Improving English Subcategorization Acquisition with Diathesis Alternations as Heuristic Information

Xiwu Han
Institute of Computational Linguistics
Heilongjiang University
Harbin City 150080 China
hxw@hlju.edu.cn

Tiejun Zhao
School of Computer Science and Technology
Harbin Institute of Technology
Harbin City 150001 China
tjzhao@mtlab.hit.edu.cn

Xingshang Fu
Institute of Computational Linguistics
Heilongjiang University
Harbin City 150080 China
fxs@hlju.edu.cn

Abstract

Automatically acquired lexicons with subcategorization information have already proved accurate and useful enough for some purposes but their accuracy still shows room for improvement. By means of diathesis alternation, this paper proposes a new filtering method, which improved the performance of Korhonen's acquisition system remarkably, with the precision increased to 91.18% and recall unchanged, making the acquired lexicon much more practical for further manual proofreading and other NLP uses.

1 Introduction

Subcategorization is the process that further classifies a syntactic category into its subsets. Chomsky (1965) defines the function of strict subcategorization features as appointing a set of constraints that dominate the selection of verbs and other arguments in deep structure. Large subcategorized verbal lexicons have proved to be crucially important for many tasks of natural language processing, such as probabilistic parsers (Korhonen, 2001, 2002) and verb classifications (Schulte im Walde, 2002; Korhonen, 2003). Since Brent (1993) a considerable amount of research focusing on large-scaled automatic acquisition of subcategorization frames (SCF) has met with 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, 2004). The general objective of this research is to acquire from a given corpus the SCF types and numbers for predicate verbs. Two typical steps during the process of automatic acquisition are hypothesis generation and selection. Usually based on heuristic rules, the first step generates SCF hypotheses for involved verbs; and the second selects reliable ones via statistical methods, such as BHT (binomial hypothesis testing), LLR (log likelihood ratio) and MLE (maximum likelihood estimation). This second step is also called statistical filtering and has been widely regarded as problematic. English researchers have proposed some methods adjusting the corpus hypothesis frequencies before or while filtering. These methods are often called backoff techniques for SCF acquisition. Some of them represent a remarkable improvement in the acquisition performance, for example diathesis alternation and semantic motivation (Korhonen, 1998, 2001, 2002).

For the convenience of comparison between performances of different SCF acquisition methods, we define absolute and relative recall in this paper. By absolute recall, we mean the figure computed against the background of input corpus, while relative recall is against the set of generated hypotheses.

At present, automatically acquired verb lexicons with SCF information have already proved accurate and useful enough for some NLP purposes (Korhonen, 2001; Han et al, 2004). As for English, Korhonen (2002) reported that semantically motivated SCF acquisition achieved a precision of 87.1%, an absolute recall of 71.2% and a relative recall of 85.27%, thus making the acquired lexicon much more accurate and useful. However, the accuracy still shows room for improvement, especially for those SCF hypotheses with low frequencies. Detailed analysis on the acquisition system and some resulting data shows that three main causes should account for the comparatively unsatisfactory performance: a. the imperfect hypothesis generator, b. the Zipfian

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 331–336,
Sydney, July 2006.

distribution of syntactic patterns, c. the incomplete partition over SCF types of a given verb. The first problem mainly comes from the inadequate parsing performance and noises existing in the corpus, while the other two problems are inherent to natural languages and should be solved in terms of acquisition techniques particularly during the process of hypothesis selection.

2 Related Work

The empirical background of this paper is the public resource for subcategorization acquisition of English verbs, provided by Anna Korhonen (2005) in her personal home page. The data include 30 verbs, as shown in Table 1, and their unfiltered SCF hypotheses, which were automatically generated via Briscoe and Carroll's (1997) SCF acquisition system, and the manually established standard.

Table 1. English Verbs in Use.

add	agree	attach
bring	carry	carve
chop	cling	clip
fly	cut	travel
drag	communicate	give
lend	lock	marry
meet	mix	move
offer	provide	visit
push	sail	send
slice	supply	swing

For each verb, there is a corpus of 1000 sentences extracted from the BNC, and all together 42 SCF types are involved in the corpus. The framework of Briscoe and Carroll's system consists of six overall components, which are applied in sequence to sentences containing a specific predicate in order to retrieve a set of SCFs for that verb:

- **A tagger**, a first-order Hidden Markov Model POS and punctuation tag disambiguator.
- **A lemmatizer**, an enhanced version of the General Architecture for Text Engineering project stemmer.
- **A probabilistic LR parser**, trained on a tree-bank derived semi-automatically from the SUSANNE corpus, returns ranked analyses using a feature-based unification grammar.
- **A pattern extractor**, which extracts subcategorization patterns, i.e. local syntactic frames, including the syntactic frames, including the syntactic categories and head lemmas.
- **A pattern classifier**, which assigns patterns to SCFs or rejects them as unclassifiable.
- **A SCF filter**, which evaluates sets of SCFs gathered for a predicate verb.

Nowadays, in most related researches, the performances of subcategorization acquisition systems are often evaluated in terms of precision, recall and F measure of SCF types (Korhonen, 2001, 2002). Generally, precision is the percentage of SCFs that the system proposes correctly, while recall is the percentage of SCFs in the gold standard that the system proposes:

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

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

$$\text{F-measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

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.

3 The MLE Filtering Method

The present SCF acquisition system for English verbs employs a MLE filter to test the automatically generated SCF hypotheses. Due to noises accumulated while tagging, lemmatizing and parsing the corpus, even though correction is implemented for some typical errors when classifying the extracted patterns, the hypothesis generator does not perform as efficiently as hoped. Sampling analysis on the unfiltered hypotheses in Korhonen's evaluation corpus indicates that about 74% incorrectly proposed and rejected SCF types come from the defects of the MLE filtering method.

Performance of the MLE filter is closely related to the actual distributions $p(scf_i|v)$ over predicates and SCF types in the input corpus. First, from the overall corpus a training set is drawn randomly; it must be large enough to ensure a similar distribution. Then, the frequency of a subcategorization frame scf_i occurring with a verb v is recorded and used to estimate the probability $p(scf_i|v)$. Thirdly, an empirical threshold θ 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 from the total set those SCF hypotheses with frequencies lower than θ.

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:

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) = 0 \quad \ldots (1)$$

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

Here, i and j are natural numbers, scf_i and scf_j are two SCF types that verb v enters, and variables in formulas henceforth will hold the same meanings. In actual application, the probability $p(scf_i|v)$ is estimated from the observed frequency $f(scf_i, v)$, and the conditional probability $p(scf_i|scf_j, v)$ is assumed to be zero. This means any two SCF types entered by a given verb are taken for granted to be probabilistically independent from each other. However, this assumption can sometimes be far from appropriate.

4 Diathesis Alternations and Filtering

Much linguistic research focusing on child language acquisition has revealed that many children are able to produce new grammatical sentences from what they have learned (Peters, 1983; Ellis, 1985). This 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 should be removed, a possible heuristic could be the information of diathesis alternations, which is also another convincing counterargument. 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[1].
(他 把 苹果 吃 了。)

In the above examples, 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 preposition *ba3* (把) to the location between the subject and the predicate, as illustrated in Figure 1.

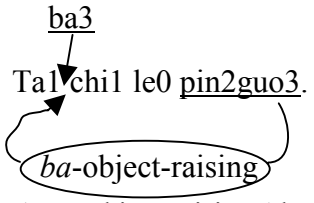

Figure 1. Ba-object-raising Alternation.

Subcategorization of verbs has much to do with diathesis alternations, and most SCF researchers regard information of diathesis alternation as an indispensable part of subcategorization (Korhonen, 2001; McCarthy, 2001). Therefore, one may 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 will be appropriate and sufficient if and only if there is no diathesis alternation between all the SCFs it enters, and formula (1) and (2) in Section 3 are efficient enough to serve as a foundation for the MLE filtering method. Otherwise, if there are diathesis alternations between some of the SCFs that a verb enters, then formula (1) and (2) must be modified as illustrated in formula (3) and (4). In either case, for the sake of convenience, it would be better to combine the formulas as shown in (5) and (6).

$$\exists i, \exists j, i \neq j, p(scf_i \mid scf_j, v) > 0 \quad \ldots (3)$$

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

$$\forall i, \forall j, i \neq j, p(scf_i \mid scf_j, v) \geq 0 \quad \ldots (5)$$

$$\sum_{i=1}^{n} p(scf_i \mid v) \geq 1 \quad \ldots (6)$$

For English verbs, previous research has achieved great progress in diathesis alternation and relative applications, such as the work of Levin (1993) and McCarthy (2001). Besides, Korhonen (1998) has proved that diathesis alternation could be used as heuristic information for backoff estimates to improve the general performance of subcategorization acquisition. However, determining where and how to seed the heuristic remains difficult.

Korhonen (1998) employed 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

[1] The numbers in sentences c and d, which are pinyin notations, show tones of the Chinese syllables, and the two sentences, in English, generally mean *He ate an apple*.

69.42% and the recall rate from 44.70% to 50.81%, the results were still not accurate enough for possible practical NLP uses.

Korhonen obtained her one-way diathesis alternations from the ANLT dictionary (Boguraev and Briscoe, 1987), calculated the alternating probability $p(scf_j|scf_i)$ according to the number of common verbs that took the alternation ($scf_i \rightarrow scf_j$), and used formula (7) and (8), where w is an empirical weight, to adjust the previously estimated $p(scf_i|v)$:

$$\text{If } p(scf_i|scf_j, v)>0,$$
$$p(scf_i|v) = p(scf_i|v) - w(p(scf_i|v) \bullet p(scf_j|scf_i)) \quad \ldots(7)$$
$$\text{If } p(scf_i|v)>0 \ \&\ p(scf_j|v)=0,$$
$$p(scf_i|v) = p(scf_i|v) + w(p(scf_i|v) \bullet p(scf_j|scf_i)) \quad \ldots(8)^2$$

Following the adjustment, 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 assumes 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.

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. Therefore, the newly established assumption doesn't fit the actual situation either, and the probability sums $\Sigma_i p(scf_i|v)$ and $\Sigma_{i,j} p(scf_i|scf_j, v)$ neither need or can be normalized.

Based on the above methodology, we formed a new filtering method with diathesis alternations as heuristic information, which is, in fact, derived from the simple MLE filter and based on formula (5) and (6). The algorithm can be briefly expressed as shown in Table 2.

Table 2. The New Filtering Method.

For hypotheses of a given verb v,
1. if $p(scf_i|v) > \theta_1$,
   accept scf_i into the output set $\boldsymbol{S}$;
2. else
   if $p(scf_i|v) > \theta_2$,
   & $p(scf_i|scf_j, v) > 0$,
   & $scf_j \in \boldsymbol{S}$,
   accept scf_i into set $\boldsymbol{S}$;
3. Go to step 1 until $\boldsymbol{S}$ doesn't increase.

In our method, two filters are employed. For each verb involved, first a common MLE filter is used, but 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 remainder of the hypotheses are checked by another MLE filter 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 and it is an alternative of an SCF type scf_j that has been accepted by the first filter, which means that $p(scf_i|scf_j, v)>0$ and $scf_j \in \boldsymbol{S}$. The filtering process will be performed repeatedly for those unaccepted hypotheses until no more hypotheses can be accepted for the verb.

5 Experimental Evaluation

We implemented an acquisition experiment on Korhonen's evaluation resources with the above-mentioned filtering method.

The diathesis alternations in use are also those provided by Korhonen, except that we used them in a two-way manner ($scf_i \leftrightarrow scf_j$) instead of one-way ($scf_i \rightarrow scf_j$), because the two involved SCF types are usually alternative pragmatic formats of the concerned verb, as shown in examples in Section 3 and 4.

In the experiment we empirically set $\theta_1 = 0.2$, which is ten times of Korhonen's threshold for her MLE filter; $\theta_2 = 0.002$, which is one tenth of Korhonen's. Thus, in a token set of hypotheses no more than 1000, an SCF type scf_i will be accepted if it occurs two times or more and has a diathesis alternative type scf_j already accepted for the verb.

The gold standard was the manually analysed results by Korhonen. Precision, recall and F-measure were calculated via expressions given in Section 2.

Table 3 lists the performances of the baseline method of non-filtering (No_f), MLE filtering with $\theta = 0.02$, and our filtering method on the

[2] For the sake of consistency in this paper and for the convenience of understanding, formulae formats here are modified. They may look different from those of Korhonen (1998), but they are actually the same.

evaluation corpus, and also gives the best results of Korhonen's method that is using extra semantic information (Kor) to make a comparison. Here, Ab_R is the absolute recall ratio, Re_R the relative recall ratio, Ab_F the absolute F-measure that is calculated from Precision and Ab_R, and Re_F the relative F-measure that is from Precision and Re_R.

Table 3. Performance Comparison.

Methods	No-f	MLE	ours	Kor
P(%)	47.85	67.89	91.18	87.1
Ab_R(%)	34.62	32.52	32.52	71.2
Re_R(%)	100	93.93	93.93	85.27
Ab_F	40.17	43.98	47.94	78.35
Re_F	64.73	78.81	92.53	86.18

The evaluation shows that our new filtering method improved the acquisition performance remarkably: a. Compared with MLE, precision increased by 23.29%, recall ratio remained unchanged, absolute F-measure increased by 3.96, and relative F-measure increased by 13.72; b. Compared with Korhonen's best results, precision, Re_R and Re_F also increased respectively[3]. Thus, the general performance of our filtering method makes the acquired lexicon much more practical for further manual proofreading and other NLP uses.

What's more, the data shown in Table 3 implies that there is little room left for improvement of the statistical filter, since the absolute recall ratio is only 2.1% lower than that of the non-filtering method. Whereas, detailed analysis of the evaluation corpus shows that the hypothesis generator accounts for about 95% of those unrecalled and wrongly recalled SCF types, which indicates, for the present time, more improvement efforts need to be made on the first step of subcategorization acquisition, i.e. hypothesis generation.

6 Conclusion

Our new filtering method removed the inappropriate assumptions and takes much more advantage of what can be observed in the corpus by drawing on the alternative relationship between SCF hypotheses with higher and lower frequencies. Unlike the semantically motivated method (Korhonen, 2001, 2002), which is dependent on verb classifications that linguistic resources are able to provide, our filter needs no prior 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 general performance of the English subcategorization acquisition system, and leaves only a little room for further improvement in statistical filtering methods. However, approaches that are more complicated still exist theoretically, for instance, some SCF types unseen by the hypothesis generator may be recalled by integrating semantic 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 lexicons in some concrete NLP tasks.

Acknowledgement This research has been jointly sponsored by the NSFC project No. 60373101 and the post-doctor scholarship of foreign linguistics and literature in Heilongjiang University. And at the same time, our great thanks go to Dr. Anna Korhonen for her public evaluation resources, and Dr. Chrys Chrystello for his helpful advice on the English writing of this paper.

References

Boguraev B. K., E. J. Briscoe. Large lexicons for natural language processing utilizing the grammar coding system of the Longman Dictionary of Contemporary English. *Computational Linguistics*, 1987: 219-240

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

Briscoe, Ted and John Carroll, Automatic extraction of subcategorization from corpora, *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, Washington, DC, 1997: 356-363.

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

Chrupala, Grzegorz, Acquiring Verb Subcategorization from Spanish Corpora, *PhD program "Cogni-*

[3] Korhonen (2002) reported the non-filtering absolute recall ratio of her experiment was about 83.5%. She didn't give any explanation with her evaluation resources why here non-filtering Ab_R was so much lower. Therefore, the Ab_R and Ab_F figures are not comparable here.

tive Science and Language", Universitat de Barcelona, 2003: 67-68.

Ellis, R. *Understanding Second language Acquisition*, Oxford University Press.1985

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

Han, Xiwu, Tiejun Zhao, Haoliang Qi, and Hao Yu, Subcategorization Acquisition and Evaluation for Chinese Verbs, *Proceedings of the COLING 2004*, 2004: 723-728.

Korhonen, Anna, 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, Anna, *Subcategorization Acquisition*, Dissertation for PhD, Trinity Hall University of Cambridge, 2001.

Korhonen, Anna, *Subcategorization Acquisition*, Technical Report Number 530, Trinity Hall University of Cambridge, 2002.

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

Korhonen, Anna. Subcategorization Evaluation Resources. http://www.cl.cam.ac.uk/users/alk23/subcat/subcat.html. 2005

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.

Peters, A. The Unit of Language Acquisition, Cambridge University Press. 1983.

Sarkar, A. and Zeman, D., Automatic Extraction of Subcategorization Frames for Czech, *Proceedings of the 19th International Conference on Computational Linguistics*, Saarbrucken, Germany, 2000. Please refer to http://www.sfu.ca/~anoop/papers/pdf/coling0final.pdf

Shulte im Walde, Sabine, Inducing German Semantic Verb Classes from Purely Syntactic Subcategorization Information, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002: 223-230.