Annotation and Classification of Light Verbs and Light Verb Variations in Mandarin Chinese

Jingxia Lin[1] **Hongzhi Xu**[2] **Menghan Jiang**[2] **Chu-Ren Huang**[2]

[1]Nanyang Technological University
[2]Department of CBS, The Hong Kong Polytechnic University
jingxialin@ntu.edu.sg, hongz.xu@gmail.com,
menghan.jiang@connect.polyu.hk, churen.huang@polyu.edu.hk

Abstract

Light verbs pose an a challenge in linguistics because of its syntactic and semantic versatility and its unique distribution different from regular verbs with higher semantic content and selectional resrictions. Due to its light grammatical content, earlier natural language processing studies typically put light verbs in a stop word list and ignore them. Recently, however, classification and identification of light verbs and light verb construction have become a focus of study in computational linguistics, especially in the context of multi-word expression, information retrieval, disambiguation, and parsing. Past linguistic and computational studies on light verbs had very different foci. Linguistic studies tend to focus on the status of light verbs and its various selectional constraints. While NLP studies have focused on light verbs in the context of either a multi-word expression (MWE) or a construction to be identified, classified, or translated, trying to overcome the apparent poverty of semantic content of light verbs. There has been nearly no work attempting to bridge these two lines of research. This paper takes this challenge by proposing a corpus-bases study which classifies and captures syntactic-semantic difference among all light verbs. In this study, we first incorporate results from past linguistic studies to create annotated light verb corpora with syntactic-semantics features. We next adopt a statistic method for automatic identification of light verbs based on this annotated corpora. Our results show that a language resource based methodology optimally incorporating linguistic information can resolve challenges posed by light verbs in NLP.

1 Introduction

Identification of Light Verb Construction (LVC) plays an important role and poses a special challenge in many Natural Language Processing (NLP) applications, e.g. information retrieval and machine translation. In addition to addressing issues related to LVC as a contributing factor to errors for various applications, a few computational linguistics studies have targeted LVC in English specifically (e.g., Tu and Roth, 2011; Nagy et al., 2013). To the best of our knowledge, however, there has been no computational linguistic study dealing with LVCs in Chinese specifically. It is important to know that, due to their lack of semantic content, light verbs can behave rather idiosyncratically in each language. Chinese LVC, in particular, has the characteristic that allows many different light verbs to share similar usage and be interchangeable in some context. We should also note that light verbs in Chinese can take both verbs, deverabal nouns, and eventive nouns, while the morphological status of these categories are typically unmarked, Hence, it is often difficult to differentiate a light verb from its non-light verb uses without careful analysis of the data.

It has been observed that some Chinese light verbs can be used interchangeably but will have different selectional restrictions in some (and generally more limited) contexts. For example, the five light verbs *congshi*, *gao*, *jiayi*, *jinxing*, *zuo* (these words originally meant 'engage', 'do', 'inflict', 'proceed', 'do' respectively) can all take *yanjiu* 'to do research' as their complement and form a LVC. However, only the light verbs *gao* and *jinxing* can take *bisai* 'to play games' as complements, whereas the other light verbs *congshi*, *jiayi*, and *zuo* cannot. Since light verbs are often interchangeable yet

Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing, pages 75–82,
Coling 2014, Dublin, Ireland, August 24 2014.

each also has its own selectional restrictions, it makes the identification of light verbs themselves both a challenging and necessary task. It is also observed that this kind of selectional versatility actually led to variations among different variants of Mandarin Chinese, such as Mainland and Taiwan. The versatility of Chinese light verbs makes the identification of LVCs more complicated than English.

Therefore, to study the differences among different light verbs and different variants of Chinese is important but challenging in both linguistic studies and computational applications. With annotated data from comparable corpora of Mainland and Taiwan Mandarin Chinese, this paper proposes both statistical and machine learning approaches to differentiate five most frequently used light verbs in both variants based on their syntactic and semantic features. The experimental results of our approach show that we can reliably differentiate different light verbs from each other in each variety of Mandarin Chinese.

There are several contributions in our work. Firstly, rather than focusing on only two light verbs *jiayi* and *jinxing* as in previous linguistic studies, we extended the study to more light verbs that are frequently used in Chinese. Actually, we will show that although *jiayi* and *jinxing* were often discussed in a pair in previous literature, the two are quite different from each other. Secondly, we show that statistical analysis and machine learning approaches are effective to identify the differences of light verbs and the variations demonstrated by the same light verb in different variants of Chinese. Thirdly, we provide a corpus that covers all typical uses of Chinese light verbs. Finally, the feature set we used in our study could be potentially used in the identification of Chinese LVCs in NLP applications.

This paper is organized as follows. Section 2 describes the data and annotation of the data. In Section 3, we conducted both statistical and machine learning methodologies to classify the five light verbs in both Mainland and Taiwan Mandarin. We discussed the implications and applications of our methodologies and the findings of our study in Section 4. Section 5 presents the conclusion and our future work.

2 Corpus Annotation

2.1 Data Collection

The data for this study is extracted from Annotated Chinese Gigaword corpus (Huang, 2009) which was collected and available from LDC and contains over 1.1 billion Chinese words, with 700 million characters from Taiwan Central News Agency and 400 million characters from Mainland Xinhua News Agency.

The light verbs to be studied are *congshi*, *gao*, *jiayi*, *jinxing*, *zuo*; these five are among the most frequently used light verbs in Chinese (Diao, 2004). 400 sentences are randomly selected for each light verb, half from the Mainland Gigaword subcorpus and the other from the Taiwan Gigaword subcorpus, which resulted in 2,000 sentences in total. The selection follows the principle that it could cover the different uses of each light verb.

2.2 Feature Annotation

Previous studies (Zhu, 1985; Zhou, 1987; Cai, 1982; Huang et al., 1995; Huang et al., 2013, among others) have proposed several syntactic and semantic features to identify the similarities and differences among light verbs, especially between the two most typical ones, i.e. *jinxing* (originally 'proceed') and *jiayi* (originally 'inflict'). For example, *jinxing* can take aspectual markers like *zhe* 'progressive marker', *le* 'aspect marker', and *guo* 'experiential aspect marker' while *jiayi* cannot (Zhou, 1987); *congshi* can take nominal phrases such as *disan chanye* 'the tertiary industry' as its complement while *jiayi* cannot. A few features are also found to be variant-specific; for example, Huang and Lin (2013) find that only the *congshi* in Taiwan, but not in Mainland Mandarin, can take informal and negative event complements like *xingjiaoyi* 'sexual trade'.

In our study, we selected 11 features which may help to differentiate different light verbs in each Mandarin variant as well as light verb variations among Mandarin variants, as in Table 1. All 2,000 examples collected for analysis were manually annotated based on the 11 features. The annotator is a trained expert on Chinese linguistics. Any ambiguous cases were discussed with another two experts in order to reach an agreement.

Feature ID	Explanation	Values (example)
1. OTHERLV	Whether a light verb co-occurs with another light verbs	Yes (***kaishi*** *jinxing taolun* Start proceed discuss 'start to discuss') No (*jinxing taolun* proceed discuss 'to discuss')
2. ASP	Whether a light verb is affixed with an aspectual marker (e.g., perfective *le*, durative *zhe*, experiential *guo*)	ASP.le (*jinxing-**le** zhandou* 'fighted') ASP.zhe (*jinxing-**zhe** zhandou* 'is fighting') ASP.guo (*jinxing-**guo** zhandou* 'fighted') ASP.none (*jinxing zhandou* 'fight')
3. EVECOMP	Event complement of a light verb is in subject position	Yes (***bisai** zai xuexiao jinxing* game at school proceed 'The game was held at the school') No (*zai xuexiao jinxing **bisai*** at school proceed game 'the game was held at the school')
4. POS	The part-of-speech of the complement taken by a light verb	Noun (*jinxing **zhanzheng*** proceed fight 'to fight') Verb (*jinxing **zhandou*** proceed fight 'to fight')
5. ARGSTR	The argument structure of the complement of a light verb, i.e. the number of arguments (subject and/or objects) that can be taken by the complement	One (*jinxing **zhandou*** proceed fight 'to fight') Two (*jinxing **piping*** proceed criticize 'to criticize') Zero (*jinxing **zhanzheng*** proceed fight 'to fight')
6. VOCOMP	Whether the complement of a light verb is in the V(erb)-O(bject) form	Yes (*jinxing **tou-piao*** proceed cast-ticket 'to vote') No (*jinxing **zhan-dou*** proceed fight-fight 'to fight')
7. DUREVT	Whether the event denoted by the complement of a light verb is durative	Yes (*jinxing **zhandou*** proceed fight-fight 'to fight') No (*jiayi **jujue*** inflict reject 'to reject')
8. FOREVT	Whether the event denoted by the complement of a light verb is formal or official	Yes (*jinxing **guoshi fangwen*** proceed state visit 'to pay a state visit') No (*zuo **xiao maimai*** do small business 'run a small business')
9. PSYEVT	Whether the event denoted by the complement of a light verb is mental or psychological activity	Yes (*jiayi **fanxing*** inflict retrospect 'to retrospect') No (*jiayi **diaocha*** inflict investigate 'to investigate')
10. INTEREVT	Whether the event denoted by the complement of a light verb involves interaction among participants	Yes (*jinxing **taolun*** proceed discuss 'to discuss') No (*jiayi **piping*** inflict criticize 'to criticize')
11. ACCOMPEVT	Whether the event denoted by the complement of a light verb is an accomplishment	Yes (*jinxing **jiejue*** proceed solve 'to solve') No (*jinxing **zhandou*** proceed fight-fight 'to fight')

Table 1: Features used to differentiate five Chinese light verbs.

3 Identification of light verbs based on annotated corpora

In this section, we adopted both statistical analysis and machine learning approaches to identify the five light verbs (*jiayi*, *jinxing*, *congshi*, *gao* and *zuo*) on the corpora with 2,000 annotated examples. The results of all approaches show that the five light verbs can be differentiated from each other in both Mainland and Taiwan Mandarin.

3.1 Identifying light verbs by statistical analysis

Both univariate analysis and multivariate analysis were used in our study for the identification. The tool we used is the Polytomous Package in R (Arppe, 2008).

3.1.1 Univariate analysis

Among the 11 independent features, one was found with only one level in both Mainland and Taiwan variants, i.e. all five light verbs in the two variants show the same preference over the features and thus excluded from the analysis. The feature is *OTHERLV* (all light verbs do not co-occur with another light verb in a sentence). Chi-squared tests were conducted for the significance of the co-occurrence of the remaining ten features with individual light verbs in both Mainland and Taiwan variants. The chisq.posthoc() function in the Polytoumous Package (Arppe, 2008) in R was used for the tests. The results are presented in Table 2, where the "+" and "-" signs indicate respectively a statistically significant overuse and underuse of a light verb with a feature, and "0" refers to a lack of statistical significance.

		Mainland Mandarin					Taiwan Mandarin				
Feature	N	*congshi*	*gao*	*jiayi*	*jinxing*	*zuo*	*congshi*	*gao*	*jiayi*	*jinxing*	*zuo*
POS.N	585	+	+	-	*0*	*0*	+	+	-	-	-
POS.V	1415	-	-	+	*0*	*0*	-	-	+	+	+
ARGSTR.one	376	*0*	-	-	*0*	+	+	-	-	+	*0*
ARGSTR.two	1039	-	*0*	+	*0*	-	-	-	+	-	+
ARGSTR.zero	585	+	+	-	*0*	*0*	+	+	-	-	-
VOCOMP.no	1939	0	0	*0*	*0*	0	0	0	+	-	0
VOCOMP.yes	61	0	0	*0*	*0*	0	0	0	-	+	0
EVECOMP.no	1919	+	-	+	-	-	+	*0*	+	-	*0*
EVECOMP.yes	81	-	+	-	+	+	-	*0*	-	+	*0*
ASP.guo	9	0	0	0	0	0	0	0	0	0	0
ASP.le	155	-	-	-	+	+	-	-	-	-	+
ASP.no	1835	+	+	+	-	-	+	+	+	+	-
ASP.zhe	1	0	0	0	+	0					
DUREVT.no	35	-	0	+	-	-	*0*	0	+	*0*	*0*
DUREVT.yes	1965	+	0	-	+	+	*0*	0	-	*0*	*0*
FOREVT.no	66	*0*	*0*	-	0	+	+	-	-	0	*0*
FOREVT.yes	1934	*0*	*0*	+	0	-	-	+	+	0	*0*
PSYEVT.no	1981	0	0	-	0	*0*	0	0	*0*	0	-
PSYEVT.yes	19	0	0	+	0	*0*	0	0	*0*	0	+
INTEREVT.no	1870	+	*0*	+	-	+	+	+	*0*	-	*0*
INTEREVT.yes	130	-	*0*	-	+	-	-	-	*0*	+	*0*
ACCOMPEVT.no	1904	+	+	-	+	+	+	+	-	+	*0*
ACCOMPEVT.yes	96	-	-	+	-	-	-	-	+	-	*0*

Table 2: Identifying light verbs in Mainland and Taiwan Mandarin via univariate analysis.

Table 2 suggests that in both Mainland and Taiwan Mandarin, each light verb shows significant preference for certain features, and thus can be distinguished from each other. For example, in Mainland Mandarin, although both *congshi* and *gao* show significant preference for the features *POS.N* and *ACCOMPEVT.no*, *congshi* differs from *gao* in that it also significantly prefers *DUREVT.yes* (taking complements denoting durative events, e.g., *yanjiu* 'to research'), *EVECOMP.no* (event complements do not occur in subject position), and *INTEREVT.no* (not taking complements denoting events involving interaction among participants, e.g., *taolun* 'to discuss'), whereas *gao* shows either a dispreference or no significant preference over these features. Take *gao* and *zuo* in Taiwan Mandarin as another example. While both light verbs literally means 'to do', there is no single feature preferred by both: *gao* prefers *POS.N*, *ARGSTR.zero*, *FOREVT.yes*, *INTEREVT.no*, *ACCOMPEVT.no*, whereas *zuo* shows significant preferences for *POS.V*, *ARGSTR.two*, *ASP.le*, and *PSYEVT.yes*.

3.1.2 Multivariate analysis

As shown in Table 2, in both Mainland and Taiwan Mandarin, some of the five light verbs share some features, which thus explains why sometimes they can be interchangeably used. This also indicates (a) that a particular feature is unlikely to be preferred by only one light verb and thus differentiates the verb from the others; (b) a certain context may allow the occurrence of more than one light verb. In

this sense, a multivariate analysis was adopted to better classify the five light verbs in each variant. The multivariate analysis used in the current study is polytomous logistic regression (Arppe, 2008), and the tool we used is the Polytomous() function in the Polytoumous Package (Arppe, 2008) in R.

The results from the multivariate analysis were summarized in Table 3. The numbers shown in the table are the odds for the features in favor of or against the occurrence of each light verb: when the estimated odd is larger than 1, the chance of the occurrence of a light verb is significantly increased by the feature, e.g., the chance of Mainland *jiayi* occurring is significantly increased by *ARGSTRtwo* (76.47:1), followed by *ACCOMPEVTyes* (56:1), *VOCOMPyes* (23.54: 1), and *PSYEVTyes* (19.87: 1). When the estimated odd is smaller than 1, the chance of the occurrence of a light verb is significantly decreased by the feature, e.g., the chance of Mainland *jinxing* occurring is significantly decreased by *ACCOMPEVTyes* (0.1849: 1); in addition, "inf" and "1/inf" refer to odds larger than 10,000 and smaller than 1/10,000 respectively, whereas non-significant odds (p-value < 0.05) are given in parentheses.

	Mainland Mandarin					Taiwan Mandarin				
	congshi	*gao*	*jiayi*	*jinxing*	*zuo*	*congshi*	*gao*	*jiayi*	*jinxing*	*zuo*
(Intercept)	(1/Inf)	*0.02271*	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)	(1/Inf)
ACCOMPEVTyes	(1/Inf)	*0.09863*	***56.25***	*0.1849*	(1/Inf)	(0.3419)	(1/Inf)	***11.33***	(0.1607)	*0.2272*
ARGSTRtwo	*0.2652*	***2.895***	***76.47***	(1.481)	*0.2177*	*0.1283*	(0.7613)	(Inf)	(0.7062)	(1.217)
ARGSTRzero	(1.097)	***3.584***	(1/Inf)	(1.179)	*0.245*	(0.6219)	***7.228***	(4.396)	*0.5393*	*0.2068*
ASPle	(0.7487)	(0.1767)	(0.8257)	(0.9196)	(1.853)	(1/Inf)	(1/Inf)	(0.3027)	(Inf)	***32.98***
ASPno	(Inf)	(1.499)	(Inf)	(0.2307)	(0.2389)	(0.9273)	(0.6967)	(Inf)	(Inf)	(0.2385)
ASPzhe	(1.603)	(1/Inf)	(0.4571)	(Inf)	(1/Inf)					
DUREVTyes	(Inf)	(2.958)	(1/Inf)	(Inf)	(Inf)	(Inf)	(Inf)	(1/Inf)	(0.9575)	(Inf)
EVECOMPyes	(1/Inf)	(1.726)	(1/Inf)	***3.975***	(1.772)	(1/Inf)	(0.8491)	(1/Inf)	***8.113***	(0.5019)
FOREVTyes	(2.744)	(1.227)	(Inf)	(0.7457)	*0.2679*	*0.0867*	(Inf)	(Inf)	(1.437)	(1.467)
INTEREVTyes	*0.03255*	(0.5281)	(0.5432)	***18.67***	*0.08902*	*0.1896*	(1/Inf)	(0.951)	***10.47***	(0.398)
PSYEVTyes	(1/Inf)	(1/Inf)	***19.87***	(1/Inf)	(0.9619)	(1/Inf)	(1/Inf)	(1.395)	(1/Inf)	(3.323)
VOCOMPyes	(0.1346)	(3.043)	***23.54***	(1.086)	(0.5344)	*0.18*	(2.35)	(Inf)	***3.161***	(0.5956)

Table 3: identifying light verbs in Mainland and Taiwan Mandarin via multivariate analysis.

As shown in Table 3, each of the light verbs in each Mandarin variant shows its favor and disfavor of certain features. Take Mainland Mandarin for example: although *congshi* has no feature significantly in its favor, but it is significantly disfavored by *ARGSTRtwo* (0.27:1) and *ITEREVTyes* (0.03:1); *gao* is disfavored by the aggregate of default variable values (0.02:1), and *ACCOMPEVTyes* (0.1:1), but is significantly favored by *ARGSTRtwo* and *ARGSTRzero*; the chance of *jiayi*'s ocucrrence is significantly increased by *ARGSTRtwo*(76.47:1), *ACCOMPEVTyes* (56.25:1), *VOCOMPyes* (23.54:1), and *PSYEVTyes* (19:87:1); *jinxing* has *INTEREVTyes* and *EVECOMPyes* in its favor, but *ACOMPEVTyes* in its disfavor; no feature is significantly in the favor of *zuo*, but this light verb is significantly disfavored by *ARGSTRtwo, ARGSTRzero, FOREVTyes* and *INTEREVTyes*.

The results in Table 3 also show that sometimes one key feature is able to identify two light verbs from each other, although not all five light verbs. Take Mainland Mandarin again for example. Most combinations of two light verbs from the five can be effectively differentiated by one feature. For instance, the feature *ARGSTRtwo* can differentiate *congshi/gao*, *congshi/jiayi*, *jiayi/zuo* and *gao/zuo*; the feature *INTEREVTyes* can differentiate *congshi/jinxing* and *jinxing/zuo*; the feature AC*COMPEVTyes* can differentiate the pairs *gao/jiayi* and *jinxing/jiayi*.

3.2 Identifying light verbs by classification

In this section, we resorted to machine learning technologies to study the same issue. Different classifiers were adopted to discriminate the five light verbs with the annotated corpora: ID3, Logistic Regression, Naïve Bayesian and SVM that are implemented in WEKA (Hall et al., 2009) and 10-fold cross validations were performed separately on the Taiwan and Mainland corpora.

The results were presented in Table 4. We can see that different classifiers provide similar results on both corpora, which means that the classification results are reliable and the features we annotated are effective in identifying the five light verbs. Overall, ID3 out-performs SVM slightly, with Logistic and NB not far behind. ID3 performs the best since the data is in low dimension. The detailed results including precision, recall and F-measure by ID3 on both corpora are shown in Table 5. The corresponding confusion matrixes are presented in Table 6. The confusion matrixes suggest two very important generalizations: (a) all five verbs can be classified with good confidence, and (b) the overall classification patterns of the Mainland and Taiwan Mandarin are very similar, which is consistent with the fact that Mainland and Taiwan Mandarin are two variants. However, we also observe that the confusion matrixes between various light verb pairs may differ between Mainland and Taiwan Chineses. This is the difference we would like to explore in the next section to propose a way to automatically predict these two variants. In addition, it is worth noting that all classifiers identify *jiayi* more effectively than other light verbs, which thus shows a potential different usage of *jiayi* from the others.

	ID3		Logistic		NB		SVM	
	TW	ML	TW	ML	TW	ML	TW	ML
jingxing	0.365	0.494	0.372	0.455	0.411	0.444	0.422	0.485
gao	0.612	0.391	0.609	0.364	0.598	0.377	0.575	0.354
zuo	0.571	0.566	0.568	0.582	0.525	0.576	0.574	0.561
jiayi	0.759	0.800	0.758	0.807	0.752	0.794	0.759	0.767
congshi	0.552	0.646	0.526	0.643	0.486	0.648	0.523	0.633
Average	0.574	0.585	0.567	0.576	0.555	0.573	0.571	0.565

Table 4: Result in F1-score of 10-fold cross validation of the classification of the five light verbs with different classifiers on the Taiwan (TW) and Mainland (ML) Corpora.

	Precision		Recall		F-Measure	
	TW	ML	TW	ML	TW	ML
jingxing	0.442	0.593	0.311	0.423	0.365	0.494
gao	0.681	0.449	0.557	0.347	0.612	0.391
zuo	0.610	0.570	0.537	0.562	0.571	0.566
jiayi	0.634	0.720	0.946	0.900	0.759	0.800
congshi	0.528	0.583	0.579	0.724	0.552	0.646
Average	0.580	0.586	0.588	0.599	0.574	0.585

Table 5: 10-fold cross validation result of ID3 algorithm on both corpora.

	jingxing		*gao*		*zuo*		*jiayi*		*congshi*	
	TW	ML	TW	ML	TW	ML	TW	ML	TW	ML
jingxing	**61**	**83**	15	27	36	40	38	11	46	35
gao	20	16	**113**	**70**	13	23	24	39	33	54
zuo	24	25	8	28	**108**	**118**	39	25	22	14
jiayi	5	11	0	6	5	6	**192**	**206**	1	0
congshi	28	5	30	25	15	20	10	5	**114**	**144**

Table 6: Confusion matrix of the classification with ID3 algorithm on both corpora.

3.3 Identifying light verbs by automatic clustering

We further used the clustering algorithm to test the differentiability of the five light verbs in both Mainland and Taiwan Mandarin. The results using the simple K-Means clustering algorithm on Taiwan and Mainland corpora are shown in Table 7. The results show that the light verb *jiayi* behaves

quite differently from the other four light verbs in both Mainland and Taiwan corpora, which is similar to the analysis based on statistical methods in Section 3.1 and classification methods in Section 3.2. In both corpora, *jiayi* has a narrower usage than the other light verbs. Meanwhile, we can also find a cluster which is mainly formed by instances of *jiayi* from the Mainland corpus (i.e. cluster 0). After closer examination of the examples in this cluster, we found that it mainly includes sentences where *jiayi* takes complements denoting accomplishment events, e.g. *gaizheng* 'to correct' and *jiejue* 'to solve'. However, *jiayi* in Taiwan corpus mainly takes complements denoting activity events, and thus almost all instances of Taiwan *jiayi* are mixed with those of the other light verbs. Meanwhile, our results show a tendency that all other light verbs (*jinxing*, *congshi*, *zuo*, and *gao*) mostly take activity complements but fewer accomplishment complements in both Taiwan and Mainland corpora. More discussion on the light verb variations between Mainland and Taiwan Mandarin can be found in (Huang et al., 2014).

	Mainland					Taiwan				
	0	1	2	3	4	0	1	2	3	4
jinxing	2	32	110	23	37	30	10	77	20	64
gao	2	33	116	41	11	120	23	30	0	31
zuo	0	36	80	14	81	19	4	47	5	132
jiayi	68	0	161	0	0	0	0	1	6	196
congshi	0	67	66	21	46	90	20	68	0	22

Table 7: Clustering results on Mainland and Taiwan corpora.

4 Applications and Implications

4.1 Implications for Future Studies

In the study above, we were able to annotate a corpus with all the types of significant context and, based on this annotated corpus, we were able to use statistic model to differentiate the use of different light verbs in different contexts. Such a module of generic linguistic tools can have several potentially very useful applications. First, in translation, LVC is one of the most difficult constructions as there is less grammatical or contextual information to make the correct translation. Our approach is especially promising. As we encode contextual selection information for all light verbs, the same approach can be applied to the other languages in the target-source pair to produce optimal pair. Second, in information extraction, selection of different light verbs often conveys subtle difference in meanings. Our ability to differentiate similar light verbs in the same context could have great potential in extracting the subtle information change/increase in the same context. Lastly, in second language learning as well as error detection, light verbs have been one of the most challenging ones. Our studies can be readily applied to either error detection or second language learning environment to provide the correct context where a certain light very is preferred over another.

4.2 From light verb variations to variants for the same language

One of the biggest challenges in computational processing of languages is probably to identify newly emergent variants, such as the cross-strait variations of Mandarin Chinese. For these two variants, the most commonly cited ones were on lexical differences. Systematic grammatical differences were much more difficult to study and hence rarely reported (comp. Huang et al., 2009). As these are two newly divergent variants, their main grammars are almost all identical, except for some subtle differences, such as the selection between different light verbs and their complements. Our preliminary results of univariate and multivariate analysis can be found in Table 2 and 3. It shows not only the similarities/differences among the light verbs in each variety (e.g., both ML and TW *congshi* and *gao* show preferences over *POS.N*, whereas both ML and TW *jiayi* show dispreference), but also the similarities/differences of the corresponding light verbs in Mainland and Taiwan Mandarin. For instance, *jinxing* in TW tends to take VO compounds as its complements e.g., *jinxing toupiao* "cast a vote",

which is consistent with the analysis in (Huang et al., 2013) (see more in Huang et al., 2014). But one thing should be pointed out is the difference is more between a significant and non-significant feature, rather than between a significant positive and significant negative feature.

5 Conclusion

In this paper, we addressed the issue of automatic classification of Chinese light verbs based on their usage distribution, based on an annotated corpus marking relevant contextual information for light verbs. We used both statistical methods and machine learning technologies to address this issue. It is found that our approaches are effective in identifying light verbs and their variations. The automatic generated semantic and syntactic features can also be used for future studies on other light verbs as well as other lexical categories. The result suggested that richly annotated language resources paired with appropriate tool can lead to effective general solution for some common issues faced by linguistics and natural language processing.

Acknowledgements

The work is supported by a General Research Fund (GRF) sponsored by the Research Grants Council (Project no. 543512) and NTU Grant NO. M4081117.100.500000.

Reference

Antti Arppe. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonymy. *Publications of the Department of General Linguistics*, University of Helsinki, volume 44.

Wenlan Cai. (1982). Issues on the complement of jinxing (“進行”帶賓問題). *Chinese Language Learning (漢語學習)* (3), 7-11.

Yanbin Diao. 2004. *Research on Delexical Verb in Modern Chinese (現代漢語虛義動詞研究)*. Dalian: Liaoning Normal University Press.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann and Ian H. Witten. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, 11(1):10-18.

Chu-ren Huang, Meili Yeh, and Li-ping Chang. 1995. *Two light verbs in Mandarin Chinese*. A corpus-based study of nominalization and verbal semantics. Proceedings of NACCL6, 1: 100-112.

Chu-Ren Huang. 2009. *Tagged Chinese Gigaword Version 2.0*. Philadelphia: Lexical Data Consortium, University of Pennsylvania. ISBN 1-58563-516-2

Chu-Ren Huang and Jingxia Lin. 2013. The ordering of Mandarin Chinese light verbs. In *Proceedings of the 13th Chinese Lexical Semantics Workshop*. D. Ji and G. Xiao (Eds.): CLSW 2012, LNAI 7717, pages 728-735. Heidelberg: Springer.

Chu-Ren Huang, Jingxia Lin, and Huarui Zhang. 2013. *World Chineses based on comparable corpus: The case of grammatical variations of jinxing.* 《澳门语言文化研究》, pages 397-414.

Chu-Ren Huang, Jingxia Lin, Menghan Jiang and Hongzhi Xu. 2014. Corpus-based Study and Identification of Mandarin Chinese Light Verb Variations. *COLING Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. Dublin, August 23.

István Nagy, Veronika Vincze, and Richárd Farkas. 2013. Full-coverage Identification of English Light Verb Constructions. *In Proceedings of the International Joint Conference on Natural Language Processing*, pages 329-337.

Yuancheng Tu and Dan Roth. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*. Association for Computational Linguistics.

Gang Zhou. 1987. *Subdivision of Dummy Verbs (形式動詞的次分類)*. Chinese Language Learning (漢語學習), volume 1, pages 11-14.

Dexi Zhu. (1985). *Dummy Verbs and NV in Modern Chinese (現代書面漢語里的虛化動詞和名動詞)*. Journal of Peking University (Humanities and Social Sciences) (北京大學學報(哲學社會科學版)), volume 5, pages 1-6.