An Easier and Efficient Framework to Annotate Semantic Roles:

Evidence from the Chinese AMR Corpus

Li Song[1], Yuan Wen[1], Sijia Ge[1], Bin Li[1], Junsheng Zhou[2], Weiguang Qu[2, 3], Nianwen Xue[4]

1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, 210024, China

2. School of Computer Science and Technology, Nanjing Normal University, Nanjing, 210023, China

3. Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, 350108, China

4. Computer Science Department, Brandeis University, Waltham, 02453, USA

songli.njnu@gmail.com

Abstract

Semantic role labeling (SRL) is one of fundamental tasks in Chinese language processing. At present, it has three major problems on the construction of the SRL corpus. First, there are disagreements over the definition of the number and frame of semantic roles. Second, static predicate frames are hard to cover dynamic predicate usages. Third, it is unable to annotate the dropped semantic roles. The newly designed Abstract Meaning Representation (AMR) is a novel method of representing the meaning of sentences, which offers dynamic mechanisms to provide better solutions to the above three problems. We use the Chinese AMR corpus of 5,000 sentences to make a detailed comparison between AMR and other SRL resources. Data analysis shows that in AMR, it is easier to annotate the semantic roles of a predicate with the simplified distinction between core roles and non-core roles. And 1,045 tokens of dropped roles are annotated under this new framework. It indicates that AMR offers a better solution for Chinese SRL and sentence meaning processing.

Keywords: Abstract Meaning Representation, predicate framework, semantic role, language knowledgebase

1 Introduction

Automatic semantic analysis is one of the core tasks in Natural Language Processing (NLP). Therefore, building the semantic resources is the first step for machine learning based NLP systems. In semantic representation, semantic relations between predicates and their semantic roles form the backbone of the sentence structure. Thus, building the predicate frames which describe such information becomes an important issue in linguistics and NLP. There have been many semantic role labeling (SRL) systems and SRL resources in different languages, but there are several problems in these SRL corpus.

First, the number of the semantic role labels of predicates is still to be discussed in linguistics. VerbNet uses 30 general thematic role labels to represent semantic relations (Kipper et al., 2000). Sinica Treebank distinguishes necessary and unnecessary arguments and uses 60 semantic role labels, 12 of which can represent necessary arguments (Chen et al. 2003). FrameNet defines semantic roles on a per-frame basis (Baker et al., 1998), so it avoids determining how many semantic roles are needed for a language, and there are 1224 frames in FrameNet and 323 frames in Chinese FrameNet (CFN). PropBank (Palmer et al., 2005) and Chinese Proposition Bank (CPB) (Xue & Palmer, 2009) both define 5 predicate-specific semantic roles for the core arguments and 13 semantic roles that are consistent across predicates for non-core arguments. It can be seen that the number of role labels used by different SRL resources is quite different. This is mainly because these resources are based on different theoretical backgrounds.

Second, it is hard for static predicate frames to cover dynamic predicate usages. Predicate frames which do not distinguish core and non-core roles are difficult to represent whether a semantic role is necessary for the predicate. And resources that define core roles in a predicate-independent manner just as non-core roles neither could solve the collision between core and non-core roles nor could represent multi-functional semantic roles.

Third, limited to the annotating mechanism, most SRL systems are unable to annotate the dropped semantic roles of the predicates. For example, it is hard for most SRL systems to represent correctly the meaning of the nominal phrase *the injured* whose central words are dropped and *one of which…* which drops the noun that appeared in the preceding clause.

Abstract Meaning Representation (AMR), a new method to represent meaning of sentences, defines semantic roles in a manner different from other SRL systems (Banarescu et al., 2013). It deals with core and non-core roles in different specialized ways. AMR annotates core arguments using the same five core role labels as in PropBank, which are predicate-specific, and adopts the predicate frame lexicon extracted from PropBank. But the number of non-core role labels that are general to all the predicates is up to 40. At the same time, AMR allows to add back dropped semantic roles in the sentences. Through the dynamic mechanisms, AMR can provide better solutions to the above three problems. The English AMR Sembank[1] has included 39,260 sentences and become an important semantic resource.

Referring to the guidelines of English AMR, Li et al. (2016) has developed annotation specifications for Chinese AMR (CAMR), taking linguistic characteristics of the Chinese language into account. CAMR uses the same 5 core role labels (arg0-arg4) and 44 non-core role labels (time, location, cause, etc., four of which are added based on the needs of Chinese annotation) as AMR. The predicate frame lexicon of CAMR is extracted from the corpus (Bai & Xue, 2016) of Chinese Proposition Bank (CPB) (Xue & Palmer, 2009). In addition, Li et al. (2017) designs a framework for aligning the concepts and relations to word

[1] https://catalog.ldc.upenn.edu/LDC2017T10

tokens in a sentence for CAMR, which is helpful for annotating dropped semantic roles. Since English AMR can provide better solutions to the above three problems, we try to discuss whether CAMR can provide better solutions to these problems in Chinese.

The rest of this paper is organized as follows. In Section 2, we discuss the related work. In Section 3, we introduce the core and non-core role labels of CAMR and the basic information of the CAMR corpus. In Section 4 and Section 5, we discuss the rationality of the core and non-core role labels of CAMR based on data analysis. Section 6 discusses the advantages of the permission of adding back dropped roles of AMR. The conclusions and future work can be found in Section 7.

2 Related Work

Constructing a predicate frame lexicon combining with labeling semantic roles of predicates in corpus has become a research paradigm. There are many methods to define semantic roles, but the granularity of the semantic roles of predicates are still disputed in the linguistics field. Xue (2006) argues that the specific semantic roles in different SRL resources range from very general role labels to labels that are meaningful to a specific situation to predicate-specific labels in terms of levels of abstraction.

VerbNet uses 30 general thematic role labels such as agent, theme and beneficiary to represent semantic relations (Kipper et al., 2000). Similarly, Sinica Treebank which is a semantic treebank in traditional Chinese defines 60 semantic role labels in a predicate-independent manner. Additionally, Sinica Treebank distinguishes necessary and unnecessary arguments, and uses 12 of the 60 labels to represent necessary arguments (Chen et al. 2003). There are also similar resources in simplified Chinese such as NetBank, which defines 8 kernel thematic roles (agent, patient, recipient, etc.) and 18 circumstantial thematic roles (time, location, reason, etc.), all of which are general for predicates (Yuan, 2007).

FrameNet defines semantic roles on a per-frame basis, so it avoids determining how many semantic roles are needed for a language, leading to a large quantity of semantic role labels. These labels are extracted from specific predicates and applied to the same category of verbs and nouns which have arguments. Chinese FrameNet (CFN) follows the system of FrameNet. There are 1,224 frames in FrameNet and 323 frames in CFN.

PropBank defines semantic roles for the core arguments in a predicate-specific manner. Each sense of each verb has a specific set of roles, which are given only numbers (0-5) rather than names: Arg0-Arg4. Bai & Xue (2016) argues that Core arguments have three main attributes: (1) obligate, meaning of a predicate will be incomplete if it lacks a core argument; (2) different, the core argument frames of predicates differ from one another, so each sense of each predicate has a specific set of roles; (3) exclusive, multiple core arguments do not serve as the same semantic role. Different from core roles, its semantic roles for non-core arguments are consistent across predicates, and there are 13 non-core role labels (ADV, TMP, LOC, etc.) adopted by PropBank. Following the system of PropBank, Chinese Propsition Bank (CPB) adopts the same 5 core roles and 13 non-core roles.

AMR is a novel method of meaning representation which deals with core and non-core roles in different specialized ways. It annotates core arguments using the same five core role labels as in PropBank, which are predicate-specific, and adopts the predicate frame lexicon extracted from PropBank. But the number of non-core role labels (time, location, cause, etc.) which are general to all the predicates is up to 40. Chinese Abstract Meaning Representation (CAMR) uses the same 5 core role labels (arg0-arg4) and 44 non-core role labels, four of which are added for the needs of Chinese AMR annotation.

It can be seen that there are many SRL resources in Chinese as well as English, but their granularity of semantic role labels differs from each other. Table 1 summarizes the main SRL resources in English and Chinese that differ in the granularity of semantic role labels.

Resources	Language	Role Labels
VerbNet	English	30 general role labels
Sinica Treebank	Traditional Chinese	60 general role labels (5 for nouns, 12 for core roles, 43 for non-core roles)
NetBank	Simplified Chinese	8 general core labels and 18 general non-core labels
FrameNet	English	1,224 frames (role labels are frame-specific)
CFN	Chinese	323 frames (role labels are frame-specific)
PropBank	English	5 predicate-specific core labels and 13 general non-core labels
CPB	Chinese	5 predicate-specific core labels and 13 general non-core labels
AMR	English	5 predicate-specific core labels and 40 general non-core labels
CAMR	Chinese	5 predicate-specific core labels and 44 general non-core labels

Table 1: Main SRL Resources in English and Chinese

3 Chinese AMR

3.1 Core and Non-core Roles of Chinese AMR

Following the annotation scheme of OntoNotes adopted by English AMR, CAMR uses predicate senses and core argument frames in CPB, and annotates semantic relations with core and non-core semantic relation labels. Core semantic relations refer to the inevitable semantic relations in the event framework of the predicates which are predicate-specific. Table 2 shows the 5 core semantic relations adopted from CPB. Non-core semantic relations refer to the semantic relations outside the core semantic relations, which are predicate-independent. English AMR defines 40 general non-core semantic relations so that they are fine-grained, and CAMR adds 4 non-core relations taking the characteristics of Chinese into account. In order to be compatible with AMR, CAMR still uses English words to represent labels of non-core semantic relations. Table 3 shows non-core semantic relations in CAMR.

arg0	external argument (Proto-Agent)
arg1	internal argument (Proto-Patient)
arg2	indirect object / beneficiary / instrument / attribute / end state
arg3	start point / beneficiary / instrument / attribute
arg4	end point

Table 2: Core Semantic Relations in CAMR

accompanier	direction	mod	quant
*aspect	domain	mode	range
beneficiary	duration	name	source
cause	example	ord	subevent
compared-to	extent	part-of	subset
consist-of	frequency	path	superset
condition	instrument	*perspective	*tense
cost	li	polarity	time
*cunit	location	polite	topic
degree	manner	poss	unit
destination	medium	purpose	value

* are the added relations in CAMR

Table 3: Non-core Semantic Relations in CAMR

Since core semantic roles are defined with respect to an individual verb sense, AMR and CAMR need support of predicate frame lexicons. The frame lexicon of CAMR is extracted from the CPB corpus, consisting of 26,650 senses of 24,510 predicates.

3.2 The Chinese AMR Corpus

According to the CAMR annotation specifications developed by Li et al. (2016), we extracted 5,088 Chinese sentences from Penn Chinese TreeBank (CTB) 8.0[2] and annotated them. The inter-agreement smatch score of 500 randomly selected sentences between the two annotators is 0.83. The sentences we annotated in CTB are from microblog, which cover a wide range of fields and rich topics. Most sentences are long and complicated, containing rich semantic information. Before annotating, we deleted wrong sentences artificially, and then carried on automatic word segmentation and artificial proofreading. The final corpus consists of 5,000 Chinese sentences. Table 4 shows the basic data of these sentences. Compared with the Chinese version of *the Little Prince* AMR corpus (Li et al., 2017), whose average sentence length is 12.90 words and average number of concepts is 9.48, sentences in this corpus are longer and more complex.

Sentences	5,000	Characters (AVG)	34.34
Characters	171,703	Words (AVG)	22.46
Words	112,348	Concepts (AVG)	18.36
Concepts	91,808	Added Concepts[3] (AVG)	3.02

Table 4: Basic Data of the CAMR Corpus

4 Core Roles in Chinese AMR Cover Dynamic Problems

The definition of core arguments in PropBank has been controversial in linguistics field. Some scholars consider it too broad and not conducive to classification of semantic roles, the predicate frame of AMR thus failed to be approved by the entire linguistics field. Therefore, we try to explore whether the predicate framework adopted by AMR can represent core semantic roles of predicates more reasonably.

We consider that there are two inescapable problems in predicate frameworks whose core role labels are consistent across predicates: (1) the core semantic role labels are applicable for all predicates, and the core roles and non-core roles may conflict when annotating concepts of location, cause, instrument and so on, for example, a concept of location is indispensable to the meaning of *appear*. (2) It is difficult to properly annotate the multi-functional roles, for example, a concept of agent or cause can both serve as the subject of *change*.

These problems are common in Chinese and they can be solved by the predicate framework of CAMR, whose predicate-specific frame lexicon is extracted from the CPB corpus, which contains 26,650 senses of 24,510 Chinese predicates (verbs, adjective, etc.). CPB is a corpus which adds semantic roles of predicates to CTB (Xue et al., 2005), a syntactically annotated Chinese corpus that is word-segmented, POS-tagged and syntactically bracketed with phrase structures (Xue & Palmer, 2009). Here we elaborate how CAMR solves the collision between core and non-core roles and how it annotate multi-functional roles based on statistical data of the predicate frame lexicon and CAMR corpus.

4.1 Solve the Collision between Core and Non-core Roles

Each sense of each predicate in the predicate framework of CAMR has a specific set of roles. If a concept is essential for the meaning of the predicate, it serves as the core role of the predicate, even though it represents the location or cause of the predicate, which is a kind of collision between core and non-core roles. If inessential, it serves as a non-core role of the predicate. For example, the concept of location is indispensable in the meaning of 遍布-01 (the first sense of 遍布, be spread throughout somewhere), so it is a core role of 遍布-01:

遍布-01 (be spread throughout somewhere)
arg0: theme
arg1: **location**

It's even possible that 4 of the 5 roles of a predicate are in conflict, such as 引进-01 (introduce something from one place to another):

引进-01 (introduce something from one place to another)
arg0: agent / **cause**
arg1: entity imported
arg2: **location** arg1 is imported from
arg3: predicate, **purpose**
arg4: **destination**

There are many predicates whose core and non-core roles are conflicting. We count how many predicates in the CPB lexicon have collision between their core and non-core roles. Data shows that the total number of these senses is 2,453, accounting for 9.20% of all the senses in the

[2] http://amr.isi.edu/download.html

[3] There are three main kinds of added concepts in CAMR: (1) added semantic roles, (2) types of named entities which are used to identify the names of an entity, like *country* for *China*, (3) discourse relations such as *condition*, *temporal*.

lexicon of CAMR. Among them, 5.99% have collision between more than two core roles and non-core roles. Additionally, through analyzing all the description of core roles in the CPB lexicon, we find that there are 24 kinds of non-core roles may conflict with the core roles, which means that more than half of the categories of non-core roles are able to enter the core argument frame of predicates. Table 5 shows the top 10 non-core-entering-core roles in order of occurrences in the lexicon.

Roles	Freq
cause	1,454
location	934
destination	140
time	134
source	124
name	80
beneficiary	64
instrument	63
domain	33
extent	32

Table 5: Top 10 Non-core-entering-core Roles

From Table 5, we can see that *cause* is used most frequently, which usually acts as the proto-agent. It shows that concepts which represent the reason of a predicate are very easy to enter the core argument frame of the predicate. *Location* and *time* take second and fourth place. The third and fifth are *destination* and *source*, which often used to represent start and end point of location or time.

4.2 Representation of Multi-functional Roles

Although it is impossible that a predicate has more than 5 core arguments, CPB does not limit the types of concepts that can act as core roles of predicates. As long as it is an indispensable component of the meaning of a predicate, it can act as a core role of the predicate no matter what semantic relationship it has with the predicate. Take 药物缓解疼痛 (the drug relieves the pain) for example, the 药物 (drug) can serve as the agent as well as the cause of the predicate *relieve*, so the concept which represents agent and cause both can serve as the arg0 of 缓解-01 (relieve) in CPB.

缓解-01 (relieve)
arg0: **cause**, **agent**
arg1: theme

Since the description of core roles in CPB lexicon can only explain its relationship with the predicates, we cannot exactly count how many predicate frames have multi-functional core roles. However, data shows that there have been 1,146 senses whose arg0 can be acted by both concepts of agent and cause, accounting for 4.30% of all the senses. It shows that predicates in Chinese having multi-functional roles is common, and the core argument framework of CPB lexicon can represent the multi-functional roles well. That is to say, the CAMR's definition of core roles is reasonable for semantic representation.

5 Discrimination of Non-core Roles of Chinese AMR

In spite of AMR and CAMR has the same core labels as PropBank and CPB, there is a great difference between them for the quantity of non-core role labels. CAMR has 44 non-core role labels (Table 3), which are much more diversified than the 13 non-core role labels in CPB (Table 6). We calculate the using frequency of each non-core role label in CPB corpus and CAMR corpus, showed in Table 6 and Table 7. The mean deviations of them are 7,271.53 and 440.08, respectively. It means that the degree of difference in the using frequency of non-core role labels is much higher in CPB corpus than in CAMR corpus.

Labels	Description	Freq	%
ADV	adverbial	38,262	46.63
TMP	temporal	16,876	20.57
DIS	discourse maker	10,270	12.52
LOC	locative	7,104	8.66
MNR	manner	3,793	4.62
PRP	purpose or reason	2,344	2.86
DIR	direction	874	1.07
CND	condition	864	1.05
TPC	topic	605	0.74
EXT	extent	521	0.63
BNF	beneficiary	470	0.57
FRQ	frequency	49	0.06
DGR	degree	21	0.03

Table 6: Frequencies of Non-core Role Labels in CPB

Label	Freq	%	Label	Freq	%
beneficiary	2,804	19.21	accompanier	41	0.28
mod	2,098	14.38	topic	40	0.27
polarity	1,615	11.07	direction	37	0.25
*aspect	1,432	9.81	*cunit	36	0.25
manner	1,164	7.98	source	32	0.22
mode	1,097	7.52	cost	21	0.14
time	1,045	7.16	destination	18	0.12
degree	1,012	6.93	ord	17	0.12
cause	366	2.51	poss	15	0.10
purpose	362	2.48	unit	14	0.10
location	335	2.30	example	7	0.05
domain	154	1.06	path	6	0.04
duration	146	1.00	medium	2	0.01
instrument	103	0.71	name	1	0.01
frequency	99	0.68	value	1	0.01
compared-to	86	0.59	consist-of	0	0.00
condition	81	0.56	extent	0	0.00
*tense	76	0.52	part-of	0	0.00
range	73	0.50	polite	0	0.00
*perspective	57	0.39	subevent	0	0.00
li	54	0.37	subset	0	0.00
quant	46	0.32	superset	0	0.00

Table 7: Frequency of Non-core Role Labels in CAMR

It is obvious that the 13 non-core role labels of CPB is differ greatly in using frequency and they are too board to distinguish semantic roles of the predicates. From Table 6, we can see that the frequency of using *ADV* is nearly equal to the sum of the frequency of using other 12 labels. This is because they use *ADV* to represent almost all ambiguous semantic relations, such as 不 which means negation, 再 which means repeat, 首次 which represents order. In addition, *TMP* is unable to distinguish concepts of time, duration and time interval. Therefore, the granularity of the non-core role labels in CPB is too coarse, so it is unsuitable for automatic analysis of semantic relations. Nevertheless, setting too many non-core semantic role labels is also hard

for semantic analysis, and is a heavy burden for annotators, such as FrameNet. CAMR setting 44 non-core role labels is more suitable and reasonable due to the fact that it has a satisfactory discrimination.

6 AMR's Solution to Dropped Roles

Compared with other methods of meaning representation such as Dependency Graph, a big advantage of AMR is that it allows to re-analyze and add back dropped concepts in the sentences in order to represent the meaning of sentences more completely. For example, the nominal phrase *the injured* drops the agent of the predicate *injure*, AMR can add back a virtual node *person* for the phrase. Take *one of which…* for another example, it drops the noun that appeared in the preceding clause, AMR can add back a *thing* for it. Dropping semantic roles is common in Chinese. According to the statistics, CAMR annotates 1045 tokens of dropped roles for the 5,000 sentences, which cannot be annotated in other SRL resources. 619 of the added concepts have core semantic relation with the predicate, accounting for 59.23% of all the added concepts.

6.1 Adding back Core Roles for Predicates

Core roles of Predicates are of great significance for the meaning of a sentence. We try to explore whether the permission of adding back roles of AMR can help to annotate core roles of predicates more completely by comparing the CAMR corpus with the CPB corpus.

For each sense of each predicate, according to the difference between the quantity of core roles annotated in the corpus and the number of core roles in the predicate framework lexicon, the annotation of core roles can be classified into three categories: all the core roles are annotated (the difference is 0), not all the core roles are annotated (the difference is less than 0), the core roles are more than that in the lexicon (the difference is more than 0). [4] We call them core roles annotated completely, core roles annotated incompletely and the lexicon lack of core arguments, respectively.

We extract predicated frames from the CAMR corpus and the CPB corpus[5] and calculate the difference per sense. Data shows that there are 101326 tokens of senses of predicates in CPB corpus while 19823 tokens in CAMR corpus. Table 8 shows the distribution of quantity of senses in different difference between the quantity of core roles in the two corpus and the lexicon of CPB.

Corpus	Difference (corpus minus lexicon)	-4	-3	-2	-1	**0**	**1**	**2**	Total
CAMR	Tokens of senses	23	272	1,527	6,862	**11,037**	**99**	**3**	19,823
	% of senses	0.12%	1.37%	7.70%	34.62%	**55.68%**	**0.50%**	**0.02%**	100%
CPB	Tokens of senses	344	1,260	10,060	36,539	**52,735**	**383**	**5**	101,326
	% of senses	0.34%	1.24%	9.93%	36.06%	**52.04%**	**0.38%**	**0.00%**	100%

Note: if the predicate in CPB has semantic relations with multiple roles, it just counts as one tokens of sense.

Table 8: Difference between the Quantity of Core Roles in the Two Corpus and the Lexicon of CPB

From Table 8, we can see that the percentage of predicates whose core roles are annotated completely in the CAMR corpus is 3.64 more than the CPB corpus, and the percentage of senses whose core roles are short for the lexicon in CAMR is higher than that in CPB too. But the percentage of senses whose core roles are annotated incompletely is almost lower than the CPB corpus. It means that the AMR can annotate the core roles of predicates more completely. The main reason is that CAMR allows to re-analyze and add back dropped concepts, so that AMR isn't limited in the words of sentences, but can annotate core roles as complete as possible.

The proportion shows that there are also many predicates whose core roles are annotated incompletely. We consider the main reason is that AMR is a method to represent meaning of sentences, not the whole text, so that much information between sentences are missed. In the future, we will attempt to extend the AMR to the text level in order to represent meaning of texts more completely.

6.2 Adding back Dropped Roles of 3 Categories of Special Structures in Chinese

There are quite a few nominal structures dropping core roles of the predicates in Chinese. We choose three categories of special structures in Chinese to analyze: 的 structures, 所 structures and 所…的 structures. The function of adding concepts of AMR can represent their meanings completely. For example, the 的 structure 受伤的 (the injured) drops the agent of the predicate 受伤 (injure), CAMR can add a virtual node *person* and annotate the relationship between the dropped role and the predicate by *person :arg0-of 受伤*. Moreover, it is common that the patient of 的 structures is dropped. Take 我说的 ((what) I said) for example, it drops the theme of 说 (say), CAMR can add a *thing* for the structure. The 所 and the predicate in 所 structures form a nominal structure. Similar to 我说的, 所说 ((what) is said) drops the theme of 说 and CAMR can add back a *thing*. It seems impossible that a 所 structure drops its agent. A 所…的 structure is a combination of a 所 structure and a 的 structure. 所共有的 ((thing) shared by some people) drops a semantic role of 共有 (share), CAMR can also add back a *thing*.

We extracted all the 的, 所 and 所…的 structures in the corpus. According to the statistics, there are 309 的 structures, 9 所 structures and 7 所…的 structures in the 5,000 CAMR sentences. Though not very numerous, they are important and not negligible in Chinese. Data also shows that the number of dropped roles of agent and patient of 的 structures are essentially equal and the most dropped agents are *person* and most dropped patients are

[4] If the difference is less than 0, it is also possible that there are core roles being dropped. Similarly, if the difference is more than 0, it is also possible that there are core roles have not being annotated. But these two cases can be negligible because they are few in number.

[5] Because predicates which do not have core roles in CAMR corpus are difficult to be separated from other words, we ignore them for the moment.

thing. In addition, the dropped roles of these 9 所 structures and 7 所…的 structures are all patients and *thing*. Owing to the scale of the corpus is small, the data may not be able to cover all the situations, but it can also show that the dropped concepts of 所 and 所…的 structures are always the patient of the predicates, which is mainly because these two kinds of structures can represent the objects of actions by themselves.

From the data analysis of the adding semantic roles of predicates and the three types of nominal structures, we can see that the permission of re-analyzing and adding back roles of AMR can help to annotate the meanings of predicates more complete.

The CAMR's function of adding dropped roles also benefits from the framework designed by Li et al. (2017) that can align the AMR concepts and relations to word tokens in a sentence. It uses the index of a word token as the ID of its aligned concept in the AMR representation. When adding a role that is dropped, the added role will be assigned an ID which greater than the length of the sentence. Therefore, it is impossible to confuse the added roles with the words in the given sentence.

7 Conclusion and Future Work

In this paper, based on data analysis of the 5,000 sentences Chinese AMR corpus, we find that the AMR's definition of core roles can solve the collision between core and non-core roles and represent the multi-functional roles well. And the 44 non-core role labels of CAMR have a satisfactory discrimination to non-core semantic roles. In addition, benefited from the permission of re-analyzing and adding concepts, AMR can solve the problem of dropped semantic roles in the sentences, which is especially helpful for annotating special structures in Chinese such as 的 structures. Therefore, as a method of representing meaning of sentences, AMR has unique advantages in semantic role labeling and it is suitable for representing meanings of Chinese sentences, so we need to build a larger AMR corpus to serve the Chinese semantic processing.

A high-quality predicate framework lexicon is significant for ensuring the quality of the annotation. However, there are still many problems in the predicate lexicon we use at present: (1) senses of ambivalent words are not clear; (2) semantic roles do not correspond to the same core arguments, for example the concept of cause is arg0 of 压迫-01 (oppress) and arg1 of 取舍-01 (make the choice); (3) incomplete arguments and senses, for example 贴补-01 (subsidize) is lack of arguments of object and recipient and there is no adjective sense of 丰富 (abundant) in the lexicon. These problems has lowered the quality of annotation and the accuracy of automatic analysis, so we plan to modify the predicate frames manually. Moreover, some nouns have arguments like verbs such as 信心(confidence), but they are not included in the lexicon.

In the future, we will try to annotate semantic roles of nouns in Chinese AMR. And we plan to release our data for NLP applications and linguistics studies.

Acknowledgements

We thank the reviewers. This work is partially supported by projects 61772278, 61472191 under the National Science Foundation of China, project MJUKF201705 under Open Fund Project of Fujian Provincial Key Laboratory of Information Processing and Intelligent Control (Minjiang University), the Project for Jiangsu Higher Institutions' Excellent Innovative Team for Philosophy and Social Sciences, and the Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Badarau, B., Bonial, C., Georgescu, M., et al. Abstract Meaning Representation. https://amr.isi.edu/index.html.

Bai, X., & Xue, N. (2016). Generalizing the semantic roles in the Chinese Proposition Bank. *Language Resources and Evaluation*, 50(3), pp. 643-666.

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pp. 86-90.

Banarescu, L., Bonial, C., Cai, S., et al. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pp. 178-186.

Chen, K. J., Luo, C. C., Chang, M. C., et al. (2003). *Sinica Treebank*. Springer Netherlands, pp. 231-248.

Cai, S., & Knight, K. (2012). Smatch: an Evaluation Metric for Semantic Feature Structures. In *Meeting of the Association for Computational Linguistics*, pp.748-752.

Fillmore, C. J., et al. FrameNet. https://framenet.icsi.berkeley.edu/fndrupal/

Jia, J. Z. (2007). Study on the Comparison of Framenet Wordnet Verbnet. *Information Science*. 25(11), pp.1682-1686.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

Kipper, K., Dang, H. T., & Palmer, M. (2000). Class-Based Construction of a Verb Lexicon. In *Seventeenth National Conference on Artificial Intelligence & Twelfth Conference on Innovative Applications of Artificial Intelligence*, pp. 691-696.

Li, B., Wen, Y., Bu, L., et al. (2017). A Comparative Analysis of the AMR Graphs from English and Chinese Corpus of the Little Prince. *Journal of Chinese Information Processing*, 31(1), pp. 50-57.

Li, B., Wen, Y., Bu, L., et al. (2016). Annotating the Little Prince with Chinese AMRs. In *Proceedings of the 10th Linguistic Annotation Workshop*.

Li, B., Wen, Y., Song, L., et al. (2017). Construction of Chinese Abstract Meaning Representation Corpus with Concept-to-word Alignment. *Journal of Chinese Information Processing*, 31(6), pp. 93-102.

Li, R., Lv, G., Gao, J. et al. Chinese FrameNet. http://sccfn.sxu.edu.cn/portal-zh/home.aspx.

Ma, W., Chen, K., Xie, Y., et al. Sinica Treebank. http://turing.iis.sinica.edu.tw/treesearch/.

Palmer, M., Gildea, D., & Kingsbury, P. (2006). The Proposition Bank: an Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1), pp. 71-106.

Palmer, M., Kipper, K. Loper, E, et al. VerbNet. http://verbs.colorado.edu/~mpalmer/projects/verbnet/members.html.

Weischedel, R., Hovy, E., Marcus, M., et al. (2011). Ontonotes: a Large Training Corpus for Enhanced Processing. *Handbook of Natural Language Processing & Machine Translation*.

Xue, N. (2006). A Chinese Semantic Lexicon of Senses and

Roles. *Language Resources & Evaluation*, 40(3-4), pp. 395-403.

Xue, N., Palmer, M. (2009). Adding Semantic Roles to the Chinese Treebank. *Natural Language Engineering*, 15(1), pp. 143-172.

Xue, N., Wang, C., Zhang, Y., et al. Chinese Abstract Meaning Representation. http://www.cs.brandeis.edu/~clp/camr/camr.html.

Xue, N., Xia, F., Chiou, F. D., & Palmer, M. (2005). The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus. *Natural Language Engineering*, 11(2), pp. 207-238.

Yuan, Y. L. (2007). The Fineness Hierarchy of Semantic Roles and Its Application in NLP. *Journal of Chinese Information Processing*, 21(4), pp. 10-20.