AUTOMATIC TOPIC IDENTIFICATION FOR TWO-LEVEL CALL ROUTING

John Golden, Owen Kimball, Man-Hung Siu, Herbert Gish

BBN Technologies/GTE Internetworking
70 Fawcett St., Cambridge, MA 02138 USA
jgolden@bbn.com

ABSTRACT

This paper presents an approach to routing telephone calls automatically, based upon their speech content. Our data consist of a set of calls collected from a customer-service center with a two-level menu, which allows jumping past the second level, and we view the routing of these calls as a topic-identification problem. Our topic identifier employs a multinomial model for keyword occurrences. We describe the call-routing task in detail, discuss the multinomial model, and present experiments which investigate several issues that arise from using the model for this task.

1. INTRODUCTION

Topic identification has been explored for several data sets, as in [2] for conversations in the Switchboard corpus and in [1], [3], and [5] for routing telphone calls after eliciting a response to the question, "How may I help you?" In this paper we take up the problem of topic identification for a call-routing task in which prompts constrain callers' responses but occur hierarchically and allow for jumping levels of the hierarchy.

In Section 2 we discuss the call-routing problem we have considered, introducing the features of the problem that lead to system-design questions. Section 3 gives a brief overview of topic identification in general, explains our choice of the multinomial model for word generation, and reviews the theory of the model's estimation.

Section 4 gives experimental results using the model to resolve various issues presented by the problem. To begin we show that exploiting sub-classifications by using a hierarchical decision process is worse than deciding on all classifications at once. We also consider two types of utterances—those for which specific keywords have been elicited, and those for which the responses have been left open—and demonstrate that the classifier's performance on a test set of keyword-elicited utterances is greatly enhanced by the addition to the training set of open-response utterances. In addition, we show that using the N-best hypotheses from an automatic speech recognizer can bring the performance of the classifier to near its level on human transcriptions.

2. THE CALL-ROUTING PROBLEM

Our call-routing problem consists in labelling the responses to pre-recorded prompts of callers to a customer-service center. The set of labels is fixed at each prompt, and indicates in which one of an array of services the caller is interested—or that the caller's intention is not in the set of categories and must be passed to an operator. The labels then allow calls to be routed appropriately. Before discussing the general task of labelling utterances based on their topic in Section 3, we describe the pecularities of the call-routing problem at hand.

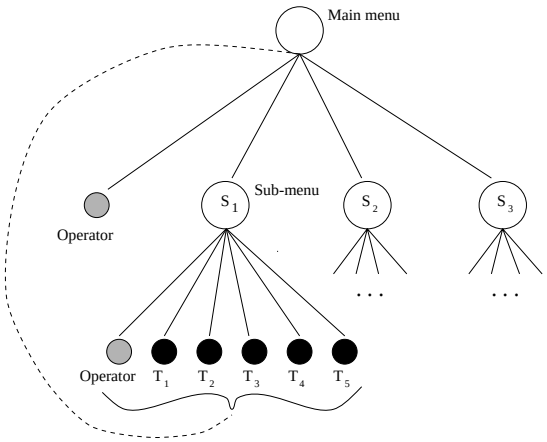

Figure 1: A tree representation of the two-level routing problem.

2.1. Two-Level Routing

Our data set includes responses to two pre-recorded prompts. At the first, *main-menu prompt*, we are to determine to which of four categories the caller's problem belongs; we denote these categories by routing labels $S_0, \ldots, S_3$. S_0 is a "rejection" label, indicating that the caller's response does not fit (or cannot be determined to fit) one of the S_1, S_2, or S_3 categories and so must be handed to an operator. Calls in the remaining three categories are passed on to further pre-recorded prompts. Of these, we limited our study to S_1-labelled calls. These S_1-calls are passed to a second, *sub-menu prompt*, which asks users for responses enabling us to refine the S_1 classification to six sub-categories $T_0, \ldots, T_5$. Again, T_0 denotes that the call must be handled by an operator.

In addition, responses may be labelled with one of $T_0, \ldots, T_5$ at the main-menu prompt if the caller gives enough information there, thus allowing "jumping" past the sub-menu. Figure 1 demonstrates the menu structure and corresponding routing labels as a decision tree.

Paying attention only to the sub-menu for responses of type S_1 allows us to study a two-level topic-identification problem but focusses the task to its minimum. If we view the labelling procedure as the decision tree in Figure 1, we are saying that $T_1, \ldots, T_5$ are the only *non-operator leaf-node labels*. For our purposes, these labels denote exactly those utterances which may be completely categorized without being sent to an operator, and thus comprise a

subset of particular interest.

Restated, then, our task is automatically to label responses to the main-menu prompt with one of $S_0, \ldots, S_3, T_0, \ldots, T_5$ and responses to the sub-menu prompt with one of $T_0, \ldots, T_5$.

2.2. Directed-Response versus Open-Response Prompts

For this study we collected responses to two types of prompts. *Directed-response prompts* include a specific keyword or phrase for each topic label to induce the caller to respond with the word or phrase appropriate to his or her request. *Open-response prompts*, by contrast, do not offer a set of words matching the routing classifications. The open-response prompt from our data set is, "Please tell me the reason for your call." This is similar to the prompt, "How may I help you?" considered in [1].

Directed-response prompts should present an easier routing problem. In the most extreme case, the problem is reduced to a speech-recognition task: if all callers repeat exactly the prompted words corresponding to their request, routing is simply a matter of figuring out which words they have said. Two complications arise with directed prompts, however, in our routing task. First, people are not so compliant as to repeat exact words or phrases from prompts and sometimes use their own words. Second, the two-level structure of our prompts creates an implicit open-response problem. This is because the keywords suggested at the main menu relate only to the high-level topics $S_0, \ldots, S_3$, and not to $T_0, \ldots, T_5$, though we allow for those classifications at this level.

3. OUR TOPIC-IDENTIFICATION SYSTEM

Topic identification is the task of correctly assigning topic labels to utterances assumed to be about one of a fixed set of topics. Our topic-identification system first subjects the spoken utterances to a HMM-based automatic speech recognizer and then passes recognized words to a classifier which models topics. Because we have separated the recognition from the modelling of topics, we are able to optimize the topic model on human transcriptions, which do not have the anomalies of recognition.

The size of the total vocabulary of our directed-response data is relatively small—roughly 450 words, of which approximately 100 were "function words" and articles, which could either be filtered out or modelled with the rest of the words. We performed a few preliminary experiments investigating alternate keyword sets, by selecting keywords by hand and filtering function words. None of these keyword sets enjoyed a material advantage over the other, so that we did not pursue more sophisticated keyword-selection algorithms. The experimental results we report below consider every word seen in training as a keyword.

Our data set consists primarily of short utterances—often two or three words—with little grammatical structure. Thus we turn to multinomial models [2], which operate at the level of words and with sparse data, rather than to higher-order models incorporating grammatical fragments, as in [5].

3.1. The Multinomial Model for Keywords

We define an *utterance* to be a sequence of words $u = \{u_i\}$, where each $u_i \in W = \{w_1, \ldots, w_M\}$, a keyword set of M words that includes a *non-keyword* which substitutes for any word not in W. Let U be the set of all utterances. We consider a set of N classes, or *topics*, $C = \{c_1, \ldots, c_n\}$; each utterance has some probability of being generated from each topic. We wish to model the probability density functions of the utterances conditioned on the topics. We use the *multinomial model* for this:

$$p(u|c_j) = \prod_{i=1}^{M} p(w_i|c_j)^{n_i(u)} \qquad (j = 1, \ldots, N) \qquad (1)$$

where $n_i(u)$ is the number of times word w_i appears in u. Given these probabilities, along with an a priori probability distribution $P(c_j)$ for the set of topics C, we can construct a Bayes classifier for utterances by maximizing $p(c_j|u)$ over all $j = 1, \ldots, N$.

To train the parameters $p(w_i|c_j)$ of the multinomial model, we employ their maximum-likelihood estimate. Specifically, we are given a finite set of labelled training utterances $X = \{(u^k, c^k) \in U \times C\}$, and for each $i = 1, \ldots M$ and $j = 1, \ldots, N$, we compute n_{ij} = the number of occurrences of w_i in all c_j-labelled training utterances. Let us denote the number of unique words which occur in topic c_j by M_j. Then the maximum-likelihood estimate for $p(w_i|c_j)$, using a Bell-Witten backoff to account for words w_i for which $n_{ij} = 0$ [4], is

$$\hat{p}(w_i|c_j) = \frac{n_{ij} + \frac{M_j}{M}}{\sum_{i=1}^{M} n_{ij} + M_j}. \qquad (2)$$

Finally, we also use the maximum-likelihood estimate for the a priori probability of each topic, by using the frequency of the topic's occurrence among the samples in X.

4. EXPERIMENTAL RESULTS

We now describe the data on which we experimented and the results along the various dimensions of the call-routing problem.

4.1. Data Sets

We consider two data sets. *Directed-response data* were collected from directed-response prompts at both the main menu and the sub-menu. We perform most of our experiments on this set. It contains 647 main-menu responses and 384 sub-menu responses. Our second set is from a scenario in which open-response prompts were allowed at the main menu but the sub-menu prompt was directed-response. Nevertheless, we refer to this set as *open-response data*. The open-response set contains 3655 main-menu responses and 855 sub-menu responses.

Our experiments on the directed-response data use a cross-validation scheme to exercise the system on a large enough test set to make the results significant. Here, we divide both the main-menu and sub-menu responses, in random order, into fourths, and we consider each fourth a test set in turn, using the remaining three-fourths as training. In this way each of the 647 main-menu and 384 sub-menu responses are used for testing in a fair manner.

Because the prompts at the two menus are different for both the directed-response set and the open-response set, we consider the classification of main-menu responses as a separate problem from the classification of the sub-menu responses. In all experiments we train separate models for these two sets of responses and assign labels independently.

4.2. Results

Most of the results we report here are an adjustment of the percent of the identifier's labels that are correct. An adjustment is necessary because a sub-menu utterance is always the second of a

System	Label accuracy	
	All topics	Leaf nodes
Two-stage	89.29	78.51
Pooled	89.70	81.26

Table 1: Classification of directed-response transcripts by two-stage and pooled decisions trained from directed transcripts.

caller's responses, and as such is conditioned on our classification accuracy of S_1 at the main-menu prompt. As a result we compute the *overall label accuracy* of a set of labels as follows: if C_m and C_s are the numbers of utterances labelled correctly at the main and sub-menus, if N_m and N_s are the total number of utterances from the main and sub-menus, and if acc(S_1) is the fraction of S_1 utterances labelled correctly, then the label accuracy for this set of labels is

$$\text{LA} = \frac{C_m + \text{acc}(S_1) \cdot C_s}{N_m + N_s}. \quad (3)$$

We are reducing the number of correctly classified sub-menu responses by our failure at the main menu to pass callers to the sub-menu. This is an approximation, which does not take into account any correlation between our success with a particular caller's main-menu response and our success with his or her sub-menu response, and hence it is probably a low estimate of the true accuracy. But this is only a mild approximation, since the classifier performs well on the S_1-labelled utterances and hence has acc(S_1) $\approx$ 1; unless otherwise indicated it is above 0.95 in the experiments to follow.

We give results for both the *all-topic task*, in which we score the topic-identifier's labels for all utterances, and the *leaf-node task*, for which we score only the labels for utterances whose true topics are among the non-operator leaf-node topics ($T_1, \ldots, T_5$). The system always models and identifies all topics at each step; we are simply looking at its label accuracy on both the entire test set and a particular subset.

4.2.1. Two-Stage versus Pooled Decisions

In view of the two-layered structure of our problem, two methods for main-menu label assignment are possible: a *two-stage decision*, in which we first categorize a call among the S_i, and, if the label is S_1, we make a second decision among S_1 and the T_j; or a single *pooled decision*, in which we decide once and for all among $S_0, \ldots, S_3$ and $T_0, \ldots, T_5$. The idea of the two-stage system is to obviate the irrelevant decision between each T_i and S_0, S_2, or S_3 by filtering S_0, S_2, and S_3 from consideration before looking at the T_is.

Table 1 compares the performance of these two methods on human transcriptions of the directed responses. Both systems use multinomial models for making classification decisions, but the two-stage system uses two such models for main-menu utterances, one for each decision. The two-stage system is worse for both tasks, but it is only slightly behind the pooled-decision system for all topic labels. Its larger failure is on the leaf-node labels, whose classification was the motivation for the method. For both systems, this more interesting task appears harder.

4.2.2. Open-Response versus Directed-Response Data

We wish to compare the performance of our system in a directed setting to its success with open-response data. Because only the main menu of the open data invites open responses, we abandon for the moment our label-accuracy measure and look simply at the percent of main-menu utterances correctly classified.

Data set		Percent of labels correct	
Training	Test	All topics	Leaf nodes
None	Directed	75.89	19.50
Directed	Directed	86.71	43.90
None	Open	17.62	0.00
Open	Open	63.33	63.21

Table 2: Classifications of transcripts of responses to the main menu in the two data sets, with matched training and test data.

Table 2 demonstrates how the system performs on test sets of both open and directed data with matching training sets. We obtained an open-response test set by drawing 1481 of the 3655 main-menu responses. Thus we have an open-response training set of 2174 utterances, as compared to our directed-response training set of 647 main-menu responses (using cross-validation). The entries in the table for which the training set is "none" indicate a system that uses an "untrained" models on the test sets. When presented with a test utterance, this system simply adds to a count of keywords taken from the prompts without considering the words as they occur in training; the topic with the highest number of corresponding keywords, weighted by a topic prior estimated from the training data, is chosen as the label. It is evident that these directed-response utterances can be labelled with 75.89% accuracy simply by an accounting of prompted words. Since there are no keywords in the open prompts to count, the untrained models for this data simply pick the topic with the highest prior, and are understandably poor.

As we expected, the discrepency between performance on the all-topic task and on the leaf-node task is striking for the directed data: the latter is the harder problem, both with untrained models and trained models. But also striking is that proper training from directed data improves performance from 19.5% to only 43.90%, whereas proper training on the open data improves performance from a zero correct to an impressive 63.21%. This large discrepency may be due in part to the greater amount of training data in the open-response set. But it also suggests that the open-response data is better in general for training for the leaf-node task, which at the main menu is an open-response problem in both data sets. Because our directed-response problem contains this open-response component, we next try supplementing our directed-response training with open data.

4.2.3. Boosting Directed-Response Performance with Open-Response Data

In order to fix the size of the training set and so only measure its quality, we perform the following experiment: for each cross-validation partition, and for both main-menu and sub-menu utterance sets, we replace half of the directed-response data, randomly chosen, with the same amount of open-response data, also randomly chosen. Table 3 gives the label accuracy of this experiment on transcriptions of the directed utterances. We see that for the all-topic task very little is gained from the open data; but for the harder, leaf-node problem, a substantial gain is obtained. If we investigate more deeply the performance increase on the leaf-node transcripts, we find that the open data boosts the percent of correct labels from 43.90% to 65.85% for main-menu utterances, but only from 92.8% to 94.12% for sub-menu utterances. So the open-response data are, as we suspected, aiding the implicit open-

Training set	Label accuracy	
	All topics	Leaf nodes
100% directed	89.70	81.26
50% directed, 50% open	89.78	87.26

Table 3: Classification of directed-response test transcripts from both directed and open training transcripts.

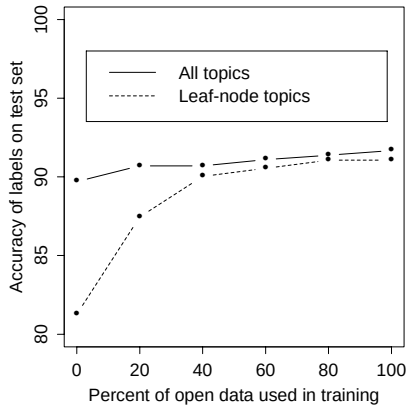

Figure 2: Label accuracy of the classifier on directed test transcriptions when trained on all directed and an increasing amount of the open transcripts.

response problem in our two-level directed-response data: the leaf-node classifications at the main menu.

Extrapolating from this substitutive experiment, we look at how the addition of open-response data to the full directed-response training set boosts performance on the directed-response test set. Figure 2 shows the label accuracy on the directed data as a function of the percent of the open data used for training (in addition to all the directed data used in training). We note that the all-topic accuracy only slightly improves as we add open data, but we can increase the leaf-node accuracy to be in line with the all-topic accuracy. We see, furthermore, that at approximately 80%, the open data ceases to afford any real improvement, indicating that the model has reached its saturation of this training data for this task.

4.2.4. Topic Identification with Automatic Speech Recognition

To this point we have reported results on human transcriptions of prompt responses. A practical system, however, must classify the errorful output of an automatic speech recognizer. We now show that we can attain nearly identical performance as above with speech-recognition output.

We trained the acoustic models of the HMM-based speech recognizer with data from the Switchboard corpus, and we trained its language model from the open-response data set, so that we might decode all directed-response data fairly. We compare using the best hypotheses from the recognizer to using its N-best list of hypotheses, in the latter case allowing all words to have equal weight. ($N = 100$ was determined optimal.) Label accuracy for these different recognizer outputs on the directed-response data appears in Table 4. We see that if we train our system with N-best lists and ask the system to classify N-best test data, we can get to within an absolute 3% of the label accuracy on human transcriptions. Training and testing on the top hypothesis do not give us this performance; we posit that the N-best lists allow for words which are relevant but do not receive high enough scores by the speech decoder to appear in the top hypothesis. In addition, the N-best list, by repeating high-scoring words in each hypothesis, implicitly weights words better modelled by the speech recognizer.

Type of data		Label accuracy	
Training	Testing	All topics	Leaf nodes
BH	BH	85.79	72.41
NB	NB	88.12	78.38
HT	HT	89.70	81.26
HT	BH	85.56	69.30
HT	NB	75.25	60.07

Table 4: Classification of directed-response data in various forms: *BH* = Best hypotheses; *NB* = N-best hypotheses; *HT* = Human transcripts. Training is on directed-response data only.

We note that training on transcripts degrades the classifier's performance on both best-hypothesis and N-best output from its performance on matched conditions. But the N-best test set, with its expanded vocabulary, suffers more from the mismatch. In this case, however, the performance of the classifier in identifying S_1 labels is, at 82.66%, significantly poorer than in other experiments; this low acc(S_1) depresses the label accuracy.

5. CONCLUSION

We have applied a multinomial model of word generation by classes to a directed-response, two-level call-routing problem. We have shown that for this application, it is more effective to account for the implicit open-response problem by using open data in modelling topics than to perform a two-level decision process after the menu structure. In addition, we demonstrated that N-best lists from a speech recognizer are, to the topic model, about as good as human transcriptions.

6. REFERENCES

[1] A.L. Gorin, B.A. Parker, R.M. Sachs, and J.G. Wilpon, "How may I help you?" *J. Acoust. Soc. Am.*, 1995, vol. 97, pp. 3441-3461.

[2] John McDonough and Herbert Gish, "Issues in Topic Identification on the Switchboard Corpus." *Proc. ICLSP*, 1994, pp. 2163–2166.

[3] G. Riccardi, A.L. Gorin, A. Ljolje, and M. Riley, "A Spoken Language System for Automated Call Routing." *Proc. ICASSP*, 1997, pp. 1143–1146.

[4] Ian H. Witten and Timothy C. Bell, "The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression." *IEEE Transactions on Information Theory*, 1991, vol. 37, no. 4, pp. 1085-1094.

[5] J.H. Wright, A.L. Gorin, and G. Riccardi, "Automatic Acquisition of Salient Grammar Fragments for Call-Type Classification." *Proc. Eurospeech*, 1997, pp. 1419–1422.