An Investigation of Chinese Selectional Preference Based on HowNet

Bin Li [1,2] Xiaohe Chen[2] Xuri Tang[2]
1. Department of Computer Science and Technology,
Nanjing University
2. School of Chinese Language and Literature,
Nanjing Normal University
Nanjing, China
libin.njnu@gmail.com, {chenxiaohe5209, xrtang}@126.com

***Abstract*—Selectional Preferences (SPs) in verb-object(VO) constructions have been widely used in NLP applications, such as WSD, metaphor comprehension etc. To estimate the number of verbs that have strong SPs, 38,119 VO types of 1,462 verbs are extracted from "Modern Chinese Cihai", tagged in HowNet sense inventory with automatic tagging algorithm, The statistics indicates that only about 50% verbs have strong SPs, while another 50% have to be well analyzed within the cognitive linguistic frames.**

Keywords- selectional preference; verb-object; semantic class; semantic feature

I. INTRODUCTION

Selectional preference(SP) is the semantic constrain on a predicate's arguments. Chomsky(1965) employed semantic features to describe SPs of verbs. For example, the verb "frighten" is always followed by nouns of animals, thus the feature [+animate] was used to describe the verb-object(VO) SP for "frighten". Selectional preference has been widely used in NLP applications. Resnik(1993) used WordNet to extract verb SPs from real texts for WSD. Mason(2004) used SPs for metaphor understanding. Jia & Yu(2008) extended its use for Chinese metaphor recognition, understanding and generation. Zapirain(2009) used it for semantic role labeling.

However, SP does not always exist. In both English and Chinese, many verbs can be followed by almost any kind of nouns, e.g. 看(see), 爱(love). So there are at least 2 kinds of SPs, strong and weak.

Strong SP: 吓唬(frighten)+animate, 吃(eat)+food

Weak SP: 看(see)+any thing, 爱(love)+any thing

Strong SPs have strict semantic restrictions on nouns. They can be easily used in metaphor detection, understanding, etc. On the contrary, weak SPs are not as useful as strong SPs. In this paper, we want to make it clear how many verbs in English/Chinese do have strong SPs on their objects.

For English, we conducted a simple statistical analysis on VerbNet[1]. VerbNet has annotated SPs of 5257 verbs under a framework of 23 semantic roles, 37 semantic classes. Statistical results show that 3 semantic classes take up about 50% of all the SPs. The 3 classes are animate, organization and concrete, which are at high level of the semantic hierarchy. The majority of verbs take weak SPs.

For Chinese, Wu et al.(2005) investigated 46 verbs' SPs in verb-object constructions in news corpus using the noun taxonomy of HowNet. The paper argued that only few of the verbs have strong SPs.

Previous works have shown that strong SP is of minority, but the concept of SP they used is the semantic class of nouns, not the semantic feature introduced by Chomsky. Thus, we want to make an investigation on Chinese SPs in VOs on the basis of more instances and a thesaurus describing semantic class and features of nouns.

The paper is organized as follows. Section 2 introduces VO data resources. Section 3 describes the semantic tagging algorithm for SPs. Section 4 shows the statistical results of SPs. Section 5 gives some explanation based on manual annotations within the cognitive linguistic theories. Section 6 makes a brief conclusion.

II. RESOURCES AND DATASET

HowNet [2] is a structured bilingual(Chinese/English) semantic resource. Quite different from WordNet, it defines a word by a set of structured semantic features, called "sememe". In HowNet(version 2007), there are about 2200 sememes, which are used to define 91000 Chinese words and 85000 English words. For example, the word 学生(student) is defined as follows in HowNet.

学生 (student) {human| 人 :{study| 学习 :agent={~},location={InstitutePlace|场所:domain={education|教育},{study|学习:location={~}},{teach|教:location={~}}}}

The definition explains the word's semantic class, related events and domains, all using sememes. But the sememes are not isolated from each other, they are structured with different kinds of relations. Every sememe is located in a hierarchy tree. They have their own definition and can inherit their father node's definition. Thus, the sememe "human|人" is not just a semantic label, but a semantic feature with full meaning and kinds of relation to other sememes(see below). In HowNet, a word's meaning is represented by many sememes.

1 http://verbs.colorado.edu/~mpalmer/projects/verbnet.html

2 http://www.keenage.com

{entity|实体}
├ **{thing|万物}** {entity|实体:{ExistAppear|存现:existent={~}}}
│ ├ **{physical|物质}** {thing|万物:HostOf={Appearance|外观},{perception|感知:content={~}}}
│ │ ├ **{animate|生物}** {physical|物质:HostOf={Age|年龄},{alive|活着:experiencer={~}}, {die|死:experiencer={~}},{metabolize|代谢:experiencer={~}},{reproduce|生殖:PatientProduct={~},agent={~}}}
│ │ │ ├ **{AnimalHuman|动物}** {animate|生物:HostOf={Sex|性别},{AlterLocation|变空间位置:agent={~}},{StateMental|精神状态:experiencer={~}}}
│ │ │ │ ├ {**human|人**} {AnimalHuman|动物:HostOf={Ability|能力}{Name|姓名}{Wisdom|智慧},{speak|说:agent={~}},{think|思考:agent={~}}}

"Modern Chinese Cihai" is a special dictionary giving mounts of collocations of a word[7]. 647,024 types of different kinds of grammatical collocations are automatically extracted from Cihai. For the purpose of statistical analysis, we only use verb-object collocations which have at least 10 nouns as their objects. Thus about 40,273 types of VO collocations are selected for the experiment. After trimming the nouns not defined in HowNet, the actual number of VO collocations extracted from Cihai is 38,119 VO collocations of 1,462 verbs.

III. NOUN SEMANTIC TAGGING ALGRITHM

Two algorithms were used to tag the nouns' semantic class and features.

A. Relative Entropy Algrithm

Resnik(1993) applied the relative entropy to compute the SP of VOs. The SP strength is defined as follows:

$$S_r(p) = D(\Pr(c \mid p) \,\|\, \Pr(c)) = \sum_c \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)} \quad (1)$$

$$A_r(p,c) = \frac{1}{S_r(p)} \Pr(c \mid p) \log \frac{\Pr(c \mid p)}{\Pr(c)} \quad (2)$$

where $Pr(c)$ is the prior probability of the noun class, $Pr(c|p)$ is the probability of class c given the predicate p.

Given the grammatical relationship r, $S_r(p)$ computes the SP strength of a predicate, $A_r(p,c)$ gets the association between predicate and semantic class. With Formula 2, we can give each object noun a semantic class $c_r(p,n)$.

$$c_r(p,n) = \max A_r(p,c_i) \qquad c_i \in class(n) \quad (3)$$

B. Naive Tagging Algorithm

HowNet defines a word's meaning by sememes. It's not hard to modify the formula by changing semantic classes with sememes. But it is time-consuming to use relative entropy because Cihai has no frequency information of VO. We designed a much simple algorithm to get the SPs.

A verb(v) has N object nouns, each noun n_i ($0<i\leqslant N$) has M sememes, and each sememe of the noun $S_v^j(n)$ ($0<j\leqslant M$) has its count in a $\{n_i|v\ ,\ 0<i\leqslant N\}$ collocation set. Then the sememe of n_i in the context of v-n_i is the sememe having max count in the collocation set.

$$\tilde{S}_v(n_i) = \max Count(S_v^j(n)) \quad (4)$$

C. Algorithm Comparison

We compared the 2 algorithms on 500 randomly selected VOs. The 2 algorithms got nearly the same results, the accuracies are about 95%, but both have few obvious errors which are caused by polysemous nouns, as each sememe of the noun occurs only once in a $\{n_i|v\ ,\ 0<i\leqslant N\}$ collocation set.

IV. STATISTICALANALYSIS

38,119 VO types of 1,462 verbs are automatically tagged by Naive Tagging Algorithm. We conducted 2 experiments using the first sememe and all sememes in the definition in HowNet. The results came to similar conclusion that the strong SP is of minority.

A. Semantic Class as SP

As introduced in section 2, the first sememe "human|人" of definition in HowNet is the semantic class of the word 学生(student). We use the first sememe to get the nouns' classes by Naive Tagging Algorithm. Figure I gives the distribution of object noun classes. The number of their object nouns varies from 2 to 67, with the average number 13.2. Most verbs have more than 5 noun classes. Most verbs' SP seems to be weak. The verbs whose nouns' most frequent 3 classes take up 50% of all classes can be divided into 2 types.

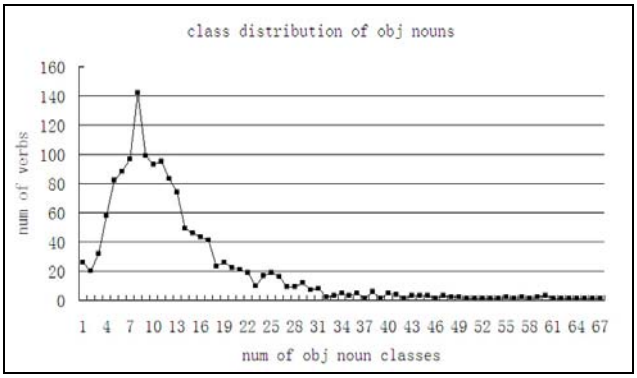

Figure 1. Distribution of object noun classes.

1) Nouns in 1 class. 26 verbs' object nouns are in 1 class. The verbs are "安顿、安慰、扮演、绑架、哺育、逮捕、吩咐、勾引、教育、虐待、聘、聘任、请、劝告、任用、杀害、提拔、推举、孝顺、熏染、熏陶、训斥、押解、优待、招待、招收". Their object nouns have the same sememe "human|人".

2) Nouns' top 3 classes take up more than 50% of all classes. Table I shows the top 3 classes's coverage. There are 122 verbs whose most frequent class(top 1) is over 50%. Among these verbs, 98 are monosemies like "哀求(beg)", and 24 are polysemies like "参加(attend)". When top3 classes are considered, there are 647 verbs. Totally it's about 44% of the verbs have strong SPs on their object nouns, including monosemies and polysemies.

TABLE I. OBJECT NOUN CLASSES TOP 3

⩾50%	Top1	Top2	Top3	SUM
# of verbs(monosemy)	98	170	194	462
# of verbs(polysemy)	24	64	97	185
SUM	122	234	291	647

Different semantic classes may have the same father node in the semantic tree, thus we compute the top3 classes' common father node in HowNet. We define a top root sememe "ALL" for all sememes. Table II shows the combined classes of nouns. Except for 3 most frequent classes "ALL, entity|实体, thing|万物" , there are only 224 verbs left.

TABLE II. FATHER NODE CLASS OF TOP 3 CLASSES

Common Class of Top3 Classes	# of Verbs	# of Poly Verbs	Example Verbs
ALL	371	96	安顿
├entity|实体	472	187	暗示
│ ├thing|万物	395	91	爱好
│ │ ├physical|物质	68	23	铲除
│ │ │ ├animate|生物	2	1	看护
│ │ │ │ ├AnimalHuman|动物	7	2	豢养
│ │ │ │ │ ├human|人	28	3	央告
│ │ │ ├inanimate|无生物	19	7	打扫
│ │ │ │ ├NaturalThing|天然物	1	1	遨游
│ │ │ │ │ ├earth|大地	2	0	遥望
│ │ │ │ ├artifact|人工物	21	4	搬运
│ │ │ │ │ ├edible|食物	1	0	品尝
│ │ │ │ │ ├wealth|钱财	2	0	核算
│ │ │ │ │ ├implement|器具	4	1	安置
│ │ │ │ │ ├building|建筑物	1	0	修筑
│ │ ├mental|精神	43	7	废除
│ │ │ ├information|信息	4	0	创作
│ │ │ ├regulation|规矩	2	0	违反
│ │ │ ├thinking|思想	1	0	应用
│ ├event|事件	0	0	
│ │ ├fact|事情	5	1	避免
├Attribute|属性	10	1	保持
│ ├Property|特性	3	1	发挥
SUM	1462	426	

B. *Semantic Feature as SP*

In this part, all the sememes in the definition of nouns in HowNet are used to compute the SP of VOs. There are 28 verbs whose nouns' sememes are of one sememe. The 2 more verbs than in section A.1 are "央告" and "召唤". Table III gives the similar results to table I, while the former has many more verbs whose top3 sememes covering more than 50% of all nouns. It seems semantic features are more useful than semantic classes.

TABLE III. OBJECT NOUN SEMEMES TOP3

⩾50%	Top1	Top2	Top3	SUM
# of verbs(monosemy)	198	306	277	781
# of verbs(polysemy)	46	117	138	301
SUM	244	423	415	1082

Table IV shows the coverage of the father node of the top3 sememes. Compared to Table II, the high level sememes take a bigger share among all sememes.

TABLE IV. FATHER NODE SEMEMES OF TOP3 SEMEMES

Common Class of Top3 Classes	# of Verbs	# of Poly Verbs
ALL	593	146
entity|实体	450	180
thing|万物	275	65
physical|物质	58	17
human|人	28	3
inanimate|无生物	14	4
mental|精神	12	3
artifact|人工物	11	3
AnimalHuman|动物	7	3
fact|事情	2	0
implement|器具	2	0
information|信息	2	0
NaturalThing|天然物	2	1
AlterKnowledge|变感知	1	0
animate|生物	1	1
building|建筑物	1	0
earth|大地	1	0
thinking|思想	1	0
wealth|钱财	1	0
SUM	1462	426

C. *Why So Many Weak SPs?*

From the above statistical results, it is obvious that SPs are not strong neither by semantic classes nor by features. This conclusion is close to the results from VerbNet and Wu et al.(2005). Wu took metonymy as one of the main causes of the weak SPs. But the VO collocations in Cihai do not have many metonymy usages. The answer needs more investigations.

V. LINGUISTIC EXPLAINATION

To explain the crucial question why strong SPs are so few, we manually annotated SPs of the VO collocations under the frame of cognitive linguistics. The annotation is just for observation. As the accuracy and qulity are very hard to control in semantic tagging, we could not give the entire statistical results but some important or good examples.

We tagged the senses of verbs by the definition in HowNet. For every sense of a verb, there are many object nouns. A SP was given in the form of semantic feature or frame elements. We borrowed some from other dictionaries when we could not find them in HowNet. From the observation, we took the causes of weak SPs as 3 types.

1) Verb-Object is too ambiguious. Every verb has semantic preferences on its arguments. These arguments are of many kinds, such as agent, patient, tool and source, etc. According to Fillmore(1977), a verb has an event frame, and

a frame has several frame elements(FE). Many FEs can serve as the object of the verb. For example, the verb 买(buy) has at least 5 FEs, "buyer, seller, goods, money, recipient". Among them, "seller, goods, recipient" can be the object of buy. Many FEs can serve as object, so the semantic SP become numerous. As a result, SPs may be extended to FEs.

2) Some FEs do not have strong SP. Take "goods" in the frame of "buy" event for example, too many things can be taken as goods as long as the speaker wants it. This kind of SP is almost weak.

3) Cognitive feature is hard to find in WordNet or HowNet. In many languages, "time" has an cognitive feature "precious", which can not be found in any dictionary, but in the mind of every member of a language community. The feature "precious" is subjective and dynamic. It is hard to discribe in the static definition of a word but exists in many speakers' mind. We call this kind of semantic features "cognitive features(CF)". Table V gives the mannual tagged SP of verbs to the noun 时间(time). The features are mostly taken from HowNet. Some SPs are traditional static semantic classes or features like "possession", some are frame elements like "resources", some are cognitive features like "precious".

TABLE V. SELECTIONAL PREFERENCES OF VERBS TO 时间(TIME)

Verbs	Selectional Preference
使用(use)	FE:tool
忘记(forget)	FE:things in memory
支配(govern)	FE:thing under control
追赶(chase)	FE:target
要(ask)、分配(assign)	FE:resources
享有(have)、丢(drop)、争(compete for)	FE:possession
买(buy)	FE:goods
找(find)	FE:goal
看(according to)	FE:condition
等(wait for)	FE:coming thing
珍惜(treasure)、浪费(waste)	CF:precious|珍
牺牲(sacrifice)、抓紧(grasp)	CF:important|重要
省(save)	CF:exhaust|消耗
推迟(put off)	CF:early-late
延长(lengthen)	CF: length|长短
取得、博得、赢得(gain)	CF: BeGood|良态

From the example of "time", we know that some weak SPs are caused by cognitive features. And the cognitive features may have more usages in NLP applications. For example, if we know that 珍惜(treasure)'s SP is "precious|珍", then the objects of it do have that feature. So we can easily get the nouns which are precious in one language.

VI. CONCLUSION AND FUTURE WORK

We made an investigation on Chinese verbs' SP on object nouns by HowNet. There are three important findings: (1) When semantic class or feature is used, most SPs of VOs are not very strong. (2)The weak SPs are caused in part by different frame elements which serve as objects. (3) Cognitive feature is a kind of strong SP, but it is not easy to get. As there are not many the verbs having strong SPs, we should be careful when applying SP in NLP tasks.

In the future, we have to find out the relation between the FE and the FE's noun, as there maybe another kind of SP. The most interesting thing we think is to design an automatic algorithm for the acquisition of cognitive features, which have potential usage in metaphor understandings and other applications.

ACKNOWLEDGMENT

This paper is supported in part by National Social Science Fund of China under contract 10CYY021, 07BYY050, National Natural Science Fund of China under contract 60773173, and Nanjing Normal University Fund under contract 1009007.

REFERENCES

[1] N. Chomsky, Aspects of the Theory of Syntax. MA: MIT Press, 1965.

[2] P. Resnik, "Selection and Information: A Class Based Approach to Lexical Relationships," University of Pennsylvania PhD. Thesis, 1993.

[3] Z. Mason, "CorMet: A Computational, Corpus-Based Conventional Metaphor Extraction System," Computational Linguistics, vol. 30(1), 2004, pp. 23-44.

[4] Yuxiang Jia, Shiwen Yu, "Unsupervised Chinese Verb Metaphor Recognition Based on Selectional Preferences," Proc. 22nd Pacific Asia Conference on Language, Information and Computation (PACLIC 22), Cebu City, Philippines, 2008, pp. 207-214.

[5] B. Zapirain, E. Agirre, and L. Mrquez, "Generalizing over Lexical Features: Selectional Preferences for Semantic Role Classification," Proc. Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing(ACL-IJCNLP09), Singapore, 2009, pp. 73-76.

[6] Yunfang Wu, Huiming Duan, Shiwen Yu, "A Study of Verb Selectional Restrictions on Objects", Applied Linguistics, China, vol. 2005(2), pp. 121-128.

[7] Ni Wenjie, Zhang Weiguo, Jixiaojun, "Modern Chinese Cihai," Beijing: People's China Publishing House, 1994.

[8] C. J. Fillmore, "Topics in Lexical Semantics". in Current Issues in LinguisticTheory, R. W. Cole, Ed. Bloomington: Indiana University Press, 1977, pp.76–138.