Technical Report

UCAM-CL-TR-491
ISSN 1476-2986

Number 491

UNIVERSITY OF CAMBRIDGE
Computer Laboratory

Is hypothesis testing useful for subcategorization acquisition?

Anna Korhonen, Genevive Gorrell,
Diana McCarthy

May 2000

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

https://www.cl.cam.ac.uk/

Technical reports published by the University of Cambridge
Computer Laboratory are freely available via the Internet:

https://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Is Hypothesis Testing Useful for Subcategorization Acquisition?

Anna Korhonen and **Genevieve Gorrell**
Computer Laboratory, University of Cambridge
Pembroke Street, Cambridge CB2 3QG, UK
alk23@cl.cam.ac.uk, g_gorrell@yahoo.com

Diana McCarthy
School of Cognitive and Computing Sciences
University of Sussex, Brighton, BN1 9QH, UK
dianam@cogs.susx.ac.uk

Abstract

Statistical filtering is often used to remove noise from automatically acquired subcategorization frames. In this paper, we compare three different approaches to filtering out spurious hypotheses. Two hypothesis tests perform poorly, compared to filtering frames on the basis of relative frequency. We discuss reasons for this and consider directions for future research.

1 Introduction

Subcategorization information is vital for successful parsing, however, manual development of large subcategorized lexicons has proved difficult because predicates change behaviour between sublanguages, domains and over time. Additionally, none of these sources provide the relative frequency of different subcategorization frames (SCFs) for a given predicate, essential in a probabilistic approach.

Over the past years acquiring subcategorization dictionaries from textual corpora has become increasingly popular. The different approaches (e.g. Brent, 1991, 1993; Ushioda *et al.*, 1993; Briscoe & Carroll, 1997; Manning, 1993; Carroll & Rooth 1998; Gahl, 1998; Lapata, 1999) vary largely according to the methods used and the number of SCFs being extracted. Regardless of this, there is a ceiling on the performance of these systems at around 80% token recall.

The approaches to extracting SCF information from corpora have frequently employed statistical methods for filtering (Brent, 1993; Manning 1993; Briscoe & Carroll, 1997; Lapata, 1999). This has been done to remove the noise that arises when dealing with naturally occurring data, and from mistakes made by the SCF acquisition system, for example, parser errors.

Filtering is usually done with a hypothesis test, and frequently with a variation of the binomial filter introduced by Brent (1991, 1993). Hypothesis testing is performed by formulating a null hypothesis, (H_0), which is assumed true unless there is evidence to the contrary. If there is evidence to the contrary, H_0 is rejected and the alternative hypothesis (H_1) is accepted. In SCF acquisition, H_0 is that there is no association between a particular verb ($verb_j$) and a SCF (SCF$_i$), meanwhile H_1 is that there is such an association. For SCF acquisition, the test is one-tailed since H_1 states the direction of the association, a positive correlation between $verb_j$ and scf_i. We compare the expected probability of scf_i occurring with $verb_j$ if H_0 is true, to the observed probability of co-occurrence obtained from the corpus data. If the observed probability is greater than the expected probability we reject H_0 and accept H_1, and if not, we retain H_0.

Despite the popularity of this method, it has been reported as problematic. According to one account (Briscoe & Carroll, 1997) the majority of errors arise because of the statistical filtering process, which is reported to be particularly unreliable for low frequency SCFs (Brent, 1993; Briscoe & Carroll, 1997; Manning, 1993; Manning & Schütze 1999).

Adopting the SCF acquisition system of Briscoe & Carroll we have experimented with an alternative hypothesis test, the binomial log-likelihood ratio (LLR) test (Dunning, 1993). This test has been recommended for use in NLP since it does not assume a normal distribution, which invalidates many other parametric tests for use with natural language phenomena. LLR can be used in a form $(-2log\lambda)$ which is χ^2 distributed. Moreover, this asymptote is appropriate at quite low frequencies, which makes the hypothesis test particularly useful when dealing with natural language phenomena, where low frequency events are commonplace.

In this paper, we compare the results of both the Brent style binomial filter of Briscoe & Carroll, and the LLR filter. In addition to these significance tests, we compare a simple method which uses a threshold on the relative frequencies of the verb and SCF combinations. We do this within the framework of the Briscoe & Carroll SCF acquisition system, which is described in section 2.1. The details of the two statistical filters are described in section 2.2, along with the details of the threshold applied to the relative frequencies output from the SCF acquisition system. The details of the experimental evaluation are supplied in section 3. We discuss our findings in section 3.3 and conclude with directions for future work (section 4).

2 Method

2.1 Framework for SCF Acquisition

Briscoe & Carroll's (1997) verbal acquisition system distinguishes 163 SCFs and returns relative frequencies for each SCF found for a given predicate. The SCFs are a superset of classes found in the Alvey NL Tools (ANLT) dictionary, Boguraev *et al.* (1987) and the COMLEX Syntax dictionary, Grishman *et al.* (1994). They incorporate information about control of predicative arguments, as well as alternations such as extraposition and particle movement. The system employs a shallow parser to obtain the subcategorization information. Potential SCF entries are filtered before the final SCF lexicon is produced. The filter is the only component of this system which we experiment with here. The three filtering methods which we compare are described below.

2.2 Filtering Methods

2.2.1 Binomial Hypothesis Test

Briscoe & Carroll (1997) used a binomial hypothesis test (BHT) to filter the acquired SCFs. They applied BHT as follows. The system recorded the total number of sets of SCF cues (n) found for a given predicate, and the number of these sets for a given SCF (m). The system estimated the error probability (p^e) that a cue for a SCF (scf_i) occurred with a verb which did not take scf_i. p^e was estimated in two stages, as shown in equation 1. Firstly, the number of verbs which are members of the target SCF in the ANLT dictionary were extracted. This number was converted to a probability of class membership by dividing by the total number of verbs in the dictionary. The complement of this probability provided an estimate for the probability of a verb not taking scf_i. Secondly, this probability was multiplied by an estimate for the probability of observing the cue for scf_i. This was estimated using the number of cues for i extracted from the Susanne corpus (Sampson, 1995), divided by the total number of cues.

$$p^e = \left(1 - \frac{|anlt\, verbs\, in\, class\, i|}{|anlt\, verbs|}\right) \frac{|cues\, for\, i|}{|cues|} \tag{1}$$

The probability of an event with probability p happening exactly m times out of n attempts is given by the following binomial distribution:

$$P(m, n, p) = \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m} \tag{2}$$

The probability of the event happening m or more times is:

$$P(m+, n, p) = \sum_{k=m}^{n} P(k, n, p) \tag{3}$$

Finally, $P(m, n, p^e)$ is the probability that m or more occurrences of cues for scf_i will occur with a verb which is not a member of scf_i, given n occurrences of that verb. A threshold on this probability, using $P(m+, n, p)$, was set at less than or equal to 0.05. This yielded a 95% or better confidence that a high enough proportion of cues for scf_i have been observed for the verb to be legitimately assigned scf_i.

Other approaches which use a binomial filter differ in respect of the calculation of the error probability. Brent (1993) estimated the error probabilities for each SCF experimentally from the behaviour of his SCF extractor, which detected simple morpho-syntactic cues in the corpus data. Manning (1993) increased the number of available cues at the expense of the reliability of these cues. To maintain high levels of accuracy, Manning applied higher bounds on the error probabilities for certain cues. These bounds were determined experimentally. A similar approach was taken by Briscoe, Carroll & Korhonen (1997) in a modification to the Briscoe & Carroll system. The overall performance was increased by changing the estimates of p^e according to the performance of the system for the target SCF. In the work described here, we use the original BHT proposed by Briscoe & Carroll.

2.2.2 The Binomial Log Likelihood Ratio as a Statistical Filter

Dunning (1993) demonstrates the benefits of the LLR statistic, compared to Pearson's chi-squared, on the task of ranking bigram data. To our knowledge, LLR has not been previously used in SCF acquisition.

The binomial log-likelihood ratio test is simple to calculate. For each verb and SCF combination four counts are required. These are the number of times that:

1. the target verb occurs with the target SCF (k_1)
2. the target verb occurs with any other SCF ($n_1 - k_1$)
3. any other verb occurs with the target SCF (k_2)

4. any other verb occurs with any other SCF $(n_2 - k_2)$

The statistic $-2log\lambda$ is calculated as follows:-

$$-2log\lambda = 2(logL(p_1, k_1, n_1) \ + logL(p_2, k_2, n_2) - logL(p, k_1, n_1) - logL(p, k_2, n_2))$$

where: $logL(p, n, k) = k * log_2 p + (n - k) * log_2(1 - p)$

and:

$$p_1 = \frac{k_1}{n_1}\ , \quad p_2 = \frac{k_2}{n_2}\ , \quad p = \frac{k_1+k_2}{n_1+n_2}$$

The LLR statistic provides a score that reflects the difference in (i) the number of bits it takes to describe the observed data, using $p1 = p(\text{SCF}|verb)$ and $p2 = p(\text{SCF}|\neg verb)$, and (ii) the number of bits it takes to describe the expected data using the probability $p = p(\text{SCF}|any\ verb)$.

The LLR statistic detects differences between $p1$ and $p2$. The difference could potentially be in either direction, but we are interested in LLRs where $p1 > p2$, i.e. where there is a positive association between the SCF and the verb. For these cases, we compared the value of $-2log\lambda$ to the threshold value obtained from Pearson's Chi-Squared table, to see if it was significant at the 95% level.

2.2.3 Using a Threshold on the Relative Frequencies as a Baseline

In order to examine the baseline performance of this system without employing any notion of the significance of the observations, we used a threshold on relative frequencies. This was done by extracting the SCFs, and ranking them in the order of the probability of their occurrence with the verb. The probabilities were estimated using a maximum likelihood estimate from the observed relative frequencies. A threshold, determined empirically, was applied to these probability estimates to filter out the low probability entries for each verb.

3 Evaluation

3.1 Method

To evaluate the different approaches, we took a sample of 10 million words of the BNC corpus (Leech, 1992). We extracted all sentences containing an occurrence of one of fourteen verbs[1]. The verbs were chosen at random, subject to the constraint that they exhibited multiple complementation patterns. After the extraction process, we retained 3000 citations, on average, for each verb. The sentences containing these verbs were processed by the SCF acquisition system, and then we applied the three filtering methods described above. We also obtained results for a baseline without any filtering.

The results were evaluated against a manual analysis of corpus data[2]. This was obtained by analysing up to a maximum of 300 occurrences for each of the 14 test verbs in LOB (Garside *et al.*, 1987), Susanne and SEC (Taylor & Knowles, 1988) corpora. Following Briscoe & Carroll (1997), we calculated *precision* (percentage of SCFs acquired which were also exemplified in the manual analysis) and *recall* (percentage of the SCFs exemplified in the manual analysis which were acquired automatically).

[1] These verbs were *ask, begin, believe, cause, expect, find, give, help, like, move, produce, provide, seem, swing.*

[2] The importance of the manual analysis is outlined in Briscoe and Carroll (1997). We use the same manual analysis as Briscoe & Carroll, i.e. one from the Susanne, LOB, and SEC corpora. A manual analysis of the BNC data might produce better results. However, since the BNC is a heterogenous corpus we felt it was reasonable to test the data on a different corpus, which is also heterogenous.

	High Freq			Medium Freq			Low Freq			Totals		
	TP	FP	FN	TP	FP	FN	TP	FP	FN	TP	FP	FN
BHT	75	29	23	11	37	31	4	23	15	90	89	69
LLR	66	30	32	9	52	33	2	23	17	77	105	82
MLE	92	31	6	0	0	42	0	0	19	92	31	67

Table 1: Raw results for 14 test verbs

Method	Precision %	Recall %
BHT	50.3	56.6
LLR	42.3	48.4
MLE	74.8	57.8
baseline	24.3	83.5

Table 2: Precision and Recall

3.2 Results

Table 1 gives the raw results for the 14 verbs using each method. It shows the number of *true positives* (TP), *false positives* (FP), and *false negatives* (FN), as determined according to the manual analysis. The results for high frequency SCFs (above 0.01 relative frequency), medium frequency (between 0.001 and 0.01) and low frequency (below 0.001) SCFs are listed respectively in the second, third and fourth columns, and the final column includes the total results for all frequency ranges.

Table 2 shows precision and recall for the 14 verbs. We also provide the baseline results, if all SCFs were accepted.

From the results given in tables 1 and 2, the MLE approach outperformed both hypothesis tests. For both BHT and LLR there was an increase in FNs at high frequencies, and an increase in FPs at medium and low frequencies, when compared to MLE. The number of errors was typically larger for LLR than BHT. The hypothesis tests reduced the number of FNs at medium and low frequencies, however, this was countered by the substantial increase in FPs that they gave. While BHT nearly always acquired the three most frequent SCFs of verbs correctly, LLR tended to reject these.

While the high number of FNs can be explained by reports which have shown LLR to be over-conservative (Ribas,1995(b); Pedersen,1996), the high number of FPs is surprising. Although theoretically, the strength of LLR lies in its suitability for low frequency data, the results displayed in table 1 do not suggest that the method performs better than BHT on low frequency frames.

MLE thresholding produced better results than the two statistical tests used. Precision improved considerably, showing that the classes occurring in the data with the highest frequency are often correct. Although MLE thresholding clearly makes no attempt to solve the sparse data problem, it performs better than BHT or LLR overall. MLE is not adept at finding low frequency SCFs, however, the other methods are problematic in that they wrongly accept more than they correctly reject. The baseline, of accepting all SCFs, obtained a high recall at the expense of precision.

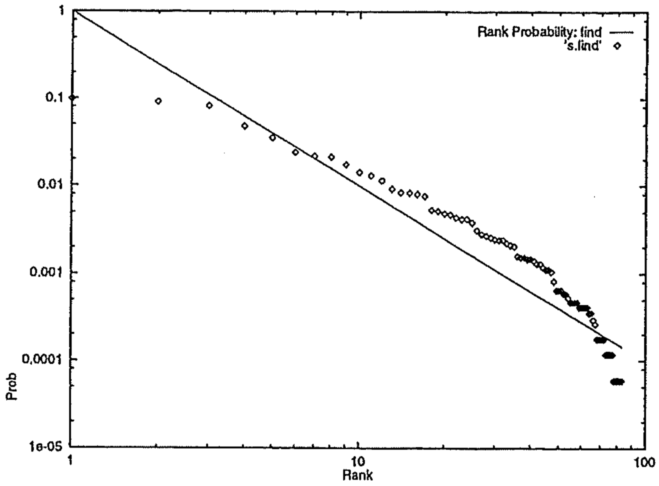

Figure 1: Hypothesised SCF distribution for *find*

3.3 Discussion

Our results indicate that MLE outperforms both hypothesis tests. There are two explanations for this, and these are jointly responsible for the results.

Firstly, the SCF distribution is zipfian. Figure 1 shows the conditional distribution for the verb *find*. This unfiltered SCF probability distribution was obtained from 20 M words of BNC data output from the SCF system. The unconditional distribution obtained from the observed distribution of SCFs in the 20 M words of BNC is shown in figure 2. The figures show SCF rank on the X-axis versus SCF frequency on the Y-axis, using logarithmic scales. The line indicates the closest Zipf-like power law fit to the data.

Secondly, the hypothesis tests make the false assumption (H_0) that the unconditional and conditional distributions are correlated. The fact that a significant improvement in performance is made by correcting the prior probabilities according to the performance of the system (Briscoe, Carroll & Korhonen, 1997) suggests the discrepancy between the unconditional and the conditional distributions.

Both LLR and BHT work by comparing the observed value of $p(scf_i|verb_j)$ to that expected by chance. They both use the observed value for $p(scf_i|verb_j)$ from the system's output, and they both use an estimate for the unconditional probability distribution ($p(scf_i)$) for estimating the expected probability. They differ in the way that the estimate for the unconditional probability is obtained, and the way that it is used in hypothesis testing.

For BHT, the null hypothesis is that the observed value of $p(scf_i|verb_j)$ arose by chance, because of noise in the data. We estimate the probability that the value observed could have arisen by chance using $p(m, n, p^e)$. p^e is calculated using:-

- the SCF acquisition system's raw (unfiltered) estimate for the unconditional distribution, which is obtained from the Susanne corpus and

- the ANLT estimate of the unconditional distribution of a verb not taking scf_i, across all SCFs

For LLR, both the conditional (p1) and unconditional distributions (p2) are estimated from the BNC data. The unconditional probability distribution uses the occurrence of scf_i with any verb other than our target.

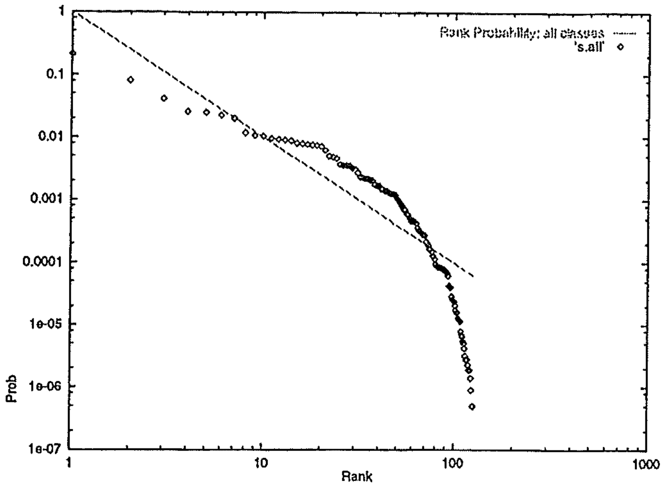

Figure 2: Hypothesised unconditional SCF distribution

The binomial tests look at one point in the SCF distribution at a time, for a given verb. The expected value is determined using the unconditional distribution, on the assumption that if the null hypothesis is true then this distribution will correlate with the conditional distribution. However, this is rarely the case. Moreover, because of the zipfian nature of the distributions, the frequency differences at any point can be substantial. In these experiments, we used one-tailed tests because we were looking for cases where there was a positive association between the SCF and verb, however, in a two-tailed test the null hypothesis would rarely be accepted, because of the substantial differences in the conditional and unconditional distributions.

A large number of false negatives occurred for high frequency SCFs because the probability we compared them to was too high. This probability was estimated from the combination of many verbs genuinely occurring with the frame in question, rather than from an estimate of background noise from verbs which did not occur with the frame. We did not use an estimate from verbs which do not take the SCF, since this would require a priori knowledge about the phenomena that we were endeavouring to acquire automatically. For LLR the unconditional probability estimate (p2) was high, simply because this SCF was a common one, rather than because the data was particularly noisy. For BHT, p^e was likewise too high as the SCF was also common in the Susanne data. The ANLT estimate went someway to compensating for this, thus we obtained fewer false negatives with BHT than LLR.

A large number of false positives occurred for low frequency SCFs because the estimate for $p(scf)$ was low. This estimate was more readily exceeded by the conditional estimate. For BHT false positives arose because of the low estimate of $p(scf)$ (from Susanne) and because the estimate of $p(\neg\text{SCF})$ from ANLT did not compensate enough for this. For LLR, there was no means to compensate for the fact that $p2$ was lower than $p1$.

In contrast, MLE did not compare two distributions. Simply rejecting the low frequency data produced better results overall by avoiding the false positives with the low frequency data, and the false negatives with the high frequency data. Interestingly, Lapata (1999) also used a threshold on the relative frequencies when establishing SCFs for diathesis alternation detection. The thresholds were determined for each SCF using the frequency of the SCF in COMLEX (Grishman et. al, 1994). She reported that these thresholds obtain slightly better results than those achieved with a Brent-style binomial filter.

4 Conclusion

This paper explored three possibilities for filtering out the SCF entries produced by a SCF acquisition system. These were (i) a version of Brent's binomial filter, commonly used for this purpose, (ii) the binomial log-likelihood ratio test, recommended for use with low frequency data and (iii) a simple method using a threshold on the MLEs of the SCFs output from the system. Surprisingly, the simple MLE thresholding method worked best. The BHT and LLR both produced an astounding number of FPs, particularly at low frequencies. Further work on handling low frequency data in SCF acquisition is warranted. A non-parametric statistical test, such as Fisher's exact test, recommended by Pedersen (1996), might improve on the results obtained using parametric tests. If the MLE thresholding still achieves better results, it would be worth investigating ways of handling the low frequency data, such as smoothing, for integration with this method. However, more sophisticated smoothing methods, which back-off to an unconditional distribution, may suffer from the lack of correlation between conditional and unconditional SCF distributions.

References

Boguraev, B., Briscoe, E., Carroll, J., Carter, D. & Grover, C. 1987. The derivation of a grammatically-indexed lexicon from the Longman Dictionary of Contemporary English. In *Proceedings of the 25th Annual Meeting of the Association for Computational Linguistics*, Stanford, CA. 193–200.

Brent, M. 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, E.J. and J. Carroll 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conf. on Applied Nat. Lg. Proc.*, Washington, DC. 356–363.

Briscoe, E., Carroll, J. & Korhonen, A. 1997. *Automatic extraction of subcategorization frames from corpora - a framework and 3 experiments.* '97 Sparkle WP5 Deliverable, available in http://www.ilc.pi.cnr.it/.

Carroll, G. & Rooth, M. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada, Spain.

Dunning, T. 1993. Accurate methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19.1: 61–74.

Gahl, S. 1998. Automatic extraction of subcorpora based on subcategorization frames from a part-of-speech tagged corpus. In *Proceedings of the COLING-ACL'98*, Montreal, Canada.

Garside, R., Leech, G. & Sampson, G. 1987. *The computational analysis of English: A corpus-based approach.* Longman, London.

Grishman, R., Macleod, C. & Meyers, A. 1994. Comlex syntax: building a computational lexicon. In *Proceedings of the International Conference on Computational Linguistics, COLING-94*, Kyoto, Japan. 268–272.

Lapata, M. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational*

Linguistics, Maryland. 397–404.

Leech, G. 1992. 100 million words of English: the British National Corpus. *Language Research* 28(1): 1–13.

Manning, C. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, Ohio. 235–242.

Manning, C. & Schütze, H. 1999. *Foundations of Statistical Natural Language Processing.* MIT Press, Cambridge MA.

Pedersen, T. 1996. Fishing for Exactness. In *Proceedings of the South-Central SAS Users Group Conference SCSUG-96*, Austin, Texas. .

Ribas, F. 1995. *On Acquiring Appropriate Selectional Restrictions from Corpora Using a Semantic Taxonomy* . Ph.D thesis, University of Catalonia.

Sampson, G. 1995. *English for the computer.* Oxford, UK: Oxford University Press.

Taylor, L. & Knowles, G. 1988. *Manual of information to accompany the SEC corpus: the machine-readable corpus of spoken English.* University of Lancaster, UK, Ms.

Ushioda, A., Evans, D., Gibson, T. & Waibel, A. 1993. The automatic acquisition of frequencies of verb subcategorization frames from tagged corpora. In Boguraev, B. & Pustejovsky, J. eds. *SIGLEX ACL Workshop on the Acquisition of Lexical Knowledge from Text.* Columbus, Ohio: 95–106.