Subcategorization Acquisition and Evaluation for Chinese Verbs

Xiwu Han, Tiejun Zhao, Haoliang Qi, Hao Yu
Department of Computer Science,
Harbin Institute of Technology, 150001 Harbin, China
{hxw, tjzhao, qhl, yh}@mtlab.hit.edu.cn

Abstract

This paper describes the technology and an experiment of subcategorization acquisition for Chinese verbs. The SCF hypotheses are generated by means of linguistic heuristic information and filtered via statistical methods. Evaluation on the acquisition of 20 multi-pattern verbs shows that our experiment achieved the similar precision and recall with former researches. Besides, simple application of the acquired lexicon to a PCFG parser indicates great potentialities of subcategorization information in the fields of NLP.

Credits

This research is sponsored by National Natural Science Foundation (Grant No. 60373101 and 603750 19), and High-Tech Research and Development Program (Grant No. 2002AA117010-09).

Introduction

Since (Brent 1991) there have been a considerable amount of researches focusing on verb lexicons with respective subcategorization information specified both in the field of traditional linguistics and that of computational linguistics. As for the former, subcategory theories illustrating the syntactic behaviors of verbal predicates are now much more systemically improved, e.g. (Korhonen 2001). And for auto-acquisition and relevant application, researchers have made great achievements not only in English, e.g. (Briscoe and Carroll 1997), (Korhonen 2003), but also in many other languages, such as Germany (Schulte im Walde 2002), Czech (Sarkar and Zeman 2000), and Portuguese (Gamallo et. al 2002).

However, relevant theoretical researches on Chinese verbs are generally limited to case grammar, valency, some semantic computation theories, and a few papers on manual acquisition or prescriptive designment of syntactic patterns. Due to irrelevant initial motivations, syntactic and semantic generalizabilities of the consequent outputs are not in such a harmony that satisfies the description granularity for SCF (Han and Zhao 2004). The only auto-acquisition work for Chinese SCF made by (Han and Zhao 2004) describes the predefinition of 152 general frames for all verbs in Chinese, but that experiment is not based on real corpus. After observing and analyzing quantity of subcategory phenomena in real Chinese corpus in the People's Daily (Jan.~June, 1998), we removed from Han & Zhao's predefinition 15 SCFs that are actually similar derivants of others, and then with this foundation and linguistic rules from (Zhao 2002) as heuristic information we generated SCF hypotheses from the corpus of People's Daily (Jan.~June, 1998), and statistically filtered the hypotheses into a Chinese verb SCF lexicon. As far as we know, this is the first attempt of Chinese SCF auto-acquisition based on real corpus.

In the rest of this paper, the second section describes a comprehensive system that builds verb SCF lexicons from large real corpus, the respective operating principles, and the knowledge coded in our SCF. The third section analyzed the acquired lexicon with two experiments: one evaluated the acquisition results of 20 verbs with multi syntactic patterns against manual gold standard; the other checked the performance of the lexicon when applied in a PCFG parser. The forth section compares and contrasts this research with related works done by others. And at last, Section 5 concludes our present achievements, disadvantages and possible future focuses.

1 SCF Acquisition

1.1 The Acquisition Method

There are generally 4 steps in the process of our auto-acquisition experiment. First, the corpus is processed with a cascaded HMM parser; second,

every possible local patterns for verbs are abstracted; and then, the verb patterns are classified into SCF hypotheses according to the predefined set; at last, hypotheses are filtered statistically and the respective frequencies are also recorded. The actual application program consists of 6 parts as shown in the following paragraphs.

a. Segmenting and tagging: The raw corpus is segmented into words and tagged with POS by the comprehensive segmenting and tagging processor developed by MTLAB of Computer Department in 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[1], developed by MTLAB of HIT, but only the intermediate parsing results are used. The training set of the parser is 20,000 sentences in the Chinese Tree Bank[2] of (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 are generally about words or POS in the corpus.
d. Pattern abstraction: Verbs with largest governing ranges are regarded as predicates, then local patterns, previous phrases and respective syntactic tags are abstracted, and isolated parts are combined, generalized or omitted according to basic phrase rules in (Zhao 2002).
e. Hypothesis generation: Based on linguistic restraining rules, e.g. no more than two NP's occurring in a series and no more than three in one pattern, and no PP TP MP occurring with NP before any predicates (Han and Zhao 2004), the patterns are coordinated and classified into the predefined SCF groups. In this part, about 5% unclassifiable patterns are removed.
f. Hypothesis filtering: According to the statistical reliability of each type of the SCF hypotheses and the linguistic principle that arguments occur more frequently with predicates than adjuncts do, the hypotheses are filtered by means of statistical methods, in this paper which are binomial hypotheses testing (BHT) and maximum likelihood estimation (MLE).

Table 1: 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}[3]

In Table 1, for example, when acquiring SCF information for "证明" (prove) and a related sentence in the corpus is a), our tagger and parser will return b), and error-driven correction will return c) with errors of NDE and the 1st BVP corrected[4]. Since the governing range of "证明" is larger than that of "追问" (ask), the other verb in this sentence, the program abstracts its local pattern BVP[vg LE] and previous phrase BNP, gen-

[1] When evaluated on auto-tagged open corpus, the parser's phrase precision if 62.3%, and phrase recall is 60.9% (Meng, 2003).
[2] A sample of the tree bank or relevant introduction could be found at http://mtlab.hit.edu.cn.
[3] {01000} projects to the Chinese syntactic morphemes {着，了，过，没，不}, 1 means the SCF may occur with the respective morpheme, while 0 may not (Han & Zhao, 2004).
[4] Note that not all errors in this example have been corrected, but this doesn't affect further procession. Also, for definitions of NDE and BVP see (Zhao, 2002).

eralizes BNP and NDE as NP, combines the second NP with isolated part "在/p" into PP, and returns d). Then the hypothesis generator returns e) as the possible SCF in which the verb may occurs. Actually in the corpus there are 621 hypothesis tokens generated, and among them 92 ones are of same arguments with e), and thus e) can pass the hypothesis testing (See also Section 1.2), so we obtain one SCF for "证明" as f).

1.2 Filtering Methods

In researches of subcategorization acquisition, statistical methods for hypothesis filtering mainly include the BHT, the Log Likelihood Ratio (LLR), the T-test and the MLE, and the most popular one is the BHT. Since (Brent 1993) began to use the method, most researchers have agreed that the BHT results in better precision and recall with SCF hypotheses of high, medium and low frequencies. Only (Korhonen 2001) reports 11.9% total performance of the MLE better than the BHT. Therefore, we applied the two statistical methods in our present experiment. This subsection chiefly illustrates the expressions of our methods and definitions of parameters in them, while performance comparison of the two will be introduced in Section 3.

When applying the BHT method, it is necessary to determine the probability of the primitive event. As for SCF acquisition, the co-occurrence of one predefined SCF $\boldsymbol{scf}_i$ with one verb $\boldsymbol{v}$ is the relevant primitive event, and the concerned probability is $p(\boldsymbol{v}|\boldsymbol{scf}i)$ here. However, the aim of filtering is to rule out those unreliable hypotheses, so it is the probability that one primitive event doesn't occur that is often used for SCF hypothesis testing, i.e. the error probability: $p^e(\boldsymbol{v}|\boldsymbol{scf}i) = 1\ p(\boldsymbol{v}|\boldsymbol{scf}i)$. (Brent 1993) estimated p^e according to the acquisition system's performance, while (Briscoe and Carroll 1997) calculated p^e from the distribution of SCF types in ANLT and SCF tokens in Susanne as shown in the following equation.

$$p^e=(1-\frac{|\text{Anlt verbs for } \boldsymbol{scf}i|}{|\text{Anlt verbs}|}) \cdot \frac{|\text{Susanne SCF tokens for } \boldsymbol{scf}i|}{|\text{all Susanne SCF tokens}|}$$

Brent's method mainly depends on the related corpus and processing program, which may cause intolerable errors. Briscoe and Carroll's method draws on both linguistic and statistical information thus leading to comparatively stable estimation, and therefore has been used by many latter researches, e.g. (Korhonen 2001). But there is no MRD proper for Chinese SCF description so we estimated p^e from the 1,775 common verbs and SCF tokens in the related corpus of 43,000 sentences used by (Han and Zhao 2004). We formed the equation as follows:

$$p^e=(1-\frac{|\text{Verbs for } \boldsymbol{scf}i|}{1775}) \frac{|\text{Tokens for } \boldsymbol{scf}i|}{43000}$$

Then the number of all hypotheses about verb $\boldsymbol{v}j$ is recorded as n, and the number of those for $\boldsymbol{scf}i$ as m. According to Bernoulli theory, the probability P that an event with probability p exactly happens m times out of n such trials is:

$$P(m, n, p)=\frac{n!}{m!(n-m)!}\ p^m(1-p)^{n-m}$$

And the probability that the event happens m or more times is:

$$P(m+, n, p)=\sum_{i=m}^{n} P(i, n, p)$$

In turn, $P(m+, n, p^e)$ is the probability that $\boldsymbol{scf}i$ wrongly occurs m or more times with a verb that doesn't match it. Therefore, a threshold of 0.05 on this probability will yield a 95% confidence that a high enough proportion of hypotheses for $\boldsymbol{scf}i$ have been observed for the verb legitimately to be assigned $\boldsymbol{scf}i$ (Korhonen 2001).

The MLE method is closely related to the general performance of the concerned SCF acquisition system. First, we randomly draw from the applied corpus a training set, which is large enough so as to ensure similar SCF frequency distribution. Then, the frequency of $\boldsymbol{scf}i$ occurring with a verb $\boldsymbol{v}j$ is recorded and used to estimate the actual probability $p(\boldsymbol{scf}i|\ \boldsymbol{v}j)$. Thirdly, an empirical threshold is determined, such that it ensures maximum value of F measure on the training set. Finally, the threshold is used to filter out those SCF hypotheses with low frequencies from the total set.

2 Experimental Evaluation

2.1 Acquisition Performance

Using the previously described theory and technology we have acquired an SCF lexicon for 3,558 common Chinese verbs from the corpus of People's Daily (Jan.~June, 1998). In the lexicon

the minimum number of SCF tokens for a verb is 30, and the maximum is 20,000. In order to check the acquisition performance of the used system, we evaluated a part of the lexicon against a manual gold standard. The testing set includes 20 verbs of multi syntactic patterns, and for each verb there are 503~2,000 SCF tokens with the total number of 18,316 (See Table 2). Table 3 gives the evaluation results for different filtering methods, including non-filtering [5], BHT, and MLE with thresholds of 0.001, 0.005, 0.008 and 0.01. We calculated the type precision and recall by the following expressions as (Korhonen 2001) did:

$$\text{Precision} = \frac{|\text{True positives}|}{|\text{True positives}| + |\text{False positives}|}$$

$$\text{Recall} = \frac{|\text{True positives}|}{|\text{True positives}| + |\text{False negatives}|}$$

In 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.

Table 2: Verbs in the Testing Set[6]

Verbs	English	Tokens	Verbs	English	Tokens
读	Read	503	希望	Hope	620
发现	Find	529	看	See	645
考虑	Reckon	543	投入	Invest	679
拉	Pull	544	认识	Know	722
反映	Report	612	送	Send	800
发展	Develop	1,006	建立	Set up	1,186
表现	Behave	1,007	坚持	Insist	1,200
决定	Decide	1,038	想	Think	1,200
结束	End	1,140	要求	Require	1,200
开始	Begin	1142	写	Write	2,000

According to Table 3, all other filtering methods outperform non-filtering, and MLE is better than BHT. Among the four MLE thresholds, 0.008 achieves the best comprehensive performance but its F-measure is only 0.74 larger than that of 0.01 while its precision drops by 2.4 percent. Hence, we chose 0.01 as the threshold for the whole experiment with purpose to meet the practical requirement of high precision and to avoid possible over-fit phenomena. Finally, with a confidence of 95% we can estimate the general performance of the acquisition system with precision of 60.6% +/- 2.39%, and recall of 51.3%+/- 2.45%.

Table 3: System Performance for Different Filtering Methods

Methods \ Measures		Precision	Recall	F-measure
Non-filtering		37.43%	85.9%	52.14
BHT		50%	57.2%	53.36
MLE	0.001	39.2%	85.9%	53.83
	0.005	40.3%	83.33%	54.33
	0.008	58.2%	54.5%	56.3
	0.01	60.6%	51.3%	55.56

2.2 Task-oriented Evaluation

In order to further analyze the practicability of the previously described technology, we performed a simple task-oriented evaluation, applying the acquired SCF lexicon in a PCFG parser helping to choose from the n-best parsing results. The concerned parser was trained from 10,000 manually parsed Chinese sentences[7]. In this experiment there are 664 verbs and their SCF information involved. The open testing set consists of 1,500 sentences, for each of which the PCFG parser outputs 5-best parsing results. Then SCF hypotheses are generated for each result by means of the formerly mentioned technology. Finally, the maximum likelihood between hypotheses and those SCF types for the related verb in the lexicon is calculated in the following way:

$$ML = \max_{i,j} \frac{|(\text{Arguments in } \boldsymbol{h}_i) \cap (\text{Arguments in } \boldsymbol{scf}_j)|}{|(\text{Arguments in } \boldsymbol{h}_i) \cup (\text{Arguments in } \boldsymbol{scf}_j)|}$$

where $i \le 5$, $\boldsymbol{h}_i$ is one of the hypotheses generated for the parsing results, and $\boldsymbol{scf}_j$ is the jth SCF type for the concerned verb. This calculation keeps the likelihood between 0 and 1. The parsing result

[5] Non-filtering means filtering with a zero threshold or not filtering at all. This method is used as baseline here.

[6] The English meanings given here are not intended to cover the whole semantic range of the respective verbs, on the contrary they are just for readers' reference.

[7] These sentences and the testing corpus mentioned latter are all taken from the Chinese Tree Bank developed by MTLAB of HIT, and a sample may be downloaded at http://mtlab.hit.edu.cn.

with maximum likelyhood is then regarded as the final choice. When two or more hypotheses hold the same likelihood, the one with larger or largest PCFG probability will be chosen.

Table 4 shows the phrase-based and sentence-based evaluation results for the parser without and with SCF heuristic information. There are three cased included: a) The output is one-best; b) The output is 5-best and the best evaluation result is recorded; c) The 5-best output is checked again for the best syntactic tree by means of SCF information. The phrased-based evaluation follows the popular method for evaluating a parser, while the sentence-based depends on the intersection of the parsed trees and those in the gold standard. Since the PCFG parser output at least one syntactic tree for every sentence in our testing corpus, the sentence-based precision and recall are equal to each other.

Table 4: Parsing Evaluation

Parsing Methods	Phrase-based		Sentence-based
	Precision	Recall	Precision = Recall
One-best	57.5%	55%	13.64%
5-best	65.28%	64.59%	26.2%
With SCF	62.86%	62.1%	21.66%

Table 4 shows that SCF information remarkably improved the performance of the PCFG parser: the phrase-based precision increased by 5.36% and recall by 7.1%, while the sentence-based precision and recall both increased by 8.04%. However, this doesn't reach the upper limit of the 5-best. The possible reasons are: a) the our present SCF lexicon remains to be improved; b) our method of applying SCF information to the parser is too simple, e.g. probabilities of PCFG parsing results haven't been exploited thoroughly.

3 Related Works

As far as we know, this is the first attempt to automatically acquire SCF information from real Chinese corpus and the first trial to apply SCF lexicon to a Chinese parser. Our research draws a lot on related works from international researches, and for the purpose of crosslingual processing, our research is kept in consistency with SCF conventions as much as possible.

Due to linguistic differences, nevertheless, not all theories, methods or experiences could adapt to Chinese. Generally, there are four aspects that our research differs from those of other languages. First, the SCF formalization of most former researches follows the Levin style, in which most SCFs omit NP before predicates, while Chinese SCFs need to depict arguments occurring before verbs. Second, except (Sarkar and Zeman 2000), most former researches are based on manual SCF predefinition, while our predefined SCF set is statistically acquired (See Han and Zhao 2004). Third, involved parsers of former researches are mostly better than Chinese parsers to some degree. Forth, our SCF information also includes 5 syntactic morphemes (See also Section 1.1).

Meanwhile, the basic purpose for Chinese SCF acquisition is also to determine the subcategory features for a verb via its argument distributions and then apply the lexicon to NLP tasks. Therefore, under similar cases the respective evaluations are comparable. And Table 5 gives the comparison between our research and the best English results without semantic backoff [8] in (Korhonen 2001).

Table 5: Performance Comparison Between Chinese and English Researches

Measures \ Filtering		Non	BHT	MLE
Precision	Ours	37.43%	50%	58.2%
	Korhonen	24.3%	50.3%	74.8%
Recall	Ours	85.9%	57.2%	54.5%
	Korhonen	83.5%	56.6%	57.8%
F-measure	Ours	52.14	53.36	56.3
	Korhonen	37.6	53.3	65.2

The comparison shows that our nonfiltering result is better than Korhonen's, both BHT results are similar, while our MLE result is much worse

[8] Semantic backoff is a method of generating SCF hypotheses according to the semantic classification of the concerned verb. Note that this paper doesn't involve verb meanings for generating hypotheses. Besides, though the evaluation for English SCF acquisition is the best, it's not the newest. For the newest, please refer to (Korhonen 2003), in which the precision is 71.8% and recall is 34.5%.

than Korhonen's. That means our hypothesis generator performs well but our filtering method remains to be improved. According to the analysis of relevant corpus, we found the main cause might be that low frequency SCF types account for 32% in our corpus while those in (Korhonen 2001) sum to nearly 21%.

Further more, (Briscoe and Carroll 1997) applied their acquired English SCF lexicon to an intermediate parser, and reported a 7% improvement of both phrase-based precision and recall. Our application of SCF lexicon to a PCFG parser leads to 5.36% improvement for phrase-based precision, 7.1% for recall, and 8.04% for sentence-based precision and recall.

4 Conclusion

This paper for the first time describes a largescale experiment of automatically acquiring SCF lexicon from real Chinese corpus. Perfor mance evaluation shows that our technology and acquiring program have achieved similar performance compared with former researches of other languages. And the application of the acquired lexicon to a PCFG parser indicates great potentialities of SCF information in the field of NLP.

However, there is still a large gap between Chinese subcategorization works and those of other languages. Our future work will focus on the optimization of linguistic heuristic information and filtering methods, the application of semantic backoff, and the exploitation of SCF lexicon for other NLP tasks.

References

Brent, M. R. 1991. *Automatic acquisition of subcategorization frames from untagged text*. In Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA. 209-214.

Brent, M. 1993. *From Grammar to Lexicon: unsupervised learning of lexical syntax*. Computational Linguistics 19.3. 243-262.

Briscoe, Ted and John Carroll, 1997. *Automatic extraction of subcategorization from corpora*. In Proceedings of the 5th ACL Conference on Applied Natural Language Processing, Washington, DC.

Dorr, B. J. Gina-Anne Levow, Dekang Lin, and Scott Thomas, 2000. *Chinese-English Semantic Resource Construction*, 2nd International Conference on Language Resources and Evaluation (LREC2000), Athens, Greece, pp. 757--760.

Gamallo, P., Agustini, A. and Lopes Gabriel P., 2002. *Using Co-Composition for Acquiring Syntactic and Semantic Subcategorisation*, ACL-02.

Han, Xiwu, Tiejun Zhao, 2004. *FML-Based SCF Predefinition Learning for Chinese Verbs*. International Joint Conference of NLP 2004.

Jin, Guangjin, 2001. *Semantic Computations for Modern Chinese Verbs*. Beijing University Press, Beijing. (in Chinese)

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

Korhonen, Anna, 2003. *Clustering Polysemic Subcategorization Frame Distributions Semantically*. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 64-71.

Meng, Yao, 2003. *Research on Global Chinese Parsing Model and Algorithm Based on Maximum Entropy.* Dissertation for Ph.D. Computer Department, HIT. 33-34.

Sabine Shulte im Walde, 2002. *Inducing German Semantic Verb Classes from Purely Syntactic Subcategorization Information*. *Proceedings of the 40st ACL*, pp. 223-230.

Sarkar, A. and Zeman, D. 2000. *Automatic Extraction of Subcategorization Frames for Czech*. In Proceedings of the 19th International Conference on Computational Linguistics, aarbrucken, Germany.

Zhan Weidong, 2000. *Valence Based Chinese Semantic Dictionary*, *Language and Character Applications*, Volume 1. (in Chinese)

Zhao Tiejun, 2002. *Knowledge Engineering Report for MTS2000*.