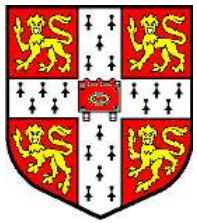
Experiments with lightly supervised discriminative training on TDT data

H.Y. Chan and P.C. Woodland

5th Sept 2003



Cambridge University Engineering Department

Overview

- Improve the HTK English Broadcast News system by adding large amount of TDT data
  - 144 hours of accurately transcribed data
  - TDT2 (450h raw data)
  - TDT4 (300h raw data)
  - Only closed-caption transcripts are available for TDT data
- Lightly supervised acoustic model training
- Investigate interactions between discriminative training and lightly supervised acoustic model training
- Compare with other sites' approaches

Lightly supervised training

- Recognize the TDT data to get the training transcriptions
  - use reasonably fast and accurate transcription system
- Language model includes closed-captions
  - construct individual LM from each text source and perform interpolation
  - bias the interpolated LM to the closed-caption sources
- Use all TDT data for training
- Compare with data selection, which filters automatically transcribed data
  - closed-captions filtering (LIMSI/BBN approach)
  - filtering based on sentence confidence score

Training and Testing Data Sets

- Acoustic training data

  bnac 144 hours broadcast news acoustic with accurate transcriptions
  TDT2 Feb 1998 - June 1998 (902 shows, $\sim$ 450h raw data)
  TDT4 Oct 2000 - Jan 2001 (448 shows, $\sim$ 300h raw data)

- Text corpora: TDT2 cc (closed captions), TDT3 cc, TDT4 cc, Marketplace and BN acoustic training transcriptions, PSM broadcast news transcriptions, CNN, commercial newswire, all before end of Jan 2001

- Test sets

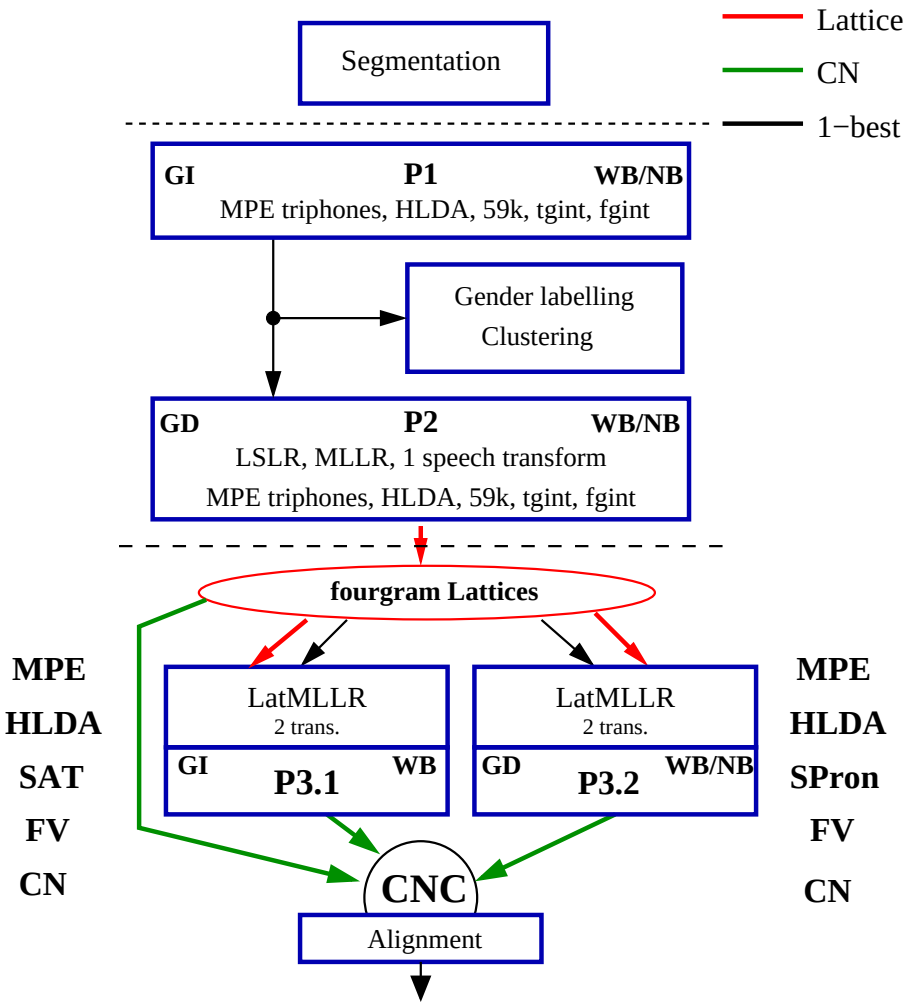
  dev03 17th Jan 2001 - 31 Jan 2001 (6 shows, 3h data)
  eval03 Feb 2001 (6 shows, 3h data)

TDT data transcription: Recognition LM

- Interpolated word LM (tg, 4g): one model for each LM data set
- 59K word-list of CU-HTK 2003 10xRT system
  - 0.76% OOV rate on TDT2 closed-captions (23K unknown words)
  - 0.85% OOV rate on TDT4 closed-captions (14K unknown words)
- Transcribing TDT2 - minimize perplexity of 10h accurate transcription set
  - trigram perplexity is 44.5, fourgram perplexity is 21.3
  - OOV rate is 0.68%
  - interpolation weight for TDT2 model is 0.92
- Transcribing TDT4 - minimize perplexity of dev03
  - trigram perplexity - 53.2, fourgram perplexity - 25.6
  - OOV rate is 0.44%
  - interpolation weight for TDT4 model is 0.90

TDT data transcription: Decoding

- Recognition of TDT data
  - Automatic segmentation
  - P1, P2 of CU-HTK 2003 10xRT system
  - Confusion network, re-alignment
  - $\sim$ 5xRT
- WER on 10h TDT2 data (P2+CN)
  - biased TDT2 LM: 9.3%
- TDT2 closed-captions WER is 10.3%
- WER on dev03 (P2+CN)
  - biased TDT4 LM: 8.3%
  - CU-HTK 2003 10xRT system LM: 12.4%



Acoustic model training

- Wide-band model training data
  - recognized TDT2: 420h (370h wb, 50h nb)
  - recognized TDT4: 255h (234h wb, 21h nb)
  - 144h bnac
- Narrow-band model training data: 144h bnac only (NB analysis)
- MLE, HLDA, MMI, MPE acoustic modeling
- cross-word triphone, $\sim$ 7000 tied states, 16 Gaussian mixture components
- Discriminative training
  - I-smoothing for both MMI and MPE
  - numerator: accurate bnac transcriptions / recognized TDT transcriptions
  - denominator: re-recognize with MLE+HLDA model
  - Heavily pruned bigram for both word lattices and phone-marked lattices

Testing

- Test set - dev03, eval03
- Single Pass decoding system
  - Gender Independent
  - Trigram LM
  - No adaptation
  - Controlled pruning beam width for $\sim$ 5xRT
- CU-HTK P1-P2 system
  - P1, P2 architecture of CU-HTK 2003 10xRT evaluation system
  - Confusion network, re-alignment
  - overall $\sim$ 5xRT include adaptation
- Dictionary, trigram LM, fourgram LM are the same as CU-HTK 2003 10xRT evaluation system

Unadapted single pass decoding WER - dev03

Wide-band data	MLE	MLE+HLDA	MMI+HLDA	MPE+HLDA
bnac (144h)	19.7	17.9	15.5	15.3
370h wb TDT2	19.5	17.7	15.0	14.9
bnac+370h wb TDT2	19.3	17.4	14.5	14.2
bnac+420h TDT2	19.3	17.4	14.7	14.4
bnac+255h TDT4	18.7	16.9	14.2	13.7
bnac+370h wbTDT2 +230h TDT4		16.8		13.6

- Using 370h wb TDT2 alone outperforms bnac with 144 hours of data
- Adding TDT2/TDT4 to bnac improve the performance
- TDT4 data is more useful than TDT2 data
- Further adding TDT2 on bnac+TDT4 only gives very small improvement

Unadapted single pass decoding WER - eval03

Wide-band data	MLE	MLE+HLDA	MMI+HLDA	MPE+HLDA
bnac	17.8	15.9	14.4	13.8
370h wb TDT2	17.1	16.1	13.9	13.7
bnac+370h wb TDT2	17.1	15.5	13.4	13.0
bnac+420h TDT2	17.2	15.8	13.4	13.1
bnac+255h TDT4	17.1	15.1	13.2	12.6
bnac+370h wb TDT2 +230h TDT4		15.1		12.4

- MPE always outperforms MMI
- More gains are obtained in discriminative stage than maximum likelihood
- Adding narrow-band TDT2 data (wide-band analysis) harms the model 0.1%-0.3%

CU-HTK P1-P2 System WER - dev03/eval03

Acoustic model	dev03		eval03	
	P1	P2	P1	P2
bnac	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+255h TDT4	14.5	11.4	13.6	10.7

- More WER reduction in P1 (GI, unadapted, tight beam-widths) than in P2
- Adding 255h TDT4 to bnac
  - 1.1% (dev03) and 0.8% (eval03) WER reduction in P2 output
  - 10.7 WER on eval03, the same as full CU-HTK 2003 10xRT evaluation system
- Adding TDT2 to bnac obtain much less gain, 0.6% (dev03) and 0.2% (eval03) WER reduction in P2 output

Data selection experiments with TDT4

- Closed-captions filtering
  - align the recognized transcriptions with the CC on a whole show basis
  - CC match: only retain segments which match best with CC
  - CC mismatch: only retain segments which match worst with CC
- Confidence measure filtering
  - word posterior from confusion network as confidence score for a word
  - averaging the word posterior to get the sentence confidence (per frame)
  - remove training sentences with low confidence
- Remove the last two weeks of TDT4 data which cover the time period of dev03, 230h TDT4 data is remaining (new baseline)

Data selection: unadapted single pass decoding - dev03

Wide-band data	MLE+HLDA	MPE+HLDA
bnac	17.8	15.0
bnac+80h TDT4(94% CC match)	17.0	14.4
bnac+115h TDT4(90% CC match)	16.9	14.2
bnac+115h TDT4(CC mismatch)	17.1	14.3
bnac+213h TDT4(0.85 CM)	16.7	13.9
bnac+230h TDT4	16.8	13.8

- Very small difference in performance on CC match and CC mismatch!!
- Confidence measure filtering doesn't appear to reduce WER
- Using all data is the best for MPE
- bnac+230h TDT4 - 1% and 1.2% WER reduction in MLE and MPE respectively

Data selection: unadapted single pass decoding - eval03

Wide-band data	MLE+HLDA	MPE+HLDA
bnac	15.6	13.5
bnac+80h TDT4(94% CC match)	15.1	12.9
bnac+115h TDT4(90% CC match)	15.0	12.9
bnac+115h TDT4(CC mismatch)	15.2	13.0
bnac+213h TDT4(0.85 CM)	15.0	12.5
bnac+230h TDT4	15.1	12.5

- Similar pattern as in the results on dev03
- bnac+115h TDT4(90% CC match) - reduces 0.6% WER for both MLE and MPE over bnac baseline
- bnac+230h TDT4 - reduces 0.5% and 1.0% WER for MLE and MPE respectively

Data selection: CU-HTK P1-P2 System - dev03/eval03

Acoustic model	dev03		eval03	
	P1	P2	P1	P2
bnac	15.9	12.6	14.9	11.5
bnac+80h TDT4 94% CC match	15.1	12.2	14.0	11.0
bnac+115h TDT4 90% CC match	15.1	11.9	13.9	10.9
bnac+115h TDT4 CC mismatch	15.2	11.9	13.9	11.0
bnac+230h TDT4	14.5	11.8	13.6	10.9

- Adding 230h TDT4 to bnac
  - 0.8% (dev03) and 0.6% (eval03) WER reduction in P2 output
  - much better performance than all CC filtering in P1 output
- CC filtering
  - no difference in performance between CC match and CC mismatch
  - adding 115h TDT4 to bnac give P2 results only slightly worse than adding 230h TDT4 to bnac data

Conclusion

- Successfully apply MPE and MMI for lightly supervised discriminative training
- MPE outperforms MMI in both supervised and lightly supervised training
- Closed-captions filtering and sentence based confidence measure filtering don't appear useful for improving recognition accuracy for MPE
- The best MPE result comes from the model trained with all data
- By adding 255h TDT4 data to 144h broadcast news acoustic training data, on eval03
  - 0.8% absolute WER rate decrease on the 5xRT P1-P2 CU-HTK system
  - 10.7% WER, which is the same as the CU-HTK 2003 10xRT Broadcast News Evaluation system

Conclusion

- By adding 230h TDT4 data, on dev03
  - 0.8% absolute WER rate decrease in the 5xRT P1-P2 CU-HTK system
  - 11.8% WER, compare with 11.6% WER of the CU-HTK 2003 10xRT Broadcast News Evaluation system
- Adding outdated TDT2 on bnac+TDT4 doesn't give much further WER reduction with current set-up
- Future work
  - Use more complex models for more data
  - Improve lightly supervised discriminative training procedures