IMPROVING BROADCAST NEWS TRANSCRIPTION BY LIGHTLY SUPERVISED DISCRIMINATIVE TRAINING

H.Y. Chan & P.C. Woodland

Cambridge University Engineering Dept, Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {hyc27,pcw}@eng.cam.ac.uk

ABSTRACT

In this paper, we present our experiments on lightly supervised discriminative training with large amounts of broadcast news data for which only closed caption transcriptions are available (TDT data). In particular, we use language models biased to the closed-caption transcripts to recognise the audio data, and the recognised transcripts are then used as the training transcriptions for acoustic model training. A range of experiments that use maximum likelihood (ML) training as well as discriminative training based on either maximum mutual information (MMI) or minimum phone error (MPE) are presented. In a 5xRT broadcast news transcription system that includes adaptation, it is shown that reductions in word error rate (WER) in the range of 1% absolute can be achieved. Finally, some experiments on training data selection are presented to compare different methods of "filtering" the transcripts.

1. INTRODUCTION

In state-of-the-art large vocabulary continuous speech recognition (LVCSR) systems, large amounts of acoustic training data is required for estimating the model parameters robustly. It is also widely believed that using more training data can reduce recognition errors and allow building more complex models. To produce accurate manual transcriptions for the acoustic training data, however, is very time consuming and thus limits the audio data that can be used. In English broadcast news, large amounts of raw audio data are available from the television and radio channels. Closed-captions and commercial transcripts, which are partially correct manual transcripts, are also available for certain broadcasts. Although these transcripts contain a number of errors and can not be used directly as the training transcriptions, Lamel et al. has shown that we can use these transcripts as the supervision data sources for acoustic model training - lightly supervised training [6]. In their experiments, they included the closed-captions in the language model (LM) training materials, and then used the constructed language model to recognise the audio data. The recognised transcripts were then used as the training transcriptions for Maximum Likelihood (ML) training. Their best results were obtained by "filtering" (removing) the recognised transcripts which disagree with the closed-captions. In the Rich Transcription 2003 (RT-03) evaluation, BBN and LIMSI reported word error rate (WER) reductions in their systems by adding subsets of automatically transcribed broadcast news data (TDT4 corpus) that were carefully chosen by closed-caption filtering [3] [7].

In this paper, we also present our experiments on lightly supervised training. In particular, we use language models biased to the closed-caption transcripts to recognise the audio data. Since discriminative training techniques such as Maximum Mutual Information (MMI) and Minimum Phone Error (MPE) have been shown to outperform ML training in LVCSR tasks [8] [12], we investigate the interactions between lightly supervised training and discriminative training. We also aim to improve our English broadcast news transcription system by using the automatically recognised transcripts. As MMI and MPE training can better exploit large amounts of training data than ML training [10], we use all the recognised transcripts for lightly supervised discriminative training, and compare this approach with training data selection schemes such as closed-caption filtering [6].

The rest of the paper is organised as follows. In Section 2, we describe the English broadcast news corpora that used in this work. Then, our lightly supervised discriminative training approach is presented in Section 3. In Section 4, a range of experiments on the TDT data are presented. Different methods of "filtering" the transcripts for training data selection are also compared. Finally, conclusions are given in Section 5.

2. ENGLISH BROADCAST NEWS DATA

2.1. Accurately transcribed acoustic training data

The accurately transcribed broadcast news acoustic training data used in our experiments was released by the LDC in 1997 and 1998. The 1997 data was annotated to ensure that each segment was acoustically homogeneous. The 1998 data was similarly transcribed at the speaker turn level but didn't distinguish between background conditions. The combined set of 1997 and 1998 data contains approximately 143 hours of usable data [4].

2.2. TDT2/TDT4 audio data

In this work, the TDT2 and TDT4 audio corpora were used for lightly supervised training. There are no accurate manual transcriptions for these TDT data but closed-captions are available. The closed-captions, however, are only partially correct and contain a number of errors such as insertions, deletions and changes in the word order [6]. The TDT2 corpus contains 500 hours of audio data broadcast in the first six months of 1998. In our experiments, only the data broadcast between February and June 1998 were used. These include broadcast news shows from the CNN Headline News, ABC World News Tonight, PRI The World,

This work was supported by DARPA grant MDA972-02-1-0013. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

VOA Today and VOA World Report. The TDT4 corpus contains about 300 hours of audio data broadcast between October 2000 and January 2001. These include CNN Headline News, ABC World News Tonight, PRI The World, VOA English news programs, NBC Nightly News and MS-NBC News with Brian Williams.

2.3. Testing data

Two sets of data were used for testing, each of them contains six 30 minutes broadcast news shows. The first set, `dev03`, contains shows which were chosen from the last two weeks of January 2001 of the TDT4 data. The second set, `eval03`, was the RT-03 evaluation data set. It contains shows which were broadcast during February 2001.

2.4. Text corpora

The text corpora used for LM training in this work were the same as the CU-HTK 2003 10xRT system [5]. These include the TDT2, TDT3 and TDT4 closed captions, broadcast news acoustic training data transcriptions, commercial broadcast news transcripts and newswire texts. To conform with the epoch restriction for both the `dev03` and `eval03` test sets, the language models used for testing contained no data from dates after mid January 2001. The language models used for transcribing the TDT4 audio, however, included all the TDT4 closed captions in order to build a biased LM. Approximately one billion words were used in total, in which only 5.8 millions words and 2.5 millions words were contributed from the TDT2 and TDT4 closed captions respectively.

3. LIGHTLY SUPERVISED DISCRIMINATIVE TRAINING

In our lightly supervised discriminative training approach, we used language models biased to the closed-caption transcripts to recognise the TDT audio data. All the recognised transcripts were then used for either maximum likelihood training or to provide the correct transcriptions in discriminative training. We also compared this approach with data selection based on closed-caption filtering and sentence based confidence measure filtering. The details of each process are as follows.

3.1. Biased language models

To build a biased language model, we first constructed one word-based LM for each training data set. No cut-off was applied when we constructed the TDT2 closed-caption LM and the TDT4 closed-caption LM because we wanted to use as much information from the closed-caption transcripts as possible. The individual LMs were then linear interpolated and merged to form a single language model. The interpolation weights were computed to minimise the perplexity of some correct transcriptions corresponding to the TDT audio data. This led the interpolated language model to be biased to the closed-caption LM. The 59k entry wordlist of CU-HTK 2003 10xRT system was used for building fourgram and trigram language models. The out-of-vocabulary (OOV) rates of this wordlist on the TDT2 closed captions and the TDT4 closed captions were 0.76% and 0.85% respectively. For transcribing TDT2 audio data, we used a 10 hour set of accurately transcribed data for perplexity minimisation. The resulting language model had an interpolation weight of 0.92 on the TDT2 closed-caption LM. The OOV rate of the wordlist, fourgram and trigram perplexities on the 10h data set were 0.68%, 21.3 and 44.5 respectively. For transcribing TDT4 audio data, we used the `dev03` accurate transcriptions for perplexity minimisation. The resulting language model had an interpolation weight of 0.90 on the TDT4 closed-caption LM. The OOV rate of the wordlist, fourgram and trigram perplexities on the `dev03` set were 0.47%, 25.6 and 53.2 respectively.

3.2. TDT data transcription

The recognition of the TDT data was performed by a reasonably fast and accurate transcription system. Commercial removal and automatic segmentation were first performed to remove non-speech such as music and noise [11]. Then, the first pass (P1) and second pass (P2) of the CU-HTK 2003 10xRT system [2] [5] were run with our biased LMs. Bandwidth dependent models were used in both passes. In P1, an initial transcription was generated for each segment. A time-synchronous decoder using gender independent (GI) MPE triphones and a word-based trigram LM was used for decoding. The output trigram lattices were then rescored with the fourgram LM. Gender labelling and clustering of the segments were then performed for adaptation purposes in P2. In P2, gender dependent (GD) MPE-MAP [9] triphones were adapted using transforms estimated based on global least squares linear regression and MLLR variance transforms with supervision came from the P1 transcriptions. Decoding were then performed again on each segment to generate trigram lattices using the adapted models. These lattices were further expanded to fourgram and then converted to confusion networks (CN) [1]. The final transcriptions were obtained from the alignments of the confusion network outputs. Approximately 5xRT of computer time was required for the overall transcription process.

Using the biased LMs, the WER of this system on the 10h TDT2 set and the `dev03` set were 9.3% and 8.3% respectively. For comparison, the WER of the TDT2 closed-captions was 10.3% and using the LMs from the CU-HTK 2003 10xRT system gave 12.4% WER on the `dev03` set. 420 hours of TDT2 data were transcribed, in which 370h were classified as wideband data and 50h were classified as narrow-band data. For the TDT4 data, 230 hours of data were transcribed (211h wideband data and 19h narrow-band data, no data from dates after mid January 2001). These TDT data, together with the 143h accurate transcribed broadcast news training data (called bnac in the following sections), were used for acoustic model training.

3.3. Training data selection

In closed-caption filtering [6], the recognised transcripts are first aligned with the closed-caption transcripts by standard dynamic programming. Then, the segments which contain recognition outputs different from the closed-captions are removed from the training set. The motivation of doing this is to remove the segments that may contain recognition errors. In the RT-03 evaluation, BBN and LIMSI chose 90h and 80h subsets of TDT4 data respectively for their systems by this method. In this work, similarly, a 80h subset of TDT4 data which best matches the closed-captions (CC match) was obtained by allowing very small differences between the recognised transcripts and the closed-captions. For comparison, we also produced three 115h subsets of TDT4 data. The first set contained the 115h segments which best matches the closed-captions. The second set contained segments which match the

closed-captions least well (CC mismatch) and the final set was obtained by random selection.

We also investigated data filtering based on confidence measure (CM). In this approach, the confidence score of a word was obtained from the word posterior probability in the confusion network [1]. The sentence confidence (per frame) for each segment was then calculated by averaging the word confidences. Finally, a threshold was set to remove the segments with low sentence confidence. Using this approach, 213h TDT4 data were retained.

3.4. Acoustic model training

Our system used tied-state cross-word triphone HMMs that were constructed by decision tree clustering. There were about 7000 tied states in total and each state contained 16 Gaussian mixture components. Each frame of input speech was represented by a 52 dimensional feature vector that included 13 MF-PLP cepstral parameters (including C0) and their 1st, 2nd and 3rd order derivatives. Cepstral mean normalisation (CMN) was applied on each speech segment. A HLDA transform was also used to project each feature vector down to 39 dimensions.

Discriminative training using MMI or MPE criteria were done in a lattice-based framework [8] [12]. The training data was first re-recognised to generate word lattices. This was done by a single pass decoder using a ML model and an appropriate heavily pruned bigram LM. For bnac data, the pruned bigram LM was trained on the accurate bnac transcriptions. For the TDT2 and TDT4 data, the pruned bigrams of our previously described biased LMs were used. The generated word lattices, as well as the "correct" transcriptions of the training data, were then aligned to find the phone model boundaries with the appropriate model set and produced denominator and numerator model-marked lattices. Suitable statistics were calculated from the lattices so that the same re-estimation formulae of the Extended Baum-Welch (EBW) algorithm could be used to give the parameter updates. I-smoothing [8], which smoothes between the discriminative and the ML estimates, was also applied for both MMI and MPE training to improve the generalisation of the discriminative trained models. A MAP-style adaptation method for MPE training (MPE-MAP) [9] was used to adapt the gender independent MPE models to gender dependent MPE-MAP models. Bandwidth specific models were trained on the data using either wideband analysis or narrow-band analysis (125-3750Hz).

4. EXPERIMENTS

4.1. Test platforms

Two systems were used for testing. The first one was a single pass decoding system without adaptation. It used a trigram language model and GI acoustic model for decoding. The other system, the `P1-P2` system, used the P1 and P2 architecture of CU-HTK 2003 10xRT system [2]. Adaptation was applied to the GD MPE-MAP models in the P2 of the system and final transcriptions were obtained from the alignments of the confusion network outputs. Both systems were bandwidth dependent and ran in a total of ∼5xRT. The dictionary and language models used for testing were taken from the CU-HTK 2003 10xRT system.

4.2. Experimental results of different training data set

Table 1 gives the unadapted single pass decoding results for different training data sets. For simplicity, only wideband models were constructed for each training data set. All decoding of the narrow-band data were done using the narrow-band models which only trained on the 143h bnac data with narrow-band analysis. From the table, it is observed that compared to using 143h bnac data with accurate transcriptions, using 370h automatically recognised TDT2 data alone achieved comparable performances for ML model and better performances for MMI and MPE models. By adding these TDT2 data to the bnac data, WER reductions were obtained. Further adding 50h of narrow-band classified TDT2 data (using wideband analysis) harmed the models. The TDT4 data was more useful than the TDT2 data. More gains were obtained by using TDT4 data as the extra data, even though the amounts of TDT4 data was smaller than TDT2 data. This means that using training data which is close to the time period of the test set can give larger improvements. Further adding the TDT2 data to bnac+TDT4 data gave no improvement for ML and only small improvements for MMI and MPE. For the same training set, MPE always outperforms MMI. This means that MPE is a better discriminative training technique than MMI for both supervised and lightly supervised training. Compared with using bnac data alone, using TDT data together with bnac data obtained greater WER reductions by discriminative training (for both MMI and MPE). This is because discriminative training can better exploit large amounts of training data than ML.

Training data set	ML	MMI	MPE
bnac (143h)	17.9	15.5	15.3
370h wb TDT2	17.7	15.0	14.9
bnac+370h wb TDT2	17.4	14.5	14.2
bnac+420h TDT2	17.4	14.7	14.4
bnac+230h TDT4	16.8	14.4	13.8
bnac+370h wb TDT2 +230h TDT4	16.8	14.1	13.6

(a)

Training data set	ML	MMI	MPE
bnac (143h)	15.9	14.4	13.8
370h wb TDT2	16.1	13.9	13.7
bnac+370h wb TDT2	15.5	13.4	13.0
bnac+420h TDT2	15.8	13.4	13.1
bnac+230h TDT4	15.1	13.3	12.5
bnac+370h wb TDT2 +230h TDT4	15.1	12.9	12.4

(b)

Table 1. %WER on (a) `dev03` and (b) `eval03` for different training data sets. GI unadapted single pass decoding system.

Table 2 gives WER on the test sets using the ∼5xRT `P1-P2` system. From the table, we can see that there was much greater WER reduction for P1 (GI, unadapted, tight beam-widths) than for P2. It seems that some of the gains by using TDT data were diminished after applying adaptation. The best results were obtained from the training set that uses all the data (bnac, TDT2, TDT4), where 1.1% and 0.9% absolute WER reductions were achieved in the P2 output for the `dev03` and `eval03` test sets respectively. These results were even slightly better than the full 10xRT CU-

HTK 2003 system used in the RT-03 evaluation (which gave 11.6% and 10.7% WER on the dev03 and eval03 test sets respectively [5]).

	dev03		eval03	
Training data set	P1	P2	P1	P2
bnac (143h)	16.2	12.5	14.8	11.5
370h wb TDT2	15.8	12.3	14.7	11.8
bnac+370h wb TDT2	15.1	11.9	14.0	11.3
bnac+420h TDT2	15.5	12.0	14.2	11.4
bnac+230h TDT4	14.5	11.8	13.6	10.9
bnac+370h wb TDT2 +230h TDT4	14.5	11.4	13.3	10.6

Table 2. %WER on dev03 and eval03 for different training data sets. P1-P2 system.

4.3. Experimental results of data selection

Table 3 compares different methods of filtering the recognised transcripts. It can be seen that closed-caption filtering was not useful for data selection. Compare with the unfiltered case (bnac+230h TDT4), closed-caption filtering removed large amounts of TDT4 data (80h TDT4 data remained) and the resulting MPE model obtained less gain from using TDT4 data. Moreover, from the results of the three different bnac+115h TDT4 data sets, it seems that choosing data based on the best match with the closed-captions doesn't outperform random selection or even selection based on the poorest match with the closed-captions for both ML and MPE training. The sentence based confidence measure (CM) filtering removed only small amounts of TDT4 data with low confidence, the corresponding models gave comparable performance to the unfiltered models for both ML and MPE. Therefore, we believe that using all or most of the recognised transcripts is the best for MPE in lightly supervised training.

Training data set	ML	MPE
bnac (143h)	17.8	15.0
bnac+80h TDT4 CC match	17.0	14.4
bnac+115h TDT4 CC match	16.9	14.2
bnac+115h TDT4 CC mismatch	17.1	14.3
bnac+115h TDT4 random	16.9	14.3
bnac+213h TDT4 CM	16.7	13.9
bnac+230h TDT4	16.8	13.8

(a)

Training data set	ML	MPE
bnac (143h)	15.6	13.5
bnac+80h TDT4 CC match	15.1	12.9
bnac+115h TDT4 CC match	15.0	12.9
bnac+115h TDT4 CC mismatch	15.2	13.0
bnac+115h TDT4 random	15.2	12.8
bnac+213h TDT4 CM	15.0	12.5
bnac+230h TDT4	15.1	12.5

(b)

Table 3. %WER on (a) dev03 and (b) eval03 for different data selection schemes. GI unadapted single pass decoding system.

5. CONCLUSIONS

This paper has investigated the use of automatically recognised transcripts for MMI and MPE discriminative training to improve the English broadcast news transcription. In particular, these transcripts were recognised by a reasonably fast ($\sim$ 5xRT) state-of-the-art transcription system, with supervision provided by language models biased to the closed captions. The experimental results have shown that reductions in word error rate can be achieved by using these transcripts for acoustic model training. Furthermore, we have also investigated training data selection by closed-caption filtering and sentence based confidence measure filtering. The experimental results have suggested that these data filtering schemes don't appear useful for improving accuracy for MPE.

6. REFERENCES

[1] G. Evermann & P.C. Woodland (2000). "Posterior Probability Decoding, Confidence Estimation and System Combination." *Proc. Speech Transcription Workshop*, College Park.

[2] G. Evermann & P.C. Woodland (2003). "Design of Fast LVCSR Systems." To appear *Proc. ASRU'2003* , St. Thomas.

[3] J.L. Gauvain, L. Lamel, G. Adda, L. Chen & H.Schwenk (2003). "The LIMSI RT03 BN Systems." *Proc. Rich Transcription Workshop*, 2003.

[4] D. Graff (2002). "An Overview of Broadcast News Corpora." *Speech Communication*, Vol.37, pp. 15-26.

[5] D.Y. Kim, G. Evermann, T. Hain, D. Mrva, S.E. Tranter, L. Wang & P.C. Woodland (2003). "Recent Advances in Broadcast News Transcription." To appear *Proc. ASRU'2003* , St. Thomas.

[6] L. Lamel, J.L. Gauvain & G. Adda (2002). "Lightly Supervised and Unsupervised Acoustic Model Training." *Computer Speech and Language* , Vol.16, pp. 115-129.

[7] L. Nguyen, N. Duta, J. Makhoul, S. Matsoukas, R. Schwartz, B. Xiang & D. Xu (2003). "The BBN RT03 BN English System." *Proc. Rich Transcription Workshop*, 2003.

[8] D. Povey & P.C. Woodland (2002). "Minimum Phone Error and I-Smoothing for Improved Discriminative Training." *Proc. ICASSP'02*, pp. I-105-108, Orlando.

[9] D. Povey, P.C. Woodland & M.J.F. Gales (2003). "Discriminative MAP for Acoustic Model Adaptation." *Proc. ICASSP'03*, pp. I-312-315, Hong Kong.

[10] D. Povey (2003). "Discriminative Training for Large Vocabulary Speech Recognition." *Ph. D. Dissertation (draft)*, Department of Engineering, University of Cambridge.

[11] S.E. Tranter, K. Yu, D.A. Reynolds, G. Evermann, D.Y. Kim & P.C. Woodland (2003). "An Investigation into the Interactions between Speaker Diarisation Systems and Automatic Speech Transcription." *Tech. Report*, Cambridge University, CUED/F-INFENG/TR-464.

[12] P.C. Woodland & D. Povey (2002). "Large Scale Discriminative Training of Hidden Markov Models for Speech Recognition." *Computer Speech and Language*, Vol. 16 No. 1, pp. 25-47.