Some statistical methods for evaluating information extraction systems

Will Lowe
Computer Science Department
Bath University
wlowe@latte.harvard.edu

Gary King
Center for Basic Research
in the Social Sciences
Harvard University
king@harvard.edu

Abstract

We present new statistical methods for evaluating information extraction systems. The methods were developed to evaluate a system used by political scientists to extract event information from news leads about international politics. The nature of this data presents two problems for evaluators: 1) the frequency distribution of event types in international event data is strongly skewed, so a random sample of newsleads will typically fail to contain any low frequency events. 2) Manual information extraction necessary to create evaluation sets is costly, and most effort is wasted coding high frequency categories .

We present an evaluation scheme that overcomes these problems with considerably less manual effort than traditional methods, and also allows us to interpret an information extraction system as an estimator (in the statistical sense) and to estimate its bias.

1 Introduction

This paper introduces a statistical approach we developed to evaluate information extraction systems used to study international relations. Event extraction is a form of categorization, but the highly skewed frequency profile of international event categories in real data generates severe problems for evaluators. We discuss these problems in section 3, show how to circumvent using a novel sampling scheme in section 4, and briefly describe our application. Finally we discuss the advantages and disadvantages of the methods, and their relations to standard evaluation procedure. We start with a brief review of information extraction in international relations.

2 Event Analysis in International Relations

Researchers in quantitative international relations have been performing manual information extraction since the mid-1970s (McClelland, 1978; Azar, 1982). The information extracted has remained fairly simple; a researcher fills a 'who did what to whom' template, usually from historical documents, a list of countries and international organizations to describe the actors, and a more or less articulated ontology of international events to describe what occurred (McClelland, 1978). In the early 1990s automated information extraction tools mostly replaced manual coding efforts (Schrodt et al., 1994). Information extraction systems in international relations perform a similar task to those competing in early Message Understanding Competitions (Sundheim, 1991, 1992). With machine extracted events data it is now possible to do near real-time conflict forecasting with data based on newswire leads, and detailed political analysis afterwards.

3 Event Category Distributions

We wanted to evaluate an information extraction system from Virtual Research Associates[1]. This system bundles extraction and visualization software with a custom event ontology containing, at last count, about 200 categories of international event.

We found two problems with the nature of international events data. First, the frequency distribution over the system's ontology, or indeed several other ontologies we considered, is heavily skewed. A handful of mostly diplomatic event types predominate, and the frequency of other event types falls of very sharply: we ran the system over all the newsleads in Reuters' coverage of the Bosnia conflict, and of the approximately 45,000 events it extracted, 10,605 were in the category of 'neutral comment', 4 of 'apology' and 35 of 'threat of force'. Thus the relative frequencies of event categories in this data can be 2,500 to 1.

Also, as these figures suggest, the more interesting and politically relevant events tend to be of low frequency. This problem is quite general in categorization systems with reasonably articulated category systems, and not specific to international relations. But any dataset with these properties causes an immediate problem for evaluation.

Ideally we would choose a random subset of leads whose events are known with certainty (because we have coded them manually beforehand), run the system over them, and then compute various sample statistics such as precision and recall[2]. However, a small randomly chosen subset is very unlikely contain instances of most interesting events, and so the system's performance will not be evaluated on them. Given the possible frequency ratios above, the size of subset necessary to ensure reasonable coverage of lower frequency event categories is enormous. Put more concretely, to construct a test set of news leads the evaluator will on average have to code around 2,500 comments to reach a single apology and about 300 comments to find a single threat of force.

[1] `http://www.vranet.com`

[2] This paper only evaluates extraction performance on event types, though there would seem to be no reason why a similar approach would not work for actors etc.

3.1 Standard Evalution Methods

The standard evaluation methods developed over the course of the Message Understanding Competitions consist mainly in sample statistics to compute over the evaluation materials e.g. precision and recall, but do not give any guidance for choosing the materials themselves (Cowie and Lehnert, 1996; Grishman, 1997). This is just done by hand by the judges. Perhaps because the selection question is neglected, it is seldom clear what larger population the test materials are from (save that it is the same one as the training examples), and as a consequence it is unclear what the implications for generalization are when a system obtains a particular set of scores for precision and recall (Lehnert and Sundheim, 1991).

Since this literature did not help us generate a suitable evaluation sample, we approached the problem from scratch, and developed a statistical framework specific to our needs.

4 Method

One reasonable-sounding but *wrong* way to address the problem of creating a test set without having to code tens of thousands of irrelevant stories is the following:

1. Use the extraction system itself to perform an initial coding,
2. Take a sample of the output that covers all the event types in reasonable quantities,
3. Examine each coding to see whether the system assigned the correct event code.

This looks like it can guarantee a good sample of low frequency events at much lower cost to the manual coder; we can just pick a fixed number of events from each category and evaluate them. However, this method exhibits *selection bias*. To see this, let M and T be variables indicating which event category the Machine (that is, the information extraction system) codes an event into, and the True category to which the event actually belongs. Statistically, the quantity of interest to us is the probability that the machine is *correct*:

$$\mathrm{P}(M = i \mid T = i) \tag{1}$$

This is the probability that the machine classifies an event into category i given that the true event coding is indeed i. A full characterization of the success of the machine requires knowing $\mathrm{P}(M = i \mid T = i)$ for $i = 0, \ldots, J$, which includes all J event categories and where $i = 0$ denotes the situation where the machine is unable to classify an event into any category. In short, the quantity of interest is the full probability density $\mathrm{P}(M \mid T)$.

In statistical terms, this distribution is a *likelihood function* for the information extraction system. This observation allows us to treat the system like any other statistical estimator and offers the interesting possibility of analyzing generalization via its sampling properties, e.g. its bias, variance, mean squared error, or risk.

Unfortunately, the problem with the reasonable-sounding approach described above is that it does not in fact allow us to estimate $\mathrm{P}(M \mid T)$ because it is implicitly conditioning on M, not T. In particular, the proportion of events that are actually in category i among those the machine *put* in category i gives us instead an estimate of

$$\mathrm{P}(T \mid M) \tag{2}$$

which is not the quantity of interest. (2) is the probability of the truth being in some event category rather than the machine's response whereas in fact the true event category is fixed and it is the machine's response that is uncertain[3]. Worse, $\mathrm{P}(T \mid M)$ is a systematically biased estimate of $\mathrm{P}(M \mid T)$ because these two quantities are related by Bayes theorem:

$$\mathrm{P}(M \mid T) = \frac{\mathrm{P}(M,T)}{\mathrm{P}(T)} = \frac{\mathrm{P}(T \mid M)\mathrm{P}(M)}{\mathrm{P}(T)}, \tag{3}$$

and the only circumstances under which they would be equal is when $\mathrm{P}(M)$ is uniform. But the figures in section 3 suggest that $\mathrm{P}(M)$ is highly skewed.

However this last observation suggests a better method for unbiased estimation of (1).

1. Estimate $\mathrm{P}(T \mid M)$ as described above
2. Compute $\mathrm{P}(M)$ by running the system over the *entire* data set and normalizing the frequency histogram of event categories
3. Estimate $\mathrm{P}(M \mid T)$ by correcting $\mathrm{P}(T \mid M)$ with $\mathrm{P}(M)$ using Bayes theorem

Our implementation of this scheme was to first run the system over 45,000 leads about the Bosnia conflict, and normalize the frequency histogram of events extracted to create $\mathrm{P}(M)$. Then, randomly choose 5 leads assigned to each event category, and manually determine which event type the instantiate. Then normalize to estimate $\mathrm{P}(T \mid M)$. And finally, use (3) to create $\mathrm{P}(M \mid T)$. We chose four times as many uncategorized leads as from each true category in addition. A larger sample here is advisable to see what sort of categories the system misses. These sample sizes are fixed, but it may also be possible to use active learning techniques to tune them (as in e.g. Argamon-Engelson and Dagan, 1999) for even more efficient sampling.

The advantage of this roundabout route to (1) is that it requires many fewer events to be manually coded. We ran the system over 45,000 leads but only manually coded a handful of events for each category. This guaranteed us even coverage of the lowest frequency event categories whilst not biasing the end result – for an ontology with about 200 categories this is a substantial decrease in evaluator effort.

This method works by making use of the extraction system itself to produce one important marginal: $\mathrm{P}(M)$. If we assume that the aim is to evaluate the system on the Bosnia conflict, $\mathrm{P}(M)$ is not estimated, but is rather an exact population marginal[4]. Then we can guarantee that our estimate of $\mathrm{P}(M \mid T)$ is unbiased because the method for estimating $\mathrm{P}(T \mid M)$ is clearly unbiased, and $\mathrm{P}(M)$ adds no error.

4.1 Summary Measures

$\mathrm{P}(M \mid T)$ allows the computation of a number of useful summary measures[5]. For example, we

[3] This is due to changes in the journalist's choice of vocabulary and syntactic construction that are uncorrelated with the identity of the event being described.

[4] We might consider the Bosnian conflict to be a sample point from the larger population of all wars, but that population – if it exists at all – is certainly difficult to quantify.

[5] Detailed discussion of several summary measures for the system we evaluated can be found in King and Lowe (2002).

can easily compute $\mathrm{P}(M,T)$ from quantities already available, so $\sum^J \mathrm{P}(M=i,T=i)$ is the proportion of time the system extracts the correct category. Alternatively, if it is more important to extract some categories than others, then various weighted measures can be constructed e.g. $\sum^J \mathrm{P}(M=i \mid T=i)w_i$ where ws are non-negative and sum to 1, representing the relative importance of extracting each category. Some more graphical methods of evaluation using $\mathrm{P}(M \mid T)$ are presented below.

4.2 Estimator Properties

Given a likelihood function for the extraction system we can investigate its properties as an estimator. It is particularly useful to know the *bias* of an estimator, defined in this case as the difference between the expected category response from the system when the true event category is i, and i itself, where the expectation is taken of repeated information extraction tasks that instantiate the same event categories. We do not examine the corresponding variance here, and a more complete evaluation might also address the question of consistency.

4.2.1 Conflict and Cooperation

The machines response and the true category is best seen as a set of multinomial probabilities (with a unit vector with the value 1 at the index of the system's extracted category or the true category respectively. Estimator properties are cumbersome to represent in this format, so here we map the system's response to a single real value corresponding to the level of conflict or cooperation of the event category. This re-representation is usual in international relations and allows standard econometric time series methods to be applied (Schrodt and Gerner, 1994; Goldstein and Freeman, 1990; Goldstein and Pevehouse, 1997).

For our purposes it also allows the straightforward graphical presentation of the main ideas. We define the level of conflict or cooperation level of an event category i as G_i, a real number between -10 (most conflictual) to 10 (most cooperative) (see Goldstein, 1992, for the full mapping). For example, according to this scheme, when i denotes the event category 'extending economic aid', $G_i = 7.4$, 'policy endorsement' maps to 3.6, 'halt negotiations' maps to -3.8, and a 'military engagement' maps to -10, the maximally conflictual event. The mapping allows univariate, and politically relevant comparison between the true conflict level and that of the event categories the system extracts.

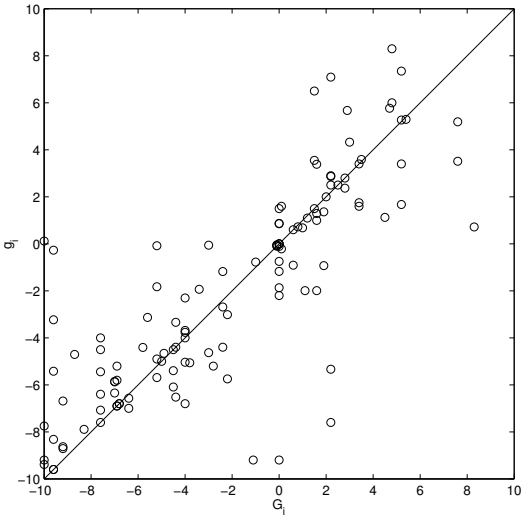

Figure 1: Expected (g_i) versus true (G_i) conflict-cooperation level for each event category.

The expected system response when the true category has conflict/cooperation level G_i is:

$$g_i = \sum^J G_j \mathrm{P}(M=j \mid T=i, M \neq 0) \qquad (4)$$

where

$$\mathrm{P}(M=j \mid T=i, M \neq 0) = \frac{\mathrm{P}(M \mid T)\mathbf{1}(M \neq 0)}{\mathrm{P}(M \neq 0 \mid T)}.$$

and $\mathbf{1}(M \neq 0)$ is an indicator function equaling 1 if $M \neq 0$ and 0 otherwise.

A plot of G_i against g_i for each event category is shown in Figure 1. An unbiased estimator would show expected values on the main diagonal. Estimator bias for event category i is simply $g_i - G_i$. Estimator variance is simply the spread around the diagonal.

4.3 Comparison

We also compared the system's performance to 3 undergraduate coders (U1-3) working on the same data set. To examine undergraduate performance requires first $\mathrm{P}(U,T)$, from which we can

get $P(U \mid T)$. However, we cannot simply count the proportion of times each undergraduate assigns a lead to category i when it is in fact in category i because this ignores the fact that we have sampled the leads themselves using the system, and must therefore condition on M. On the other hand we do have access to the relevant conditional distribution $P(U,T \mid M = i)$. This is the distribution of undergraduate and true categories, conditioned on the fact the the system assigns an event to category i. The desired $P(U,T)$ is a weighted average of these distributions:

$$P(U,T) = \sum_i P(U,T \mid M = i)P(M = i).$$

$P(U \mid T)$ is then obtained by marginalization[6]. Clearly these calculations can also be used to compare other systems with the same ontology using the same materials.

Summary statistics similar to those described above can be easily computed (King and Lowe, 2002). Here we provide graphical results: Figure 2 plots the bias of the system and that of the undergraduates over the category set (with smoothed estimates superimposed). In the figure, the bias $G_i - g_i$ is plotted against G_i, so deflections from the horizontal are systematic bias. In almost all cases we find that more conflictual (negative valued) categories are mistaken for more cooperative ones, with some suggestion of a similar effect at the cooperative end too. Of most interest is the basic similarity in performance between undergraduates and the information extraction system.

It would be helpful if the bias that appears in these plots were systematically related to the expected system response. If this was the case, in future use we could simply adjust the system's response up or down by some coefficient determined in the evaluation process and remove the bias. However, figure 3 shows that there is no systematic relation between the expected reponses and the level of bias, so no such coefficient canbe computed. This is a rather pessimistic result for this system, suggesting a level of bias that cannot be straightforwardly removed. On the other

[6]We would normally expect to use $P(U \mid T, U \neq 0)$, but the undergraduates never failed to assign categories.

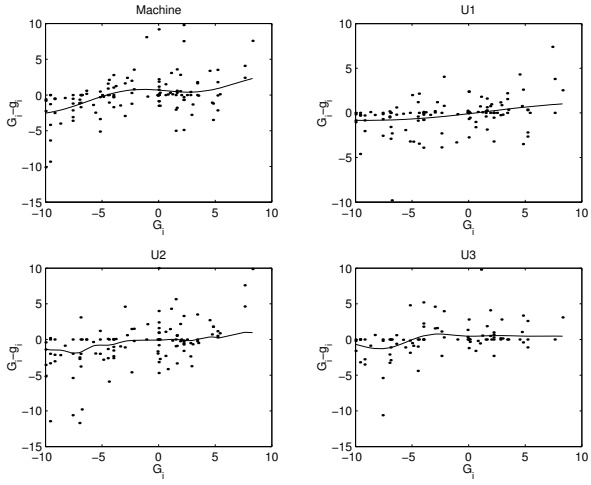

Figure 2: System (M) versus undergraduate coder (U1-3) bias. Connected lines are generated by smoothing $G_i - g_i$.

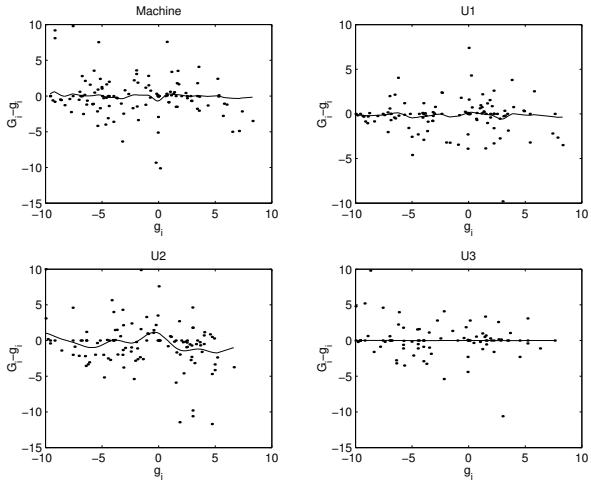

Figure 3: Bias plotted against expected system and undergraduate response. Deviations from the horizontal suggest the possibility of a post-output correction to correct for bias in subsequent application.

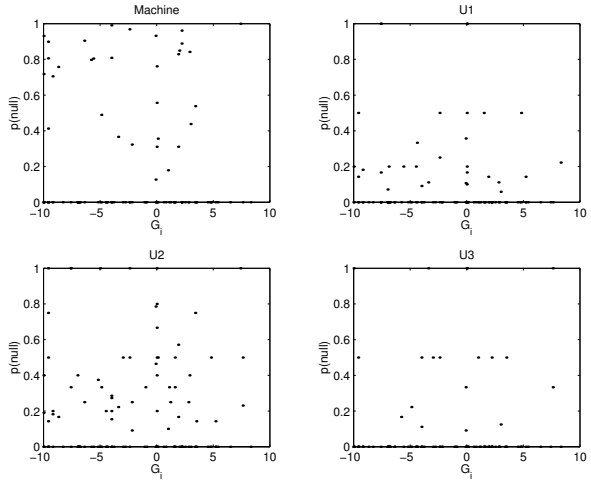

Figure 4: The probability that the system, or undergraduate fails to assign an event to a category, plotted against the level of conflict/cooperation of that category.

hand, one of the advantages of the methods presented here is that this bias is now *estimated*, and, since bias estimates are available on a category-by-category basis, redesigning effort can be directed in a way that maximizes generalization performance.

Finally, figure 4 plots the probability that the machine failed to assign an event category, $\mathrm{P}(M = 0 \mid T = i)$ (denoted p(null) in the figure), as a function of that category's conflict/cooperation value, G_i. Our interest in G_i reflects the use this data is typically put to, since we are most concerned with errors that make the world look systematically more (or less) cooperative than it really is. But we might equally have plotted $\mathrm{P}(M = 0 \mid T = i)$ against i itself, or any other property of events that might be suspected to generate difficult to categorize event descriptions.

Like the previous figures, plotting $\mathrm{P}(M = 0 \mid T = i)$ against other quantities is a useful diagnostic, indicating where future work should best be applied. In this case there appears to be no systematic relationship between the true level of conflict/cooperation and the probability that either the system or the undergraduates will fail to assign the event to a category.

5 Conclusion

We have presented a set of statistical methods for evaluating an information extraction system without unreasonable manual labour when the distribution of categories to be extracted is heavily skewed. The scheme uses a form of biased sampling and subsequent correction to estimate a probability distribution of system responses for each true category in the data. This distribution costitutes a likelihood function for the system. We then show how functions of this distribution can be used for evaluation, and estimate the system's statistical bias.

The two main ideas: using estimates of $\mathrm{P}(M \mid T)$ as the basis for evaluation, and using a non-standard sampling scheme for the estimation, are separate. Emphasis on using $\mathrm{P}(M \mid T)$ comes from standard statistical theory, and if correct, suggests how evaluation in information extraction might be integrated in to that body of theory. When a sample of leads is randomly chosen and can be expected to be reasonably representative, then the sampling machinery described above, the computation of $\mathrm{P}(M)$, and the application of Bayes theorem will not be necessary. But when the distribution of categories to be extracted is so highly skewed then our method is the only one that will make it feasible to evaluate a system on *all* of its categories in an unbiased way.

The principle difference between these and standard evaluation methods is in our explicitly statistical framework, and our consideration of how to sample in a representative way, and methods to get around cases where we cannot. The exact relationship to precision, recall etc. is the topic of current research. In the meantime we hope that the methods presented might advance understanding of effective evaluation methods in computational linguistics.

Acknowledgments

We thank Doug Bond, Craig Jenkins, Dylan Balch-Lindsay, Phil Schrodt, and two anonymous reviewers for helpful comments, and the National Science Foundation (IIS-9874747), the National Institutes of Aging (P01 AG17625-01), the Weatherhead Center for International Relations, and the

World Health Organization for research support.

References

Argamon-Engelson, S. and Dagan, I. (1999). Committee-based sample selection for probabilistic classifiers. *Journal of Artificial Intelligence Research*, 11:335–360.

Azar, E. E. (1982). *Codebook of the Conflict and Peace Databank*. Center for International Development, University of Maryland.

Cowie, J. and Lehnert, W. (1996). Information extraction. *Communications of the ACM*, 39(1):80–91.

Goldstein, J. S. (1992). A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36(2).

Goldstein, J. S. and Freeman, J. R. (1990). *Three-Way Street: Strategic Reciprocity in World Politics*. Chicago University Press.

Goldstein, J. S. and Pevehouse, J. C. (1997). Reciprocity, bullying and international conflict: Time-series analysis of the Bosnia conflict. *American Political Science Review*, 91(3):515–529.

Grishman, R. (1997). Information extraction: Techniques and challenges. In Pazienza, M. T., editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, volume 1299 of *Lecture Notes in Artificial Intelligence*, chapter 2, pages 10–27. Springer Verlag.

King, G. and Lowe, W. (2002). An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. http://gking.harvard.edu/infoex.pdf.

Lehnert, W. and Sundheim, B. (1991). A performance evaluation of text-analysis technologies. *AI Magazine*, pages 81–95.

McClelland, C. (1978). *World Event / Interaction Survey (WEIS) 1966-1978*. Inter-University Consortium for Political and Social Research, University of Southern California.

Schrodt, P. A., Davis, S. G., and Weddle, J. L. (1994). Political science: KEDS — a program for the machine coding of event data. *Social Science Computer Review*, 12.

Schrodt, P. A. and Gerner, D. J. (1994). Validity assessment of a machine-coded event data set for the Middle East, 1982-92. *American Journal of Political Science*, 38(3).

Sundheim, B. (1992). Overview of the fourth message understanding evaluation and conference. In *Proceedings of the Fourth Message Understanding Conference*, pages 3–22.

Sundheim, S., editor (1991). *Proceedings of the Third Message Understanding Conference*, San Mateo, CA. Morgan Kaufmann.