Ongoing Experiments with Lightly Supervised Discriminative Training

H.Y. Chan, M.J.F. Gales and P.C. Woodland

May 2004

Cambridge University Engineering Department

Overview

- Experiments on using TDT4 corrected closed-captions
  - Data processing
  - Acoustic modelling
  - Compare with using automatically recognised transcripts
- Lightly supervised discriminative training
  - Combine corrected closed-captions with automatic recognised transcripts
  - Incorporate confidence score into MPE
- Both unadapted and adapted systems

TDT4 corrected closed-caption data

- STT community asked LDC to create this data with the following aims:
  - obtain improved (not perfect) training transcriptions for TDT4
  - investigate impact of lightly supervised data on discriminative training
  - learn nature of lightly supervised discriminative training
- LIMSI provided segmentation and automatic transcriptions
- LDC manually resolved disagreements between cc and LIMSI transcriptions (in a few times real time)

Training and Test Data Sets

- Acoustic training data

  TDT4 Oct 2000 - Jan 2001
  - 450 shows, $\sim$ 300h recorded broadcasts
  - perform automatic segmentation
  - remove 2nd half of Jan 2001 data
  - 230h acoustic data remain

  Corrected closed-captions subset
  - 370 shows, $\sim$ 250h recorded broadcasts

- Test sets

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

Corrected Closed-Captions Processing

- Original transcriptions: 187h data from the stm files (250h recorded broadcasts from 370 shows)
- Normalise the text and apply replacement rules
  - Abbreviations, compound words, typos, ...
  - e.g. NBA → N. B. A., ACTUALY → ACTUALLY
  - About 1300 replacement rules were produced
- Produce 1200 pronunciations for unknown words with frequency greater than 1 (6 more hours of data can be retained)
- 3900 OOV words remain → remove 12h segments with unknown words
- align the segments and fix silence boundaries → remove 15h data
- remove 2nd half of Jan 2001 data → (320 shows, 142h data remain)

Training Sets and Transcriptions

- Corrected closed-captions (CC) training data
  - corrected CC 320 shows (136h)
  - corrected CC 320 shows with larger dictionary (142h)
  - corrected CC + remaining lightly supervised segments (223h): i.e. plus all the lightly supervised segments that do not overlap with the corrected CC segments

- Lightly supervised training data
  - lightly supervised best match corrected CC (136h): choose segments which overlap with the corrected CC segments in time as much as possible
  - lightly supervised 320 shows (180h)
  - lightly supervised all shows (230h)

Acoustic Modelling and Testing

- Acoustic model
  - cross-word triphone, 7000 tied states, HLDA front-end
  - 16 Gaussian mixture components per state
  - Gender independent MPE models, gender dependent MPE-MAP models
- Single Pass decoding system
  - Trigram LM
  - No adaptation
  - Gender Independent
- CU-HTK P1-P2 system
  - P1, P2 architecture of CU-HTK 2003 10xRT evaluation system
  - GI MPE model for P1, GD MPE-MAP models for P2
  - Trigram decoding, fourgram lattice rescoring
  - overall $\sim$ 5xRT include adaptation

Unadapted Single Pass Decoding WER

	dev03	eval03
	MPE	MPE
corrected CC 320 shows (136h)	14.5	13.5
lightly supervised best match corrected CC (136h)	14.9	13.6
corrected CC 320 shows (142h)	14.5	13.5
lightly supervised 320 shows (180h)	14.7	13.4
corrected CC+remaining lightly supervised (223h)	14.3	13.2
lightly supervised all shows (230h)	14.6	13.1

- Corrected CC (136h) better than recognised transcripts by 0.4%/0.1%
- Corrected CC (142) better than lightly supervised (180h) by 0.2%/-0.1%
- Combination of corrected CC and lightly supervised better than lightly supervised by 0.3%/-0.1%

CU-HTK P1-P2 System WER

	dev03	eval03
	P2	P2
corrected CC 320 shows (136h)	12.2	11.7
lightly supervised best match corrected CC (136h)	12.6	11.7
corrected CC 320 shows (142h)	12.2	11.5
lightly supervised 320 shows (180h)	12.4	11.5
corrected CC+remaining lightly supervised (223h)	12.1	11.3
lightly supervised all shows (230h)	11.9	11.4

- Similar pattern as in unadapted GI results
- Training with corrected CC doesn't outperform lightly supervised training

Numerator Word-Network for MPE Training

- Method 1
  - select segments at least 98% overlap with corrected CC segments in time
  - put them in parallel and merge into word-networks
  - 102h numerator word-networks are created
  - perform MPE training with these numerator word-networks as well as the remaining automatic recognised transcripts
- Method 2
  - perform recognition in the word-networks
  - use the first best results as the transcriptions
- Using 98% as the minimum overlap percentage, the automatic recognised transcrips have 7% word tokens which disagree with the corrected CC

Results for Numerator Word-Network

- Unadapted single pass decoding WER [136h training]

	dev03	eval03
corrected CC 320 shows	14.5	13.5
lightly supervised best match corrected CC	14.9	13.6
numerator word-network	14.9	13.5
first best result recognised from word-network	14.8	13.5

- CU-HTK P1-P2 System WER [136h training]

	dev03		eval03	
	P1	P2	P1	P2
corrected CC 320 shows	15.2	12.2	14.3	11.7
lightly supervised best match corrected CC	15.7	12.6	14.6	11.7
numerator word-network	15.7	12.7	14.6	11.8
first best result recognised from word-network	15.7	12.6	14.7	11.8

- No gain is obtained by using numerator word-network

Using Confidence Scores in MPE

- Incorporate confidence scores into MPE training

- Word posterior from confusion network used as the confidence score for a word

- Two different approaches
  - MPE with confidence: multiply the phone accuracy of a lattice arc by its corresponding confidence score
  - MPE with confidence mask: set the confidence score either as 1 or 0 based on a threshold (∼10% of data with 0 confidence score)

Results for MPE with Confidence Scores

- Unadapted single pass decoding WER [230h training]

	dev03	eval03
corrected CC+remaining lightly supervised	14.3	13.2
Standard MPE	14.6	13.1
MPE with confidence	14.5	13.0
MPE with confidence mask	14.4	13.0

- CU-HTK P1-P2 System WER [230h training]

	dev03		eval03	
	P1	P2	P1	P2
corrected CC+remaining lightly supervised	15.1	12.1	14.2	11.3
Standard MPE	15.3	11.9	14.1	11.4
MPE with confidence	15.2	11.9	13.9	11.4
MPE with confidence mask	15.1	12.2	14.0	11.4

- No gain is obtained by incorporating confidence score into MPE after performing adaptation

Conclusions

- Comparison between training with corrected CC and lightly supervised training
  - Only 7% disagreement in word tokens between corrected CC and automatic recognised transcripts
  - No significant difference in performance!
- Investigate several techniques for lightly supervised discriminative training
  - They don't appear useful for improving accuracy for MPE
- Still need to find ways to improve lightly supervised discriminative training