Computational Linguistics and Chinese Language Processing
Vol. 11, No. 2, June 2006, pp. 101-114

Two-Fold Filtering for Chinese Subcategorization Acquisition with Diathesis Alternations Used as Heuristic Information[1]

Xiwu Han*and Tiejun Zhao*

Abstract

Automatically acquired lexicons with subcategorization information have been shown to be accurate and useful for some purposes, but their accuracy still shows room for improvement and their usefulness in many applications remains to be investigated. This paper proposes a two-fold filtering method, which in experiments improved the performance of a Chinese acquisition system remarkably, with an increased precision rate of 76.94% and a recall rate of 83.83%, making the acquired lexicon much more practical for further manual proofreading and other NLP uses. And as far as we know, at the present time, these figures represent the best overall performance achieved in Chinese subcategorization acquisition and in similar researches focusing on other languages.

Keywords: Filter, Chinese, SCF, Diathesis Alternation

1. Introduction

Subcategorization is a process that classifies a syntactic category into its subsets. [Chomsky 1965] defined the function of strict subcategorization features as appointing a set of constraints that dominate the selection of verbs and other arguments in deep structure. Subcategorization of verbs, as well as categorization of all words in a language, is often implemented by means of functional distributions, which constitute different environments or distributional patterns accessible for a verb or word. Such a distribution or environment is called a subcategorization frame (SCF), and is usually combined with both syntactic and semantic information. Therefore, verb subcategorization involves much more information than verb classification, which usually only classifies verbs into groups. SCFs, on the other hand,

[1] This research is sponsored by the Natural Science Foundation of China (Grant No. 60373101 and 60375019), and High-Tech Research and Development Program (Grant No. 2002AA117010-09).
* School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
E-mail: {hxw, tjzhao}@mtlab.hit.edu.cn

[Received April 4, 2005; Revised June 10, 2005; Accepted June 19, 2005]

specify the category of the main anchor (verb hereby), the number of arguments, each argument's category and position with respect to the anchor, and other information, such as feature equations or node expansions[2].

Recently, large subcategorized verbal lexicons have been shown to be crucially important for many tasks in natural language processing, such as probabilistic parsing [Korhonen 2001] and verb classifications [Schulte im Walde 2002; Korhonen 2003]. Since Brent reported his findings [Brent 1993], a considerable amount of research has focused on large-scale automatic acquisition of subcategorization frames and achieved some success, not only in English but also in many other languages, including German [Schulte im Walde 2002], Spanish [Chrupala 2003], Czech [Sarkar and Zeman 2000], Portuguese [Gamallo *et al.* 2002], and Chinese [Han *et al.* 2004ab]. However, the relevant results are still far from sufficiently accurate and indicate that most of the existing methods are not yet practical.

This is especially true for the Chinese subcategorization acquisition system, which has achieved a precision rate of 60.6%±2.39% and a recall rate of 51.3%±2.45% [Han *et al*. 2004b]. Detailed analysis of the system and acquisition results shows that besides the imperfect hypothesis generator, there are sources of both linguistic and statistical errors. Linguistic errors mainly result from the Zipfian distributions of syntactic patterns, and statistical errors derive mostly from the inappropriate assumption of independence among SCFs that verbs enter. Hence, the statistical filter of maximum likelihood estimation (MLE) performs badly with respect to lower-frequency SCF hypotheses. In this paper, the independence assumption is eliminated on the basis of diathesis alternations reported by [Han 2004], and a two-fold filtering method is introduced, which first filters the hypotheses by means of a comparatively higher threshold and secondly, filters the left-out ones by means of a much lower threshold with diathesis alternatives of those accepted SCFs seeded as heuristic information.

Experimental evaluation of the acquisition results of 48 Chinese verbs showed that the acquisition performance was improved remarkably, with the precision rate increased to 76.94% and the recall rate to 83.83%, making the acquired lexicon much more practical for further manual proofreading and other NLP uses. Although cross-lingual comparison may lack concrete significance, at the present time, these figures represent the best overall performance achieved in both Chinese subcategorization acquisition and in similar researches focusing on other languages.

Section 2 introduces and analyzes the present Chinese SCF acquisition system and, in particular, its MLE filter. Section 3 briefly discusses the diathesis alternations used. Section 4 gives a complete description of our Two-fold filtering method. In section 5, the general

[2] See also the definition of SCF at http://www.cis.upenn.edu/~xtag/tech-report/node248.html.

performance of the modified system is evaluated on the basis experiments. Finally, section 6 discusses our achievements, weak points and possible focuses for future work.

2. Subcategorization Acquisition and MLE Filtering

In the system proposed by [Han *et al*. 2004b], there are, generally, 4 steps in the auto-acquisition process of Chinese subcategorization. First, the corpus is processed with a cascaded HMM parser; second, all possible local patterns for verbs are abstracted; third, the verb patterns are classified into SCF hypotheses according to the predefined set; fourth, the hypotheses are checked statistically with an MLE filter. The actual application program consists of 6 parts, described in the following paragraphs.

a. Segmenting and tagging: The raw corpus is segmented into words and tagged with POS's by the comprehensive segmenting and tagging processor developed by MTLAB of the Computer Department in the Harbin Institute of Technology. The advantage of the POS definition is that it describes some subsets of nouns and verbs in Chinese.

b. Parsing: The tagged sentences are parsed with a cascaded HMM parser[3], developed by MTLAB of HIT, but only intermediate portion of the parsing results is used, which means that only the syntactic skeletons make difference and, thus, that the negative effects of some errors in the deep structures can be avoided. The training set of the parser consists of 20,000 sentences from the Chinese Tree Bank[4] [Zhao 2002].

c. Error-driven correction: Some key errors occurring in the former two parts are corrected according to manually obtained error-driven rules, which generally concern words or POS in the corpus.

d. Pattern abstraction: Verbs with the largest governing ranges are regarded as predicates; then, local patterns, previous phrases and syntactic tags are abstracted and generalized as argument types (see Table 1), and isolated parts are combined, generalized or omitted according to basic phrase rules presented in [Zhao 2002].

e. Hypothesis generation: Based on linguistic restraining rules e.g., no more than two nominal phrases (NP) may occur in a series and no more than three in one pattern; and no positional phrase (PP), temporal complement (TP) or quantifier complement (MP) may occur with a nominal phrase before any predicate [Han *et al*. 2004a] (see also Table 2), the patterns are coordinated and classified into the predefined SCF groups.

---

[3] When evaluated on an auto-tagged open corpus, the parser's phrase precision rate was 62.3%, and the phrase recall rate was 60.9% [Meng 2003].

[4] A sample of the tree bank or relevant introduction could be found at http://mtlab.hit.edu.cn.

Table 1. Argument types for Chinese SCFs

Type	Definition
NP	Nominal phrase
VP	Verbal phrase
QP	Tendency verbal complement
BP	Resulting verbal complement
PP	Positional phrase
BAP	Phrase headed by "ba3" (把)
BIP	Phrase headed by "bei4" (被) or other characters with the passive sense
TP	Temporal complement
MP	Quantifier complement
JP	Adjective or adverb or "de" (得) headed complement
S	Clause or sentence

Table 2. Constraints placed on predicates and arguments

Predicate *v*		Only one *v* except in repeating positions with one *v* but two slots
Argument Types	NP	No more than two in a series and no more than three in one SCF
	VP, S	No serial occurrences
	QP, BP, JP	No serial occurrences and occurrence only after a *v*
	BAP, BIP	No more than one occurrence
	TP, PP	No co-occurrences with NP before a *v*
	MP	No serial occurrences nor occurrences in adjacency before NP

f. Hypothesis filtering: According to the statistical reliability of each type of SCF hypothesis and the linguistic principle that arguments occur more frequently with predicates than adjuncts do, the hypotheses are filtered by means of maximum likelihood estimation (MLE), which has been shown to work better than other methods, such as the binomial hypothesis test (BHT), log likelihood ratio (LLR), and T-test [Korhonen 2001; Han *et al*. 2004b].

Table 3. An example of auto-acquisition

No.	Actions	Results
(a)	Input	两个人在大伙儿的追问下证明了老人的身份。
(b)	Tag and parse	BNP[BMP[两/m 个/q]人/ng]在/p NDE[大伙儿/r 的/usde]BVP[追问/vg 下/vq]BVP[证明/vg 了/ut]NP[老人/nc 的/usde 身份/ng]。/wj
(c)	Correct errors	BNP[BMP[两/m 个/q]人/ng]在/p NDE[大伙儿/r 的/usde 追问/vg 下/vq]BVP[证明/vg 了/LE]NP[老人/nc 的/usde 身份/ng]。/wj
(d)	Abstract patterns	BNP PP BVP[vg LE] NP
(e)	Generate hypothesis	NP v NP ｛01000｝
(f)	Filter hypotheses	NP v NP {01111}[5]

Table 3 shows an example of Chinese SCF acquisition performed using the proposed system. When SCF information is acquired for the verb "zheng4ming2 证明" (prove), a related sentence in the corpus is (a), our tagger and parser returns (b), and error-driven correction returns (c) with NDE errors and with the first BVP corrected[6]. Since the governing range of "证明" is larger than that of the verb "zhui1wen4 追问" (ask), the other verb in this sentence, the program abstracts its local pattern BVP[vg LE] and previous phrase BNP, generalizes BNP and NDE as NP, combines the second NP with the isolated part "在/p" in PP, and returns (d). Then, the hypothesis generator returns (e) as the possible SCF in which the verb may occur. Actually, in the corpus, 621 hypothesis tokens are generated, and among them, 92 ones are of same argument structures with (e); and thus, (e) can pass the MLE hypothesis test, so we obtain one SCF for "zheng4ming2 证明" as (f).

Due to noises that accumulate during segmenting, tagging, and parsing of the corpus, even though error-driven correction is implemented, the hypothesis generator does not perform as efficiently as hoped. Experimental results show that its imperfect performance accounts for about 12% of the falsely accepted SCFs and 15% of the unrecalled ones. However, detailed analysis of a considerable amount of data indicates that a larger source of

[5] {01000} projects to the Chinese syntactic morphemes {"zhe0 着", "le0 了", "guo4 过", "mei2 没", "bu4 不"}, where 1 means that the SCF may occur with the respective morpheme, while 0 means that it may not [Han *et al.* 2004a].

[6] Note that not all of the errors in this example have been corrected, but this does not affect further procession. Also, NDE refers to phrases ending with "de4 的", BVP to basic verbal phrases [Zhao 2002], and LE to the Chinese syntactic morpheme "le0 了" [Han *et al.* 2004a].

errors is the MLE filter.

The MLE method is closely related to the general distributional situation of the corpus. First, from the applied corpus a training set is drawn randomly; it must be large enough to ensure a similar SCF frequency distribution. Then, the frequency of a subcategorization frame scf_i occurring with a verb v is recorded and used to estimate the possible probability $p(scf_i|v)$. Thirdly, an empirical threshold is determined, which ensures that a maximum value of the F measure will result for the training set. Finally, the threshold is used to filter out those SCF hypotheses with lower frequencies from the total set. Therefore, the statistical foundation of this filtering method is the assumption of independence among the SCFs that a verb enters, which can be probabilistically expressed in two formulas as follows:

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) = 0, \tag{1}$$

$$\sum_{i=1}^{n} p(scf_i \mid v) = 1. \tag{2}$$

In actual application, the probability $p(scf_i|v)$ is estimated from the observed frequency, and the conditional probability $p(scf_i|scf_j, v)$ is assumed to be zero. However, this assumption can sometimes be far from appropriate.

3. Diathesis Alternations

Much linguistic research focusing on child language acquisition has revealed that many children are able to create grammatical sentences previously unseen by them according to what they have learned, which implies that the widely-used independence assumption in the field of NLP may not be very appropriate, at least for syntactic patterns. If this assumption is removed, a possible heuristic could be the information of diathesis alternations, which is also another convincing anti-proof. Diathesis alternations are generally regarded as alternative ways, in which verbs express their arguments. Examples are as follows:

a. He broke the glass.
b. The glass broke.
c. Ta1 chi1 le0 pin2guo3.
   (他 吃 了 苹果。)
d. Ta1 ba3 pin2guo3 chi1 le0[7].
   (他 把 苹果 吃 了。)

[7] Sentences c and d generally mean *He ate an apple*.

Here, the English verb *break* takes the causative-inchoative alternation as shown in sentences a and b, while sentences c and d indicate that the Chinese verb *chi1* (吃, eat) may enter the *ba*-object-raising alternation where the object is shifted forward by the syntactic morpheme *ba3* (把) to the location between the subject and the predicate, as illustrated in Figure 1.

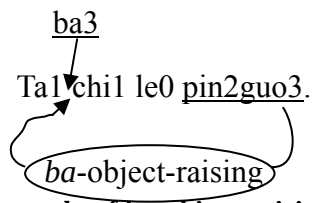

Figure 1. An example of ba-object-raising alternation

Therefore, we can conclude that for subcategorization acquisition, the independence assumption supporting the MLE filter is not as appropriate as previously thought. For a given verb, the assumption holds if and only if there is no diathesis alternation among all the SCFs it enters, and formulas (1) and (2) in Section 2 are efficient enough to serve as a foundation for an MLE method. Otherwise, if there are diathesis alternations among some of the SCFs that a verb enters, then formulas (1) and (2) must be modified as illustrated in formulas (3) and (4). In either case, for the sake of convenience, it is even better to combine the formulas as shown in (5) and (6).

$$\exists i, \exists j, i \neq j, p(scf_i \mid scf_j, v) > 0\,, \tag{3}$$

$$\sum_{i=1}^{n} p(scf_i \mid v) > 1\,, \tag{4}$$

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) \geq 0\,, \tag{5}$$

$$\sum_{i=1}^{n} p(scf_i \mid v) \geq 1\,. \tag{6}$$

For English verbs, much research has focused on diathesis alternation and relative applications [Levin 1993; Korhonen 1998; McCarthy 2001], whereas for Chinese verbs, only a comprehensive set of 82 diathesis alternations that seem suitable for NLP tasks has been reported [Han 2004]. Han's diathesis alternations are defined on the basis of verb subcategorization for Chinese described in [Han *et al.* 2004b]; among them, the arguments and SCFs are briefly defined in Table 1 and Table 2[8] in Section 2. Table 1 gives the definitions of argument types in Chinese SCFs, and Table 3 lists some constraints placed on both predicate verbs and their arguments.

[8] Detailed descriptions of the SCFs and their arguments can be found in [Han *et al.* 2004a].

From a corpus of 42,000 Chinese sentences automatically tagged with such SCFs, Han's alternation information was acquired via a combined approach, which makes use of linguistic knowledge and statistical methods. First, a set of candidates was generated according to the semantic and syntactic similarities between each pair of related sentences with the same predicate verb. Then, the candidates were checked by means of a frequency-based MLE filter. Finally, 67 SCF alternatives were automatically acquired, and 15 complemented, resulting in a statistically and linguistically reliable syntactic alternation set, a part of which is shown in Table 4.

Table 4. Some examples of Chinese diathesis alternations

scf_i	⟷	scf_j
NP BAP V	⟷	NP BAP V BP
NP NP V VP	⟷	NP V VP
NP V MP VP	⟷	NP V VP
NP BAP V VP	⟷	NP NP V VP
NP BIP V JP	⟷	NP BIP V NP
NP BIP V JP	⟷	NP BIP V QP
NP BIP V JP	⟷	NP V JP MP
NP BIP V MP	⟷	NP BIP V NP
NP BIP V MP	⟷	NP BIP V QP
NP V NP	⟷	NP NP V
NP V JP NP	⟷	NP NP V JP
.....	⟷	.....

SCFs listed in the first and the third columns are alternatives of each other, and our analysis of the verbs that take certain alternation pairs shows that one alternative SCF almost always ensures the existence of the other. This means that the value of $p(scf_i|scf_j, v)$ is much larger than zero if scf_i and scf_j form an alternation pair for a given verb.

4. Two-Fold Filtering Method

We can see from Section 3 that Han's diathesis alternations may well play a useful role as heuristic information for Chinese subcategorization acquisition. However, determining where and how to seed the heuristic remains difficult. [Korhonen 1998] applied diathesis alternations in Briscoe and Carroll's system to improve the performance of their BHT filter. Although the precision rate increased from 61.22% to 69.42% and the recall rate from 44.70% to 50.81%, the results were still not very accurate for possible practical NLP uses. Korhonen generated her one-way diathesis alternations from the ANLT dictionary, calculated the alternating

probability $p(scf_j|scfi)$ according to the number of common verbs that took the alternation ($scf_i \rightarrow scf_j$), and used formulas (7) and (8), where w is an empirical weight, to revise the observed $p(scf_i|v)$:

if $p(scf_i|scf_j, v) > 0$,

$$p(scf_i|v) = p(scf_i|v) - w(p(scf_i|v)p(scf_j|\, scf_i)); \quad (7)$$

if $p(scf_i|v) > 0$ and $p(scf_j|v) = 0$,

$$p(scf_i|v) = p(scf_i|v) + w(p(scf_i|v)p(scf_j|scfi)). \quad (8)$$[9]

Following the revision, a BHT filter with a confidence rate of 95% was used to check the SCF hypotheses.

This method removes the assumption of independence among SCF types but establishes another assumption of independence between $p(scf_j|scf_i)$ and certain verbs, which means that all verbs take each diathesis alternation with the same probability. Nevertheless, linguistic knowledge tells us that verbs often enter different diathesis alternations and can be classified accordingly. Consider the following examples:

e. He broke the glass. / The glass broke.

f. The police dispersed the crowd. / The crowd dispersed.

g. Mum cut the bread. / *The bread cut.

h. Ta1 chi1 le0 pin2guo3.(他吃了苹果。) / Ta1 ba3 pin2guo3 chi1 le0.(他把苹果吃了。)

i. Ta1 xie3 le0 ben3 shu1.(她写了本书。)[10] / *Ta1 ba3 shu1 xie3 le0.(她把书写了。)

Both of the English verbs “break” and “disperse” can take the causative-inchoative alternation and, hence, may be classified together, while the verb “cut” does not take this alternation. Also, the Chinese verb “chi1 吃” can take the *ba*-object-raising alternation, while the verb “xie3 写”(write) cannot. Therefore, this newly established assumption does not hold either, and the probabilistic sum of $p(scf_i|v)$ need not and cannot be normalized.

For dealing with this problem, our basic principle is that enough exploitation should be made on the observable data, yet no more than what can be observed. If both sentences in e, f or h are observed in the corpus, and if the SCF type of the first one has a high enough frequency to pass the MLE testing, while that of the second type does not, then both SCF

[9] For the sake of consistency in this paper and for the convenience to understand, the formats of formulas here are different from those of [Korhonen 1998], but they are actually the same.

[10] The Chinese sentence means *She wrote a book*.

types should be taken into consideration. Otherwise, the one with lower frequency might be falsely rejected. On the other hand, if the first sentence in i or g has a satisfactory SCF type frequency, while the SCF type of the second sentence does not occur in the input corpus, then the SCF type of the sentence may well be rejected.

Based on the above methodology, we formed our two-fold filtering method, which is, in fact, derived from the simple MLE filter and based on formulas (5) and (6). In our method, two filters are employed. First, a common MLE filter is used, except that it employs a threshold θ_1 that is much higher than usual, and those SCF hypotheses that satisfy the requirement are accepted. Then, all of the rest hypotheses are checked by another MLE filter that is seeded with diathesis alternations as heuristic information and equipped with a much lower threshold θ_2. Any hypothesis scf_i left out by the first filter will be accepted if its probability exceeds θ_2, which means that $p(scf_i|scf_j, v) > 0$, and if it is an alternative of any SCF type accepted by the first filter, which means that the verb v almost surely enters scf_j. The algorithm can be briefly expressed as shown in Table 5.

Table 5. Two-fold filtering algorithm

For hypotheses of a given verb v,
if $p(scf_i|v) > \theta_1$, scf_i is accepted;
else
  if $p(scf_i|v) > \theta_2$,
    $p(scf_i|scf_j, v) > 0$,
    and $p(scf_j|v) > \theta_1$,
  scf_i is accepted for v.

5. Experimental Evaluation and Analysis

The testing set included 48 verbs, as shown in Table 6. Thirty of them were of multiple syntactic patterns, while the rest were syntactically simple.

In the experiment, SCF hypotheses for the 48 verbs were generated from a corpus of the People's Daily from January to June of 1998 as described in Section 2. The resulting minimum number of SCF tokens for a verb was 86, and the maximum was 3200. The thresholds were experientially set as follows: θ_1= 0.017, which is much larger than the 0.008 threshold used by [Han *et al.* 2004b]; θ_2= 0.0004, which generally means a hypothesis would have a chance to check its diathesis alternations if it occurs even just one time in a token set no larger than 2,500. The probabilities that verbs take SCF types were also estimated according to the observed frequencies.

Table 6. The investigated Chinese verbs[11]

Chinese Verbs	English	Chinese Verbs	English
jie4 jian4(借鉴)	refer	chao1(抄)	copy
biao3 xian4(表现)	behave	du2(读)	read
jue2 ding4(决定)	decide	fang4(放)	put
cui1 can2(摧残)	torture	kan4(看)	see
dong4 jie2(冻结)	freeze	la1(拉)	pull
fa1 xian4(发现)	find	mo2(磨)	grind
fa1 zhan3(发展)	develop	shan3(闪)	flash
fan3 kang4(反抗)	rebel	song4(送)	send
fan3 ying4(反映)	reflect	tai2(抬)	carry
fen1 san4(分散)	disperse	tun1(吞)	devour
feng1 suo3(封锁)	blank	xi1(吸)	sock
shou1 fu4(收复)	reoccupy	xiang3(想)	Think
jian1 chi2(坚持)	insist	xiao4(笑)	laugh
jian4 li4(建立)	set up	xie3(写)	write
jie2 shu4(结束)	end	yong4(用)	use
jie3 fang4(解放)	release	zhe1(遮)	cover
xi1 wang4(希望)	wish	tao2 tai4(淘汰)	reject
yao1 qiu2(要求)	require	cai3 na4(采纳)	adopt
zeng1 qiang2(增强)	enforce	tou2 ru4(投入)	invest
zheng3 dun4(整顿)	neaten	bi1 jin4(逼近)	approach
zhu3 guan3(主管)	charge	gu3 wu3(鼓舞)	encourage
tong3 yi1(统一)	unify	kai1 shi3(开始)	begin
suo1 duan3(缩短)	shorten	kao3 lv4(考虑)	consider
tan4 wang4(探望)	visit	ren4 shi5(认识)	know

The evaluation standard was the manually analyzed results obtained from the applied corpus, and the precision and recall rates were calculated based on the following expressions used by [Korhonen 2001] and [Han *et al*. 2004b].

[11] The second and third columns give the relevant English meanings for the Chinese verbs, but they are far from being equivalents in English; they are just provided for reference for readers who don't know Chinese.

$$\text{Precision} = |\text{True positives}| / (|\text{True positives}| + |\text{False positives}|); \quad (9)$$

$$\text{Recall} = |\text{True positives}| / (|\text{True positives}| + |\text{False negatives}|). \quad (10)$$

Here, true positives are correct SCF types proposed by the system, false positives are incorrect SCF types proposed by system, and false negatives are correct SCF types not proposed by the system. For comparison, the performance of the system without any filter, with the simple MLE filter of a 0.008 threshold, and with a two-fold filter applied to the above-mentioned data is shown in Table 7.

Table 7. Comparison of performance

Method	Precision	Recall	F-measure
No-filter	37.64%	86.55%	52.46
MLE	60.3%	57.52%	58.89
Two-fold	76.94%	83.83%	80.24

The comparison shows that acquisition performance of the two-fold filter was remarkably improved, with a precision rate 16.64% better and a recall rate 26.31% better than that of the simple MLE, making the acquired lexicon much more practical for further manual proofreading and other NLP uses.

Meanwhile, the data shown in Table 7 imply that there is little room left for improvement of the statistical filter, since the precision rate achieved by the two-fold method is more than double that for the unfiltered results, and the recall rate is only 2.72% lower than that of the no-filter method. As far as we know, for English subcategorization, the best F-measure result previously reported by [Korhonen 2001], which used semantic backoff, was 78.4, while the best F-measure result for German obtained by [Shulte im Walde 2002] was 72.05, and that for Spanish by [Chrupala 2003] was 74. Therefore, although cross-lingual comparison may lack concrete significance, at present, ours is the best result obtained for Chinese and other languages.

6. Conclusions

Our two-fold filtering method makes more exploitation of what can be observed in the corpus by drawing on the alternative relationship between SCF hypotheses with higher and lower frequencies. Unlike the semantic motivated method [Korhonen 2001], which is dependent on verb classifications that linguistic resources are able to provide, two-fold filtering assumes no pre-knowledge other than reasonable diathesis alternation information and may work well for

most verbs in other languages with sufficient predicative tokens.

Our experimental results suggest that the proposed technique improves the Chinese subcategorization acquisition system, and leaves only a little room for further improvement in statistical filtering methods. Certainly, more sophisticated approaches still exist theoretically; for instance, some unseen SCFs found by a generator may be recalled by integrating verb-classification information into the system. More essential aspects of our future work, however, will focus on improving the performance of the hypothesis generator, and testing and applying the acquired subcategorization information in some common NLP tasks.

References

Brent, M., "From Grammar to Lexicon: unsupervised learning of lexical syntax," *Computational Linguistics,* 19(3), 1993, pp. 243-262.

Briscoe, T., and J. Carroll, "Automatic extraction of subcategorization from corpora," In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC, 1997, pp. 356-363.

Chomsky, N., *Aspects of the Theory of Syntax*, MIT Press, Cambridge, 1965.

Chrupala, G., "Acquiring Verb Subcategorization from Spanish Corpora", *PhD program "Cognitive Science and Language,"* Universitat de Barcelona, 2003, pp. 67-68.

Gamallo, P., A. Agustini, and P. Lopes Gabriel, "Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation," In *Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, Philadelphia, 2002, pp. 34-41.

Han, X., T. Zhao, and M. Yang, "FML-Based SCF Predefinition Learning for Chinese Verbs," In *Proceedings of the International Joint Conference of NLP 2004*, 2004a, pp. 115-122.

Han, X., T. Zhao, H. Qi, and H. Yu, "Subcategorization Acquisition and Evaluation for Chinese Verbs," In *Proceedings of the COLING 2004*, 2004b pp. 723-728.

Han, X., "Chinese Syntactic Alternation Acquisition Based on Verb Subcategorization Frames," In *Proceedings of the Chinese SWCL 2004*, 2004, pp. 197-202. [in Chinese]

Korhonen, A., "Automatic Extraction of Subcategorization Frames from Corpora–Improving Filtering with Diathesis Alternations," 1998. Please refer to http://www.folli. uva.nl/CD/1998/ pdf/keller /korhonen.pdf

Korhonen, A., *Subcategorization Acquistion*, Dissertation for Ph.D, Trinity Hall University of Cambridge, 2001, pp. 29-77.

Korhonen, A., Y. Krymolowski, and Z. Marx, "Clustering Polysemic Subcategorization Frame Distributions Semantically," In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 64-71.

Levin, B., *English Verb Classes and Alternations*, Chicago University Press, Chicago, 1993.

McCarthy, D., *Lexical Acquisition at the Syntax-Semantics Interface: Diathesis Alternations, Subcategorization Frames and Selectional Preferences*, PhD thesis, University of Sussex, 2001.

Meng, Y., *Research on Global Chinese Parsing Model and Algorithm Based on Maximum Entropy*, Dissertation for Ph.D. of Computer Department, HIT. 2003, pp. 33-34.[in Chinese]

Sarkar, A., and D. Zeman, "Automatic Extraction of Subcategorization Frames for Czech," In *Proceedings of the 19th International Conference on Computational Linguistics*, aarbrucken, Germany, 2000. Please refer to http://www.sfu.ca/~anoop /papers/pdf/coling 00_final.pdf

Shulte im Walde, S., "Inducing German Semantic Verb Classes from Pure- ly Syntactic Subcategorization Information," In *Proceedings of the 40st Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 223-230.

Zhao, T., *Knowledge Engineering Report for MTS2000*. Machine Translation Laboratory, Harbin Institute of Technology, Harbin, 2002. [in Chinese]