Advances in Structural Metadata at CUED

Marcus Tomalin and Phil Woodland

1st May 2004

Cambridge University Engineering Department

Contents

Advances in Structural Metadata (SMD):

- The CUED CTS SU-Detection system.
  - overview of the system.
  - down-sampling PFM training data.
  - ensembles of PFMs.
- The CUED BN SU-Detection system.
  - overview of the system.
  - down-sampling PFM training data.
  - using additional SULM training data.

CUED CTS SU-Detection System

CUED CTS SU-Detection System Overview:

- **Training Data** = LDC data (30 hrs).
- **Test Data** = dev03f data (3 hrs) eval03f data (3 hrs).
- RT-03 CU-HTK CTS STT 187×RT system output (with optionally deletable tokens retained) used as input to MDE system.
- **Prosodic Feature Model (PFM)**:
  - 10 prosodic features (1 pause, 1 duration, 5 F0, 3 energy).
  - PFMs = CART decision trees.
- **Slash Unit Language Model (SULM)**:
  - N-gram and Class-based SULMs built.
  - Interpolation Weights and Perplexities calculated using stream info for SU tokens only.
  - SULM = Interpolated trigram, 40-class trigram and 40-class fourgram.
- **Lattice-based Decoder**:
  - Decoder = 1-Best Posterior Decoding.

Down-Sampling CTS Training Data

The distribution of SU and non-SU tokens in the PFM training data (td):

LDC Training Data	Total # Toks	% Non-SU Toks
td_14-86 (EVAL03F-SYS)	465,000	86%

The PFM sample space can be modified to reduce the non-SU token percentage:

LDC Training Data	Total # Toks	% Non-SU Toks
td_30-70	254,950	70%
td_40-60	191,215	60%
td_50-50	152,972	50%
td_60-40	127,480	40%
td_70-30	109,270	30%

PFMs can be constructed in the usual way using these modified sample spaces. The PFMs were built using the R software.

Down-Sampling CTS Training Data

For a single PFM, the following results were obtained for down-sampling:

SYSTEM	DEL		INS		%Err	
	dev03f	eval03f	dev03f	eval03f	dev03f	eval03f
Baseline (EVAL03F-SYS)†	33.0	32.0	15.0	17.9	**48.0**	**49.9**
#PFMs_1 td_30-70 + SULM	38.8	40.5	11.0	10.5	49.9	51.0
#PFMs_1 td_40-60 + SULM	35.9	37.3	12.4	12.6	48.3	49.9
#PFMs_1 td_50-50 + SULM	33.4	34.6	14.0	15.0	47.3	49.6
#PFMs_1 td_60-40 + SULM	30.6	32.2	16.4	17.0	**47.0**	**49.3**
#PFMs_1 td_70-30 + SULM	29.1	29.7	18.8	19.8	47.8	49.5

† The PFM used in EVAL03F-SYS was built using CUED-internal code.

Down-sampling can reduce the Err by c.0.8% abs.

NB: all SU results in these slides were obtained using exact end detection statistics output by mdeval-v08.pl with the settings '-w -W -t 1.00' specified.

Ensembles of PFMs

A single PFM was used in EVAL03F-SYS, but an ensemble of PFMs can be used:

1. Partition the PFM training data into into two sets:
   the set of all SU tokens, S, and the set of all non-SU tokens, L.

2. Select N subsets, $D_{1...N}$, from L using random sampling.

3. Combine S with each of the D_is to create N sets of training data.

4. Construct a separate PFM using each of the N sets of training data.

The probabilities obtained from the N PFMs are combined without weights.

NB: **#PFMs_N** $=$ an ensemble of N PFMs.

Results for Ensembles of PFMs

The results for different ensembles of PFMs are as follows:

SYSTEM	DEL		INS		%Err	
	dev03f	eval03f	dev03f	eval03f	dev03f	eval03f
#PFMs_1 td_50-50	33.4	34.6	14.0	15.0	47.3	49.6
#PFMs_1 td_60-40	30.6	32.2	16.4	17.0	**47.0**	**49.3**
#PFMs_1 td_70-30	29.1	29.7	18.8	19.8	47.8	49.5
#PFMs_10 td_50-50	33.2	34.9	13.9	14.9	47.1	49.8
#PFMs_10 td_60-40	30.8	32.0	16.0	16.8	**46.9**	48.8
#PFMs_10 td_70-30	28.8	29.1	18.5	19.5	47.3	**48.6**
#PFMs_20 td_50-50	33.2	34.8	13.9	14.8	47.1	49.7
#PFMs_20 td_60-40	30.8	32.0	16.1	16.8	**46.9**	48.8
#PFMs_20 td_70-30	28.9	29.1	18.6	19.4	47.6	**48.5**

There are some small gains using ensemble techniques, but the gains are not consistent across the dev03f and eval03f test sets.

BN SU-Detection System

Since Feb 2004 we have built a BN SU-Detection System

The basic stages in the process are:

- Classify segments in the training data into gender subtypes (M, F) and bandwidth subtypes (WB, NB).
- Generate forced alignments for gender/bandwidth data subsets.
- Construct PFMs using the forced alignments.
- Construct SULMs using training data.
- Combine the PFM and SULM information using a decoder.

[NB: This is still 'work in progress'!]

BN SU-Detection System

CUED BN SU-Detection System Overview:

- **Training Data** = LDC BN Data (c.20hrs).
- **Test Data** = dev03f data (1.5 hrs), eval03f data (1.5 hrs).
- RT-03 CU-HTK BN STT 10×RT system output (with optionally deletable tokens retained) used as input to MDE system.
- **Prosodic Feature Model (PFM)**:
  - 10 prosodic features (1 pause, 1 duration, 5 F0, 3 energy).
  - PFM = CART decision tree.
- **Slash Unit Language Model (SULM)**:
  - N-gram and Class-based SULMs built (e.g., tg = trigram, cl40-tg = 40 class trigram).
  - Interpolation Weights and Perplexities calculated using stream info for SU tokens only.
- **Lattice-based Decoder**:
  - Decoder = 1-Best Posterior Decoding.

BN SU-Detection System

Initially, all the LDC training data was used (without down-sampling).

Results were obtained for SULMs only, and also for a single PFM + SULMs:

SYSTEM	DEL		INS		%Err	
	dev03f	**eval03f**	**dev03f**	**eval03f**	**dev03f**	**eval03f**
tg	67.9	67.5	17.1	13.2	85.0	**80.6**
tg+cl40-tg	63.8	64.5	17.4	16.2	**81.1**	**80.6**
PFM + tg	73.8	73.8	12.9	10.4	86.7	84.2
PFM + (tg+cl40-tg)	71.5	69.5	14.7	13.4	**86.3**	**82.9**

When all the LDC training data is used, the PFM degrades the performance of the system (!).

Down-Sampling BN Training Data

Down-sampling was used to improve the performance of the PFMs.

The distribution of SU and non-SU tokens in the PFM training data (td) is:

LDC Training Data	Total # Toks	% Non-SU Toks
td_08-92 (no down-sampling)	185,940	92%

Down-sampling can reduce the non-SU token percentage:

LDC Training Data	Total # Toks	% Non-SU Toks
td_30-70	50,757	70%
td_40-60	38,068	60%
td_50-50	30,454	50%
td_60-40	25,378	40%
td_70-30	21,753	30%

The subsets were selected from the set of all non-SU tokens using sampling without replacement.

Down-Sampling BN Training Data

Single PFMs were constructed using the reduced sample spaces.

For the single PFMs, the following results were obtained for down-sampling (with SULM = tg+cl40-tg):

SYSTEM	DEL		INS		%Err	
	dev03f	eval03f	dev03f	eval03f	dev03f	eval03f
(PFM td_08-92) + SULM	71.5	69.5	14.7	13.4	**86.3**	**82.9**
(PFM td_30-70) + SULM	61.9	62.5	17.6	16.0	79.6	78.5
(PFM td_40-60) + SULM	58.3	58.6	19.6	16.7	77.9	75.3
(PFM td_50-50) + SULM	58.2	56.4	18.6	18.0	76.8	74.4
(PFM td_60-40) + SULM	56.0	55.5	21.0	19.6	76.9	75.1
(PFM td_70-30) + SULM	55.8	51.3	19.9	22.5	**75.7**	**73.8**

These results show that down-sampling improves the performance of the PFM, lowering SU Err by c.10% abs.

Using Additional BN Training Data in SULM

Since SUs appear so infrequently in the LDC training data, it is necessary to consider additional training data:

BN Corpus: DB98 (100 hrs of Hub-4 data, 1998)

The DB98 data contains punctuation marks (full-stops, commas, question marks).

The intention was to overgenerate SUs in the SULM to reduce the DEL error.

This data was processed as follows:

1. Map punctuation marks in DB98 to SU tokens:
   full-stops → statement, commas → statement, question marks → question.
2. Convert DB98 data into SULM training data files.
3. Build SULMs for the DB98 data.
4. Build interpolated SULMs using the LDC and DB98 SULMs (i.e., LDC+DB98 SULMs).

Although acoustic data is available for the DB98 data, so far it has only been included in the BN SULMs.

Using Additional BN Training Data

Results for the LDC and DB98 SULMs (with no PFM):

SYSTEM	DEL		INS		%Err	
	dev03f	eval03f	dev03f	eval03f	dev03f	eval03f
LDC tg	67.9	67.5	17.1	13.2	85.0	**80.6**
LDC tg+cl40-tg	63.8	64.5	17.4	16.2	**81.1**	**80.6**
DB98 tg	46.3	43.0	40.3	41.7	87.2	88.1
DB98 tg+cl40-tg	44.3	43.3	39.6	43.8	**83.9**	**87.0**
LDC+DB98 tg	50.3	46.5	32.6	34.5	**82.9**	80.9
LDC+DB98 tg+cl40-tg	49.2	48.1	34.0	32.0	83.2	**80.1**

The DB98 SULMs reduce the DEL error by c.20% abs (while increasing the INS error by c.20% abs) compared to the LDC SULMs.

Using Additional BN Training Data

Results for LDC and DB98 SULMs when combined with a PFM:

SYSTEM	DEL		INS		%Err	
	dev03f	eval03f	dev03f	eval03f	dev03f	eval03f
(PFM td_50-50) + (LDC tg)	65.7	62.5	15.4	14.2	81.1	76.8
(PFM td_50-50) + (LDC tg+cl40-tg)	58.2	56.4	18.6	18.0	76.8	74.4
(PFM td_70-30) + (LDC tg)	61.0	57.0	16.3	19.6	77.3	76.6
(PFM td_70-30) + (LDC tg+cl40-tg)	55.8	51.3	19.9	22.5	**75.7**	**73.8**
(PFM td_50-50) + (DB98 tg)	38.5	35.9	42.2	45.1	80.7	80.9
(PFM td_50-50) + (DB98 tg+cl40-tg)	35.3	32.8	42.1	45.4	77.4	**78.2**
(PFM td_70-30) + (DB98 tg)	34.7	34.4	42.6	46.9	77.4	81.3
(PFM td_70-30) + (DB98 tg+cl40-tg)	32.0	32.4	42.6	50.5	**74.7**	83.0
(PFM td_50-50) + (LDC+DB98 tg)	43.1	38.9	34.8	36.3	77.9	75.2
(PFM td_50-50) + (LDC+DB98 tg+cl40-tg)	41.0	35.8	35.4	35.7	76.4	**71.4**
(PFM td_70-30) + (LDC+DB98 tg)	38.6	36.5	36.4	40.3	75.0	76.8
(PFM td_70-30) + (LDC+DB98 tg+cl40-tg)	35.9	34.6	36.5	41.4	**72.5**	76.0

The PFM combined with LDC+DB98 SULMs can give some small gains, but the patterns are not consistent across the dev03f and eval03f test sets.

CUED SMD Plans

Current SMD research plans include the following:

- Try to optimise interpolation weights for LDC and DB98 SULMs.
- Include DB98 data in BN PFMs.
- Explore ensembles of PFMs for BN system.
- Modify the posterior decoding strategy so that SU subtypes are modelled in the decoder.
- Build free-standing IP detection system.
- Build combined SU and IP detection system.
- Explore interactions between SUs and IPs.
- Start to build Edit Disfluency Detection system.

CUED MDE Issues

The following issues need to be considered:

- When will the development data for diarisation be released (for both the eval03 data set and the 'new' dev04 data set)?
- The scoring tool has known problems and has still not been satisfactorily verified (see http://macears.ll.mit.edu/mactech_mail/0293.html).
- When will the final versions of the MDE scoring tools be released?