Construction of the Contemporary Chinese Common Verbs' Semantic Framework Dictionary

Tongfeng Guan[1], Kunli Zhang[1(✉)], Xuemin Duan[1], Hongying Zan[1], and Zhifang Sui[2]

[1] School of Information Engineering, Zhengzhou University, Zhengzhou, China
guantf@gs.zzu.edu.cn, {ieklzhang,iehyzan}@zzu.edu.cn, xueminduan@stu.zzu.edu.cn
[2] Institute of Computational Linguistics, Peking University, Beijing, China
szf@pku.edu.cn

Abstract. Semantic lexicon and semantic framework are the primary support of natural language processing tasks such as information extraction, sentiment analysis, and machine translation. Therefore, it is essential to construct the contemporary Chinese common verbs' semantic framework dictionary that covers rich semantic knowledge. Based on an analysis of current research results, this paper defines the lexical framework of common Chinese verbs. According to the predicate thematic roles, the semantic framework is divided into the basic semantic framework and extended semantic framework. Frameworks which are automatically extracted, taking semantics as the processing unit, and summarized based on large-scale lexical and thematic roles labeling corpus. The complete and simplified versions of the verb framework is constructed with the help of manual proofreading. The final verb framework contains a detailed description and corresponding example sentences of 2,782 common verbs with 4,516 meanings.

Keywords: Contemporary Chinese common verbs · Semantic framework dictionary · Automatic extraction · Basic semantic framework · Extended semantic framework

1 Introduction

Natural Language Processing (NLP) is the technology used to aid computers to understand natural language statements and texts, and extract essential information for analysis, retrieval, reading, question and answer, machine translation and text generation. To achieve this goal, scholars have conducted in-depth research in the fields such as part-of-speech tagging, syntactic analysis, and semantic dependency parsing. Significant advances have also been made in related research areas.

There have been many lexical-semantic knowledge bases, such as WordNet [1], FrameNet [2], MindNet [3]. Representative achievements in Chinese language include the Chinese Semantics Dictionary (CSD) [4], the Chinese Concept Dictionary (CCD) [5], the Chinese Function words Knowledge Base (CFKB) [6, 7] and the

J.-F. Hong et al. (Eds.): CLSW 2019, LNAI 11831, pp. 546–553, 2020.
https://doi.org/10.1007/978-3-030-38189-9_56

Chinese Large-Scale Knowledge Base (CLSKB) [8]. In Chinese FrameNet (CFN) [9] and Mandarin VerbNet (MV) [10], verbs are taken as the primary research object. CFN is a FrameNet-style Chinese framework semantic frame net, which translates or creates a framework suitable for Chinese semantic content, and defines framework-framework relationships. However, it defines a small number of verb frames, only 309 frames. MV is the result of linguistic semantics research using a framework-based constructional approach. It has completed research on major verb categories such as 'caused-motion', 'cognition', 'communication', 'emotion', 'motion', 'perception' and 'social-interaction'. It proves that framework-based Chinese verb and verb class framework analysis methods have proper motivation and effect in language. However, due to the artificial construction method, the classification is limited. Currently, the construction of a semantic knowledge base mostly adopts the method of artificial construction and fails to put the semantic of words in a particular combination framework for observation. The static aggregation classification method is often used in the construction. But few attribute descriptions are added, so the construction scale and update speed cannot meet the needs of current NLP development.

The lexical semantic frame of verbs is a semantic relationship framework system composed of verbs and their thematic roles in the dictionary. Verbs exist in every complete sentence, expressing the movement, development, change, existence, or demise of a person or thing. Their thematic roles are divided into core roles and peripheral roles. The core roles, including an agent of the action, a patient of the action, and both. The peripheral roles that complement the sentence context and add a sense of picture to the sentence include instrument, material, manner, reason, purpose, a point of departure, a point of arrival, start and end time, etc.

This paper is supported by the National Key Basic Research and Development Program 973 (2014CB340504), in the context of the Broad-Spectrum semantic Dictionary of Contemporary Chinese (BCSD). It has integrated and absorbed current research results, and it defines the verb semantic framework dictionary structure. The paper separates the semantic framework into a Basic Semantic Framework (BSF) and an Extended Semantic Framework (ESF). With the semantics as the processing unit, the BSF and the ESF are automatically extracted and summarized from the large-scale lexical and thematic roles labeling corpus. Supplemented by manual proofreading, a lexical semantic framework dictionary has been constructed. This paper simplifies the framework when studying the differences in the verb implementation framework. We have constructed a detailed description of the semantic framework containing a total of 2,782 commonly used verbs (4,516 meanings) with corresponding example sentences.

2 Verbs Semantic Framework

Predicate verbs play a cohesive role in sentences. The construction of sentences must surround the core of verbs. Therefore, verbs are considered the soul of sentences. On the other hand, verbs have constraints on the semantic classes of their arguments, so constructing the BSF of verbs is conducive to grasping the human cognitive model of linguistic knowledge. In the construction of the lexical semantic dictionary, current

research results are fully absorbed and borrowed. On this basis, the semantic attribute description framework of the existing word classes is expanded and improved.

The construction of the verb semantic framework refers to the principle of the collection of words and the semantic of the fifth edition of the Modern Chinese Dictionary (XH5). Meanwhile, fusion and inheritance of the modern Grammar Knowledge Base (GKB) [11], The Modern Chinese Semantic Dictionary and the word class system, semantic classification system and related semantic attribute description information of the large-scale lexical semantic database are involved to reduce the time and manpower required for the construction process, and to ensure the high quality and high credibility of the dictionary. The lexical semantic dictionary selects 2782 of the intersection of the fifth edition of Modern Chinese Dictionary and the Modern Chinese Grammar Information Dictionary, including 1865 single senses and 917 polysemous words, according to Modern Chinese predicate Semantic Role Labeling Corpus Specification (the 973 specification) in the 'predicate thematic roles hierarchy classification system' as shown in Fig. 1. It is a system of semantic relationship framework constructed by combining the verbs with their theoretic roles.

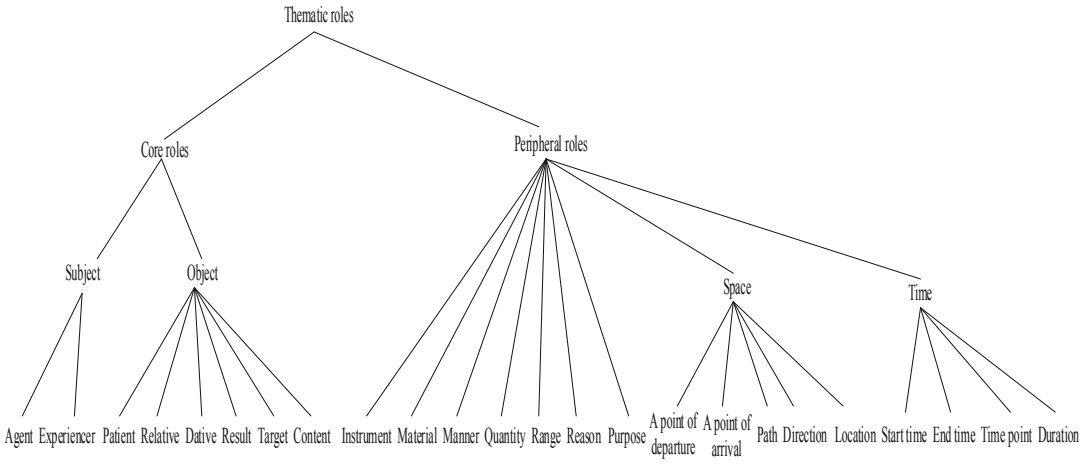

Fig. 1. The hierarchical classification system of predicate thematic roles

According to the different relationships between the roles of the verbs, roles can be divided into two types: core roles and peripheral roles. Core roles represent an essential component closely related to the verb. Without this part, the verb cannot be completely described. The core roles include agent, experiencer, patient, relative, dative, result, target, content. Peripheral roles represent a component that is more distant from the verb. If this part is missing, the verb of the core roles can still describe the event, but the content is not comprehensive. Adding the peripheral components makes the narrative vivid and stereoscopic. Peripheral roles consist of two subordinate themes of space and time and sixteen three-level theoretic roles.

2.1 Basic Framework of the Dictionary

The basic framework of the verb library includes three main parts: verb entity, semantic framework, and thematic roles. The verb entity includes verb term and phonetic, definition, and example of the word. The semantic framework contains the BSF and ESF. Moreover, the thematic roles consist of 8 core roles and 18 peripheral roles (a total of 24 three-level thematic roles and 2 secondary thematic roles).

2.2 Framework Description

The verb entity part is extracted directly from the fifth edition of the Modern Chinese Dictionary. Semantic framework refers to the semantic relationship framework composed of verbs and their thematic roles, including BSF and ESF. Verb semantic framework dictionary needs to match the corresponding example sentences with the verbs, and then automatically extracts the BSF and the ESF from the example sentences. Therefore, the dictionary needs to sort the verb necessary information, the semantic framework, and the thematic roles. Some verbs may be empty due to insufficient corpus size. The BSF only involves the semantic relationship framework that contains core roles, and the ESF involves the peripheral roles.

3 Semantic Framework Extraction

Both the BSF and the ESF have various implementation frameworks. Therefore, it is necessary to select a typical structural framework as a representative. Verb semantic framework dictionary should extract a typical structural framework.

The building of the verb semantic framework dictionary and semantic roles labeling corpus is an iterative process. Firstly, develop a verb framework and semantic roles labeling specifications for characteristics. Secondly, the automatic framework extraction method is determined. Then, the annotations are discussed weekly. As problems arise, adjust the labeling specification, and guide the labeling of the next corpus to sort out the new verb framework.

The semantic framework is extracted from Peking University Chinese Tree Library [12] and People's Daily's corpus. Among 55,764 sentences in the Chinese tree library of Peking University, 10,634 sentences were extracted according to rules of Chinese grammar. The semantic roles were marked in the corpus (referred to as 10,000 sentences). From the corpus of People's Daily on January 2000 and January 1998, we have extracted two batches of corpus (referred to as People's Daily and new, total 42,000 sentences) that have been marked with the target verb meaning and thematic roles.

In the process of processing the semantic framework, four main steps are included: data preprocessing, BSF processing, ESF processing, and simplification, as shown in Fig. 2. Monosyllabic words can be treated as polysemous words with only one meaning. In the process, when there are multiple example sentences, the default example sentences are selected according to the priority order of '10,000 sentences -> People's Daily -> new'.

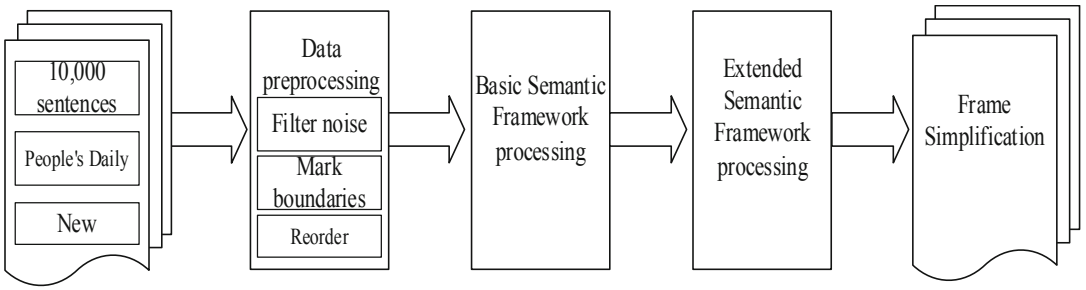

Fig. 2. Frame processing

3.1 Data Preprocessing

The main task of the data preprocessing is deleted useless example sentences, mark boundaries, reorder, and filter noise. Reordering is to align the same implementation framework. Noise is filtered according to the rationality of the framework before calculating the typical framework. In the 973 specifications, the correspondence between the subject and the object is constrained, as shown in Table 1. After processing, monosyllabic words only retain the words with more than three example sentences, and polysemous words delete the meanings without the example sentences.

Table 1. Subject and object correspondence relation

Object	Subject	Relation
Relative	Experiencer	Can't correspond to agent
Patient	Agent	Can't correspond to experiencer
Dative	Agent	Usually does not correspond to the experiencer
Result	Agent | experiencer	Can correspond to agent and experiencer
Content	Agent | experiencer	Can correspond to agent and experiencer
Target	Agent | experiencer	Can correspond to agent and experiencer

3.2 Basic Semantic Framework Processing

The position of the thematic roles in the actual statement is flexible. A semantic term of a verb may have multiple semantic structures that differ in the order in which the roles are located. The verb library regards it as the semantic framework of the implementation. The semantic framework abstracted from the implemented semantic framework is considered as a typical semantic framework. The implemented semantic framework and typical semantic framework are general and typical of the semantic framework. The typical semantic frameworks are arranged in the logical order in which the thematic roles appear.

The implemented framework can be regarded as a group member, and the typical framework is the squad leader selected from the group members. With the typical framework, the meaning item can be easily recognized. The processing of the BSF is divided into three steps. The first step is to select all the monosyllabic words to identify

the typical framework. The second step is to replace the typical structure that is automatically identified in the first step by using a typical structure that has been manually selected. And the third step determines the example sentences of the corresponding implementation framework and the merged semantic roles for the results of the first step and the second step.

3.3 Extended Semantic Framework Processing

The ESF contains 18 peripheral roles, so there are multiple implemented semantic frameworks, from which the typical semantic framework needs to be extracted. The typical ESF is based on the typical BSF but contains peripheral roles. Also, some adjustments are made according to the actual situation.

Begin the judgment of whether the typical BSF is a subset of some implementation frameworks. If it is a subset, the next step is to consider whether there are multiple implemented semantic frameworks. If yes, merge these frameworks, and if not, choose the only framework. Intersections also need to consider whether the core roles are the same when merging. If they are different, consider choosing the one with the longest implementation structure and the largest number of examples. Eventually, all verbs can reach the end state.

3.4 Frame Simplification

After completing the automatic extraction of the BSF and the ESF, to compare the differences in the implementation framework of the verbs in different sentences, and explore the internal rules of the semantic framework, we need to simplify the results. The implementation framework is merged into a single cell, separated by '|', leaving only typical example sentences. The semantic roles are deduplicated and merged, and finally, one verb retains only two rows of records.

4 Results and Analysis

Statistical analysis of the results of a typical framework for the extraction of monosyllabic and polysemous words, as shown in Table 2. The total number of meanings in the table is the number of all the meanings contained in each type of verb. Since each polysemous word contains multiple meanings, 917 polysemous words have a total of 2651 meanings.

Among these meanings, 785 show semantic frameworks and extended semantic frameworks, 672 show basic semantic frameworks only, 208 show extended semantic frameworks only, and 986 meanings do not show any framework. Extracting parts from 986 meanings for analysis leads to the identification of two reasons. The first is that some verbs appear in single words, such as '上[*shàng*] (*to board; to attend; to go to*)', '下[*xià*] (*to fall; to go down*)', '交[*jiāo*] (*exchange, communicate; to turn over; to make*)', '令[*lìng*] (*to order; to cause*)', etc. Some of the meanings of these verbs are not used often enough, so it does not appear in the corpus, there are also verbs containing these words, such as '上台[*shàngtái*] (*to go on stage (in the theater); to rise to power (in politics)*)',

‘上学[*shàngxué*] (*to go to school; to attend school*)’, ‘下船[*xiàchuán*] (*disembark; boarding*)’, ‘下岗[*xiàgǎng*] (*to lay off; to come off sentry duty*)’, ‘交换[*jiāohuàn*] (*to exchange; to switch*)’, ‘交替[*jiāotì*] (*to replace; alternately; in turn*)’, etc., respectively covering the meaning of ‘上[*shàng*] (*to board; to attend; to go to*)’, ‘下[*xià*] (*to fall; to go down*)’, and ‘交[*jiāo*] (*exchange, communicate; to turn over; to make*)’. When the corpus is segmented, it is taken as a whole. The second reason is that the corpus used is mainly news. Some words do not appear in the news of that period, such as ‘上 升[*shàngshēng*] (*to rise; to go up*)’, ‘临门[*línmén*] (*facing the goalmouth (soccer); to be at the door*)’, ‘修养[*xiūyǎng*] (*accomplishment; training*)’, etc. Subsequent work should consider more types of corpora, thus increasing the number of examples.

Table 2. Statistical semantic framework extraction results

Category	Num	Filter noise	Total of meanings	Only BSF	Only ESF	Both	Neither
Polysemous	917	–	2651	672	208	785	986
Monosyllabic	1865	190 (The number of examples is less than 3)	1865	236	37	1592	0
Total	2782	190	4516	908	245	2377	986

5 Conclusion

This paper defines a verb semantic framework dictionary, divides the framework into BSF and ESF, and automatically extracts and summarizes the BSF and ESF based on the large scale labeled lexical and thematic corpus. We have constructed a dictionary of contemporary Chinese common verbs semantic framework containing 917 polysemous words and 1865 monosyllabic words, which provide support for natural language processing applications such as machine translation, sentiment analysis, and information extraction. By analyzing the dictionary, we find that some meanings of polysemous words have neither the BSF nor the ESF. The main reason is that the source of the corpus is news. For the next step, coverage of other corpora with different data sources should be considered to expand the diversity of example sentences, extract a more representative typical framework from the diversified implementation frameworks to improve the quality of the verb semantic framework dictionary. Moreover, the scale of the verb semantic framework should be expanded to include more verb entries.

Acknowledgments. We thank the anonymous reviewers for their constructive comments, and gratefully acknowledge the support of the National Key Basic Research and Development Program under Grant No. 2014CB340504; The National Social Science Fund of China under Grant No. 18ZDA315; the Key Scientific Research Program of Higher Education of Henan under Grant No. 20A520038; the science and technology project of Science and Technology Department of Henan Province under Grant No. 192102210260; and the international cooperation project of Science and Technology Department of Henan Province under Grant No. 172102410065.

References

1. Fellbaum, C., Miller, G.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge (1998)
2. Bake, C.-F., Fillmore, C.-J., Lowe, J.-B.: The Berkeley FrameNet project. In: Proceedings of COLING 1998, pp. 86–90 (1998)
3. Richardson, S.-D., Dolan, W.-B., Vanderwende, L.: MindNet: acquiring and structuring semantic information from text. In: ACL, pp. 1098–1102 (1998)
4. Hui, W., Zhan, W.-D., Yu, S.-W.: The specification of the semantic knowledge-base of contemporary Chinese. J. Chin. Lang. Comput. **13**(2), 159–176 (2003). (in Chinese)
5. Liu, Y., Yu, S.-W., Yu, J.-S.: A study on the construction of CCD. J. Chin. Comput. Syst. **26** (8), 1411–1415 (2005). (in Chinese)
6. Zan, H.-Y., Zhang, K.-L., Zhu, X.-F., et al.: Research on the Chinese function word usage knowledge base. Int. J. Asian Lang. Process. **21**(4), 185–198 (2011). (in Chinese)
7. Zhang, K.-L., Zan, H.-Y., Chai, Y.-M., et al.: Survey of the Chinese function word usage knowledge base. J. Chin. Inf. Process. **29**(3), 1–8 (2015). (in Chinese)
8. Shi, J.-M., Zan, H.-Y., Han, Y.-J.: Specification of the large-scale Chinese lexical semantic knowledge base building. J. Shanxi Univ. Nat. Sci. Ed. **38**(4), 581–587 (2015). (in Chinese)
9. You, L.-P., Liu, K.-Y.: Building Chinese FrameNet database. In: Proceedings of 2005 IEEE NLP-KE, pp. 301–306 (2005)
10. Liu, M.-C., Chang, J.-C.: Semantic annotation for Mandarin verbal Lexicon: a frame-based constructional approach. In: Proceedings of the 2016 International Conference on Asian Language Processing, pp. 30–36 (2016)
11. Yu, S.-W., Zhu, X.-F., Wang, H., et al.: The Grammatical Knowledge-Base of Contemporary Chinese—A Complete Specification. Tsinghua University Press, Beijing (2003). (in Chinese)
12. Zhan, W.D.: Encyclopedia of Chinese Language and Linguistics, vol. 3, pp. 332–336. Brill Publishing House, Leiden (2016)