IMPROVED TOPIC SPOTTING THROUGH STATISTICAL MODELLING OF KEYWORD DEPENDENCIES

Jerry H.Wright, Michael J.Carey and Eluned S.Parris

Ensigma Ltd, Turing House, Station Road, Chepstow, NP6 5PB, U.K.

ABSTRACT

Keywords are chosen on the basis of their usefulness for discriminating a topic from background speech. Good topic recognition can be achieved with a small set of well-chosen keywords, but particular combinations of keywords often achieve better discrimination than can be accounted for by regarding them as independent. This paper describes a higher-order statistical approach involving models of keyword-topic interdependence. A linear-logistic model brings some improvement in performance, but better results are obtained using log-linear contingency table models. Although the potential number of these is very large, good models tend to be simple and are suggested by heuristic measures inferred from the training data. The approach is tested using a broadcast radio database.

1. INTRODUCTION AND KEYWORD SELECTION

Our aim is to use a small set of well-chosen keywords for discriminating a topic from background speech, taking account of the fact that keywords in combination are more powerful than they are independently. Criteria for keyword selection are plausibly based on measures of frequency, information or both [1,2]. The measure used here is simple to apply and has also been employed for selecting phonemes for speaker identification [3]. Distributions of counts of the chosen keywords in windows of speech can then be trained and put to use. A multinomial distribution or mixture [2] treats the keywords as occurring independently, and this limitation is partially lifted by logistic discrimination models [2]. In this paper we compare this with more powerful models of dependence between keywords and topic.

When selecting the keywords we treat them as independent. Let T denote topic and consider the string of spotted keyword occurrences for the topic training data, of total length N. If this contains n_k occurrences of the kth keyword w_k (some may be false positives), then the relative probability for the topic data is $P(w_k \mid \mathrm{T}) = n_k/N$. Similarly the probabilities for non-topic data $P(w_k \mid \overline{\mathrm{T}})$ can be found, and the log-likelihood ratio for the keyword data (given N) follows from the multinomial distribution:

$$\log \frac{P(n_1, n_2, \cdots \mid \mathrm{T})}{P(n_1, n_2, \cdots \mid \overline{\mathrm{T}})} = \sum_k n_k \log \frac{P(w_k \mid \mathrm{T})}{P(w_k \mid \overline{\mathrm{T}})}$$

$$= N \sum_k P(w_k \mid \mathrm{T}) \log \frac{P(w_k \mid \mathrm{T})}{P(w_k \mid \overline{\mathrm{T}})}$$

Keywords are ranked on the basis of their relative contribution to the discrimination for the topic, described as "usefulness" [3]:

$$P(w_k \mid \mathrm{T}) \log \frac{P(w_k \mid \mathrm{T})}{P(w_k \mid \overline{\mathrm{T}})}$$

Keyword selection therefore depends both on the relative frequency of occurrence in topic and non-topic speech and on the absolute frequency of occurrence in topic speech. The highest-ranked keywords are taken forward for training and scoring.

2. SCORING DISTRIBUTION

The speech is segmented into windows, of typically 60 sec duration with 30 sec overlap. A window-scoring criterion can be based on an accumulation of likelihood ratio over all spotted keywords within the window [4], where the probabilities are derived from actual word spotting rates rather than from text [3], and the decision between topic and non-topic is controlled by a threshold to generate an operating characteristic (ROC) curve.

An improved criterion is based on the smoothed one-dimensional distributions of keyword counts within the window: $P_{t,k}(n_k)$ for count $n_k = 0,1,\cdots$ occurrences of the kth keyword, and from now on we use the subscript $t = 0$ or 1 to denote conditionalisation on non-topic or topic respectively. The likelihood ratio for a new window assuming independent keywords is then given by

$$\frac{P_1(n_1, \cdots, n_K)}{P_0(n_1, \cdots, n_K)} = \prod_{k=1}^{K} \frac{P_{1,k}(n_k)}{P_{0,k}(n_k)} \qquad (1)$$

where K is the number of keywords engaged. Once again, the topic/non-topic decision is controlled by applying a threshold to this quantity, yielding an ROC curve.

0-7803-2431-5/95 $4.00

The keyword count distributions are estimated from training data, and smoothed using a Poisson mixture distribution [5] given by

$$\left(\frac{c_{t,k}}{c_{t,k}+1}\right)^{a_{t,k}} \frac{a_{t,k}(a_{t,k}+1)\cdots(a_{t,k}+n-1)}{n!(c_{t,k}+1)^n}, \quad n=0,1,2,\cdots$$

with mean and variance

$$\mu_{t,k}=\frac{a_{t,k}}{c_{t,k}}, \quad \sigma^2_{t,k}=\frac{a_{t,k}}{c_{t,k}}+\frac{a_{t,k}}{c^2_{t,k}}$$

The parameters $a_{t,k}, c_{t,k}$ are easily identified from the measured mean and variance. The usual Poisson distribution is a limiting case, but the mixture achieves a far superior fit to the keyword data because the mean and variance are separately controllable, and typically the variance of the keyword count is higher than the mean. The smoothing parameters are assigned optimally without reference to test data, using a minimum mean-square error criterion [6], for each keyword and for $t = 0,1$. This model (independent keywords) is the baseline model for comparison with the higher-order procedures.

3. MODEL DEVELOPMENT

One way to produce a likelihood ratio directly from keyword counts, and that permits keyword dependence, is to use the linear-logistic model

$$\log\frac{P(\mathrm{T}\,|\,n_1,\cdots,n_K)}{P(\overline{\mathrm{T}}\,|\,n_1,\cdots,n_K)}=\sum_{k=0}^{K}\beta_k n_k$$

from which

$$\frac{P_1(n_1,\cdots,n_K)}{P_0(n_1,\cdots,n_K)}=\exp\left(\sum_{k=0}^{K}\beta_k n_k\right)\frac{P(\overline{\mathrm{T}})}{P(\mathrm{T})} \qquad (2)$$

where $n_0=1$, and $\beta_0,\cdots,\beta_K$ are coefficients that can be estimated using the Newton-Raphson algorithm to maximise the training data likelihood [7]. This model is lacking in structure, so we aim for a framework in which a comprehensive set of dependencies can be explored.

Because of the very large number of potential combinations of keyword counts, it is useful to employ a set of reduced variables $r_1,\cdots,r_K$ with smaller range, based on quantiles of the distributions. We then have

$$\frac{P_1(n_1,\cdots,n_K)}{P_0(n_1,\cdots,n_K)}\approx\frac{P_1(r_1,\cdots,r_K)}{P_0(r_1,\cdots,r_K)}\prod_{k=1}^{K}\frac{P_{1,k}(n_k\,|\,r_k)}{P_{0,k}(n_k\,|\,r_k)}$$

This separates the broad structure of dependence between keywords and topic from the fine structure of the individual counts. As the reduced variables become coarser, some of the dependence information is lost but the training problem is eased.

During training, a table of counts of windows for each combination of values of the reduced variables is accumulated for $t = 0,1$. Let $f_t(r_1,\cdots,r_K)$ be the measured count, and $F_t(r_1,\cdots,r_K)$ be the predicted count on the basis of a fitted model. A standard methodology [6,7] for the analysis of tables of this kind is to equate the log of the latter quantity to a linear sum of parameters (log-linear models), the parameters are then identified and the fit of the model assessed. Let the K keywords be labelled A,B,..., and the topic T, and consider the following model:

$$\log F_t(r_1,\cdots,r_K)=u+u^{\mathrm{T}}_t+u^{\mathrm{A},\mathrm{T}}_{r_1,t}+u^{\mathrm{B},\mathrm{T}}_{r_2,t}+\cdots+u^{K,\mathrm{T}}_{r_K,t}+ \\ +u^{\mathrm{A},\cdots,K}_{r_1,\cdots,r_K}+\text{implied lower-order terms} \qquad (3)$$

Here, u is an overall mean, u^{T}_t allows for different frequencies of topic and non-topic windows such that $u^{\mathrm{T}}_0+u^{\mathrm{T}}_1=0$, each term $u^{X,\mathrm{T}}_{r_k,t}$ (where keyword X has count value r_k) allows for dependence between that keyword and topic, and $u^{\mathrm{A},\cdots,K}_{r_1,\cdots,r_K}$ allows for full interdependence between keywords. All terms except u sum to zero over each of their subscripts. The implication of lower-order terms refers to the requirement that for each term involving two or more variables, all terms involving subsets of those variables must also be present. This particular model corresponds to the linear-logistic model (2) modified to the reduced variables, because

$$\frac{P_1(r_1,\cdots,r_K)}{P_0(r_1,\cdots,r_K)}=\frac{F_1(r_1,\cdots,r_K)}{F_0(r_1,\cdots,r_K)}\times\frac{W_0}{W_1} \\ =\frac{W_0}{W_1}\exp\left(u^{\mathrm{T}}_1-u^{\mathrm{T}}_0+u^{\mathrm{A},\mathrm{T}}_{r_1,1}-u^{\mathrm{A},\mathrm{T}}_{r_1,0}+\cdots+u^{K,\mathrm{T}}_{r_K,1}-u^{K,\mathrm{T}}_{r_K,0}\right)$$

Thus, $\beta_0=2u^{\mathrm{T}}_1$, $\beta_1 r_1=2u^{\mathrm{A},\mathrm{T}}_{r_1,1}$, ..., $\beta_K r_K=2u^{K,\mathrm{T}}_{r_K,1}$. W_1 and W_0 are the total numbers of topic and non-topic windows seen in training.

Log-linear models can range between total independence in which the only term on the RHS is u, and saturation in which a term $u^{\mathrm{A},\cdots,K,\mathrm{T}}_{r_1,\cdots,r_K,t}$ plus all lower-order terms ensure that the model fits exactly to the table. We use the iterative proportional fitting algorithm [6,7] to get the best fit for each model and find the parameter values, and the fit is assessed by the likelihood-ratio statistic

$$G^2=2\sum_{r_1}\cdots\sum_{r_K}\sum_t f_t(r_1,\cdots,r_K)\log\frac{f_t(r_1,\cdots,r_K)}{\hat{F}_t(r_1,\cdots,r_K)}$$

($\hat{F}_t(r_1,\cdots,r_K)$ is the expected count for the fitted model).

Because the keyword count distributions differ between topic and non-topic situations, we actually use different

sets of reduced variables $r_{1,1},\cdots,r_{1,K}$ and $r_{0,1},\cdots,r_{0,K}$ respectively, and then the likelihood ratio is given by

$$\frac{P_1(n_1,\cdots,n_K)}{P_0(n_1,\cdots,n_K)} \approx \frac{P_1(r_{1,1},\cdots,r_{1,K})}{P_0(r_{0,1},\cdots,r_{0,K})} \prod_{k=1}^{K} \frac{P_{1,k}(n_k \mid r_{1,k})}{P_{0,k}(n_k \mid r_{0,k})} \qquad (4)$$

where the first term on the RHS is obtained directly from the estimated log-linear model, and the remaining terms from the smoothed independent distributions.

4. TEST ON BROADCAST RADIO DATABASE

4.1 Weather forecast keywords

To test the topic-spotting models we use a database of 48 hours of speech, recorded in-house from BBC Radio 4. Weather forecasts typically occur as 2 to 3 minute items every few hours, and the ten best keywords (in terms of the criterion described in section 1) were found to be A:temperature, B:northern, C:showers, D:weather, E:England, F:Scotland, G:Ireland, H:sunshine, I:degrees, J:tomorrow. Using the first half of the data for training, and running windows of 60 sec duration (with 30 sec overlap) over the spotted keywords, we found 82 topic and 3526 non-topic windows. In view of the small number of topic windows we use a binary reduced variable for each keyword:

$$r_{t,k} = \begin{cases} 1 & \text{if } n_k > m_{t,k} \\ 0 & \text{otherwise} \end{cases}$$

where $m_{t,k}$ is the median count for keyword k, topic $t = 0,1$. Even then, with 10 keywords there are 1024 topic cells for all combinations in the table, and most are empty. In order to fit the log-linear models we do a small amount of smoothing using the independent distributions.

A convenient way to specify models is by groups of items, with each group corresponding to a term plus all implied lower-order terms in the log-linear model. The linear-logistic type model (3) is then specified as

AT/BT/ ... /JT/ABCDEFGHIJ

indicating that the topic depends upon each keyword separately (the first ten groups) and that the keywords have full interdependence (final group). Here there is no higher-order interaction between keywords and topic, and we explore this interaction by specifying different models. The number of possible models, where the groups span the variables and are allowed to overlap but with none a subset of any other, is very large: we have devised a lower bound which implies that there exist at least 2×10^{26} possible models for 10 keywords. A heuristic procedure is therefore needed to search for good models.

The term AB...J (or in general AB...K) detaches the keyword interdependence, so that the rest of the model can focus on the keyword-topic dependence. One natural way to search for higher-order effects is to test models of the form XYT/AB...J for pairs of keywords X,Y. Although the statistic G^2 has an asymptotic chi-square distribution (as for the Pearson statistic X^2), this hardly applies here because of the sparseness of the table, and its skewness, with a small number of cells containing most of the observations. If instead we compare two models by

$$G^2(\mathrm{XT/YT/AB\cdots J} \mid \mathrm{XYT/AB\cdots J}) = G^2(\mathrm{XT/YT/AB\cdots J}) - G^2(\mathrm{XYT/AB\cdots J})$$

where the first model is a special case of the second (the third-order term being absent) then this has two advantages [7]. First, the chi-square approximation is much better because the figure effectively depends only on the marginal totals in the table, which tend to have higher values than the cell entries, and second the test is more powerful because there are less degrees of freedom. In fact by the general collapsibility conditions for multi-dimensional tables [6] it is equivalent (and easier) to perform this test on the $2\times2\times2$ table for X,Y,T.

The top-ranked keyword pairs by this criterion are found to be BG, AB, BF, FG, IJ, FI, FJ and so on, but it turns out that these are not the most useful combinations in scoring. Note that the combination BG is "Northern Ireland", and this occurs even more frequently in news and current-affairs broadcasts than in weather forecasts, and is not therefore a good discriminator for the latter. A better guide is given by the statistic

$$G^2(\mathrm{AT/BT/\cdots/JT/AB\cdots J} \mid \mathrm{XYT/AT/\cdots/JT/AB\cdots J})$$

This tests the pair X,Y with the remaining keywords also present, independently associated with topic, in comparison with the linear-logistic model. This time the top-ranked pairs include AJ ("temperature tomorrow", chi-square P-value 0.046) and FH ("Scotland sunshine", P-value 0.141), and these prompt a good scoring model.

4.2 Spotting performance

The models are tested on the held-out half of the data, using a minimum length of 120 sec both for a correctly-spotted topic and for a false alarm. Figure 1 contains ROC curves for

1. the baseline independent case (equation (1), corresponding to the log-linear model AT/BT/.../JT),
2. the linear-logistic model (2),
3. the table of combinations of reduced counts, with light smoothing applied (equivalent to the saturated log-linear model ABC...JT),
4. the log-linear model
   AJT/CFHT/BT/DT/ET/GT/IT/AB...J (5)

Equation (4) is used for 3 and 4. The curves are not

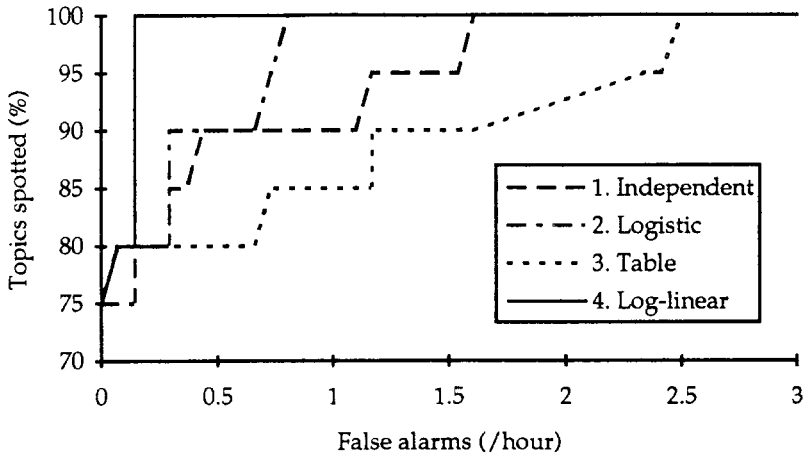

Figure 1: ROC curves for weather topic spotting.

smooth because of the finite number of topics in the test data. The poor performance for the smoothed data table (curve 3) is attributable to lack of training data. Of the reduced-variable keyword combinations seen in the test data, 96.1% for non-topic occur at least once in training but only 17.2% for the topic. Window-scoring using topic statistics is therefore based primarily (83% of the time) on smoothing. On the other hand, the independence model is better trained but ignores the higher-order effects.

A carefully-chosen log-linear model intermediate between independence and exact fit can represent the dominant higher-order effects while still being sufficiently thin in parameters to permit adequate training. The table cell probability estimators, obtained from the model and used for scoring windows, have lower mean square error than the sample proportions inferred from the data table. This lifts the performance above the other models, including the linear-logistic model.

The model (5) also works well when training and test data are interchanged, and various other models also give similar results.

With less than about 80 topic windows in the training data, performance of the log-linear model declines more rapidly than that of the linear-logistic model, but good results can still occur. For example there are 14 financial reports of at least one minute duration in the database, (21 60-sec topic windows, best keywords include Hundred, Market, Sterling and Wallstreet). Using 8 of these reports for training (8 topic windows), the best log-linear model spots the remaining 6 topics with a false-alarm rate of 0.32 per hour compared with 2.03 per hour for the independence model. However, this performance is very fragile with such light training and the training data heuristics are not a reliable guide in this case.

5. CONCLUSIONS

Creating a table of keyword-count combinations (in reduced form) enables the methodology of log-linear contingency table models to be put to use. This first reveals the combinations of keywords that are most indicative for discriminating a topic from general background, and then provides a convenient decision procedure for windows of speech.

Although the potential number of models is enormous, the best models tend to be relatively simple and can be discovered by heuristic methods, assuming statistical consistency between training and test data. Scoring is very fast, and performance is far superior to that obtained assuming independent keywords, with much less training data than would be required for accurate estimation of multivariate keyword count distributions.

Clearly this approach is best suited to relatively homogeneous topics characterised by small numbers of good keywords. Other topics are more heterogeneous, but the approach described here may still be applied if the topic is first partitioned into homogeneous sub-topics.

Acknowledgement

This work is supported by the Royal Society.

References

[1] A.L.Gorin, S.E.Levinson and A.Sankar, "An experiment in spoken language acquisition", IEEE Trans. on Speech and Audio, vol 2, 1994, pp 224-240.

[2] J.McDonough and H.Gish, "Issues in topic identification on the Switchboard corpus", Proc. ICSLP-94, Yokohama, pp 2163-2166.

[3] E.S.Parris and M.J.Carey, "Discriminative phonemes for speaker identification", Proc. ICSLP-94, Yokohama, pp 1843-1846.

[4] R.C.Rose, E.I.Chang and R.P.Lippman, "Techniques for information retrieval from voice messages", Proc. ICASSP-91, pp 317-320.

[5] M.G.Kendall and A.Stuart, "The Advanced Theory of Statistics, vol. 1", 3rd ed., Charles Griffin & Co., 1969.

[6] Y.M.Bishop, S.E.Feinberg and P.W.Holland, "Discrete Multivariable Analysis: Theory and Practice", MIT Press, 1975.

[7] A.Agresti, "Categorical Data Analysis", John Wiley and Sons, 1990.