AUTOMATIC ACQUISITION OF SALIENT GRAMMAR FRAGMENTS FOR CALL-TYPE CLASSIFICATION

J.H.Wright, A.L.Gorin and G.Riccardi

AT&T Labs – Research, 180 Park Avenue, Florham Park, NJ 07932, USA
{jwright,algor,dsp3}@research.att.com

ABSTRACT

We present an algorithm for the automatic acquisition of salient grammar fragments in the form of finite state machines (FSMs). Salient phrase fragments are selected using a significance test, then clustered using a combination of string and semantic distortion measures. Each cluster is then compactly represented as an FSM. Flexibility is enhanced by permitting approximate matches to paths through each FSM. Multiple fragment detections are exploited by means of a neural network. The methodology is applied to the "*How may I help you?*" (*HMIHY*) call-type classification task.

1. INTRODUCTION

We are interested in spoken language understanding and dialogue systems, in particular for providing automated services to non-expert users. In previous work [1-6] we have considered the problem of automatic call routing in response to the open-ended prompt "*How may I help you?*". The aim is to infer an appropriate machine action from the caller's utterance, and the issues of large-vocabulary speech recognition and dialogue management are addressed in [4,5]. In this paper we consider the automatic acquisition of salient grammar fragments in order to learn the mapping between the speech channel and the set of machine actions. We show that by applying these grammar fragments we can achieve improved performance in call classification, as compared to the results in [2].

We use a database of 10k spoken transactions between callers and human agents, separated into training and test sets. The caller's first utterance in each transaction has been transcribed and labelled with one or more of 14 call types such as *collect*, *person-to-person*, and *calling-card*, plus a 15th class denoted *other* which is intended to subsume the remainder [3]. Previously we have used a large set of automatically-acquired salient phrase fragments, individually associated with the call types via an estimated posterior distribution. Detected occurrences of these phrase fragments within a speech-recognized utterance provide a basis for classification of that utterance. Three issues that naturally arise are the following:

- most fragments occur rather rarely in the training data, so the estimated posterior distributions tend to be erratic,
- many fragments are similar to each other and occur in similar semantic contexts,
- most test utterances contain more than one detected fragment.

The purpose of the grammar fragments and the new decision rule described below is to take advantage of the second and third of these points while allowing for the first. Using these new methods, we achieve a useful operating point with 87% correct classification rate at rank 2, for a 40% false rejection rate.

2. SALIENT GRAMMAR FRAGMENTS

2.1 Significance test for phrase fragments

We want to select phrase fragments that are meaningful for the task. Previously we have used a measure of *salience* [1,2] to assess (for a particular fragment) the distortion between the prior and posterior distributions over the call types, but this does not take into account the frequency with which a fragment occurs. A fortunate conjunction of events can give a low-frequency fragment a high apparent salience purely by chance. Here we avoid this shortcoming by testing, for each fragment, the null hypothesis that its behaviour is governed by the prior distribution over the call types, and that it therefore occurs at random. Suppose that a fragment f has a total of n occurrences of call-type labels in training, and let $\{r_1, r_2, \cdots\}$ denote the set of all possible partitions of n occurrences into $K = 15$ classes. Let the actual observed distribution of counts for f be r_f, and the prior distribution be denoted $\{p_k\}_{k=1,\cdots,K}$. Under the null hypothesis, the probability of any particular partition $r_i = n_{i1}, \cdots, n_{iK}$ of n occurrences over the classes is given by the multinomial distribution:

$$P(r_i \mid n) = n! \prod_{k=1}^{K} \frac{p_k^{n_{ik}}}{n_{ik}!}$$

A fragment f of frequency n is accepted at significance level α if

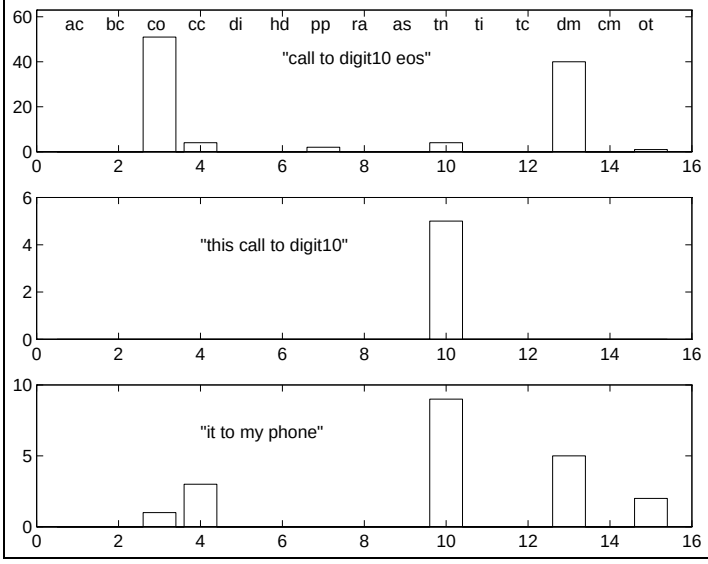
$$\sum_{r_i \in A(f)} P(r_i \mid n) \le \alpha$$

where $A(f) = \{r_i : P(r_i \mid n) \le P(r_f \mid n)\}$ is the set of partitions that have probability not exceeding that for the observed distribution. Any fragment for which the observed distribution can be seen to be a relatively likely random sample from the prior is therefore rejected.

This is an exact test of significance, and is valid even for fragments with very small occurrence counts. For the "*How may I help you?*" (*HMIHY*) task, the median frequency within 7884 transcribed training utterances for the fragments derived in [2] is 6. Imposing a significance

level of 5% reduces the total number of phrase fragments by about 30%. While this appears to be a drastic reduction, and many of the rejected fragments would appear to be meaningful to a human judge, it highlights the problem of basing inferences on specific but complex objects for which the training statistics are fragile: in this case, a posterior distribution over 15 classes based on typically only 6 observations. Because many fragments that are rejected are mild variations of others that are accepted, the final coverage is unaffected — there is no significant increase in false rejection rate for the test utterances as a result of applying the fragment significance test.

Figure 1: Occurrence count distributions for three fragments



2.2 Clustering the fragments

In order to exploit the similarity between fragments, the next step is to cluster the surviving phrase fragments by using an agglomerative clustering procedure. For this we use a Levenshtein string distance measure $d_S(f_1, f_2)$ between fragments f_1, f_2, in which the insertion, deletion and substitution penalties are weighted by the salience of the respective words. This has the effect of penalizing salient errors (such as substitution of "*collect*" for "*credit*") more than non-salient errors (such as substitution of "*this*" for "*the*").

However, fragments that are similar as strings can have different semantics, e.g. the fragments "*need a credit*" and "*a credit card*" indicate a billing credit request and credit card payment respectively. It would be undesirable for these to enter the same cluster. In assessing this we must again allow for the variability attributable to small samples. We therefore use a measure of *semantic distortion* defined as

$$d_M(f_1, f_2) = \frac{1}{K}\sum_k \frac{[\hat{P}(c_k \in C_t \mid f_1 \in F_t) - \hat{P}(c_k \in C_t \mid f_2 \in F_t)]^2}{\hat{\mathrm{Var}}[\hat{P}(c_k \in C_t \mid f_1 \in F_t) - \hat{P}(c_k \in C_t \mid f_2 \in F_t) \mid \mathrm{H}]}$$

where $\hat{P}(c_k \in C_t \mid f \in F_t)$ is the estimated posterior distribution over call types c_k for fragment *f*, and C_t, F_t are the sets of labels and observed fragments for an utterance *t*.

This is a weighted mean-square error between the estimated posterior distributions. We chose not to use the Kullback-Leiber measure because it is important to take the small sample sizes into account, and using the mean-square measure enables this. The denominator consists of an estimate of the variance of the difference between the estimated posterior values (for each call type) under the hypothesis H. This hypothesis states that the two fragments have the same true (but unknown) posterior distribution. It follows that under this hypothesis H, the expected value of $d_M(f_1, f_2)$ is equal to 1.0 regardless of the fragment occurrence frequencies, which may be small. A large value for this measure can then be taken to imply that H is false, and therefore provides evidence for divergence between the posterior distributions. Note however that this measure does not obey the triangle inequality and so is not a true measure of distance. The evaluation of $d_M(f_1, f_2)$ in terms of actual occurrence counts is given in the Appendix.

The overall measure used for clustering is the maximum of the string and semantic distortions. It is interesting to note cases where fragments that are similar as strings but semantically different are placed in different clusters, whereas fragments that are different as strings but similar semantically are placed in the same cluster. The occurrence count distribution for the fragment "*call to digit10 eos*" is shown in the top part of figure 1, ("*digit10*" is a non-terminal symbol representing a ten-digit string, and "*eos*" is the end-of-sentence marker). This fragment occurs often in *collect* and *dial-for-me* calls. As shown in the central part of figure 1, the very similar fragment "*this call to digit10*" occurs in only five training calls, all of them labelled as *third-number*. This difference in semantic behaviour keeps these two fragments apart. The bottom part of figure 1 shows the distribution for "*it to my phone*", which is also mainly a *third-number* fragment. The visible differences between this and the central fragment are explainable by the small numbers of occurrences for these two fragments, and it turns out that they end up in the same cluster.

Test results in call-type classification after using the combination of string and semantic distortions are superior to those obtained using either distortion measure alone.

2.3 Conversion into FSMs

Each of the resulting fragment clusters is converted into an FSM representing a grammar fragment. For this we use a method similar to the ECGI algorithm [7]. A very simple example is shown in figure 2. For each FSM, the posterior distribution over the call types is then obtained by searching the training data for exact matches to paths through the FSM from the start node to a final node. It is

Figure 2: A simple grammar fragment

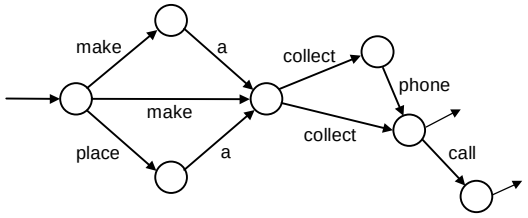
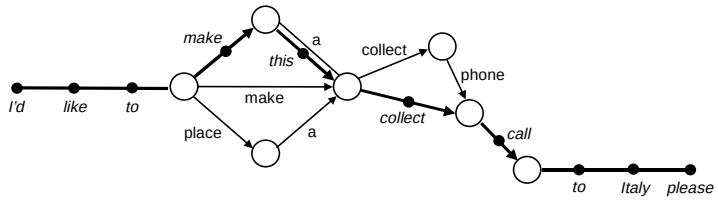


Figure 3: Approximate match of a grammar fragment



noteworthy that the median occurrence frequency for the FSMs on the *HMIHY* training data is 15, a significant improvement on the 6 observed for the phrase fragments, and the statistics for these objects are consequently more robust.

Observations in the form of exact or approximate matches to a path through the FSM can then be found for the test utterances. Approximate matches are found by using a dynamic programming algorithm in which word salience is used to weight the errors. An example of an approximate match is shown in figure 3, where the word *a* is substituted by *this*, both words having low salience.

3. CALL-TYPE CLASSIFICATION

3.1 Peak-of-fragments classifier

We need to infer a call-type from the fragments detected within the speech. More generally, we would like to assign a ranking to the call-types. A simple method for doing this involves finding, for each call type, the highest posterior probability attributed to it by any detected fragment:

$$s_k = \max_i \hat{P}(c_k \in C_t \mid f_i \in F_t), \quad k = 1, \cdots, K$$

where $\{f_i\}$ are the detected fragments for the utterance. The rank-1 decision for this utterance is then the call type which has the highest of these scores:

$$\hat{c} = \arg\max_k s_k$$

This method is used in [6].

3.2 Neural network classifier

Typically an utterance contains several detected fragments (even when the phrases are clustered and formed into FSMs) and it is advantageous to combine the evidence from these when arriving at a classification. One method for doing this is as follows. The detected fragments form a lattice which is parsed, for each call-type separately, to find the path generating the highest cumulative score — summing the posterior probability for that call-type along the path. The resulting set of scores is passed through a feed-forward neural network in order to generate a final output score for each call-type, which serves for a ranking. The neural network is trained by back-propagation using the transcribed and labelled training data. For each training utterance, the inputs to the network are the scores found from the text in the manner just described, and the desired outputs are set to 1 for those (one or more) call-types that are correct and 0 for the remainder. We have used single and two-layer networks, and the performance advantage from having the second layer was found to be negligible.

The neural net has two main purposes:

- each output score is in the range (0,1) and may reasonably be interpreted as the posterior probability of the call-type, given the cumulative evidence for the call-types expressed in this simple way as sum of posteriors,
- it embodies the information that certain conjunctions of call-types are possible (e.g. *collect* and *dial-for-me*) whereas others are not (e.g. *collect* and *billing-credit*).

We have found that this method for exploiting multiple fragments gives better results than a bag-of-fragments model.

4. EXPERIMENTS

4.1 Evaluation method

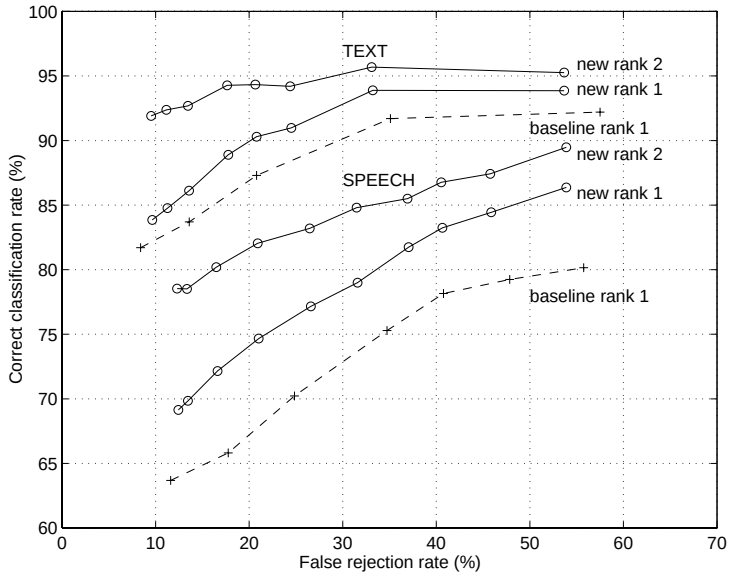
In the *HMIHY* task there are 14 call types, plus a class called *other* for which the intention is that these calls be transferred immediately to a human agent. There is therefore a criterion for rejection and we can measure the true and false rejection rates for manually labelled data, as well as the true classification rate. At each rank, a call is "rejected" either if the decision for that rank is *other* or if the score for that rank is below a given threshold. By varying the threshold we can generate ROC curves. For the experiments reported here, rejection at rank-1 is taken to imply rejection also at rank-2 (although this won't necessarily be the case in the context of a dialogue).

4.2 Results

For training we use a set of 7884 utterances that have been manually transcribed and labelled. For testing we use a similar set of 1000 further utterances, processed by a large-vocabulary speech recognizer [4]. Matches of grammar fragments to the output are found, and used for classification as described in section 3.

The dashed lines in figure 4 show the ROC curves for rank-1 on both speech and text , obtained using a set of 3720 salient phrase fragments with the peak-of-fragments classifier. After applying the significance test to the phrase fragments (with α set at 5%), clustering and formation into 560 FSMs as described in section 2, and using the network classifier, we obtain the ROC curves shown by the solid lines in figure 4. A significant improvement in performance at rank-1 is seen both for

Figure 4: ROC curves for 1000 test sentences



speech and for text, and the curves for rank-2 are also shown. This suggests a useful operating point for speech with 87% correct classification rate at rank 2, for a 40% false rejection rate.

In figure 5 we show the ROC curves for rejection, for the speech-recognized utterances. Again, the new methods lead to an improvement, with a true rejection rate of 85% at rank-1, for a 40% false rejection rate.

5. CONCLUSIONS

The extraction and application of salient phrase fragments allows us to classify incoming calls using units that are larger (and usually less ambiguous) than individual words but without the need to parse an entire sentence. Many fragments turn out to be minor variations of each other, and each variation may in turn be a rare event, but by automatically selecting and clustering phrase fragments using methods that are valid for small samples we obtain grammar fragments that are both robust and highly informative. Compacting the clusters into FSMs gives us a representation that has reduced space and time demands when processing new data. Central to the work presented here is the requirement that the acquisition and deployment of phrase and grammar fragments is achieved by automatic procedures.

Appendix — Evaluation of semantic distortion measure for two fragments

Let N_1, N_2 be the number of training utterances for which the fragments f_1, f_2 occur, respectively,

X_{1k}, X_{2k} be the number of training utterances of call type c_k for which the fragments f_1, f_2 occur, respectively,

Y_{12k} be the number of training utterances of call type c_k for which both the fragments f_1, f_2 occur.

Then a simple version of the semantic distortion measure defined in section 2.2, which works well in practice, is

$$d_M(f_1, f_2) = \frac{1}{K} \sum_{k=1}^{K} \frac{(N_2 X_{1k} - N_1 X_{2k})^2}{N_1 N_2 (X_{1k} + X_{2k} - 2Y_{12k})}$$

where K is the number of classes.

References

[1] A.L.Gorin, "On automated language acquisition", Journal of the Acoustical Society of America, vol. 97(6), 1995, pp. 3441-3461.

[2] A.L.Gorin, "Processing of semantic information in fluently-spoken language", Proc. ICSLP, Philadelphia, 1996, pp. 1001-1004.

[3] A.L.Gorin, B.A.Parker, R.M.Sachs and J.G.Wilpon, "How May I Help You?", Proc. IVTTA, Basking Ridge, 1996, pp. 57-61.

[4] G.Riccardi, A.L.Gorin, A.Ljolje and M.Riley, "A spoken language system for automated call routing", Proc. ICASSP, Munich 1997, pp 1143-1146.

[5] A.Abella and A.L.Gorin, "Generating semantically consistent inputs to a dialogue manager", Proc. Eurospeech 97.

[6] A.L.Gorin, G.Riccardi and J.H.Wright, "How May I Help You?", to appear in Speech Communication.

[7] H.Rulot and E.Vidal, "Modelling (sub)string-length based constraints through a grammatical inference method", Pattern Recognition Theory and Applications, P.A.Devijver and J.Kittler (eds), Springer-Verlag, 1987.

Figure 5: ROC curves for rejection (speech)