An Investigation of Heterogeneity and Overlap in Semantic Roles

Long Chen(✉) and Weidong Zhan(✉)

Peking University, Peking, China
{chenlong,zwd}@pku.edu.cn

Abstract. An inventory of semantic roles is needed in semantic role labelling. However, no matter what are semantic roles defined as, there is always heterogeneity and overlap in semantic roles. The semantic properties of members in the same semantic role can be different, while those of members in different semantic roles can be similar. It is a widespread phenomenon in semantic role types. The paper analyzes the cause of heterogeneity and overlap, and points out the close connection between them. The severity of heterogeneity and overlap described in the article is assessed using data from the survey of Chinese PropBank, a publicly available semantic resource consisting of two parts: the Frameset and the corpus.

Keywords: Semantic roles · Heterogeneity · Overlap · Chinese PropBank

1 Introduction

Semantic role labeling is an essential task in natural language processing. It helps computers understand the meanings of natural language, and provides necessary semantic information for tasks such as machine translation, question and answering, and information extraction [1]. Based on the theory of semantic roles in linguistics, semantic role labeling requires an inventory of semantic roles.

Linguists set up the concept of semantic roles, and classify the semantic roles of verbs, in order to summarize the meanings of verbs, find out the commonality of participants of events described by some verbs, and distinguish them from other participants with different semantic properties. However, in the practice of semantic role labeling, it is often discovered that there are members of the same semantic role with different semantic properties, and members of different semantic roles with similar semantic properties.

(1a) 小明刚才哭了，现在正在擦脸。(Xiaoming cried, and is wiping his face now.)
(1b) 小明刚才哭了，现在正在擦眼泪。(Xiaoming cried, and is wiping out his tears.)
(2a) 张三想办法吵醒了李四。(Zhangsan managed to awaken Lisi.)

J.-F. Hong et al. (Eds.): CLSW 2019, LNAI 11831, pp. 821–833, 2020.
https://doi.org/10.1007/978-3-030-38189-9_82

(2b) 闹钟吵醒了李四。(The alarm clock awakened Lisi.)
(2c) 张三吵醒了李四。(Zhangsan awakened Lisi.)

In sentence (1a) and (1b), "脸" (face) and "眼泪" (tears) should both be labeled as patients according to PKU SemBank Annotation Guidelines [2] and the classification of Peking University Netbank [3]. However, they are semantically heterogeneous. The former exists after the action happens, and the latter disappears.

According to the semantic role systems listed above, agents are defined as volitional actors of actions. "张三" (Zhangsan) in (2a) matches the definition of agents, and "闹钟" (alarm clock) in (2b) does not. In (2c) it is hard to tell whether "张三" (Zhangsan) is volitional, as he might have awakened "李四" (Lisi) intentionally, or accidentally. Therefore it is hard to classify the semantic role of "张三" (Zhangsan). It can be concluded that there are overlaps between agents and other proto-agent roles like experiencers.

The heterogeneity and overlap of semantic roles cause trouble to natural language understanding. Take the task of question and answering as an example, if the computer is asked "什么被小明擦了？(What's wiped out by Xiaoming?)", the answer should be "眼泪" (tears), no matter which sentence in (1) is the input sentence. However, if the question and answering system is based on the result of semantic role labeling, it may yield a wrong answer, since "脸" (face) in (1a) and "眼泪" (tears) in (1b) are both labeled as patients. If we try to annotate them as different roles, the difficulty of semantic role labeling will increase, and when the object is in the intersection of the different roles, its annotation is still tricky.

In this paper, the heterogeneity and overlap of semantic roles in Chinese and the way Chinese PropBank reflects and deals with the problem are examined.

2 The Heterogeneity of Semantic Roles

If some participants in the events expressed by a verb have different semantic properties, but we have to label these participants as the same semantic role, this semantic role is considered heterogeneous in this paper. Heterogeneity occurs in many types of semantic roles.

2.1 The Heterogeneity of Agents

Agents are usually defined as the volitional actor of an action. In sentence (3a), "小明" (Xiaoming) fits with the definition of agents. However, "风" (wind) in (3b) has no life or volition, and cannot be viewed as a typical agent. Lin [4] suggested that natural forces can be treated as agents, and Fillmore [5] regarded winds as instruments. In fact, no matter what semantic role is labeled on "风" (wind), it is an atypical member of that semantic role, which has different semantic properties from other typical members. Some linguists noticed the different semantic properties of natural forces from those of typical agents and proposed to set up a role called "Actor" or "Force" to label natural forces [6].

(3a) 小明把门关上了。(Xiaoming closed the door.)
(3b) 风把门关上了。(The Wind closed the door.)

Generally, as for any semantic roles, the members that do not fit with the definitions well are usually heterogeneous with other members. They have different semantic properties with the typical members. In semantic role labeling, for the consistency in annotation, we can classify the atypical members through analogy. In shallow semantic parsing, the differences between the typical members and atypical ones of a semantic role can be omitted sometimes, but further semantic analyses are needed to uncover the real accurate meanings of the atypical members and the differences between them and the typical members.

Besides the atypical agents, some members that match the definition of agents well may also have different meanings with other members.

(4a) 小赵去修自行车了。(Xiaozhao is going to have his bike repaired.)
(4b) 修车师傅在给小赵修自行车。(A repairer is repairing the bike for Xiaozhao.)
(5) 小赵去洗澡了。(Xiaozhao is going to take a bath.)

In sentence (4a), "小赵" (Xiaozhao) indeed volitionally caused the event "修自行车" (repairing the bike) to happen, so "小赵" (Xiaozhao) matches the definition of agents. According to the theory of proto-roles [7], "小赵" (Xiaozhao) has volition and causality, which are two properties of proto-agents. Compared with sentence (4b), "修车师傅" (bike repairer) seems to be the "real agent", and "小赵" (Xiaozhao) in (4b) is more likely to be labeled as a beneficiary. This phenomenon is not rare in Chinese. It also appears with verbs like "上课" (take/have a lesson), "看病" (see a doctor/patient), "理发" (cut one's hair). A type of the subjects of these verbs has the semantic feature of being benefited, while another type has the semantic feature of "doing by oneself". If we label the "benefited subjects" as beneficiaries, then it will be hard to classify "小赵" (Xiaozhao) in sentence (5), because on one hand, "小赵" (Xiaozhao) is both being benefited from the action and is doing it by himself, and we cannot label two semantic roles on it; on the other hand, we can imagine an occasion when "小赵" (Xiaozhao) cannot take a bath on his own for some reasons and needs to take a bath with the assist of others. Above all, these sentences reflect the heterogeneity of agents. Typical agents have the semantic feature of volition, but there are also agents who do not have the semantic feature of volition, and agents who have the semantic feature of being benefited.

2.2 The Heterogeneity of Patients

The role patient is also a heterogeneous semantic role. Patients are usually defined as the participants affected by the events or actions. Yuan [8] pointed out that patients have the semantic feature of being affected, being changed, and existing. They are general and abstract semantic properties of patients. More meticulously, patients of different verbs change in different ways. Patients

of “吃” (eat) and “销毁” (demolish) change in existence, patients of “洗” (wash) and “污染” (pollute) change in cleanness, and patients of “扔” (throw) and “放” (put) change in positions. Hownet [9] classifies verbs according to the types of change of their patients. These indicate that patients are heterogeneous.

If heterogeneous patients appear on complementary occasions, or heterogeneity does not happen on patients of the same verb, the heterogeneity of patients will not affect natural language understanding, since the types of change can be inferred from the types of verbs. The classification of Hownet also acquiesces that patients of the same verb are not heterogeneous. However, sentences (6a) and (6b) show that the same verb can take patients of different semantic properties.

(6a) 小王正在扫地。(Xiaowang is sweeping the floor.)
(6b) 小王正在扫雪。(Xiaowang is sweeping the snow.)

“地” (floor) in (6a) and “雪” (snow) in (6b) are both patients of the verb “扫” (sweep), but the former changes in cleanness, and the latter changes in existence. The label “patient” cannot reflect the differences between their meanings.

2.3 The Heterogeneity of Other Semantic Roles

Besides core semantic roles like agents and patients, other non-core semantic roles such as places are heterogeneous. Zhu [10] and Zhan [11] pointed out that there are two types of places by constructional transformations. Some places are entity places, like “纸” (paper) in sentence (7a) is the place of the entity “论文” (essay), and some places are event places, like “教室” (classroom) in (7b) is the place of the event “写论文” (write an essay).

(7a) 张三在纸上写论文。(Zhangsan is writing an essay on a paper.)
(7b) 张三在教室里写论文。(Zhangsan is writing an essay in the classroom.)

Instruments, materials, and some other non-core semantic roles are also heterogeneous. Their heterogeneity is not discussed due to the limitation of the length of the article.

3 The Overlap Between Semantic Roles

If it is hard to classify a participant of an event, and it fits with the definitions of different semantic roles, these semantic roles overlap. Like the heterogeneity, the overlap is also common among semantic roles.

3.1 The Overlap Among Proto-Agents

In the Chinese semantic role systems, the division of proto-agents varies from one another. An agreement is only reached on the definition of agents. For example, PKU SemBank Annotation Guidelines [2] divides proto-agents into agents and

experiencers; Yuan [8] divides proto-agents into agents, experiencers, causers and themes; Lu [12] divides proto-agents into agents, experiencers, and possessors. However, no matter how proto-agents are divided, agents always overlap the rest part of the proto-agents. Agents also overlap non-core semantic roles like causes and instruments.

(8a) 马克思证明了共产主义制度的优越性。(Marx proved the superiority of communism.)
(8b) 事实证明了共产主义制度的优越性。(Facts proved the superiority of communism.)

"马克思" (Marx) in (8a) meets the definition of agents, but "事实" (facts) in (8b) are causative, but not volitional. In the sentences (2b) and (2c) listed above, "闹钟" (alarm clock) is not volitional, and it is hard to determine whether "张三" (Zhangsan) is volitional. According to PKU SemBank Annotation Guidelines [2], "马克思" (Marx) and "张三" (Zhangsan) in the sentences above are agents, "事实" (facts) is an instrument, and "闹钟" (alarm clock) is a cause; according to Yuan's definitions [3], "马克思" (Marx) is an agent, and the other three should be labeled as causers. In Lin's semantic role system [4], they are all agents. It can be concluded that it is hard to distinguish these semantic roles. According to Chen's calculation [13], among the verbs whose proto-agents are agents in the semantic role system of Hownet, agents of 85 verbs overlap experiencers. These verbs have the semantic feature of causation, but do not have the feature of volition.

3.2 The Overlap Among Proto-Patients

Proto-patients are usually divided into semantic roles such as patients, results, contents, and objects. These roles also overlap. Take patients, results, and objects as examples, according to PKU SemBank Annotation Guidelines [2] and Yuan [3], the difference between patients and results is the semantic feature of existence; the former exist before the actions or events happen and the latter come into being after the actions; the difference between patients and results is the semantic feature of being affected; patients are affected by the actions or events and results are not. However, sentence groups (9) and (10) indicate that it is sometimes hard to tell whether a proto-patient exists before the action and whether it is affected by the action.

(9a) 这位教练培养了许多运动员。(This coach cultivated many athletes.)
(9b) 这位教练培养出了许多人才。(This coach cultivated many talents.)
(9c) 这位教练培养出了姚明。(This coach cultivated Yao Ming.)
(10a) 我正在修改的论文超过一万字。(The essay I am revising is more than 10 thousand words.)
(10b) 我修改的论文不到一万字。(The essay I revised is less than 10 thousand words.)

In (9a), “运动员” (athletes) exist before the event, and should be annotated as patients or objects; in (9b), “人才” (talents) usually do not exist before being cultivated, and the preposition “出” (out) also indicates they are results. However, in (9c), “姚明” (Yao Ming) obviously exists before cultivation, but there is also preposition “出” (out) in the sentence, indicating that “姚明” (Yao Ming) in the sentence may not be the same one before cultivation, so it is hard to determine which role should it be labeled as.

In (10a), “论文” (essay) refers to the essay before revision finishes, so it should be labeled as an object; in (10b), “论文” (essay) refers to the essay after revision, which can be seen as a new essay, so it can be labeled as a result. Many other verbs can also cause the overlap, such as “升级” (upgrade), “优化” (improve), “翻译” (translate). The meanings of these verbs are all concerned with changing the state of existence of their proto-patients.

3.3 The Overlap Between Proto-Agents and Proto-Patients

Overlapping not only exists among the roles of proto-agents and proto-patients. Proto-agents and proto-patients overlap as well. Dowty [7] pointed out that semantic roles are not discrete classes, so it is hard to define them precisely, and therefore, he proposed to use proto-roles to describe the semantic roles. Semantic properties like volition, sentience, and movement are relevant to proto-agents, and semantic properties like the change of state and being affected are related to proto-patients. According to the proto-role theory, there is, of course, no clear boundary between proto-agents and proto-patients. Dowty mentioned that movement was also a change of state.

Semantic roles of ergative verbs are also in the overlapping areas of proto-agents and proto-patients. For example, in (11a), “中国” (China) is a proto-agent and “经济” (economy) is a proto-patient; in (11c), “中国” (China) is a proto-agent; in (11b), it is hard to tell whether “中国” (China) and “经济” (economy) are proto-agents or proto-patients.

(11a) 中国正在发展经济。(China is developing the economy.)
(11b) 中国经济正在发展。(Chinese economy is developing.)
(11c) 中国正在发展。(China is developing.)

Other semantic roles also overlap. For example, instruments and materials overlap each other, and instruments and places also overlap each other.

4 The Causes for Heterogeneity and Overlap of Semantic Roles and Their Correlations

4.1 The Causes for Heterogeneity and Overlap of Semantic Roles

From the introduction of heterogeneity and overlap, it can be inferred that there is a similar cause for the phenomenon. Semantic roles are usually defined according to the semantic properties or semantic features of a participant of the event. However, the semantic features of a semantic role of a verb do not always match

the definition perfectly. For example, as is discussed above, some of the proto-agents of the verb "吵醒" (awaken) are volitional, and some are not. It leads to the overlap of agents and experiencers.

More basically, the heterogeneity and overlap of semantic roles are caused by two properties of people's cognition of things, fuzziness and generality.

There are often no clear boundaries between different semantic features. In the case of the heterogeneity of agents, the boundary of semantic features of volition and non-volition is fuzzy. It is often hard to determine whether the person's behavior that awakened another one is volitional or not.

Moreover, because of the generality in our understanding of events, when someone woke up because of another person, Chinese use the verb "吵醒" (awaken) to describe the event, regardless of the volition of the person. Thus, the proto-agent of the verb "吵醒" (awaken) can be either volitional or non-volitional, resulting in the heterogeneity or the overlap of relevant semantic roles.

Therefore, the heterogeneity and overlap cannot be eliminated in semantic role labeling. In order to overcome the trouble the phenomena bring to natural language understanding, more in-depth semantic analysis is needed on relevant verbs.

4.2 The Heterogeneity and Overlap Are Closely Related and Cannot Be Eliminated

From the perspective of the fineness of semantic role systems, heterogeneity and overlap of semantic roles are closely related to each other. Yuan [3] pointed out that there can be three levels of fineness of semantic role systems, which are macro levels, medium levels and, micro levels. Most semantic role systems in linguistic theories belong to medium-level systems; the proto-role theory proposed by Dowty belongs to the macro-level systems. In language resources, the semantic role classification of PropBank is a macro-level system; the semantic role system of VerbNet is a micro-level system; the rest are medium-level semantic role systems.

If several semantic roles overlap each other in a medium-level semantic role system, heterogeneity will happen in a large semantic role in a macro-level system, and vice versa. For example, in medium-level semantic role systems, patients, results, and objects overlap each other, and in a macro-level system, they are all proto-patients. The overlap of the three roles turns into the heterogeneity of proto-patients. Therefore changing the fineness cannot eliminate the heterogeneity and overlap.

Even in the same level of fineness, because of the difference in the definitions of semantic roles, the heterogeneity of a semantic role in one semantic role system may appear as the overlap of some semantic roles in another semantic role system. In example (2b), some semantic role systems classify "风" (wind) as an agent, so it is a heterogeneous agent; some classify it as an instrument, so in these systems agents and instruments overlap. In a word, adjusting the definitions of semantic roles on the same fineness level can only transform heterogeneity into overlap, or transform overlap into heterogeneity; it cannot eliminate the phenomena.

Above all, heterogeneity and overlap are closely related. They are twin phenomena caused by the fuzziness and generality of language. This problem cannot be solved by adjusting the definitions of semantic roles or changing the fineness of semantic role systems.

5 The Heterogeneity and Overlap of Semantic Roles in Chinese PropBank

Based on the proto-role theory, PropBank [14] uses "ArgN" to represent semantic roles of verbs, where N is an integer from 0 to 6. It also annotates the meanings of each semantic role of verbs. Chinese PropBank [15] is a Chinese semantic role labeling corpus based on Chinese Treebank. It contains over 80000 tokens of 11000 verbs. The annotation of Chinese PropBank is the same as the one of English PropBank. It is also a macro-level semantic role system. Apart from the corpus, there is also a Frameset in PropBank, which contains the meanings of semantic roles of verbs. For example, the Frameset of the verb "培养" (cultivate) contains two arguments, named as "Arg0" and "Arg1". Their meanings are "agent" and "entity cultivated" respectively. The meanings in the Frameset can be seen as micro-level or medium-level semantic roles. In other words, there are semantic roles on different fineness levels in the Frameset.

According to the analysis above, there is always heterogeneity and overlap in any semantic role systems. In this section, the phenomena in Chinese PropBank are examined.

5.1 The Heterogeneity and Overlap in the Frameset

The heterogeneity and overlap of some semantic roles can be discovered by only examining the Frameset. If an ArgN of a verb has more than one meaning label in the Frameset, the ArgN of the verb is heterogeneous; if a meaning label is used to describe Arg0s of some verbs, and Arg1s of other verbs, the label reveals the overlap of Arg0 and Arg1. In this paper, the heterogeneity and overlap of Arg0s and Arg1s is examined.

The Heterogeneity of Arg0s and Arg1s. In the Frameset of Chinese PropBank, the meanings of semantic roles can reveal the heterogeneity of semantic roles. After traversing the Arg0s and Arg1s of all verbs in the Frameset of Chinese PropBank, many instances of heterogeneity are discovered. The meaning of the Arg0 or Arg1 of a verb may be described as "Label1/Label2", where there are two meanings separated by a "/" or ",". This implies the Arg0 or Arg1 is heterogeneous. For example, the meaning of the Arg0 of the verb "破坏" (destroy) is described as "agent/cause", indicating that the Arg0 of the verb is sometimes volitional, and sometimes causative.

There are 1345 heterogeneous Arg0s in the Frameset, accounting for 5.07% of all 26534 Arg0s in the Frameset. Some heterogeneous Arg0s with high frequency are listed in Table 1.

Table 1. Examples of heterogeneity of Arg0s

Heterogeneous Arg0s	Frequency	Examples
agent/cause	1070	破坏(destroy), 振兴(revitalize)
agent/entity	20	残害(cruelly injure), 吓唬(threaten)
agent/causer	16	放大(magnify), 打破(break)
theme/entity	8	下坠(fall), 上火(inflame)
agent/organization	7	编撰(compile), 转载(reprint)
agent/experiencer	5	中意(like), 侧目(glance)

Among all the heterogeneous Arg0s, 1186 of them are tagged as agents and other semantic roles. As is shown in Table 1, 1070 Arg0s are tagged as agents and causes. There are also some Arg0s tagged as agents and entities, causers, organizations or fine-grained semantic labels like drivers, guardians, helpers. These data show that the main reason for the heterogeneity of Arg0s is the overlap of agents and other roles like causes and entities. Causes and entities are usually causative but not volitional.

Similar phenomena occur in other languages. In English PropBank, there are 67 Arg0s tagged as "agent/cause", and 125 Arg0s tagged as "agent/causer". Thus, the heterogeneity in Arg0s may be a universal problem across the languages.

In fact, there should be more heterogeneous Arg0s. First, in addition to "," and "/", "or" is also used in heterogeneous Arg0s. However, in the Frameset, some homogeneous arguments may also contain "or", like the Arg1 of the verb "逗" (amuse) is tagged as "person Arg0 amuses or teases". Thus, in this paper, heterogeneous arguments represented by "or" are not calculated. Second, not all heterogeneous Arg0s are tagged correctly with both meanings. According to Chen [13], among the heterogeneous Arg0s of the 85 verbs, some are only tagged as "agent", and fine-grained meaning labels conceal the heterogeneity of the Arg0s of some verbs. For instance, the Arg0 of "违反" (violate) is tagged as "violator", and the Arg0 of "促使" (prod) is tagged as "entity prodding".

There are 659 heterogeneous Arg1s in the Frameset, accounting for 4.72% of all 13968 Arg1s in the Frameset. Some high-frequency heterogeneous Arg1s are listed in Table 2.

Table 2. Examples of heterogeneity of Arg1s

Heterogeneous Arg1s	Frequency	Examples
theme/entity	29	推开(push away), 孵化(hatch)
time/place	12	进入(enter), 建于(build on)
locative/destination	11	折返(turn back), 流落(wander)
theme/locative	3	留驻(be stationed), 穴居(live in caves)

Some of the heterogeneity of Arg1s is caused by non-core roles like time and place, because if a verb has only one core semantic role, its non-core role may become its Arg1 in the semantic role system of PropBank. It revealed the heterogeneity of place, since a "locative" or "place" is the static location of an entity, and a "destination" is the location into which an entity moves.

In the heterogeneity caused by core roles, there are 57 heterogeneous Arg1s concerning themes, and 29 of them overlap entities.

The heterogeneity of Arg1s is not fully revealed in the Frameset as well. For example, according to the observation above, the Arg1s of "培养" (cultivate) and "修改" (revise) are heterogeneous, but in the Frameset of "培养" (cultivate), its Arg1 is tagged as "entity cultivated", which is close to a patient; there are two Framesets for "修改" (revise), its Arg1s are both tagged as "thing revised", which is close to a patient, and one of its Frameset contains an Arg2 tagged as "thing Arg1 becomes after revision", which is close to a result. Neither of the two representations reveals the heterogeneity of their Arg1s.

The Overlap of Arg0 and Arg1. The Frameset also reveals the overlap of Arg0 and Arg1. One hundred and fifty meaning labels are used to tag both Arg0s and Arg1s. The most frequent one is the theme, which is used 505 times in Arg0s and 3291 times in Arg1s. Some high-frequent medium-level semantic roles in the overlapping areas of Arg0 and Arg1 are listed in Table 3.

Table 3. Semantic roles in the intersection of Arg0 and Arg1 and their frequencies

Semantic roles	Frequency in Arg0	Examples	Frequency in Arg1	Examples
theme	505	翻滚(roll)	3291	划分(divide)
agent	13624	目击(collect)	20	减缓(relieve)
experiencer	124	担心(worry)	12	煎熬(suffer)
cause	117	惊吓(frighten)	43	得意(benefit)
source	6	产生(produce)	23	来自(come from)
beneficiary	3	获益(benefit from)	31	撑腰(support)
recipient	10	启蒙(enlighten)	159	拍马(flatter)

These data show heterogeneity and overlap is common among semantic role types, but verbs whose arguments are heterogeneous account for about 5%. Therefore, the current semantic role labeling task can express the meanings of most verbs accurately.

5.2 The Heterogeneity of Semantic Roles in the Corpus

We can also observe the heterogeneity of semantic roles from the corpus of the Chinese PropBank, and examine whether the annotation of the Chinese PropBank reflected the heterogeneity.

According to the examination above, the Frameset may not reflect the heterogeneity of Arg1s. This will influence the semantic role labeling of relevant

verbs. The study examined the sentences containing verbs "培养" (cultivate) and "修改" (revise), and discovered that their heterogeneous Arg1s are not distinguished in the annotation.

In the corpus of Chinese PropBank, 39 sentences contain the verb "培养". In 31 of them, the Arg1s of "培养" do not have the semantic feature of existence. These Arg1s include "人才" (talent), "能力" (ability), "兴趣" (interest). In three sentences, the Arg1s of "培养" exist before cultivation. The Arg1s are "小孩" (child), "我们" (we), and "病毒" (virus). There are also four sentences where it is hard to tell whether the Arg1s of the verb exist before cultivation. The Arg1s are "警犬" (police dog), "律师" (lawyer), "接班人" (successor), and "将士" (soldiers). Taking "警犬" (police dog) as an example, before cultivation, it may be a police dog, and it may also be an ordinary dog.

Besides, there is a sentence in the corpus "把他们培养成为美国服务的精英" (cultivate them into elites serving the USA), where the existent proto-patient "他们" (them) co-occurs with the non-existent proto-patient "为美国服务的精英" (elites serving the USA). According to the Frameset, "他们" (them) corresponds to the meaning of the Arg1 of the verb, and "为美国服务的精英" (elites serving the USA) cannot be correctly annotated. In Chinese PropBank, this sentence is not annotated.

Forty-five sentences contain the verb "修改" (revise). In 34 of them, the Arg1s exist before revision. They correspond to the meaning of the Arg1 in the Frameset. There are only two sentences where the proto-patients of the verb do not have the semantic feature of existence. Although the proto-patients in these two sentences are closer to the meaning of the Arg2 in the Frameset, they are still labeled as Arg1s. There are two sentences where the existent proto-patients co-occur with the non-existent proto-patients. The existent proto-patients are labeled as Arg1s, and the non-existent proto-patients are labeled as Arg2s. In the rest seven sentences, the Arg1s of the verb do not appear.

According to our examination, the Frameset of "培养" and "修改" cannot help represent the heterogeneity of their proto-patients in the annotation of the corpus. If the meanings of their proto-patients are described as the meaning of the Arg1 of "培养", which is closer to a patient, it can only represent the existent proto-patients accurately, and the ones that do not have the semantic roles of existence cannot be expressed precisely; If the meanings of their proto-patients are described closer to a result, the existent proto-patients cannot be represented accurately; If the meanings of their proto-patients are described in the form of "patient/result", it enabled further semantic analysis on relevant heterogeneous proto-patients, but this type of representation fails when the existent and non-existent proto-patients co-occur in one sentence. If the existent proto-patient and the non-existent proto-patient are labeled as Arg1 and Arg2 respectively, like the Frameset of "修改", it can cope with the problems mentioned above, but it will be challenging to annotate a sentence when the existence of the proto-patient is hard to determine. Therefore in the task of semantic role labeling, there is no perfect solution for the heterogeneity and overlap of semantic roles.

Above all, according to the examination of sentences containing “培养” and “修改”, heterogeneity of Arg1s exists in the corpus of Chinese PropBank. The Frameset of Chinese PropBank fails to represent the meanings of the Arg1s accurately and fails to reflect the heterogeneity.

6 Conclusive Remarks

The phenomena of heterogeneity and overlap in semantic roles occur among all semantic roles types. From the perspective of the fineness of semantic roles, the study discovers that heterogeneity and overlap are twin problems caused by the generality and fuzziness of language. The problems cannot be solved by adjusting the definitions or the fineness of the semantic roles.

The study examines relevant language resources, and discovers the heterogeneity and overlap in Chinese PropBank. In the Frameset of Chinese PropBank, 1345 Arg0s and 659 Arg1s exhibit heterogeneity, each accounting for 5% of all Arg0s and Arg1s. Meanwhile, some heterogeneous arguments are not accurately labeled in the Frameset, like the Frameset of “培养” (cultivate) and “修改”(revise). The semantic role labeling of these verbs in the corpus cannot reflect the precise meanings of their proto-patients. The heterogeneity of these proto-patients needs to be discovered by deeper semantic analysis.

There are other problems in semantic role labeling, such as the difficulties in representing the meanings of syntactic constructions and irregular compositions. These problems need further investigation.

Acknowledgment. This paper was supported by Major Project of Humanities and Social Sciences of Ministry of Education, P.R.China (Project NO.15JJD740002).

References

1. Palmer, M., Gildea, D., Xue, N.: Semantic Role Labeling. Morgan & Claypool Publishers, San Rafael (2010)
2. PKU SemBank Annotation Guidelines (2015). http://klcl.pku.edu.cn/xwdt/231664.htm
3. Yuan, Y.: The fineness hierarchy of semantic roles and its application in NLP. J. Inf. Proc. **21**(4), 10–20 (2007)
4. Lin, X.: Lexical Semantics and Computational Linguistics. Language & Culture Press, Beijing (1999)
5. Fillmore, C.J.: The case for case. In: Bach, E., Harms, R.T. (eds.) Universals in Linguistic Theory. Holt, Rinehart & Winston, New York (1968)
6. Saeed, J.I.: Semantics. Wiley-Blackwell, Hoboken (2009)
7. Dowty, D.: Thematic proto-roles and argument selection. Language **67**(3), 547–619 (1991)
8. Yuan, Y.: On the hierarchical relation and semantic features of the thematic roles in Chinese. Chin. Teach. World **61**(3), 10–22 (2002)
9. Dong, Z., Dong, Q.: Construction of a knowledge system and its impact on Chinese research. Contemp. Linguist. **3**(1), 33–44 (2001)

10. Zhu, D.: "Zai heiban shang xiezi" and relevant constructions. Lang. Teach. Linguist. Stud. **3**, 4–18 (1978)
11. Zhan, W.: Argument structure and constructional transformation. Stud. Chin. Lang. **300**, 209–221 (2004)
12. Lu, C.: Semotactic network of Chinese. Appl. Linguist. **26**, 82–88 (1998)
13. Chen, L., Zhan, W.: An investigation of inconsistency in semantic role labeling: a case study of agent. J. Inf. Proc. **33**(1), 1–10 (2019)
14. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (2005)
15. Xue, N., Palmer, M.: Adding semantic roles to Chinese Treebank. Nat. Lang. Eng. **15**(1), 143–172 (2009)