DISCRIMINATIVE TRAINING OF ACOUSTIC MODELS APPLIED TO DOMAINS WITH UNRELIABLE TRANSCRIPTS

Lambert Mathias *

Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218

Girija Yegnanarayanan, Juergen Fritsch

Multimodal Technologies, Inc.
319 South Craig St.
Pittsburgh, PA 15213

ABSTRACT

Training Automatic Speech Recognition (ASR) systems require availability of training transcripts for the speech data. Obtaining these transcripts is a time consuming and costly process, especially for the medical domain. On the other hand, medical reports which are generated as a by-product of the normal medical transcription workflow are available easily. However, they only partially represent the acoustic data . In this paper, we present a method for the automatic generation of transcripts from these medical reports [1]. In particular, we identify "reliable" regions in the transcript that can be used for training acoustic models. Experiments based on maximum likelihood (ML) and lattice-based discriminative training with frame filtering are presented. It is shown that discriminative training gives us word error rate (WER) reductions of 8-15% relative to the baseline.

1. INTRODUCTION

Building ASR systems requires availability of speech corpora with accurate orthographic transcriptions. Furthermore, large amounts of acoustic training data can potentially help reduce recognition errors, by estimating model parameters more robustly. However, accurately transcribed training data are not always available. Manually generating transcripts for the vast amounts of raw acoustic data is both time consuming and prohibitively expensive and hence, not a feasible option. This is especially true in the medical transcription domain, which is the primary domain of our application. The medical transcription domain is a very interesting domain as there is potentially an unlimited amount of speech data available for each speaker. However, there are no verbatim transcripts for supervised training of ASR systems - only final medical reports accompanying each of the speech recordings. Medical transcriptionists listen to dictated recordings made by physicians and other healthcare professionals and transcribe them into medical reports. The final medical reports with the grammatical error corrections, removal of disfluencies and repetitions, addition of non-dictated sentence and paragraph boundaries, rearranged order of dictated paragraphs, formatting and other edits, reflect only partially what was being spoken in the original audio recordings. Depending on the speaker, this could account for a significant portion of the speech. However, the medical report can still be explored as an information source for generating training transcripts. Besides, the medical reports are easy to obtain as it is a normal by-product of the medical transcription workflow.

In this paper, we present a method for training acoustic models for the medical domain using automatically generated transcripts. The central idea is to transform the medical reports to spoken form transcripts, then identify reliable regions in the transcripts that can be used for acoustic model (AM) training. Since, discriminative training techniques such as Maximum Mutual Information (MMI) have been shown to outperform ML training in speech recognition tasks [1, 2], we investigate the efficacy of using the automatically generated transcripts in MMI training. More specifically, we present an approach of frame-based filtering for lattice-based MMI training that takes advantage of the reliability information available at the frame level.

The rest of the paper is organized as follows. In Section 2 we present related work. Then in Sections 3 and 4 we describe the automatic transcript generation and discriminative training method. In Section 5 a range of experiments on the medical domain data are presented. Finally, the conclusions and future work are presented in Section 6.

2. RELATED WORK

Recently, there has been considerable interest in lightly supervised acoustic model training [3, 4, 5]. In [4], the acoustic models are decoded using a closed caption (CC) trained language model for supervision. The confusion network (CN) derived from the hypothesized lattices is used to obtain the

*The work presented in this paper was the result of an internship of the first author at Multimodal Technologies, Inc.

[1] Patent pending - application entitled "Document Transcription System Training"

word posteriors. The word posteriors are then averaged per frame of each utterance to get sentence level posteriors. A threshold is then used to filter the utterances. In addition, filtering strategies based on CC matching are investigated. The filtering technique used here works at the sentence level and so tends to throw away significant portion of the training data that might be reliable. In [5], a similar approach is used at the word level. Again, the CC is aligned with the CN. The word hypotheses are selected (using some thresholding scheme) based on their posterior score either from the CN or the CC. Although this approach retains more data, by filtering at the word level it does not account for the fact that an inferior acoustic model may cause a CN hypothesis, that is not the truth, to have a high posterior score. In our work, we explore a frame-level approach for lightly supervised training.

3. AUTOMATICALLY GENERATING "RELIABLE" TRAINING TRANSCRIPTS

The medical transcription domain exhibits the full spectrum of effects seen in conversational/spontaneous speech - non-grammatical sentences, false starts, disfluencies, repetitions and hesitation. The final medical report transcribed by the humans is formatted to ignore or correct such effects and as a result contains insertions (INS), deletions (DEL) and substitutions (SUB) that may or may not represent what was actually spoken. Acoustic model training however, relies on availability of good time alignments of the underlying phone sequence with the acoustic data, which in turn relies on the availability of verbatim transcripts. In order to use the medical reports we need to devise a method to transform the medical reports into trainable transcripts and then identify reliable segments that can be used for model parameter estimation.

3.1. Partially Reliable Transcripts (PRT)

Generating training transcripts from the medical reports can be seen as an iterative procedure, where at each iteration the currently available best AMs are used to generate the orthographic transcriptions, and the medical reports associated with the decoded speech are then used to filter out the training transcripts to be used in the next iteration. The detailed procedure is as follows:

1. Normalize the medical reports to a common format.
2. Generate a report-specific finite state grammar(FSG) for all the available medical reports.
3. Use the normalized medical reports across all speakers to train a language model (LM)
4. Use the best available AMs along with the LM to decode the audio corresponding to the medical reports and generate the orthographic transcriptions.
5. Annotate the automatically generated transcripts by aligning it against the corresponding report-specific FSG and markup the reliable segments in each utterance (explained in Section 3.2).
6. Use the reliable segments to retrain the AMs.
7. Repeat from Step 4.

Note that it is not necessary to train the initial AMs on domain dependent data. For example, the initial models can be trained on English Broadcast news corpus.

3.2. Identifying RELIABLE segments in the PRT

Decoder search errors, lack of coverage of the LM and inferior acoustic models all contribute to the errors in the PRT. As a result, we cannot use the entire PRT for training the AMs. This is especially of concern to us since MMI training is more sensitive to transcription errors than ML training. Hence it is important to identify segments in the transcripts that can be used for training. Such segments are marked as RELIABLE and the rest are marked as UNRELIABLE. The annotated transcripts are generated as follows:

1. Parse the PRT using the report-specific FSG using a robust parser that allows for INS, DEL and SUB.
2. If the parser output matches the PRT at a given position then mark the word as RELIABLE. If the word is in the parser output but not in the PRT then mark it as an INS. If the word is in the PRT but not in the parser output then mark it as a DEL. Finally, if the word at a given position is different than that in the PRT, then mark it as a SUB.
3. Project the word markers onto the frame level. If the word is an INS, DEL or SUB then mark the frames of the underlying phone sequence as "unreliable". Also, allow for crossword context effects by marking the frames of the immediate preceding and succeeding phone as "unreliable". All the other frames are marked as "reliable".

4. MMI TRAINING WITH FRAME FILTERING

MMI training was first proposed in [6] as an alternative to ML training. MMI training attempts to maximize the *a posteriori* probability of the model sequence corresponding to the training data given the training data. It does so by optimizing the objective function that is the log of the ratio of the probability of the training data given the correct models to the probability of the same training data given all the general models.

Since, we cannot use the entire PRT for training, the MMI training framework has to be slightly modified to exploit the reliable/unreliable annotations on the PRT. The obvious approach would be to mark each arc on the MMI

training lattices (both numerator and denominator) as RELIABLE (or UNRELIABLE) depending on the annotation of the corresponding word hypothesis in the PRT. Counts are then accumulated only on the RELIABLE arcs during the estimation procedure. However, this method tends to exclude words which may not be entirely unreliable. A more refined approach is to make use of the frame level reliability markers. In this case, during score accumulation we process only those frames which are marked "reliable". This allows for inclusion of partially reliable words in the training, thereby allowing us to retain more training data. Also, one can account for crossword context effects in the neighborhood of unreliable frames. In this paper, we follow the latter approach.

5. EXPERIMENTS AND RESULTS

The domain of application is the radiology domain, although, the techniques described in this paper can easily be applied to other domains. The aim of the experiments is to evaluate whether speaker-adapted MMI (SA-MMI) trained models can outperform speaker-adapted ML (SA-ML) training when trained on the automatically generated transcripts. Also, of interest is how much data is needed to obtain performance improvements from MMI. The evaluation is done using two different LMs, the speaker-independent language models (SI-LMs) and speaker-dependent language models (SD-LMs).

5.1. Experimental Setup

The radiology training corpus used in this study consists of about 180 hours of recorded speech across 31 speakers (both male and female). The data was recorded at 11kHz using desktop microphones. The speech data is segmented so that for each formatted medical report there is an associated speech utterance. The utterances are typically about 2-20 minutes in length. The 5 speakers chosen for acoustic model adaptation consist of 4 male speakers and 1 female speaker. All the speakers are native English speakers. Approximately 8-12 hours of data are used for acoustic model adaptation of each speaker. All the evaluations were performed on an independent test set.

5.2. Training

ML estimation was used to train speaker-independent (SI) acoustic models from the PRT generated for the 180 hours of training data. Only the reliable frames were used during the count accumulation.The final SI system consisted of about 2000 context-dependent models and 24 mixtures per Gaussian. The vocabulary size was approximately 26k. Since, ASR systems perform best when adapted to a speaker's speech, the SI models were adapted using the PRT for each of the 5 randomly chosen speakers. These final SA-ML models for each of the 5 speakers formed our baseline. Background trigram and unigram SI-LMs were trained on all the training text available for the 31 speakers. In addition, SD-LMs were generated for each of the 5 speakers by interpolating the SI-LM with a speaker-specific LM trained on the corresponding speaker's available training text.

Table 1. WER for SA-ML and SA-MMI models using SI-LM for the 5 speakers(Speakers 1-4 are male and Speaker 5 is female)

Speaker	SA-ML WER	SA-MMI WER	Reliable acoustic data (hours)	%reliable frames
Speaker 1	12.3	10.4	13hrs	69%
Speaker 2	14.5	12.8	7.2hrs	52%
Speaker 3	8.6	7.5	7.5hrs	77%
Speaker 4	12.2	11.3	7.1hrs	76%
Speaker 5	6.2	6.4	7.1hrs	83%

Discriminative training was done using the lattice-based MMI framework as described in [7]. The SA-ML models for each of the speakers and the trigram SI-LM were used to decode the training set and generate word level phone boundary marked numerator and denominator lattices. To improve the generalization of the SA-MMI models, the lattice trigram SI-LM scores were replaced by unigram SI-LM scores and acoustic scaling was also employed to get a broader posterior probability distribution. Prior to the model parameter reestimation step, counts were gathered only for the reliable frames. The unreliable frames were skipped during the count accumulation. Typically, 1 to 2 iterations of MMI were performed to get the final SA-MMI acoustic models.

5.3. Results

All the acoustic models were evaluated on an independent test set. The test set consists of 20 utterances per speaker. "True" manual transcriptions were available for each of these test utterances.

5.3.1. Experiments with SI-LM

The first task was to evaluate the WER of the SA-MMI models and the baseline SA-ML models on an independent test using the SI-LMs. From Table 1 we can see that after MMI the WER reduces by a significant amount (8-15% relative gains) for the first 4 speakers. However, there is no improvement for Speaker 5. This may be due to the fact that Speaker 5 is a high accuracy speaker and there is not enough discriminative information in the lattices for MMI training to benefit from. Column 4 shows the number of hours of acoustic training data that was marked as reliable for each

Table 2. WER for the SA-MMI models using SI-LM trained with different amounts of acoustic data for the 4 male speakers

Speaker	SA-MMI WER			
	0hours	0.6hours	2.5hours	$\geq$7hours
Speaker 1	12.3	10.9	10.5	10.4
Speaker 2	14.5	13.6	12.7	12.8
Speaker 3	8.6	8.8	8.0	7.5
Speaker 4	12.2	11.8	11.2	11.3

Table 3. WER for SA-ML and SA-MMI models using SD-LM for the 4 male speakers

Speaker	SA-ML WER	SA-MMI WER
Speaker 1	11.7	9.6
Speaker 2	14.4	13.0
Speaker 3	7.9	6.8
Speaker 4	11.8	11.5

speaker and column 5 shows what percentage of the total frames per speaker is reliable. Even when only 52% of the frames are reliable, we still get reasonable gains from the training procedure. Of course, the more hours of reliable data we have the better the gains from MMI.

It is also interesting to note how much reliable training data is needed to obtain gains from MMI, for this particular task. Row 2 in Table 2 represents the hours of training data used for MMI. The 0hours indicate that MMI training hasn't been performed for the speaker (our SA-ML baseline). If we compare column 4 in Table 2 with column 2, approximately 2 hours of data per speaker gets us 6-14% relative gains over the baseline SA-ML system.

5.3.2. Experiments with the SD-LM

It is also of interest to evaluate if the gains obtained for the 4 Speakers in Table 1 carry over if we replace the SI-LM with an appropriate SD-LM while decoding using the SA-MMI models for each speaker. For the new baseline, from Table 3 we observe that even though the SD-LM results in the best performance for the SA-ML baseline models, we still manage to get as much as 17% relative reductions in WER for the 4 male speakers.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we investigated the use of automatically generated transcripts for MMI training of speaker-adapted acoustic models. Experimental results show that the MMI training with frame filtering is effective in reducing the WER by as much as 15% relative to the baseline. Furthermore, the gains from MMI carry over when we use a speaker-specific LM for decoding. The main advantage of this approach is that it does not need any kind of transcripts to seed the initial acoustic model training. The techniques can be easily extended to other domains such as broadcast news where the corresponding CCs are available.

Additional improvements can be obtained by generating confusion networks from the MMI models and then do MMI on this confusion network for only those segments with high confusion [8]. Also, instead of simply skipping unreliable frames, we need to explore data re-weighting approaches with weights that reflect the reliability of each frame. Finally, we can explore the use of CNs for generating the PRT. The intuition is that using the CN might help us identify more reliable segments in the PRT.

7. REFERENCES

[1] Y. Normandin, *Hidden Markov Models, Maximum Mutual Information Estimation and the Speech Recognition Problem*, Ph.D. thesis, McGill University, Montreal, 1991.

[2] P. C. Woodland and D. Povey, "Large scale discriminative training for speech recognition," in *Proc. ITW ASR, ISCA*, 2000.

[3] Lori Lamel, Jean-Luc Gauvain, and Gilles Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, pp. 115–129, 2002.

[4] H. Y. Chan and P. C. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training," in *Proc. ICASSP*, 2004, vol. 1, pp. 737–740.

[5] Chen Langzhou, Lori Lamel, and Jean-Luc Gauvain, "Lightly supervised acoustic model training using consensus networks," in *Proc. ICASSP*, 2004, vol. 1, pp. 189–192.

[6] L. R. Bahl, P. F. Brown, P. V. de Souza, and R. L. Mercer, "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," in *Proc. ICASSP*, Tokyo, 1986, pp. 49–52.

[7] P. C. Woodland and D. Povey, "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition," *Computer Speech and Language*, vol. 16, no. 1, pp. 25–47, 2002.

[8] V. Doumpiotis, S. Tsakalidis, and W. Byrne, "Discriminative training for segmental minimum Bayes risk decoding," in *Proc. ICASSP*, Hong Kong, 2003.