A Glue-like Semantics for Mandarin Chinese: Annotation and Formalism

Zi Lin
Department of Chinese Language and Literature
zi.lin@pku.edu.cn

1 Introduction

1.1 FrameNet

The FrameNet project is to design a lexical database of English based on a linguistic theory called Frame Semantics. The meaning of most words can best be understood on the basis of a semantic frame, and the basic assumption on which the frames are built is that each word evokes a particular situation with particular participants (Fillmore, 1968). For example, the concept of cooking usually involves a person doing the cooking (Cook), the food that is to be cooked (Food), something to hold the food while cooking (Container) and a source of heat (Heating_instrument). In FrameNet, this event is represented as a frame called Apply_heat, while the Cook, Food, Heating_instrument and Container are all frame elements (FEs), and predicates evoking frames are called lexical unites (LUs).

FrameNet provides a description of more than 1,200 semantic frames underlying the meaning of the words described, and valence representation (semantic and syntactic) of several thousand words and phrases, each accompanied by a representative collection of more than 200,000 manually annotated sentences (Baker et al., 1998). However, FrameNet fails to provide precise syntax-semantics interface as the annotation for frames and FEs is directly based on running sentences without the interaction with syntactic entities.

Many of the FrameNet frames, including the constructions, apply equally well to other languages, as evidences by the various efforts to develop FrameNets in other languages such as French (Candito et al., 2014), Chinese (Liu, 2011), German (Burchardt et al., 2009) and so on. Among the previous work, the Chinese FrameNet (CFN) is most relevant to our work. However, the CFN is still under construction, and like many other FrameNet project, they take Berkeley's FrameNet Project as the reference and annotate the frame semantics on running sentences, which would be really time-consuming and labor-intensive. And similar to English FrameNet, they are dependent from syntactic structure, while other semantic representations, like the Proposition Bank (PropBank), provide information about basic semantic propositions and predicate-argument relations based on syntactic trees.

1.2 PropBank

The purpose of the Chinese PropBank (CPB) project is to add a layer of annotation to the hand-parsed sentences in the Chinese Treebank (CTB) (Xue et al., 2005). This layer of annotation assigns predicate-specific argument labels to the constituents in parse trees.

Argument labels in the form of Arg$N$ are assigned to arguments of each predicate in sentences, where N is an integer between 0 and 5, and the predicate are limited to verbs and their nominalizations. These numbered arguments refer to core arguments that are defined with respect to the predicate. The major senses of the predicate is called frameset. For example, in the sentence "警方正在详细调查事故原因" (the police are investigating the cause of the accident in detail), the verb "调查" has the following argument set:

Arg0: 警方("police")
Arg1: 事故("accident")原因("cause")

CPB addresses the syntax-semantics mismatches as syntactic entities do not always have a unique mapping to an entity in the predicate-argument structure and syntactic trees do not always transparently reflect the argument structure of a predicate. And CPB also takes special problems of Chinese such as BA-construction and BEIconstruction into consideration. However, as for deep semantic representation, the label of $ArgN$ is not sufficient compared with FEs in FrameNet. We propose that to offer an efficient deep semantic presentation for Chinese, the construction should be grounded in both the lexical semantics and syntactic structure. Therefore, taking advantages of FrameNet and Propbank, it is necessary to link the entities in both representations as to build a syntax-semantics interface.

2 SemBanking: Annotation

The annotation procedure aims to provide type-to-type mappings between the lexical units for each framework. For Chinese Propbank (CPB) these are the framesets and arguments for verbs; for English FrameNet they are the frames and FEs.

There are over 1,200 frames in FrameNet, making the range of frame selections extremely wide. However, high-quality classification can be induced for new languages by concentrating on translation pairs of source and target language lemmas which are especially likely to be frame-preserving (Burchardt et al., 2009).

Therefore, before the annotation, 200,751 English sentences in FrameNet are translated into Chinese to obtain the corresponding Chinese verbs for each LU within FrameNet frames. By doing so, each Chinese verb can be linked to one or more candidate frames.

The verbs which have candidates are limited to translated ones, to expand our Chinese verb set, we clustered verbs by associating them with the same conceptual domains using a Chinese thesaurus - the Tongyici Cilin (Jiaju et al., 1983). If in the same conceptual domain, there is any verb that has been linked to FrameNet, other verbs in CPB can inherit all candidate frames of this verb. In this way, we finally get 7,996 verbs with totally 901 possible FrameNet frames. These candidate frames will obtain higher ranks on the option list in the subsequent annotation.

To evaluate the efficiency of our methodology, we list average candidate frames and the hit ratio with regard to different ranges of word frequency in CTB corpus, as shown in Table 1. Based on our annotation result, totally, 55.35% of the framesets have the exact FrameNet frames in our candidate set, demonstrating that our collection of candidate frames would definitely narrow the searching space for annotators.

word frequency	**# word**	**# candidate per word**	**hit ratio**
>=500	24	11.54	70.97
400-500	7	9.29	100.00
300-400	15	6.53	79.17
200-300	36	6.53	80.39
100-200	103	5.33	63.00
50-100	183	4.25	74.54
10-50	1172	2.59	58.09
<10	1087	2.12	44.17
total	2627	2.79	55.35

Table 1: Statistics of candidate frame and the corresponding hit ratio

An annotation tool has been developed for linking from CPB to FrameNet, accessible by web interface. In this step, we aims to annotate 3,728 high frequency framesets in CPB, listed in descending frequency order.

It's better for annotators to get a good understanding for the full range of frames within a shorter period and double-check the frames to achieve high accuracy. We thus split the procedure into two steps: (1) Linking between predicates in CPB and frames in FrameNet: offered a list of FrameNet frames in descending order of ranks, annotators have to select the most suitable one. (2) Linking between arguments in CPB and FEs in FrameNet: annotators will determine corresponding FEs for each argument based on the chosen frames in the first step. Revision should be made if the chosen frames is not appropriate.

For example, for CPB frameset “住” (live) we first mapped it to Frame `Residence` in FrameNet. “住” has two arguments - $Arg0$ (entity described) and $Arg1$ (location arg0 lives), corresponding to `Resident` and `Location` in FEs. Annotators would set the degree of confidence for their answers and search for frames or FEs by inputting the name if necessary.

Most of the English FrameNet frames can be re-used for the semantics analysis of Chinese. In our annotation work, 3,270 framesets in CPB have equivalents in FrameNet, covering 627 Frames. Nonetheless, four central types of problems should be discussed.

Frame: Be	
Definition: 联系两种事物，表明后者说明前者的种类、属性、情况、值 (Relating two constituents, indicating the latter is the Category, Attribute, Situation or Value of the former Item)	
Item	**藏羚羊、野牦牛、野驴、盘羊** 都是珍稀动物 **Tibetan antelope, wild yak, wild donkey, argli and etc.** are all rare animals.
Category	藏羚羊、野牦牛、野驴、盘羊都是**珍稀动物** Tibetan antelope, wild yak, wild donkey, argli and etc. are all **rare animals**.
Attribute	盆地中大量恐龙蛋化石的发现，已属**世界罕见** The discovery of a large number of dinosaur egg fossils in the basin has become **rare in the world**.
Situation	假如他加盟纽卡斯尔队，也算是 **回家乡球队效力** If join Newcastle, he can be considered as **playing for his hometown team**.
Value	乡镇企业贷款增幅为 **61.83%** The growing rate of township enterprise loans is **61.83%**.

Table 2: Example of a proto-frame

(1) Polysemy: For word sense disambiguation, CPB defines different framesets for different meanings within same lemma. Identical lemmas can have multiple framesets and can be in several FrameNet frames, thus users should hand-correct each frameset with identical lemma.

(2) Differences in lexicalization of frames: Languages show strong preferences as to what kinds of semantic components they lexicalize (Talmy, 1985). The meanings of Chinese verbs sometimes cut across the frame distinctions designed on the basis of English data. We notice that some framesets may be linked to multiple frames, thus users are allowed to assign more than one frames. According to our result, 109 CPB framesets have multi-frames.

(3) Missing frames: A major problem is that FrameNet is still under development, and thus does not yet cover all senses of the framesets we annotate. Another, more significant problem, is that there exists gap between Chinese and English in terms of verb frames. For those 458 framesets which cannot be described in terms of existing frames. We group them into coarse-grained groups and construct a proto-frame for each group. We further linked them to definitions given by the Contemporary Chinese Dictionary and list their FEs, as shown in Table 2.

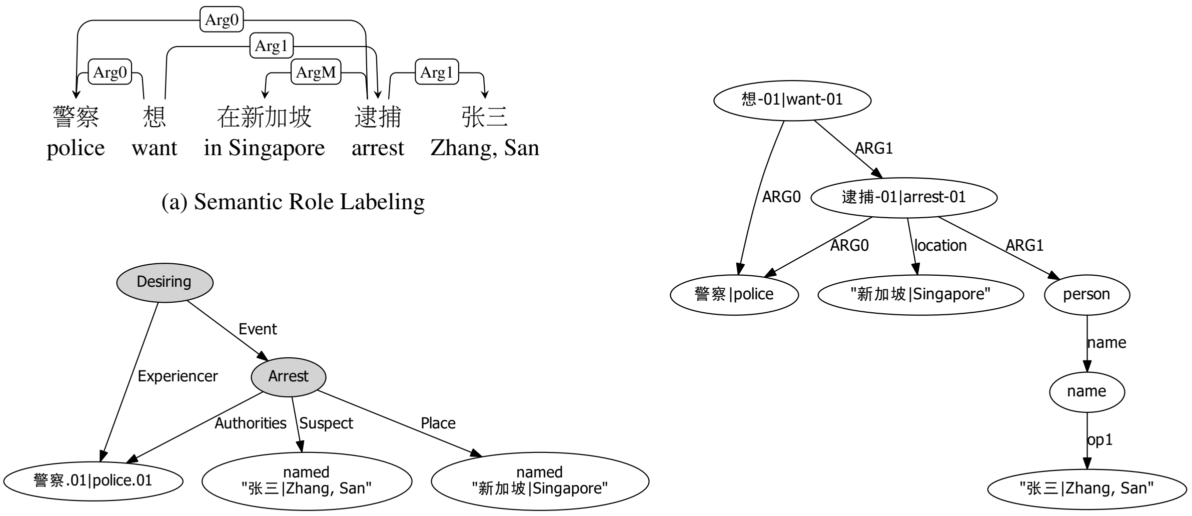
(4) Difficult role distinctions: FrameNet uses ontological criteria to differentiate closely related but mutually exclusive FEs, but these criteria may not necessarily apply to different arguments in CPB. There are two situations - arguments can be mapped to several FEs (19 cases) and no appropriate FE is defined (56 cases). Similar to multi-frames, we allow annotators to choose multiple FEs. As for the other one, we set default FE for the moment.

3 SemBanking: Formalism

Graphs have a long history in Computer Science as representational devices for relational data. In Natural Language Processing, however, linguistic representations of syntax and semantics have traditionally been limited to strings and trees. While directed graphs are an intuitive and versatile representation of natural language meaning because they can capture relationships between instances of events and entities, including cases where entities play multiple roles. There have been many efforts in philosophy of language, linguistics, and NLP to develop more advanced meaning representations, such as Abstract Meaning Representation (AMR) (Langkilde and Knight, 1998; Banarescu et al., 2013), Boxer Discourse Representation Structures (DRS) (Bos, 2014) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005), etc.

AMR is a graph-based semantic representation that aims to **abstract away from the syntactic strucutre of natural langauge sentences**, representing the sentences as rooted, labeled, directed, acyclic graphs (DAGs). AMR frequently assigns the same graph to different sentences that mean the same thing. For example, the following sentences should have the same AMR graph:

1. Their voting early surprised me.
2. Their early voting surprised me.
3. That they voted early surprised me.
4. Them voting early surprised me.



(a) Semantic Role Labeling

(b) SemBank Graph

(c) AMR Graph

Figure 1: SRL, SemBank and AMR Graph for the sentence, “警察想在新加坡逮捕张三”(The police want to arrest Zhang, San in Singapore)

Figure 1 shows different semantic graphs for the sentence “警察想在新加波逮捕张三” (The police want to arrest Zhang, San in Singapore), among which we give the simple SRL, AMR as well as our sembanking graph.

References

Collin F Baker, Charles J Fillmore, and John B Lowe. 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-Volume 1*. Association for Computational Linguistics, pages 86–90.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. pages 178–186.

Johan Bos. 2014. Comparative computational semantics: Capturing meaning by boxes at the berlin workshop .

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009. Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. *Multilingual FrameNets in Computational Lexicography: methods and applications* pages 209–244.

Marie Candito, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël De Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, et al. 2014. Developing a French Framenet: Methodology and First Results. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A Sag. 2005. Minimal recursion semantics: An introduction. *Research on language and computation* 3(2-3):281–332.

Charles J Fillmore. 1968. The Case for Case. *Universals in Linguistic Theory* .

Mei Jiaju, Zhu Yiming, Gao Yunqi, and Yin Hong-Xiang. 1983. Tongyici Cilin. *Shanghai Cishu Publisher* .

Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, pages 704–710.

Kaiying Liu. 2011. Research on Chinese FrameNet Construction and Application Technology. *Journal of Chinese Information Processing* 25(6):46–52.

Leonard Talmy. 1985. Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description* 3(99):36–149.

Naiwen Xue, Fei Xia, Fu-Dong Chiou, and Marta Palmer. 2005. The Penn Chinese TreeBank: Phrase structure annotation of a large corpus. *Natural language engineering* 11(2):207–238.