Annotating Chinese Light Verb Constructions according to PARSEME guidelines

Menghan Jiang, Natalia Klyueva, Hongzhi Xu and Chu-Ren Huang
The Hong Kong Polytechnic University
University of Pennsylvania
menghan.jiang, natalia.klyueva, churen.huang@polyu.edu.hk
xh@cis.upenn.edu

Abstract

In this paper we present a preliminary study on application of PARSEME guidelines of annotating mutiword expressions to Chinese language. We focus on one specific category - light verb constructions (LVCs). We make use of an existing resource containing Chinese light verbs and examine whether this resource fulfill the requirements of the guidelines. We make a preliminary annotation of a Chinese UD treebank in two steps: first automatically identifying potential light verbs and then manually assigning the corresponding nouns or correcting false positives.

Keywords: verbal multiword expressions, light verb constructions, Chinese, corpus linguistics

1. Introduction

Multiword Expressions (MWEs) present a challenge in a bunch of areas of Natural Language Processing like Machine Translation or Information Extraction. They are idiosyncratic in their nature - the meaning of the whole can not be derived from the meaning of its parts. The exact definition of MWEs varies across different linguistic theories and can not be necessarily universal for all languages.

PARSEME (Savary et al., 2017) is a European project which aims at processing multiword expressions from different perspectives and for various languages. Universal guidelines were created to annotate verbal multiword expressions distinguishing several subtypes unique for a language: idioms, LVCs, verb-particle constructions, inherently reflexive verbs and others (miscellaneous category). Corpora in 18 languages have been annotated which result in a multilingual resource (Savary et al., 2017). One of the long-term plans of the project is to extend the set of (mostly European) languages to Asian languages, including Chinese Mandarin. Chinese linguistic tradition is different from the European which presents a challenge when trying to apply the guidelines created under the European project to Chinese. In this paper we make a pilot study of one particular type of verbal MWEs - LVCs and see how the existing resources for Chinese can be adjusted to the PARSEME annotation schema.

We consider two possible corpora for annotation: the Sinica Balanced Corpus of Modern Chinese(Huang et al., 2000) and Universal Dependencies (Nivre et al., 2017). Finally, we have chosen Chinese Universal Dependencies treebank[1] (wiki data) because it features syntactic annotation in dependency-based format required by the shared task.Unlike other languages such as Czech (Bejček et al., 2017), there is no resources we can use to generate MWEs automatically for Chinese. But we don't do the annotation from scratch as well. We use a list of Chinese light verbs to pre-process the data, and then manual work is performed to assign the corresponding nouns of the MWEs and exclude the false ones.

This paper is structured as follows. In Section 2 we describe which types of verbal MWEs can occur in Chinese, in Section 3 we give more extensive information on Chinese LVCs, and in Section 4 we explore how the annotation guidelines created under the PARSEME shared task can be applied to the already existing resource of Chinese LVCs examining the tests. Section 5 presents the statistical information of the annotation result.

2. Verbal Multiword Expressions in Chinese

Following the concept of VMWEs (Verbal Multiword Expressions) that was set in PRASEME project, we can explore two categories of VMWEs in Chinese: verbal idioms (ID),such as 吃醋 *eat vinegar* 'to be jealous' and light-verb constructions (LVC).

In addition to light verb construction, Chinese **Verb+Obj1+Obj2** construction can also be considered as VMWE. The VO compound (e.g., 帮忙 *bangmang* 'do favor'/入籍 *ruji* 'naturalize') can be used transitively taking an external object, such as 入籍中国 *ruji zhongguo* 'naturalize China'. Furthermore, the external object Obj2 can also be placed between Verb and Obj1 (i.e. Verb+Obj2+Obj1), such as 入中国籍 *ru zhongguo ji* 'to naturalize China'.

In this work we will concentrate on LVCs.

3. Light Verbs in Chinese

In modern Chinese, Light verbs are generally defined as the semantically bleached verbs in the sense that the predicative content mainly comes from its taken complement (Zhu, 1985) while the light verb itself may only serve as a syntactic operator, without containing any eventive information. For example, for the construction 加以讨论 *jiayi taolun* 'to discuss', the predicative information all comes from the complement 讨论 *taolun* 'discuss' while the light verb 加以 *jiayi* 'inflict' adds no semantic to the LVC. The most typically used Chinese light verbs are 进行/加以/做/搞/从事 *jinxing/jiayi/zuo/gao/congshi*'proceed/inflict/do/do/engage'.

[1] http://universaldependencies.org/zh/overview/introduction.html

Chinese Light verbs pose a challenge in both traditional linguistics and computational linguistics because of its syntactic and semantic versatility and its unique distribution different from regular verbs with higher semantic content and selectional restrictions. (Butt, 2010) examines Chinese light verbs in the paper, but deals with directional complements and aspectual markers only, without mentioning any of the more typical usage of light verbs in Chinese as dealt with in literature on LVC in Chinese (e.g., (Zhu, 1985)). Moreover, the Propbank (Xue and Palmer, 2005) also treats light verbs as verb with zero argument. However, the issues is, these previous work cannot be directly transferred to the PARSEME framework, hence we further refer to (Lin et al., 2014). In (Lin et al., 2014), the authors specifically dealt with annotation of light verbs in Chinese, as well as automatic classification of different Chinese light verbs, therefore is the directly relevant resource we can make reference to.

4. Adjusting PARSEME guidelines to the database of Light Verbs

The annotation guidelines of the PARSEME shared task[2] define light verb constructions with several key characteristics. The first one is that LVCs are formed by a verb and a (single or compound) noun, which either directly depends on verb (and possibly contains a case marker or a postposition) (e.g., *give a lecture*), or is introduced by a preposition (e.g., *come into bloom*). The second one is the (single or compound) noun is predicative, often referring to an event (e.g. *to make a decision*) or a state (e.g. *to have fear*). The third characteristic is that the verb is "light", in the sense that it contributes to the meaning of the whole only by bearing morphology: person, number, tense, mood, as well as morphological aspect (perfective/imperfective) in some languages. It may be "light" either per se, or when used in the specific context of the noun.

Based on these three main characteristics, the annotation guideline further proposes five specific tests to help determine whether a MWE is LVC or not. In this section, we are going to discuss how these tests can be applied to the determination of Chinese light verbs. A list of Chinese light verbs has been compiled and collected from previous studies (Hu and Fan, 1995), (Diao, 2004), it contains 34 verbs in total. We also include the 5 light verbs mentioned in (Lin et al., 2014). Each of them is put through the tests to decide whether it is a light verb or not. By doing this, we can also examine to what extent the annotation guidelines apply in Chinese.

4.1. Test 1

Test 1 examines whether the taken complement of the light verb (i.e. the n) is a predicative noun or not. For example, *'pay a visit'*, the *visit* represents an event with two arguments (the visitor and the visitee) while the *cake* in *make a cake* represents a simple noun, without any eventive information. Therefore the former one can continue to the next test while the latter one is excluded. In terms of Chinese data, this standard can also be used to make the distinction. Among all the 34 verbs, 26 (e.g., 进行/加以/予以/开展 *jinxing/jiayi/yuyi/kaizhan* 'proceed/inflict/give/carry out') of them pass test 1 while the other 8 verbs (有/做/搞/干/处　置 *you/zuo/gao/gan/chuzhi* 'proceed/inflict/give/carry out') are ambiguous and need to rely on the context for further decision (i.e. they can take both eventive complements and simple nouns). For example, the light verb 进行 *jinxing* 'proceed' and 加以 *jiayi* 'inflict' can only take eventive noun (either deverbal noun as in 加以研究 *jiayi yanjiu* 'conduct research' or event noun as in 进行比赛 *jinxing bisai* 'to have a competition') under all contexts. While the light verb 做 *zuo* 'do' and 搞 *gao* 'do' can have both light verb (e.g., 做治理 *zuo zhili* 'provide governance'/搞竞赛 *gao jingsai* 'to have competition') and non-light verb usage (e.g., 做蛋糕 *zuo dangao* 'make a cake'/搞形式主义 *gao xingshizhuyi* 'do formalism').

4.2. Test 2

Test 2 is to investigate whether the noun is used in one of its original sense. Examples like 'pay a visit' can pass the test since the noun is literally understood while 'kitten' in 'have kittens' is not used in one of its normal senses. With respect to Chinese light verbs, all of the 34 verbs in our wordlist can pass the test.

4.3. Test 3

Test 3 is to check whether a light verb bears morphology (tense, mood etc.) and adds no semantic that is not already present in a noun, other than pointing to which semantic role is played by verb's subject with respect to noun's predicate. For constructions like 'take a walk', 'make a decision' and 'perform a check', the light verb 'take/make/perform' add no meaning to the whole construction, while 'start' in 'to start a walk' does add an aspectual meaning to the noun. For Chinese light verb, we use a simple test to examine this property. That is, if the light verb can be omitted without changing the proposition/semantic meaning of the construction, we consider the verb itself adds no semantic and is semantically bleached. For 10 out of 34 verbs, this light semantics of the verb is usual (i.e. the verb is used as a pure syntactic operator under different contexts, like 进行 *jinxing* 'proceed', 加以 *jiayi* 'inflict', 开展 *kaizhan* 'carry out', 作出 *zuochu* 'make'). For example, the function of 进行/加以研究 *jinxing/jiayi yan jiu* 'conduct research' is the same as 研究 *yanjiu* 'research'.

For another 10 verbs, they fail in this test in the sense that they do contribute to the construction and cannot be omitted. For example, 禁得起/禁不起诱惑 *jindeqi/jinbuqu youhuo* 'be/not be able to stand temptation' is certainly different from 诱惑 *youhuo* 'temp'. 遭到批评 *zaodao piping* 'be criticized' is also different from 批评 *piping* 'criticize' in meaning, in the sense that the former constructions contains the passive reading. However, for the other 14 verbs, this light verb semantics happens in the context of the particular noun. For example, constructions like 做点心 *zuo dianxin* 'make dessert' and 对环境做整治 *dui huanjing zuo zhengzhi* 'to renovate the environment' can

[2] http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.0/?page=040_Annotation_process_-_decision_tree

both be found in corpus. In the construction 做整治 *zuo zhengzhi* 'to renovate', 整治 *zhengzhi* 'renovate' as a deverbal noun contains the eventive information. 做 *zuo* 'do' in this case can be omitted without change proposition of the construction (e.g., 对环境(做) 整治 *dui huanjing (zuo) zhengzhi* 'environment'), although sometimes the whole construction/sentence needs to be re-written to ensure the grammaticality. In contrast, 点心 *dianxin* 'dessert' is a common NP which refers to a concrete entity and does not contain any predicative information. Therefore the eventive information all comes from the verb 做 *zuo* 'do'. In this case, 做 *zuo* 'do' in 做点心 *zuo dianxin* 'make dessert' may represent a series of actions including 'stir' 'blend' 'knead' 'bake' and so on. Hence it cannot be omitted in the construction.

4.4. Test 4

Test 4 is a syntactic test, aiming at testing whether a NP in which verb's subject becomes a noun's dependent evoke the same event. For example, 'Paul had a walk' and 'Paul's walk' both refer to the same walking event, while 'Paul made a good impression' and 'Paul's impression on his wife' refer to different semantics. In terms of Chinese data, the remaining 24 verbs all pass this test. For example, 政府进行改革 *zhengfu jinxing gaige* 'government carry out reform' and 政府的改革 *zhengfu de gaige* 'government's reform' refer to the same 'reform' activity.

4.5. Test 5

Test 5 is focused on the noun's prohibited argument, aiming at examining whether the (single or compound) noun, in the presence of a verb, prohibit at least one syntactic argument a which it normally licensed in the absence of a verb (except when a is in the whole–part relation with verb's subject). In English, for example, *The Queen paid a visit to the Prime Minister* + *a visit of the Lady to the Prime Minister* →**The Queen paid a visit of the Lady to the Prime Minister*. In other words, the visitor cannot be a modifier of visit. To be specific, the noun *visit* takes two semantic arguments, the visitor and the visited entity, as in 'the visit of the Queen to the Prime Minister'. When used in to *pay a visit*, the semantic argument, *visitor*, is realized as the subject of *to pay* (*The Queen paid a visit to the Prime Minister*), and cannot be realized at the same time within the NP headed by *visit (*The Queen paid a visit of the Lady to the Prime Minister)*.In contrast, *Paul transmitted the advice to his sister* + *Peter's advice* →*Paul transmitted Peter's advice to his sister*. The advice can be complemented by its author. Therefore this one is excluded from light verb construction. For Chinese light verbs, all the remaining 24 verbs can pass the tests. The semantic argument of n cannot be realized as its syntactic dependent, since it is already realized as verb's syntactic dependent instead (usually verb's subject). For example, 上海市进行对税收制度的改革 *shanghai shi jinxing dui shuishouzhidu de gaige* Shanghai proceed for tax system DE reform 'Shanghai carry out reform on tax system' + 财政局的改革 *caizhengju de gaige* Bureau of Finance DE reform Bureau of Finance's reform = *上海市进行对税收制度的财政局的改革* *shanghaishi jinxing dui shuishouzhidu de caizhengju de gaige*Shanghai proceed for tax system DE Bureau of Finance DE reform. The reformer cannot be a modifier of the 改革 *gaige* 'reform'.

It can be summarized as the annotation guidelines are efficient and effective in underlying Chinese data. In terms of the determination of Chinese light verbs, only a small part of the verbs can be considered as light verb per se, while most of them rely on the context. Only under certain context as well as with the co-occurrence of certain nouns, these verbs can be considered as light verb.

5. Tagging Light Verbs in a corpus

Provided the selection of potential candidate of light verbs described above, we use this list to pre-tag the UD Chinese corpus, and manually checked and corrected it afterwards. The UD Chinese corpus is in two parts, namely train set and development set. We finally got 836 hits of light verbs in the train set and 108 hits in the development set. Among them, the most frequent light verb is 有 *you* 'have', which gets 483 and 65 hits in the train set and development set respectively. The second and third frequent light verbs are 进行 *jinxing* 'proceed' and 做 *zuo* 'do'. On the other hand, there are 11 light verbs that didn't appear in the corpus. There are 8 light verbs only appear in the train set.

The pre-annotated data as described above were manually checked as follows:

- If the verb was not LVC, it was substituted by '_'
- If a verb was true LVC, the tag was left and the respective noun was assigned with a tag 'cont' - continuation.

In Table 1 below we summarize the changes to the file (comparing tags before the annotation and after)

corpus	lvc-auto	lvc-corrected	cont-inserted
train	836	176	184
dev	108	13	13

Table 1: Summary for a number of automatically assigned tags (lvc-auto), lvc tags left after the manual correction (lvc-corrected), and the number of nouns tagged as a part of LVC (cont-inserted)

The initial format of data will contain only four attributes: form, lemma, POS tag and an LVC tag. Following is an example of an annotated sentence:

```
此後    此後    NOUN    _
，      ，      PUNCT   _
廣東    廣東    PROPN   _
的      的      PART    _
動蕩    動蕩    ADJ     _
局面    局面    NOUN    _
得到    得到    VERB    lvc
基本    基本    NOUN    _
扼止    扼止    NOUN    cont
。      。      PUNCT   _
```

6. Conclusion and future work

In this paper, we demonstrated a pilot study on annotating light verbs in Chinese Universal Dependencies treebank adjusted to PARSEME annotation guidelines. First, a corpus was tagged with a list of potential light verbs, and then it was manually checked and the respective noun were tagged as noun components of LVCs. For the two sets of UD data – training and dev – 189 LVC instances were annotated in total in about 4000 sentence corpus. This corresponds to the LVC-per-sentence ratio of some other corpora annotated under PARSEME. Out future work will include revising the annotation of LVC and including more light verbs as well as other VMWE constructions. When the whole data is ready we can plan to train a classifier to automatically tag Chinese text with LVCs/VMWEs.

7. Bibliographical References

Bejček, E., Hajič, J., Straňák, P., and Urešová, Z. (2017). Extracting verbal multiword data from rich treebank annotation. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories (TLT 15)*, pages 13–24. Indiana University, Bloomington, Indiana University, Bloomington.

Butt, M. (2010). The light verb jungle: still hacking away. *Complex predicates in cross-linguistic perspective*, pages 48–78.

Diao, Y. B. (2004). Research on delexical verb in modern chinese. *Henan University Press*.

Hu, Y. S. and Fan, X. (1995). Research on verbs. *Dalian: Liaoning Normal University Press*.

Huang, C.-R., Chen, F.-Y., Chen, K.-J., Gao, Z.-m., and Chen, K.-Y. (2000). Sinica treebank: Design criteria, annotation guidelines, and on-line interface. In *Proceedings of the Second Workshop on Chinese Language Processing: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 12*, CLPW '00, pages 29–37, Stroudsburg, PA, USA. Association for Computational Linguistics.

Lin, J., Xu, H., JIANG, M., and Huang, C.-R. (2014). Annotation and classification of light verbs and light verb variations in mandarin chinese. In *Proceedings of Workshop on Lexical and Grammatical Resources for Language Processing*, pages 75–82, Dublin, Ireland, August. Association for Computational Linguistics and Dublin City University.

Savary, A., Ramisch, C., Cordeiro, S., Sangati, F., Vincze, V., QasemiZadeh, B., Candito, M., Cap, F., Giouli, V., Stoyanova, I., and Doucet, A. (2017). The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain.

Xue, N. and Palmer, M. (2005). Automatic semantic role labeling for chinese verbs. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 1160–1165, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhu, D. X. (1985). Dummy verbs and nv in modern chinese. *Journal of Peking University (Humanities and Social Sciences)*, pages 1–6.

8. Language Resource References

Nivre, Joakim and Agić, Željko and Ahrenberg, Lars and Aranzabe, Maria Jesus and Asahara, Masayuki and Atutxa, Aitziber and Ballesteros, Miguel and Bauer, John and Bengoetxea, Kepa and Bhat, Riyaz Ahmad and Bick, Eckhard and Bosco, Cristina and Bouma, Gosse and Bowman, Sam and Candito, Marie and Cebiroğlu Eryiğit, Gülşen and Celano, Giuseppe G. A. and Chalub, Fabricio and Choi, Jinho and Çöltekin, Çağrı and Connor, Miriam and Davidson, Elizabeth and de Marneffe, Marie-Catherine and de Paiva, Valeria and Diaz de Ilarraza, Arantza and Dobrovoljc, Kaja and Dozat, Timothy and Droganova, Kira and Dwivedi, Puneet and Eli, Marhaba and Erjavec, Tomaž and Farkas, Richárd and Foster, Jennifer and Freitas, Cláudia and Gajdošová, Katarína and Galbraith, Daniel and Garcia, Marcos and Ginter, Filip and Goenaga, Iakes and Gojenola, Koldo and Gökırmak, Memduh and Goldberg, Yoav and Gómez Guinovart, Xavier and Gonzáles Saavedra, Berta and Grioni, Matias and Grūzītis, Normunds and Guillaume, Bruno and Habash, Nizar and Hajič, Jan and Hà My, Linh and Haug, Dag and Hladká, Barbora and Hohle, Petter and Ion, Radu and Irimia, Elena and Johannsen, Anders and Jørgensen, Fredrik and Kaşıkara, Hüner and Kanayama, Hiroshi and Kanerva, Jenna and Kotsyba, Natalia and Krek, Simon and Laippala, Veronika and Lê Hong, Phuong and Lenci, Alessandro and Ljubešić, Nikola and Lyashevskaya, Olga and Lynn, Teresa and Makazhanov, Aibek and Manning, Christopher and Mărănduc, Cătălina and Mareček, David and Martínez Alonso, Héctor and Martins, André and Mašek, Jan and Matsumoto, Yuji and McDonald, Ryan and Missilä, Anna and Mititelu, Verginica and Miyao, Yusuke and Montemagni, Simonetta and More, Amir and Mori, Shunsuke and Moskalevskyi, Bohdan and Muischnek, Kadri and Mustafina, Nina and Müürisep, Kaili and Nguyen, Luong and Nguyen Thi Minh, Huyen and Nurmi, Hanna and Ojala, Stina and Osenova, Petya and Øvrelid, Lilja and Pascual, Elena and Passarotti, Marco and Perez, Cenel-Augusto and Perrier, Guy and Petrov, Slav and Piitulainen, Jussi and Plank, Barbara and Popel, Martin and Pretkalniņa, Lauma and Prokopidis, Prokopis and Puolakainen, Tiina and Pyysalo, Sampo and Rademaker, Alexandre and Ramasamy, Loganathan and Real, Livy and Rituma, Laura and Rosa, Rudolf and Saleh, Shadi and Sanguinetti, Manuela and Saulīte, Baiba and Schuster, Sebastian and Seddah, Djamé and Seeker, Wolfgang and Seraji, Mojgan and Shakurova, Lena and Shen, Mo and Sichinava, Dmitry and Silveira, Natalia and Simi, Maria and Simionescu, Radu and Simkó, Katalin and Šimková, Mária and Simov, Kiril and Smith, Aaron and Suhr, Alane and Sulubacak, Umut and Szántó, Zsolt and Taji, Dima and Tanaka, Takaaki and Tsarfaty, Reut and Tyers, Francis and Uematsu, Sumire and Uria, Larraitz and van Noord, Gertjan and Varga, Viktor and Vincze, Veronika and Washington, Jonathan North and Žabokrtský, Zdeněk and Zeldes, Amir and Zeman, Daniel and Zhu, Hanzhi. (2017). *Universal Dependencies 2.0*.

Savary, Agata and Ramisch, Carlos and Cordeiro, Silvio

Ricardo and Sangati, Federico and Vincze, Veronika and QasemiZadeh, Behrang and Candito, Marie and Cap, Fabienne and Giouli, Voula and Stoyanova, Ivelina and Doucet, Antoine and Adalı, Kübra and Barbu Mititelu, Verginica and Bejček, Eduard and El Maarouf, Ismail and Eryiğit, Gülşen and Galea, Luke and Ha-Cohen Kerner, Yaakov and Liebeskind, Chaya and Monti, Johanna and Parra Escartín, Carla and Kovalevskaitė, Jolanta and Krek, Simon and van der Plas, Lonneke and Aceta, Cristina and Aduriz, Itziar and Antoine, Jean-Yves and Attard, Greta and Azzopardi, Kirsty and Boizou, Loic and Bonnici, Janice and Boz, Mert and Bumbulienė, Ieva and Busuttil, Jael and Caruso, Valeria and Cherchi, Manuela and Constant, Matthieu and Czerepowicka, Monika and De Santis, Anna and Dimitrova, Tsvetana and Dinç, Tutkum and Elyovich, Hevi and Fabri, Ray and Farrugia, Alison and Findlay, Jamie and Fotopoulou, Aggeliki and Foufi, Vassiliki and Galea, Sara Anne and Gantar, Polona and Gatt, Albert and Gatt, Anabelle and Herrero, Carlos and Iñurrieta, Uxoa and Jagfeld, Glorianna and Hnátková, Milena and Ionescu, Mihaela and Klyueva, Natalia and Koeva, Svetla and Kovács, Viktória and Kuzman, Taja and Leseva, Svetlozara and Louisou, Sevi and Lynn, Teresa and Malka, Ruth and Martínez Alonso, Héctor and McCrae, John and de Medeiros Caseli, Helena and Miral, Ayşenur and Muscat, Amanda and Nivre, Joakim and Oakes, Michael and Onofrei, Mihaela and Parmentier, Yannick and Pasquer, Caroline and Pia di Buono, Maria and Priego Sanchez, Belem and Raffone, Annalisa and Ramisch, Renata and Rimkutė, Erika and Rizea, Monica-Mihaela and Simkó, Katalin and Spagnol, Michael and Stefanova, Valentina and Stymne, Sara and Sulubacak, Umut and Tabone, Nicole and Tanti, Marc and Todorova, Maria and Urešová, Zdenka and Villavicencio, Aline and Zilio, Leonardo. (2017). *Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.0)*.