Task Adaptation of Acoustic and Language Models Based on Large Quantities of Data

K. Visweswariah, R. Gopinath, V. Goel

IBM T. J. Watson Research Center, Yorktown Heights, NY 10598
kv1, rameshg, vgoel@us.ibm.com

Abstract

We investigate use of large amounts, over 1500 hours, of untranscribed data recorded from a deployed conversational system to improve the acoustic and language models. The system that we considered allows users to perform transactions on their retirement accounts. Using all the untranscribed data we get over 19% relative improvement in word error rate over a baseline system. In contrast, a system built using 70 hours of transcribed data results in over 31% relative improvement.

1. Introduction

In this paper we investigate the use of large amounts of data available from a deployed conversational system to improve the performance of that system. As soon as a system is deployed (possibly as a pilot) and a few callers call in, obvious bugs and some tuning of various parameters can improve the system performance drastically. We are interested in the situation where this initial phase is complete, and we are actually able to collect large amounts of data from the deployment of the system. In this paper we investigate the use of this data to improve the performance of both language and acoustic models. Various papers have investigated unsupervised adaptation/training of acoustic models [1], [2] and language models [3] separately. The system we deal with is a natural language understanding system that is based on a statistical language model.

As large amounts of data become quickly available, the bottleneck in using this data becomes the lack of transcriptions for this acoustic data. Ideally we would like to be able to achieve performance improvements without having to transcribe the data. This paper describes our initial attempts at using this untranscribed data. To contrast this with what is possible with supervised methods we also had a small fraction of the data transcribed.

The outline for the remainder of the paper is as follows. Section 2 describes the application that we collected data from and describes how the baseline acoustic and language models (that were used in the deployment) were built. Section 3 describes the data that we had available to us. Section 5 describes our experimental setup and the our experimental results. Finally in Section 6 we discuss the results of this paper and some directions that we will be pursuing in the future.

2. Application Description

The application that we consider is a conversational system where users perform transactions on their retirement account. Interaction with the system happens in natural language under a mixed initiative dialogue. A sample dialogue with the system is (‘S’ denotes system turn and ‘U’ denotes user turn)

S: How may I help you?
U: Balance, balances please.
S: You have four investments in your account. To stop playback say stop at any time. Your balance in equity income fund is ... What next?
U: Closing price of XYZ plan equity income fund.
S: Closing price of XYZ plan equity income fund as of April one is twenty eight ninety four. What else?
U: (silence)
S: Please speak one of the following. Information or transaction.
U: Price of Spectrum growth funds.
S: Closing price of Spectrum growth fund is fourteen thirty four. How may I help you?
U: I’d like to end this call.
S: Thank you for calling.

A schematic diagram showing the flow of information through the system is presented in Figure 1. The user utterance is first transcribed into text which is then “parsed” by a natural language understanding module which extracts relevant information and passes it to the dialogue manager. The dialogue manager then decides on its next action and sends an appropriate response to language generator which in turn creates a prompt to be played back to the user. The dialogue manager also sends the new dialogue state to the speech recognition module which uses this information for recognition of next user utterance.

Figure 1: A schematic diagram of our NLU based transactional system.

2.1. Speech Recognition Component

For its speech recognition component the system uses an acoustic model that was built on about 1000 hours acoustic data of which about 750 hours was read speech and 250 hours was spontaneous speech data. The acoustic feature vectors were obtained by first computing 13 Mel-cepstral coefficients (including energy) for each time slice under a 25.0 msec. window with a 10 msec. shift. Nine such vectors were concatenated and projected to a 39 dimensional space using LDA, and transformed to match the diagonal Gaussian assumption using MLLT.
The baseline acoustic models were built on these features with a phone set containing 50 phones. Each phone was modeled with a three state left to right HMM. This, in addition to six states of two silence phones results in 156 context independent states which were decision tree clustered to 2198 context dependent states. Each context dependent phone state was modeled with a mixture of 18 Gaussians. The silence states are context independent, they are modeled with a mixture of 120 Gaussians each.

The data to build language model was collected in a dialogue state specific manner with the intention of building state specific language models. People were given a description of the system and some scenarios, specific to some dialogue state, based on which they were asked to write down what they would say. The resulting database had about 150,000 sentences. The baseline language model was a deleted interpolation trigram [4] model. It was built using a vocabulary of 3228 words. The dialogue state specificity was introduced by using the dialogue state as the sentence start context for the model. Then at system run time the dialogue state was used to prime the language model.

3. Data Sets
The data used in our experiments was collected over several months. This came from a system which had already been deployed for a long time before we started our collection effort, so we may presume that the system had reached a stable state. Altogether we collected 2000 hours of audio. The from our collection was manually transcribed. This set had approximately 138,000 sentences and 550,000 words. There were 1820 unique words that were not covered by the vocabulary. However, these out-of-vocabulary words accounted for less than 1% of the total words in the transcribed data.

From the 75 hours of transcribed data, the first five hours collected was set aside for testing and the rest was used for training. The test set contained 3418 sentences and 7760 words.

The word error rate on this test set with our baseline system is 13.83%.

4. Acoustic and Language Model Parameter Estimation
For the amount of data that we have to adapt on it seems like MAP adaptation is the method of choice. Assume we are given adaptation data \(x_1, x_2, ..., x_T\) where \(x_t\) denotes the \(t^{th}\) utterance. Let the parameters in the language model be \(\theta_{LM}\) and the parameters of the acoustic model be \(\theta_{AM}\). We would like to find \(\theta_{LM}\) and \(\theta_{AM}\) that maximize:

\[
\log P(\theta_{LM}, \theta_{AM}) + \sum_w \log \sum_{\gamma(w)} P(x_1, w|\gamma(w)),
\]

where \(P(\theta_{LM}, \theta_{AM})\) is the prior on the parameters and \(\gamma\) is the represents the set of allowed word sequences. Assuming the language model and acoustic model parameters are independent the prior can be written as the product of two distributions \(P(\theta_{LM}, \theta_{AM}) = P_{LM}(\theta_{LM})P_{AM}(\theta_{AM})\). We can then write an auxiliary function to maximize:

\[
\log P_{LM}(\theta_{LM}) + \log P_{AM}(\theta_{AM}) + \sum_{t} \sum_{\gamma} \gamma_i(w) \log P(w)P(x_t|w),
\]

where \(\gamma_i(w)\) is the posterior of the sentence \(w\) given the acoustics \(x_t\).

This splits into the sum of a function of \(\theta_{LM}\) and a function of \(\theta_{AM}\) which can be optimized independently:

\[
\log P_{LM}(\theta_{LM}) + \sum_t \sum_{w \in G} \gamma_t(w) \log P(w)
\]

and

\[
\log P_{AM}(\theta_{AM}) + \sum_t \sum_{w \in \hat{G}} \gamma_t(w) \log P(x_t|w).
\]

We make the assumption, that we intend to remove in the future, that \(\gamma_t(w)\) is 1 for the most likely word sequence and zero for all others.

For the acoustic model we use conjugate priors which gives us the usual updates [5]:

\[
\mu_k = \frac{x > k + \tau \mu_k}{N_k + \tau}
\]

and

\[
\hat{\mu}_k = \frac{x > k + \tau (\hat{\mu}_k + \mu_k)}{N_k + \tau} - \mu_k.
\]

For the acoustic model it is shown in [3] that both model interpolation and count merging are MAP estimates with appropriate choice of prior. In this paper we use count merging or simply retraining the data on the adaptation data.

5. Experimental Results
5.1. Confidence Measure
We decide which data to use to adapt our acoustic and language models based on a confidence measure. The confidence we use is a word based measure which is then averaged over words in a sentence to get a sentence level score. The word level confidence is obtained by a linear combination an acoustic score, a language model score and a posterior score based on an nbest list. The confidence score ranges from -50 to 50 with 50 being the most confident. The method used was a version of algorithms described in [6]. We could use the word level scores and decide to keep some and drop other words in a sentence. This would require fairly accurate segmentation of the acoustic data, to deal with the acoustic and language model context properly in the case of dropped words. In this paper we keep or reject whole sentences.

To evaluate the goodness of our confidence measure, we used the data that was manually transcribed. This data was decoded with the baseline system and for any given confidence threshold, a false acceptance rate (fraction of total sentences that is wrongly decoded but has confidence higher than threshold) and a correct acceptance rate (fraction of total sentences that is correctly decoded and has confidence higher than threshold) are computed. A plot of correct acceptance rate as a function of false acceptance rate for the baseline system is shown in Figure 2 using ‘+’ marks.

From Figure 2 we note that the system built with 70 hours of transcribed data significantly outperforms the system built with untranscribed data only, which in turn is significantly better than the baseline system. The “knee” in the three curves - the midpoint of significant change in their slopes - appears to be happening around a false acceptance rate of 5-7%. However, the confidence thresholds that result in a given false acceptance rate are quite different for the three systems, e.g. for the false acceptance rate of 7% is achieved with a threshold of 0 for the baseline, 5 for unsupervised adapted system and 4 for the system adapted with transcribed data.
5.2. Use of Untranscribed Data

The confidence threshold we use to keep or reject an utterance for use in adaptation potentially has a significant impact on the performance of the adapted system. Clearly we want to keep correctly decoded sentences, but if we set the threshold very high and keep only sentences we are extremely confident about this may not actually help the system since these are sentences we are getting right anyway. To measure the effect of the confidence threshold on the performance of the adapted system, we built several language models using data selected according to several different confidence thresholds. Based on Table 1 we choose to keep sentences that are decoded with confidence above zero. Note that the absolute value of this threshold is system dependent.

<table>
<thead>
<tr>
<th>Confidence threshold</th>
<th>-10</th>
<th>-5</th>
<th>0</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>11.68</td>
<td>11.61</td>
<td>11.64</td>
<td>11.56</td>
<td>12.16</td>
</tr>
</tbody>
</table>

Table 1: Effect of confidence thresholds on unsupervised language model adaptation

Figure 3 shows the WER of LMs built with varying amounts of untranscribed data decoded with the baseline system and filtered with confidence threshold of 0.0. Note that filtering with this threshold we discard roughly half the data. The number of hours in the graph shows the total number of hours before filtering with the confidence threshold. The solid line in the figure corresponds to adding the decoded text to the data that was used to build the baseline LM and the dashed line corresponds to using only the decoded text. Removal of the baseline language modeling data was done under the assumption that that data may have a mismatch from the task specific data and may be hurting the language model. However, from the two curves it appears that use of baseline data is helpful, even with all the untranscribed data that we had. From the figure we note that the first 70 hours of untranscribed data has the largest impact on WER after which the gains are only marginal.

Figure 4: Performance with unsupervised language model and acoustic model adaptation.

5.3. Transcribed Vs Untranscribed Data

Our next set of experiments were carried out using the 70 hours of transcribed data to study how it compares with the untranscribed data.

Figure 5 shows the word error rates with language models
built using varying amounts of transcribed data. We note that the performance gains have not saturated even at 70 hours. We also note that in the case of supervised data it does not seem to matter if we keep the data collected artificially to build the baseline language model.

Our next experiment was to use the transcribed data to build acoustic models with MAP adaptation. In Figure 6 we plot the performance of the systems with acoustic model adaptation on transcribed data as a function of the number of hours of data used. Note that we use a language model adapted with the corresponding amount of data. We see that we get significant gains over the system with adaptation only in the language model. In all of the MAP experiments we used a weight $\tau = 3200$. We did experiment with finding an optimal weight, and although there were improvements with different weights $\tau = 3200$ did fairly well across a variety of adaptation data sizes.

Since the gap between the models adapted on the transcribed data and untranscribed data is significant we investigated if further iterations of unsupervised adaptation would give us further improvements. We started with language model ex-

6. Conclusions

In this paper we considered unsupervised adaptation of language and acoustic models for large amounts of data available from a real deployment of a conversational system. We found that significant improvements (19% relative reduction in WER) can be obtained. We found that this improvement is significantly smaller than the improvement obtained with supervised adaptation (31% relative reduction in WER) with much smaller amount of data. Furthermore the unsupervised adaptation improvements seem to have saturated with respect to the amount of data, whereas for supervised adaptation it seems that significant further gains would be obtained by transcribing more data. For this task, it seems that adapting the language model is far more effective (as compared to the acoustic model) and should be done as a first step.

In future work we hope to explore techniques that would reduce this gap between supervised and unsupervised adaptation and also investigate active learning methods [7] to reduce the amount of transcribed data required.

7. References