UNSUPERVISED AND ACTIVE LEARNING IN AUTOMATIC SPEECH RECOGNITION FOR CALL CLASSIFICATION

Dilek Hakkani-Tür Gokhan Tur Mazin Rahim Giuseppe Riccardi

AT&T Labs-Research,
180 Park Avenue, Florham Park, NJ, USA
{dtur,gtur,mazin,dsp3}@research.att.com

ABSTRACT

A key challenge in rapidly building spoken natural language dialog applications is minimizing the manual effort required in transcribing and labeling speech data. This task is not only expensive but also time consuming. In this paper, we present a novel approach that aims at reducing the amount of manually transcribed in-domain data required for building automatic speech recognition (ASR) models in spoken language dialog systems. Our method is based on mining relevant text from various conversational systems and web sites. An iterative process is employed where the performance of the models can be improved through both unsupervised and active learning of the ASR models. We have evaluated the robustness of our approach on a call classification task that has been selected from AT&T VoiceTone^{SM} customer care. Our results indicate that with unsupervised learning it is possible to achieve a call classification performance that is only 1.5% lower than the upper bound set when using all available in-domain transcribed data.

1. INTRODUCTION

Spoken natural-language understanding (SLU) plays an important role in automating complex transactional requests, such as those for customer care and help desk. SLU provides callers with the flexibility to speak naturally without having to follow laboriously a directed set of prompts. One central technology of the SLU in AT&T VoiceTone is the use of a semantic classifier for detecting callers' requests or their intent. This is clearly important especially in call routing applications. In [1], we have framed this problem as a classification task. A set of requests (or *call-types*) are initially defined using manual labeling, and a classifier is trained to learn a mapping from spoken language into call-types. Labeling involves associating each transcribed utterance with one or more semantic call type. For example, if the user says "I would like to get my balance", then the corresponding semantic label is "Request(Balance)."

A key challenge when using statistical classifiers for call routing is the need for an extensive set of in-domain data that is manually transcribed and labeled, a process that is rather expensive and noisy. In [2], prior knowledge of the application is used to alleviate the reliance on labeled data by balancing the conditional likelihood of the data against the distance of the data-generated model from the model provided by the human. In [3], an iterative approach is proposed using phone-based units to reduce the dependency on transcribed data at the cost of having labelers listen to speech, rather than reading text, to determine the semantic labels. In [4], maximum likelihood linear regression was applied for adapting ASR models and bootstrapping call routing applications. Other related studies include [5, 6].

In this paper, we measure the performance impact of recognition accuracy on call classification as we reduce the amount of in-domain transcribed speech. We present a novel approach that aim at reducing the amount of transcribed data required for building ASR models. Our method involves constructing a bootstrapped model through mining relevant text from various conversational systems and web sites. An iterative process is then employed where the performance of the ASR models can be improved through both unsupervised and active learning. For unsupervised learning, a two step method is adopted that involves decoding followed by model building. For active learning, a confidence score is computed and is used to identify problematic utterances that need to be manually transcribed [7].

The organization of this paper is as follows: Section 2 presents our approach for unsupervised and active learning. Section 3 presents experimental results for call classification. We summarize this paper in Section 4.

2. APPROACH

We consider the problem of identifying a caller's request as a multi-class multi-label problem. Given a set of semantic call types (or semantic classes) $C = \{C_1, ..., C_n\}$ and a sequence of input words $W = \{W_1, ..., W_m\}$, the objective is to compute the posterior probability of each class, $P(C_i|W)$ and retain those that are above a predetermined threshold. In this paper, we adopt BoosTexter [8], a member of the boosting family of learning algorithms to compute $P(C_i|W)$. BoosTexter combines many simple and moder-

ately accurate categorization rules that are trained sequentially into a single, highly accurate model that can reliably predict a class.

Although state-of-the-art performance has been reported using BoosTexter for identifying callers request (e.g., [8]), these classifiers are known to degrade in performance when presented with an output from a speech recognizer. One goal of this paper is to measure this robustness effect when transcribed data is either unavailable or limited during training of the ASR models. Let's first examine the ASR process. Given a set of observations X, a hypothesized sequence of words $\hat{W}$ can be obtained using a maximum a posteriori (MAP) decoder:

$$\hat{W} = \arg\max_{W} P_{\Theta}(X|W) \cdot P_{\Phi}(W)^{\eta}, \qquad (1)$$

where $P_{\Theta}(X|W)$ is the acoustic observation probability that is modeled by a hidden Markov model Θ. $P_{\Phi}(W)$ is the n-gram language model probability with underlying set of parameters Φ. The factor η is the grammar scale.

Although $P_{\Theta}(X|W)$ can be used across different applications without the need for in-domain speech data, $P_{\Phi}(W)$ requires extensive in-domain conversational data to reliably compute the n-gram statistics. Even when speech data is available, transcribing it manually is an expensive process, errorful and generally delays the application creation cycle. If sufficient transcribed data is available, then the natural solution is to apply MAP adaptation so that a new model $\hat{\Phi}$ is computed such that:

$$\hat{\Phi} = \arg\max_{\Phi}[f(W|\Phi) \cdot g(\Phi)],$$

where $f(W|\Phi)$ is the discrete density function of W and $g(\Phi)$ is the prior distribution which is typically modeled using a Dirichlet density [9]. With some simplification, the MAP estimate can be reduced to a weighted linear interpolation of the out-of-domain prior model and in-domain samples. Another approach to language model adaptation is the mixture modeling. While MAP adaptation preserves the model structure of the background language models, mixture models incorporate the parameters from all sources:

$$P(w_i|w_{i-n+1}\ldots w_{i-1}) = \sum_j \gamma_j P_j(w_i|w_{i-n+1}\ldots w_{i-1}),$$

where $P_j(.)$ is the j^{th} mixture probability estimate and γ_j is the mixture weight, estimated through held out data, such that $\sum_j \gamma_j = 1$.

In this paper we consider the impact of using three scenarios while creating spoken language models on call classification. The first scenario assumes no in-domain transcription or speech data is available, thus relies solely on the prior. In this scenario a bootstrapped language model Φ is formed based on mining relevant material from various sources of data. This includes (a) Human/human conversational data that was taken from the Switchboard corpus, (b) Human/machine conversational data that was collected from various spoken dialog applications, and (c) Text data that was mined from the world-wide-web of relevant web sites. We have found that including web data reduces the out-of-vocabulary rate and provides a sizable improvement in accuracy.

The second scenario assumes that speech data is available but is untranscribed. This is typically the case when a Wizard trial is performed prior to application deployment. In this scenario, we adopted an iterative two-step method. In the first step the bootstrapped model, Φ, in Equation 1 is used to generate the word sequences $\hat{W}$. Given that Φ is universal, a lower grammar scale was used to strengthen the effect of $P_{\Theta}(X|W)$. In the second step, a new language model $\hat{\Phi}$ is computed using the ASR output of the in-domain speech data and other available transcribed data.

The third scenario assumes that limited data can be manually transcribed. In this scenario, we apply active learning to intelligently select and then transcribe a small fraction of the data that is most informative. We use word and utterance confidence scores computed from ASR output word lattices during the selection. We use the rest of the data that is not yet transcribed in unsupervised learning. The transcribed data is used in conjunction with $\hat{W}$ for building $\hat{\Phi}$. The details of this scenario are explained in [10]. Having high-quality ASR output is essential for labelers to generate high-quality labels. Active learning reduces the labeling effort as well as improves the labeling accuracy since it identifies utterances with low confidence scores for manual transcription. The rest of the data with high confidence scores can be labeled directly using recognized speech.

3. EXPERIMENTS AND RESULTS

In order to evaluate these scenarios, we carried out experiments using human-machine dialogs as collected by the AT&T *VoiceTone*^{SM} natural spoken dialog system. We first describe our test domain and data, and define the evaluation metrics. We then give the results obtained by the semantic classifier using the output of the ASR, which is trained using unsupervised and active learning.

3.1. Data

Our dataset includes human/machine dialogs collected from a pharmaceutical domain. In this application, callers are greeted by an open ended prompt "*How May I Help You?*" In the event the system is unable to understand the caller then the conversation will proceed with either a clarification or a confirmation prompt. Otherwise, a caller is routed to a specific destination.

Table 1 summarizes the characteristics of our application including amount of training and test data, total number of call-types, average utterance length, and call-type perplexity. Perplexity is computed using the prior distribution over all the call-types in the training data.

3.2. Evaluation Metrics

The ASR performance is measured in terms of word accuracy on the test set. Inspired by the information retrieval

Training Data Size	29,561 utterances
Test Data Size	5,537 utterances
Number of Call-Types	97
Call-Type Perplexity	32.81
Average Length	10.13 words

Table 1. Data characteristics used in the experiments.

community, the classification performance is measured in terms of the *F-Measure* metric. F-Measure is a combination of *recall* and *precision*:

$$F - Measure = \frac{2 \times recall \times precision}{recall + precision}$$

where recall is defined as the proportion of all the true call-types that are correctly deduced by the classifier. It is obtained by dividing the number of true positives by the sum of true positives and false negatives. Precision is defined as the proportion of all the accepted call-types that are also true. It is obtained by dividing true positives by the sum of true positives and false positives. True (False) positives are the number of call-types for an utterance for which the deduced call-type has a confidence above a given threshold, hence accepted, and is (not) among the correct call-types. False (True) negatives are the number of call-types for an utterance for which the deduced call-type has a confidence less than a threshold, hence rejected, and is (not) among the true call-types. The best F-Measure value is selected by scanning over all thresholds between 0 and 1.

3.3. Results

There are several sources of text data for bootstrapping language models. In this paper, our dataset included 3.6M words of human-human conversations that were taken from the Switchboard corpus, 72,000 words from relevant pharmaceutical web-sites, and around 1M words from human-machine dialogs that were collected from different customer care applications. Four trigram language models were generated. The first one, which is referred to as "bootstrap model", was built from the above dataset. The second and third models, which are referred to as "adaptation model" and "mixture model" are created by following the two step process that we outlined in Section 2. In the adaptation model, we just use the ASR output of the in-domain data, and in the mixture model, we also use the transcribed data used for the bootstrap model. The fourth model, which is referred to as "transcription model", is trained from manually transcribed data. This is considered as the upper bound. In all these experiments, we used an off-the-shelf acoustic model, trained from over 100 hours of telephone speech. Figure 1 shows the test set word accuracy (WA) versus recognition run-time for these four language models. The run-time speed was adjusted by varying the beam width of the decoder. The bootstrap model gave the worst performance, in terms of both word accuracy and run-time. The

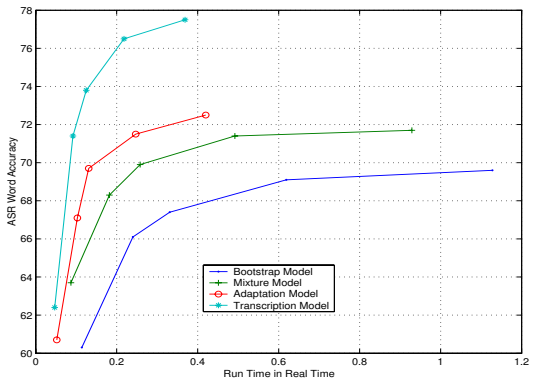

Fig. 1. Word accuracy vs. recognition run-time in real time.

Training Set	**n=1**	**n=2**	**n=3**
Bootstrap model	11.19	29.91	54.99
Adaptation model	11.69	29.63	51.22
Transcription model	1.88	9.51	27.27

Table 2. The test set out-of-vocabulary n-gram rates.

adaptation model shows a 4-6% absolute improvement in word accuracy without the need for transcribed in-domain data. This model is about 4-6% worse than the upper bound set by the transcription model. The mixture model is also better than the bootstrap model, but due to the large size of the language model, it results in a longer recognition time.

We conducted an analysis of the out-of-vocabulary word (OOV) rates for the above three models. Table 2 displays the n-gram OOV rate which is defined as the percentage of n-grams in the test data that do not appear in the training data. As expected, the table shows that the OOV rate for the transcription model is significantly less than the other two models. Interestingly enough, the adaptation model is able to maintain similar unigram statistics and better trigram coverage than the bootstrap model. The increase in speed of the adaptation model over the bootstrap model is justified by its lower perplexity.

The impact of using different language models on the call classification performance is shown in Figure 2. All curves were computed using the best ASR output at the highest beam for the training set. Given that the focus of this paper is on the impact of ASR on call classification performance, we assume that all data are labeled through either listening to audio or reading their transcriptions. We have conducted experiments to demonstrate the call classification accuracy when labeling from recognized speech which we will report in future publications. Figure 2 clearly shows that there is a strong correlation between ASR word accuracy and F-Measure. One striking result in this figure is that the call classification performance when using the adaptation model is only 1.5%-2% inferior to the upper bound set by the transcription model. The adaptation model was trained without any in-domain transcription data.

In the final scenario, we evaluated the impact of vary-

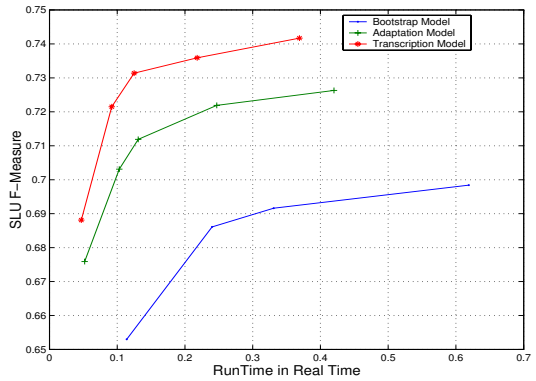

Fig. 2. SLU performances using ASR outputs in Figure 1.

ing the amount of manually transcribed data on call classification. Figure 3 shows how the word accuracy changes when utterances are selected either randomly or through active learning. These plots were generated at a run-time of 0.11 times real time. At equal number of transcribed utterances, the ASR accuracy clearly rises faster with active learning than random selection. Figure 4 shows the corresponding call classification performances. Even though the maximum gain in F-measure when using the transcription model is only 1.5-2%, the combination of active and unsupervised learning is superior to combining random sampling of the data with unsupervised learning.

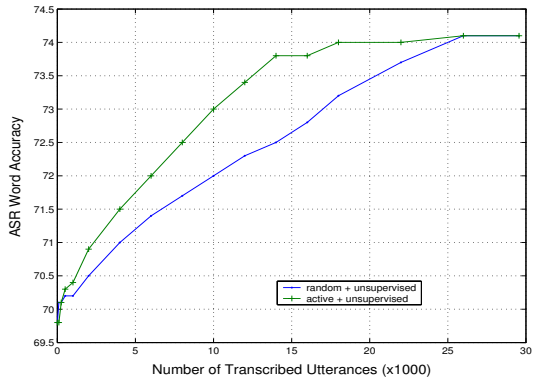

Fig. 3. ASR performances using active and unsupervised learning.

4. CONCLUSIONS

Unsupervised and active learning of ASR models is essential for reducing the reliance on transcribed data. In this paper, we presented an iterative approach for bootstrapping language models by first mining relevant conversational data and web-sites, applying standard MAP decoding, and then regenerating a new model. We evaluated the impact of different learning methods on call classification in spoken dialog systems. At any run-time operating point, unsupervised learning can achieve a call classification performance that is only 1.5%-2% lower than the upper bound set when using all available in-domain transcribed data. Future work will include extending our approach to SLU and studying

Fig. 4. SLU performances using ASR outputs in Figure 3.

the effect of human labeling from ASR output as opposed to transcription.

ACKNOWLEDGMENTS: We would like to thank Junlan Feng for providing us the web data for our experiments.

5. REFERENCES

[1] A. L. Gorin, G. Riccardi, and J. H. Wright, "How May I Help You?," *Speech Communication*, vol. 23, pp. 113–127, 1997.

[2] G. Di Fabbrizio, D. Dutton, N. Gupta, B. Hollister, M. Rahim, G. Riccardi, R. Schapire, and J. Schroeter, "AT&T help desk," in *Proceedings of the ICSLP*, Denver, CO, September 2002.

[3] H. Alshawi, "Effective utterance classification with unsupervised phonotactic models," in *Proceedings of the 2003 NAACL-HLT*, Edmonton, Canada, May 2003.

[4] R. Iyer, H. Gish, and D. McCarthy, "Unsupervised training techniques for natural language call routing," in *Proceedings of the ICASSP*, Orlando, FL, 2002.

[5] M. Bacchiani and B. Roark, "Unsupervised language model adaptation," in *Proceedings of the ICASSP*, Hong Kong, May 2003.

[6] P. S. Rao, M. D. Monkowski, and S. Roukos, "Language model adaptation via minimum discrimination information," in *Proceedings of the ICASSP*, Detroit, MI, May 1995.

[7] D. Hakkani-Tür and G. Riccardi, "A general algorithm for word graph matrix decomposition," in *Proceedings of the ICASSP*, Hong Kong, May 2003.

[8] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," *Machine Learning*, vol. 39, no. 2/3, pp. 135–168, 2000.

[9] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observation of markov chains," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 2, pp. 291–298, 1994.

[10] G. Riccardi and D. Hakkani-Tür, "Active and unsupervised learning for automatic speech recognition," in *Proceedings of the Eurospeech*, Geneva, Switzerland, September 2003.