzhi
UN industry development organization
近期] [**Arg1** 对 图们江 地区 工业 项目
jinqi dui tumenjiang diqu gongye xiangmu
recently regarding Tumenjiang region industrial project
现状] [**Sup** 进行] 了 [**ArgM-MNR** 充分][**Rel** 调查]
xianzhuang jinxing le chongfen diaocha
status conduct ASP thorough survey
"The UN Industry Development Organization recently conducted a through survey regarding the status of the industrial projects in Tumenjiang Region."

b. 此间 最新 一 项 [**Arg1** 投资 环境] [**Rel** 调查]
cijian zuixin yi xiang touzi huanjing diaocha
here latest one CL investment environment survey
显示 . . .
xianshi
show . . .
"A latest local survey on investment environment shows . . . "

Related to the event/result distinction is the phenomenon of incorporation, where the nominalized predicate absorbs one of its arguments. When this happens the

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

nominalized predicate takes on the meaning of both the predicate and the argument and loses its event interpretation. In this sense it is similar to the result interpretation of a nominalized predicate. Unlike the result interpretation, however, incorporation does change the argument structure. When an argument is incorporated into the nominalized predicate, the nominalized predicate is labeled as both the REL and the label of the argument it incorporates. For any given predicate, incorporation may occur in some cases but not in others. (25a) illustrates the incorporation of an argument of "投资/invest" into the nominalized predicate while (25b) is an example where no incorporation has taken place.

(25) a. 到 去年 十二月 底 , 总 [**Rel-Arg2** 投资] 达
dao qunian shieryue di zong touzi da
by last year December end , total investment reach
十二亿 美元 。
shieryi meiyuan .
one point two billion dollar .
"By the end of December of last year, total investment reached 1.2 billion dollars."

b. 最近 一 个 时期 , 山东省 大力 改善
zuijin yi ge shiqi shandongsheng dali gaishan
recent one CL period , Shangdong Province diligently improve
投资 环境 。
touzi huanjing .
investment environment .
"During a recent period, Shangdong Province has been diligently improving its investment environment."

5 Bootstrapping the semantic role labeling with subcategorizaiton frame extraction

The semantic role labels added to the CTB provide a way to represent the predicate-argument structure consistently despite the different ways in which the argument structure of a predicate is realized. Due to these syntactic variations, the mapping from the syntactic structure to the predicate-argument structure is a nontrivial problem. Still, the syntactic annotation of the CTB provides valuable structural information that can be exploited to bootstrap the predicate-argument annotation.[8] We describe in this section the basis for a semantic tagger, which provides a first approximation of the propbank semantic annotation, to be later corrected by human annotators. This tagger therefore plays a role in the creation of the Chinese Propbank similar to that of the system described in (Palmer, Rosenzwieg and Cotton 2001) for the English Propbank. A major difference however is that the tool for the English Propbank takes advantage of lexical resources such as the Susanne corpus and WordNet, while such resources are not available for

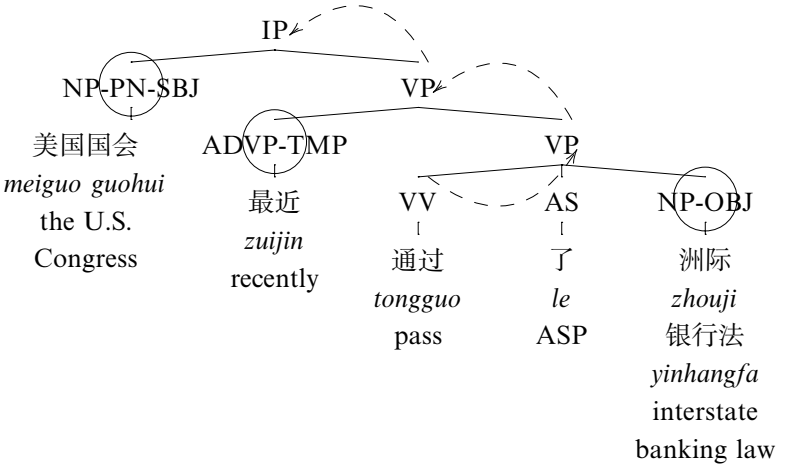
[8] The discussion here is limited to verbs, as this is more difficult for nominalized predicates due to the lack of NP-internal structures.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Chinese. Since it is also not a simple matter to induce lexical diathesis information from the Treebank, we manually provide some such information for the semantic tagger.

To carry out this first approximation to mapping the arguments to semantic roles, we extract out of the CTB subcategorization frames for each predicate instance. In most cases this is straightforward, consisting of traversing the parse tree from each verb to the top of that verb's clause, picking up all the complements, the modifiers, and the subject of the verb. The algorithm starts from the verb and iteratively moves up on level to its parent and collects its sisters at each level. This process is illustrated in (26):

(26)



"The U.S. Congress recently passed the interstate banking law."

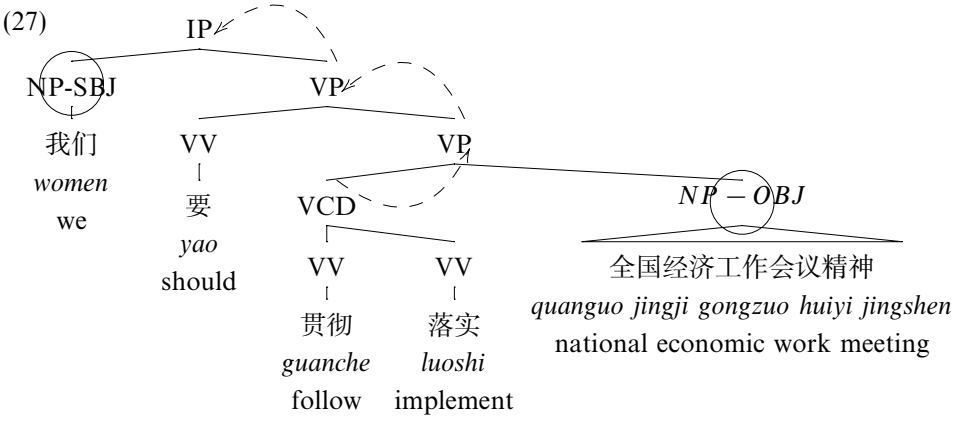
Notice that in this algorithm both potential arguments or adjuncts are collected. For example, the sentence in (26) would produce the output:

NP-PN-SBJ: 美国/the U.S. 国会/Congress
REL: 通过/pass
ADVP-TMP: 最近/recently
NP-OBJ: 州际/interstate 银行法/banking law

Using this sort of simple structural approach to recover the predicate-argument structure allowed us to capture the simple syntactic cases, namely all the ones in which there is sufficient syntactic annotation. This includes cases of syntactic movement, such as wh-movement or topicalization. The extraction of the sub-categorization frames is not always so straightforward and there are syntactic constructions that need to be handled with a little more effort. One such case is verb compounds, which are fairly frequent in the CTB. In the CTB, verb compounds can be loosely defined as several verbs sharing the same arguments. For example, the VCD compound indicates that the verb constituents of the compound share the same arguments, and so have the same subcategorization frame. This is illustrated in (27). The sentence actually produces two subcategorization frames, one for each of the verbs in the VCD compounds:

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

(27)



"We should follow and implement the spirit of the national economic work meeting."

Subcat Frame 1:
NP-SBJ: 我们/we
REL: 贯彻/follow
NP-OBJ: 全/whole 国/nation 经济/economic 工作/work 精神/spirit
Subcat Frame 2:
NP-SBJ: 我们/we
REL: 落实/implement
NP-OBJ: 全/whole 国/nation 经济/economic 工作/work 精神/spirit

There are a variety of other verb compounds, which we cannot cover in detail here, but we have determined for each of the different compounds, whether or not and how to break up the constituents of the compounds into different subcat frames.

Having extracted the subcat frames of the verbs, we are ready to map them to predicate-argument structures. For each element in the subcat frame, we first of all determine heuristically whether it is an argument or adjunct. To do this we rely heavily on the functional tags in the CTB. Constituents marked -SBJ (subject), -OBJ (object), and -DIR (direction) are considered to be potential semantic arguments, while constituents marked -TMP (temporal), -LOC (location), -MNR (manner), -PRP (purpose and reason), -CND (condition), etc. are considered to be potential adjuncts. As the set of functional tags of the Chinese PropBank to a large extent overlap with that of the CTB, the adjuncts are simply assigned its CTB functional tag prefixed by *ArgM-*. To map the semantic arguments, we rely on a mapping table specified for each subcategorization frame in the frame file of that predicate. For example, the mapping specified for the subcat frame in (27) is illustrated in Table 2. Needless to say, there are ambiguities in the mapping and in this case the mapping algorithm would simply arbitrarily choose one. The mistakes in the mapping are then corrected by the annotators.

Evaluated against the adjudicated final annotation results, this heuristic semantic tagger achieved a precision and recall of 0.76 and 0.63, respectively, for all arguments. For the core arguments, the precision and recall are 0.87 and 0.61, respectively. The

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Table 2. *Mapping table for the core arguments of 贯彻 and 落实*

verb	SBJ	OBJ	verb	SBJ	OBJ
贯彻	Arg0	Arg1	落实	Arg0	Arg1

higher precision and lower recall generally helps the manual correction in the next stage because the human annotator does not have to undo the arguments that the system wrongly proposes. It is also interesting to point out that statistical semantic role labeling systems trained on this data achieve a much higher accuracy (e.g., Xue and Palmer 2005 and Xue 2008 reported system accuracy of over 90 per cent (F1) on combined argument identification and classification in which treebank parses are used as input) so that future annotation efforts can be done at a much faster pace.

6 Ensuring annotation consistency

Consistency is critically important for the success of any annotation project. There are two main factors in our efforts to ensure consistency in annotation: guideline design and a fairly standard linguistic annotation procedure that involves double-blind annotation and adjudication. We view guidelines to be more than just a way of specifying information that needs to be annotated. Properly designed guidelines also contribute to consistency in annotation. We will describe our annotation guidelines in Section 6.1 and our annotation procedure as well as consistency evaluation in Section 6.2.

6.1 General versus lexical guidelines

There are two parts to our predicate-argument structure annotation guidelines, lexical guidelines (also called *frame files*) that specify the core arguments for each predicate, and general guidelines (Xue 2003) that provide working definitions for the function tags for semantic adjuncts, in keeping with the core argument versus adjunct dichotomy. The general guidelines also provide instructions on how to create a frame file and can thus be viewed as meta guidelines. As the general guidelines are very self-explanatory, our focus here would be on the anatomy of frame files. The concepts in the frame files, all in XML formats, are hierarchically organized. Each frame file has a predicate as its root that is partitioned into one or more major senses called *framesets*. One frameset is distinguished from another based on the set of arguments it expects or the set of semantic roles it is associated with. The sets of expected arguments can differ in number or type. For example, in (5), Frameset 1 differs from Frameset 2 in that while the former expects three arguments the latter has only two. Both Frameset 1 and Frameset 3 have three arguments, but they are still different framesets because they have very different arguments. It is hoped that the number and type of arguments are solid criteria based on which framesets can be consistently differentiated and annotated. It is important to note that the framesets are posited through examination of the data in the corpus to be annotated and therefore might be incomplete in terms of all the possible framesets of the predicate.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

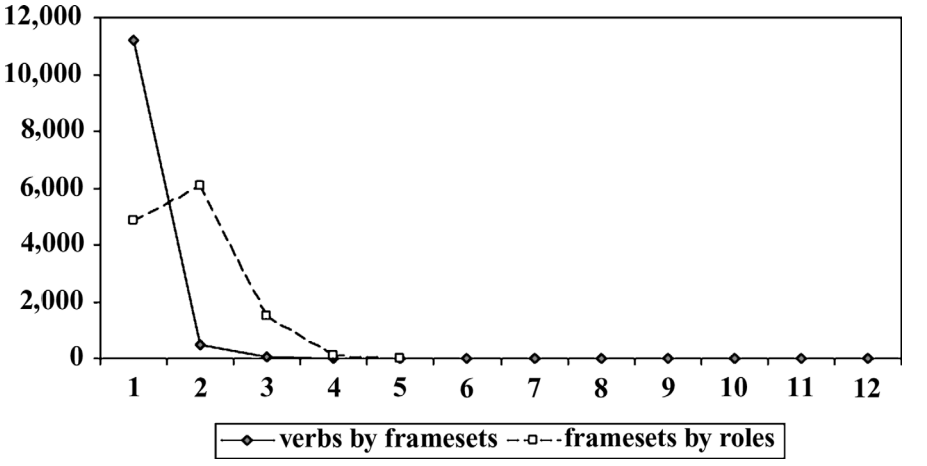


Fig. 1. Verbs by framesets and frameset by roles.

The frameset ID and the semantic roles are the information that gets recorded in the actual annotation. The frameset ID is assigned to each predicate instance in the corpus while the semantic role labels are assigned to arguments of the predicate. The mnemonic explanations for the semantic roles, such as "entity making deposit," are not part of the formal annotation. Their intended purpose is to help the annotators understand the different semantic roles of a frameset. Each instantiation of the semantic roles for a frameset is called a *frame*. Typically a frameset is realized in more than one frame. For example, (5a) and (5b) are two possible realizations of Frameset 1 of 存. These frames, annotated with the semantic role labels, are not part of the formal annotation. Rather, they are used as examples for the annotators.

All together, 11,765 frame files are created for the 500,000-word CTB, for a total of 12,555 framesets. The majority of the predicates, 11,185 of them have one frameset. The predicate that has the most framesets is 做 (do), which has 12 framesets. Each frameset has one to five semantic roles, with the majority of framesets having one, two, or three semantic roles. A breakdown of the number of verbs by the number of framesets and the number of framesets by the number of semantic roles is presented in Figure 1.

6.2 Double-blind annotation and adjudication

The annotation procedure we implemented is standard double-blind annotation plus adjudication. That is, each predicate instance in the corpus is independently annotated by two annotators who have no knowledge of the other annotator's annotation. When the two annotations for the same predicate instance are different, they are presented to an adjudicator to decide which annotation is to be chosen. In cases where the adjudicator feels both annotations are erroneous, they both are rejected and the adjudicator, who is generally a more experienced annotator, makes corrections.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Table 3. *Inter-annotator agreement, prior to adjudication*

		$P(A)$	$P(E)$	k
Including ArgM	Argument identification	0.998	0.935	0.965
	Role classification	0.938	0.232	0.919
	Combined decision	0.996	0.934	0.935
Excluding ArgM	Argument identification	0.998	0.954	0.966
	Role classification	0.956	0.410	0.925
	Combined decision	0.997	0.954	0.945

Adopting the evaluation approach for Propbank annotation described in (Palmer *et al.* 2005), we calculated the kappa (Siegel and Castellan 1988) statistics with semantic adjuncts (labeled *ArgM*) included and excluded. When the semantic adjuncts are included, the ArgM labels are treated like any other, where each secondary label of *ArgM* is considered to be a distinct tag for measuring inter-annotator agreement. When the semantic adjuncts are excluded, the ArgMs are treated as unlabeled nodes. The *kappa* statistics are calculated using equation 1, which are defined by the probability that both annotators agree, $P(A)$, and the probability that they agree by chance, $P(E)$. For both sets of statistics, we calculated kappa for argument identification, role classification, and the combined decision on argument identification and role classification. For argument identification, we only consider whether or not a node in the parse tree is identified as an argument, independently of whether the correct argument label is assigned to the node. The inter-annotator agreement probability $P(A)$ for argument identification is the number of node observation agreements divided by the total number of nodes considered, which is the total number of nodes in the parse tree multiplied by the number of predicates annotated in the sentence. The probability of agreement for argument identification is expected to be very high because for any given predicate only a few nodes close (loosely defined) to the verb are legitimate arguments for that predicate. Most nodes are not arguments to the predicate and annotators agree by default. To offset this artificial inflation of the inter-annotator agreement, we also calculated the kappa for role classification, where we factor out argument identification and only consider nodes that both annotators identify as arguments. The inter-annotator agreement probability $P(A)$ is calculated as the number of nodes that both annotators assigned the same semantic role label divided by the total number of nodes both annotators identified as arguments. Finally, we calculated the kappa for the combined decision, where the two annotators agree if they assigned the same semantic role label to a node or if they both decided that the node is not an argument.

$$k = \frac{P(A) - P(E)}{1 - P(E)} \tag{1}$$

The kappa statistics for these different calculations are presented in Table 3. As expected, the argument identification kappa is very high, but more importantly,

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Table 4. *Confusion matrix for core arguments with adjuncts considered as a whole*

	ARG0	ARG1	ARG2	ARG3	ARG4	ARGM
ARG0	0.350	0.015	0.002	0.000	0.000	0.005
ARG1		0.295	0.004	0.001	0.000	0.004
ARG2			0.031	0.000	0.000	0.002
ARG3				0.003	0.000	0.000
ARG4					0.000	0.000
ARGM						0.284

Table 5. *Differentiating arguments from adjuncts*

	ARGn	ARGM
ARGn	0.703	0.011
ARGM		0.284

the kappa for role classification is also high, indicating the annotation quality is satisfying. These numbers are comparable to the inter-annotator agreement figures reported in (Palmer *et al.* 2005) for the English Proposition Bank, attesting to the feasibility of this semantic annotation approach in general. All of these inter-annotator agreement figures are reported before adjudication is performed, which provides further assurance of the annotation quality. Given these figures and careful adjudication, we are assured of high-quality Gold Standard data. We should anticipate eventual system performance that attains or even surpasses the ITA rates, similar to English.

The confusion matrices in Tables 4, 5, and 6 provide more details on the inter-annotator agreement on individual semantic role labels. Table 4 is the confusion matrix for core arguments,[9] with all ArgM labels lumped together. It shows the highest level of confusion is between *Arg0* and *Arg1*, with a value of 0.015. Differentiating arguments from adjuncts is generally considered to be a challenging task, but Table 5 shows that annotators can generally make the distinction very consistently, with a very low confusion value (0.011) between core arguments and adjuncts. Table 6 is the confusion matrix for the secondary tags for ArgMs, which shows that the most difficult distinctions are between *ADV* and *DIS*, and between *ADV* and *TMP*.

[9] A reviewer points out that since the core arguments are defined in a predicate-specific manner, the confusion matrix for core arguments cannot be interpreted in a meaningful way. While we agree with this general assessment, there is some consistency in how the semantic roles are defined whenever possible. This is especially true for Arg0 and Arg1, with the former roughly corresponding to the prototypical agent and the latter the prototypical patient.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Table 6. *Confusion matrix for secondary tags. Entries are a fraction of the total number of arguments, including core arguments*

	ADV	BNF	CND	DIR	DIS	EXT	FRQ	LOC	PRP	TMP	TPC
ADV	0.139	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.000	0.003	0.000
BNF		0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
CND			0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DIR				0.002	0.000	0.000	0.000	0.000	0.000	0.000	0.000
DIS					0.003	0.000	0.000	0.000	0.000	0.000	0.000
EXT						0.001	0.000	0.000	0.000	0.000	0.000
FRQ							0.000	0.000	0.000	0.000	0.000
LOC								0.028	0.000	0.000	0.000
PRP									0.006	0.000	0.000
TMP										0.065	0.000
TPC											0.001

7 Conclusion

In this paper, we described the addition of semantic role labels to the CTB. We presented a descriptive framework in which semantic arguments are annotated with predicate-specific semantic role labels while the semantic adjuncts are annotated with concepts that can be interpreted globally. We believe that this framework strikes a balance between linguistic plausibility and engineering feasibility. We also discussed challenges that face the annotation of verbs and their nominalizations. For verbs, we focused on semantic role determination in the face of a wide range of syntactic variations as well as the handling of discontinuous arguments. For nominalized predicates, our discussion focused on whether a nominal modifier is an argument in the first place. We also discussed the relevance of the event/result distinction to the annotation of nominalized predicates and the phenomenon of incorporation. In any large-scale linguistic annotation project speed and consistency are always crucially important for the success of the project. We described a method of exploiting the structural information in the underlying CTB by extracting subcategorization frames. We also discussed the role of frame files in ensuring annotation consistency and provided evaluation results for annotation consistency.

Acknowledgments

We would like to thank Meiyu Chang, Zixin Jiang, Vicky Lai, and Gang Chen for their tireless annotation efforts, Zixin Jiang for creating frame files and Scott cotton for performing the original programming support for this project. This work is supported in part by the DOD via MDA904-02-C-0412 and in part by the NSF ITR via grant no. 130-1303-4-541984-XXXX-2000-1070. All findings, conclusions, and opinions are that of the authors' and do not necessarily reflect the views of the DOD or the NSF. Thanks also to three anonymous reviewers for their insightful comments.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

References

Abney, S., Schapire, R., and Singer, Y. 1999. Boosting applied to tagging and PP attachment. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, 1999*, College Park, MD, USA.

Baker, C., Fillmore, C., and Lowe, J. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Pado, S., and Pinkal, M. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 969–974.

Chen, Keh-Jiann, Huang, Chu-Ren, Chen, Feng-Yi, Luo, Chi-Ching, Chang, Ming-Chung, and Chen, Chao-Jan. 2004. Sinica Treebank: design criteria, representational issues and implementation. In Anne Abeillé (ed.), *Building and Using Parsed Corpora*, Dordrecht, the Netherlands: Kluwer.

Chierchia, G. 1984. *Topics in the Syntax and Semantics of Infinitives and Gerunds*. Ph.D. thesis, University of Massachusetts at Amherst.

Hajič, Jan, Böhmová, A., Hajicová, E., and Hladká, B. 2003. The Prague Dependency Treebank: a three level annotation scenario. In Anne Abeillé (ed.), *Treebanks: Building and Using Annotated Corpora*, Dordrecht, the Netherlands: Kluwer Academic Publishers.

Hindle, D., and Rooth, M. 1991. Structural ambiguity and lexical relations. In *The 29th Annual Meeting of the Association for Computational Linguistics*, University of California, Berkeley.

Levin, B. 1993. *English Verbs and Alternations: A Preliminary Investigation*. Chicago: The Unversity of Chicago Press.

Li, C., and Thompson, S. 1976. Subject and topic: a new typology of language. In Charles Li (ed.), *Subject and Topic*. New York: Academic Press.

Marcus, M. P., Santorini, B., and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of english: the penn treebank. *Computational Linguistics* **19**(2): 313–30.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: an interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, Boston, MA, pp. 24–31.

Palmer, M., Gildea, D., and Kingsbury, P. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics* **31**(1): 71–106.

Palmer, M., Rosenzwieg, J., and Cotton, S. 2001. Automatic predicate argument analysis of the penn treebank. In *Proceedings of the First International Conference on Human Language Technology Research*, San Francisco.

Pantel, P., and Lin, D. 2000. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *Proceedings of the 38th Meeting of the Association for Computational Linguistics*, October 2000, Hong Kong, pp. 101–8.

Siegel, S., and Castellan, N. J., Jr. 1988. *Nonparametric Statistics for the Behavioral Sciences*, 2nd ed. New York: McGraw-Hill.

Xue, N. 2003. Guidelines for the Chinese Proposition Bank.

Xue, N. 2004. Handling Dislocated and Discontinuous Constituents in Chinese Semantic Role Labeling. In *Proceedings of the 4th Workshop on Asian Language Resources, ALR04*, Hainan Island, China.

Xue, N. 2006a. A Chinese lexicon of roles and senses. *Language Resources and Evaluation* **40**(3–4): 395–403.

Xue, N. 2006b. Annotating the predicate-argument structure of Chinese nominalizations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, Genoa, Italy.

Xue, N. 2008. Labeling Chinese Predicates with Semantic Roles. *Computational Linguistics* **34**(2): 225–55.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press

Xue, N., and Palmer, M. 2003. Annotating the propositions in the Penn Chinese Treebank. In *The Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Xue, N., and Palmer, M. 2005. Automatic semantic role labeling for Chinese verbs. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, pp. 1160–5.

Xue, N., and Xia, F. 2000. The Bracketing Guidelines for Penn Chinese Treebank Project. Technical Report IRCS 00-08, University of Pennsylvania.

Xue, N., Xia, F., Chiou, F. d., and Palmer, M. 2005. The Penn Chinese TreeBank: phrase structure annotation of a large corpus. *Natural Language Engineering* **11**(2): 207–38.

Yi, S., Loper, E., and Palmer, M. 2007. Can semantic roles generalize across genres? In *Proceedings of NAACL-2007*, Rochester, NY, pp. 548–55.

https://doi.org/10.1017/S1351324908004865 Published online by Cambridge University Press