ACTIVE LEARNING FOR AUTOMATIC SPEECH RECOGNITION

Dilek Hakkani-Tür, Giuseppe Riccardi and Allen Gorin

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

ABSTRACT

State-of-the-art speech recognition systems are trained using transcribed utterances, preparation of which is labor intensive and time-consuming. In this paper, we describe a new method for reducing the transcription effort for training in automatic speech recognition (ASR). Active learning aims at reducing the number of training examples to be labeled by automatically processing the unlabeled examples, and then selecting the most *informative* ones with respect to a given cost function for a human to label. We automatically estimate a confidence score for each word of the utterance, exploiting the lattice output of a speech recognizer, which was trained on a small set of transcribed data. We compute utterance confidence scores based on these word confidence scores, then selectively sample the utterances to be transcribed using the utterance confidence scores. In our experiments, we show that we reduce the amount of labeled data needed for a given word accuracy by 27%.

1. INTRODUCTION

State-of-the-art speech recognition systems require transcribed utterances for training, and transcription is a labor intensive and time-consuming process. Active learning aims at reducing the number of training examples to be labeled by inspecting the unlabeled examples, and intelligently selecting the most *informative* ones with respect to a given cost function for a human to label [1]. The goal of the learning algorithm is to select the examples for labeling which will have the largest improvement on the performance.

In this paper, we describe a new method for reducing the transcription effort for training in ASR, by selectively sampling a subset of the data. For this purpose, we automatically label each word of the utterance with a confidence score, exploiting the lattice output of a speech recognizer, which was initially trained on a small set of transcribed data. We compute utterance confidence scores from the word-based confidence scores, and selectively sample the utterances to be transcribed using these scores.

We test our approach in the framework of AT&T's *How May I Help You?*$^{\mathrm{SM}}$ natural spoken dialog system. Transcription is an important procedure both for extending the system to other domains, and for incorporating new call-types into the existing system. The transcription capability is limited, so selective sampling over the terabytes of speech database is crucial.

In the following, we first describe the related work in the machine learning domain, as well as review some of the related work in language processing. In Section 3, we describe our algorithm, and in Section 4 we describe how we compute confidence scores using the lattice output of ASR. Section 5 describes our experiments and results.

2. RELATED WORK

The search for effective training data sampling algorithms, in order to have better systems with less annotated data by giving the system some control over the inputs on which it trains, has been studied under the title of active learning. Previous work in active learning has concentrated on two approaches: certainty-based methods and committee-based methods. In the *certainty-based methods*, an initial system is trained using a small set of annotated examples [2]. Then, the system examines and labels the unannotated examples, and determines the certainties of its predictions of them. The k examples with the lowest certainties are then presented to the labelers for annotation. In the *committee-based methods*, a distinct set of classifiers is also created using the small set of annotated examples [1, 3]. The unannotated instances, whose annotations differ most when presented to different classifiers are presented to the labelers for annotation. In both paradigms, a new system is trained using the new set of annotated examples, and this process is repeated until the system performance converges to a limit.

In the language processing framework, certainty-based methods have been used for natural language parsing and information extraction [4]. Similar sampling strategies were examined for text categorization, not to reduce the transcription cost, but to reduce the training time by using less training data [5]. While there is a wide literature on confidence score computation in ASR [6, 7, among others], to the authors' knowledge none of these works address the active

learning question for speech recognition.

3. APPROACH

Inspired by the certainty-based active learning methods to reduce the transcription effort, we select the examples that we predict that the speech recognizer has misrecognized, for transcription, and leave out the ones that it has recognized correctly.

We first train a speech recognizer, using a small set of transcribed data, S_t. Using this recognizer, we recognize the utterances that are candidates for transcription, S_u. We then use lattice based confidence measures, to predict which candidates are recognized (in)correctly [8]. We transcribe the utterances that are most likely to have recognition errors. Our algorithm is as follows:

1. Train acoustic and language models, AM_i and LM_i, for recognition, using S_t (i is the iteration number)
2. Recognize the utterances in set S_u using AM_i and LM_i, and compute the confidence scores for all the words
3. Compute confidence scores of utterances
4. Select k utterances which have the smallest confidence scores from S_u, and transcribe them. Call the new transcribed set as S_i
5. $S_t = S_t \bigcup S_i$; $S_u = S_u - S_i$
6. Stop if word accuracy has converged, otherwise go to Step 1

In order to make better decisions in the future selections with respect to the labeling cost, k should be one. However, for efficiency reasons in retraining, it is usually set higher.

4. CONFIDENCE SCORE COMPUTATION

In the literature, there are two leading methods for confidence score estimation. The first one is based on acoustic measurements [6] and the other one is on word lattices. The latter one has the advantage that the probability computation does not require training of an estimator. There are also approaches, which use features from the two types of methods.

We use Mangu *et al.*'s algorithm to compute confusion networks (sausages) from the lattice output of a speech recognizer, and use the word posterior probability estimates on the sausages as word confidence scores [9]. A sausage is a compact representation which specifies the sequence of word-level confusions, that is, the group of words, including a null word, which compete in (approximately) the same time interval, of the candidate hypotheses represented by the lattice. In Figure 1, we demonstrate the general structure of a lattice and a sausage. Each word in the confusion sets has a posterior probability, which is the sum of the probabilities of all the paths that contain that instance, and the sum of the posterior probabilities of all words in a confusion set is equal to 1.

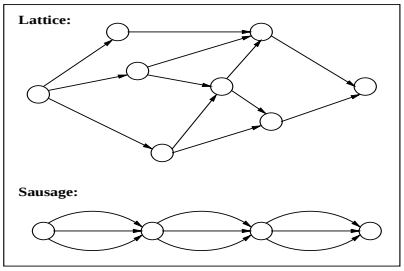

Fig. 1. General structure of lattices and sausages.

Mangu *et al.*'s algorithm takes as input a word lattice, prunes its low probability links, and computes the posterior probability for each link. It first merges different occurrences of the same word, around the same time interval (intra-word clustering), and sums their posterior probabilities. Then, it groups different words which compete around the same time interval, and forms confusion sets (inter-word clustering). The sequence of words with the lowest expected word error rate, the *consensus hypothesis*, is obtained by selecting the word that has the highest posterior probability from each confusion set. More information on the algorithm can be found in [9].

We use the word posterior probability estimates as word confidence scores, which can be interpreted as the probability of being correctly recognized for a word w, $P_{correct}(w)$, and use the notation $C(w_1, ..., w_n)$ to represent the confidence score of the word sequence $w_1, ..., w_n$.

We evaluated different approaches to obtain utterance level confidence measures from word confidence scores that we extract from sausages. One approach is to compute the confidence score of an utterance as the arithmetic mean of the confidence scores of the words that it contains:

$$C(w_1, ..., w_n) = \frac{1}{n} \sum_{i=1}^{n} P_{correct}(w_i) \qquad (1)$$

Another approach is to compute the confidence score of an utterance as the product of the confidence scores of the words that it contains:

$$C(w_1, ..., w_n) = \prod_{i=1}^{n} P_{correct}(w_i)^{\alpha_i(w_i)} \qquad (2)$$

where $\alpha_i(w_i)$ is a scaling function. We also used other func-

tions to compute the utterance confidence scores:

$$C(w_1, ..., w_n) = F(P_{correct}(w_i)) \quad (3)$$

where F can be the geometric mean or the `min` function.

5. EXPERIMENTS AND RESULTS

We performed a series of experiments to verify that the posterior probabilities of the consensus hypothesis can be used to select more informative utterances to transcribe. For these experiments, we used utterances from the database of the *How May I Help You?*$^{\mathrm{SM}}$ system for customer care [10]. The language models used in all our experiments are trigram models based on Variable Ngram Stochastic Automata [11]. The acoustic models are subword unit based, with triphone context modeling and variable number of gaussians (4-24).

5.1. Training and Test Data

The initial set of transcribed utterances, which is used to train the initial acoustic and language models consists of 4,000 utterances (70,000 words). The additional set of transcription candidate utterances consists of 37,720 utterances (664,600 words). The test data consists of 2,076 utterances (30,882 words). All utterances are the responses to the greeting prompt class (e.g., "*Hello. This is AT&T. How May I Help You?*")

5.2. Word Confidence Scores

We use the word posterior probabilities as confidence scores to determine whether they are correctly recognized or not (binary decision). According to this, a word is considered to be correctly recognized if its posterior probability is higher than some threshold, and misrecognized if not. We computed the word posterior probabilities for the utterances in our test set. Figure 2 shows the distribution of the posterior probabilities of the words that have been correctly recognized and misrecognized. The separability between the posterior probability distributions of correctly recognized and misrecognized words suggests that, the posterior probability is a good candidate for a confidence score. Figure 3 shows the receiver operating characteristic (ROC) curve of correct classification versus false rejection rates, by varying the threshold value, when we classify our test data.

Note that the estimation of these confidence scores does not require any training of any type of models (using acoustic or lexical features).

5.3. Results

For active learning in ASR, we trained language and acoustic models using the initial set of 4,000 utterances. Using these models, we then generated lattices and sausages for our additional training data, and computed the confidence scores for words and utterances, as described in Section 4. We incrementally trained language models only, every 4000 utterances ($k = 4000$) (1000 and 2000 utterances at the initial points), and generated learning curves for word accuracy and vocabulary size, which are presented in Figure 4. We plot the results using the arithmetic mean of the word confidence scores (that is, F is the mean function in equation 1), which gave the best results in our case.[1]

From these curves, we see that selective sampling is effective in reducing the need for labeled data (for a given word accuracy). The best performance with random sampling was achieved using all of the training data (7.3×10^5). We achieved the same word accuracy (67.1%) with selective sampling and using 27% less data (with 5.3×10^5 words). Therefore, by selective sampling, it is possible to speed up

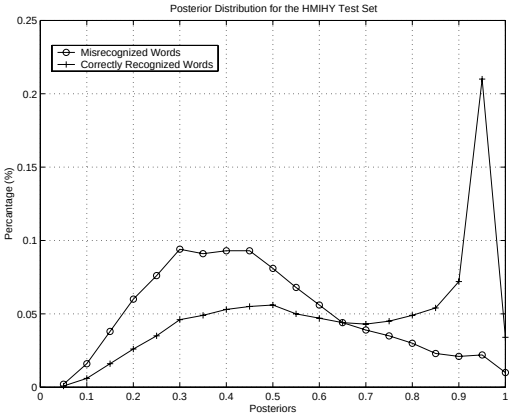

Fig. 2. Distribution of the word posterior probabilities for correctly recognized and misrecognized words.

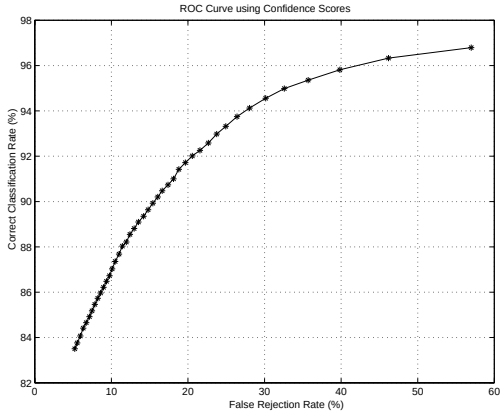

Fig. 3. ROC curves.

[1] We also used the normalized utterance likelihood as a sampling criterion, and it gave inferior performance.

the learning rate of ASR with respect to the amount of labeled transcriptions.

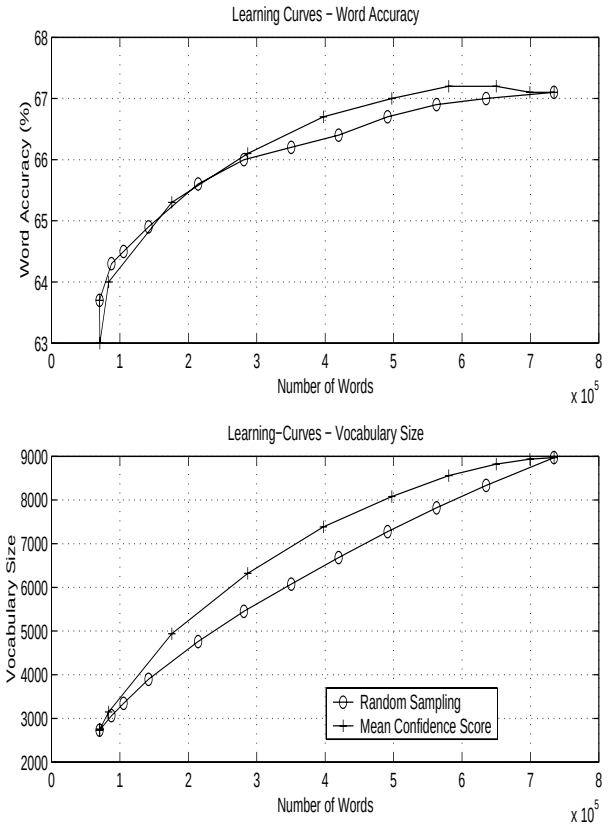

Fig. 4. Learning curves.

6. CONCLUSIONS

We described new methods for reducing the amount of labeled training examples by selectively sampling the most informative subset of data for transcription using lattice based confidence measures. By selective sampling using utterance-level confidence measures, we achieve the same word accuracy results using 27% less data. We have empirically shown that it is possible to detect utterances which have little new information when added to an initial set of utterances.

Acknowledgments We would like to thank Lidia Mangu for providing the sausage computation software, and Anne Kirkland, Murat Saraçlar, Gökhan Tür, and Roberto Gretter for their help with various software.

7. REFERENCES

[1] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning," *Machine Learning*, vol. 15, pp. 201–221, 1994.

[2] D.D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *Proc. of the* 11^{th} *International Conference on Machine Learning*, 1994, pp. 148–156.

[3] I. Dagan and S.P. Engelson, "Committee-based sampling for training probabilistic classifiers," in *Proc. of the* 12^{th} *International Conference on Machine Learning*, 1995, pp. 150–157.

[4] C. Thompson, M. E. Califf, and R.J. Mooney, "Active learning for natural language parsing and information extraction," in *Proc. of the* 16^{th} *International Conference on Machine Learning*, 1999, pp. 406–414.

[5] Y. Yang, "Sampling strategies and learning efficiency in text categorization," in *Proc. of the AAAI Spring Symposium on Machine Learning in Information Access*, M. Hearst and H. Hirsh, Eds. 1996, pp. 88–95, AAAI Press.

[6] R.C. Rose, B.H.Juang, and C.H. Lee, "A training procedure for verifying string hypotheses in continuous speech recognition," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 1995, pp. 281–284.

[7] R. Zhang and A. Rudnicky, "Word level confidence annotation using combinations of features," in *Proc. of* 7^{th} *European Conference on Speech Communication and Technology*, 2001, pp. 2105–2108.

[8] Roberto Gretter and Giuseppe Riccardi, "On-line learning of language models with word error probability distributions," in *Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2001.

[9] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.

[10] A. Gorin, J.H. Wright, G. Riccardi, A. Abella, and T. Alonso, "Semantic information processing of spoken language," in *Proc. of ATR Workshop on Multi-Lingual Speech Communication*, 2000.

[11] G. Riccardi, R. Pieraccini, and E. Bocchieri, "Stochastic automata for language modeling," *Computer Speech and Language*, vol. 10, pp. 265–293, 1996.