Metadata at CUED: Progress, Plans, and Issues

Marcus Tomalin, Sue Tranter, and Phil Woodland

4th Feb 2004

Cambridge University Engineering Department

Contents

- **Diarisation**:
  - New splitting algorithm and stopping criterion for BN speaker clustering.
  - New results and cross site experiments for BN speaker clustering.

- **Structural Metadata (SMD)**:
  - CTS SU-Detection system for RT-03F evaluation.
  - Modifying CTS Prosodic Feature Model (PFM) sample space.

- **Issues for RT-04**:
  - Issues concerning the Metadata Extraction (MDE) tasks, data, and tools

Speaker Clustering - Splitting Algorithm

RT-03s Clustering:

- This algorithm was a legacy of minimising computational effort in previous work using MLLR-based clustering.
- Split a parent node into 4 children; if split is not accepted, try recombination to 3 then 2.
- Highly unstable algorithm: the stopping criterion has to cope with different numbers of children nodes.
- Too many tunable parameters: recombination adds more parameters.

2-way Clustering (Dec 2003):

- Split a parent node into 2 children.
- Increased stability (i.e. robustness to changes in input segmentation) and performance (around 5% abs reduction in Diarisation Error Rate (DER)).
- Fewer tunable parameters.

Speaker Clustering - Stopping Criteria

2 BIC-based stopping criteria introduced : BIC $= \mathcal{L} - \alpha P$

$$
\begin{aligned}
P &= \frac{1}{2}\left(d + \frac{d(d+1)}{2}\right)\log(N_z) \\
\text{BIC}_{x+y} &= -\frac{1}{2}[N_x \log(|S_x|) + N_y \log(|S_y|)] - 2\alpha P + N_z C \\
\text{BIC}_{z} &= -\frac{1}{2} N_z \log(|S_z|) - \alpha P + N_z C \\
\Delta\text{BIC}_{split} &= \frac{1}{2}[N_z \log(|S_z|) - N_x \log(|S_x|) - N_y \log(|S_y|)] - \alpha P
\end{aligned}
$$

- Split $z \longrightarrow x + y$ if $\Delta\text{BIC}_{split} > 0$.
- Use N_z (BIC-local) or N_{total} (BIC-global) in the penalty term P.

Speaker Clustering - Diarisation Results

Stopping Criterion	Optimal Param		Diarisation Error Rate	
	bndidev03	bneval03	bndidev03	bneval03
RT-03s 4-way system	-	-	**33.29**	**32.30**
Cost-based (2-way)	0.825	-	28.51	27.24
Cost-based (2-way)	-	0.8	28.66	27.09
BIC-global (2-way)	6.25	6.25	26.13	25.21
BIC-local (2-way)	7.25	-	**25.54**	25.12
BIC-local (2-way)	-	6.75	26.47	**24.27**

- There is a single parameter to tune which generalises well.
- Approx. 8% absolute ($\sim$24% relative) reduction in DER.
- Stability also greatly increased: standard deviation when using many input segmentations reduced from 4.9% to 1.1%.

Diarisation - Cross-Site Experiments

- 'Plug and Play' experiments conducted using MIT-LL/CUED components.
- 3 stages: Advert removal (optional), Segmentation, Clustering.
- Experiments use 2003 BN diarisation dev data (LDC forced alignments).

Advert removal CUED only:

DER when penalising adverts reduced on average by ∼40% relative (compared to no advert removal) whilst only increasing the standard DER (excluding adverts from scoring) by 1.2%.

Segmentation MIT-LL and CUED attain similar standard:

Perfect clustering gives 12.0% (CUED), 12.3% (MIT-LL).

Average over all automatic clusterings 25.8% (CUED) and 26.3% (MIT-LL).

Clustering MIT-LL slightly better than CUED:

Average across segmentations is 26.6% (CUED) and 25.5% (MIT-LL).

Both systems are robust to changes in segmentation:

standard deviation = 1.1% (CUED) and 1.0% (MIT-LL).

SU-Detection System for RT-03F Evaluation

CUED CTS SU-Detection System for the RT-03F Evaluation (RT-03F-SYS):

- **Training Data** = LDC Data (30hrs) + Meteer-mapped Data (10hrs).
- 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).
  - PFM = CART decision tree (183 terminal nodes).
- **Slash Unit Language Model (SULM)**:
  - N-gram and Class-based SULMs built.
  - Interpolation Weights and Perplexities calculated using stream info for SU tokens only (mod-streams).
  - SULM = Interpolated trigram, 40-class trigram and 40-class fourgram.
- **Lattice-based Decoder**:
  - Decoder = 1-Best Posterior Decoding, rather than 1-Best Viterbi Decoding.

SU-Detection System for RT-03F Evaluation

Results for the dev03 (3hrs) and eval03 (3hrs) test sets:

SYSTEM	dev03 %Err	eval03 %Err
SULM	51.63	53.03
PFM+SULM (viterbi-decoding)	47.55	49.29
+ mod-streams	47.28	48.68
+ posterior-decoding	**45.88**	**46.04**

The # Ins errors is $\sim \frac{1}{2}$ the # Del errors:

System	%Del	%Ins	%Err
RT-03F-SYS (dev03)	31.75	14.12	**45.88**
RT-03F-SYS (eval03)	30.60	15.44	**46.04**

All scores obtained using su-eval-v15.pl with the '-w -W -t 1.00' settings.

PFM Research

Del-Ins error ratio due in part to the distribution of SU tokens in the sample space.

Given the sample space for LDC training data (all-ldc), create a modified sample space (samp-ldc).

SU tokens Key:
N = # Non-SU tokens
S = # Statements, Q = # Questions, I = # Incompletes, B = # Backchannels

The samp-ldc space is created by random sampling from all-ldc:

- Eliminate some Non-SU tokens in all-ldc sample space in order to reduce N.
- Stop when the arbitrary threshold N = (S + Q + I + B) is reached.

Sample Space	Total # Toks	% Non-SU Toks
all-ldc	465,000	86%
samp-ldc	126,000	50%

The random sampling that creates samp-ldc reduces the all-ldc space by c.73%.

PFM Research

- **CUED PFM**: PFM decision tree created using CUED software.
- **R PFM**: PFM decision tree created using R.
- **SULM**: same SULM as for RT-03F, but constructed using only the LDC training data.

System	# Terminal Nodes	Del	Ins	%Err
RT-03F-SYS†	183	31.75	14.12	45.88
all-ldc CUED PFM + SULM	167	33.02	14.45	**47.47**
all-ldc R PFM + SULM	100	46.24	5.46	51.70
samp-ldc CUED PFM + SULM	123	32.62	14.27	46.89
samp-ldc R PFM + SULM	100	32.74	11.99	**44.73**

† This system was built using the LDC training data + 10hrs Meteer-mapped training data.

- The 'samp-ldc R PFM + SULM' system reduces the SU Err by 2.5% abs compared to 'all-ldc CUED PFM + SULM' system.
- The 'samp-ldc R PFM + SULM' system reduces the SU Err by 1% abs compared to RT-03F-SYS.

MDE at CUED: Progress Summary

- **Diarisation**:
  - The new BN 2-way clustering algorithm gives a ∼5% abs reduction in DER compared to the RT-03s system.
  - The new BIC-based stopping criterion gives a further ∼3% abs reduction in DER.
  - Ongoing cross-site experiments involving CUED and MIT-LL explore the performance and robustness of the diarisation system components.

- **Structural Metadata**:
  - A CTS SU-Detection system was developed for RT-03F.
  - The 'samp-ldc PFM + SULM' system can reduce SU Err by ∼2.5% abs compared to the 'all-ldc PFM + SULM' system.
  - The 'samp-ldc R PFM + SULM' system reduces the SU Err by ∼1% abs compared to RT-03F-SYS.

MDE Plans at CUED

- **Diarisation**:
  - Ongoing research into a "cluster voting" scheme to combine information from different clusterer outputs.
  - The diarisation system will be adapted for the SA-STT RT-04 task.
  - Different features, parameters, algorithms etc. will be evaluated in more detail when new RT-04 diarisation dev data becomes available.

- **Structural Metadata**:
  - Ongoing PFM-related SU research (sample spaces, neural nets vs. decision trees).
  - Ongoing SULM-related SU research (e.g. SULM interpolation strategies).
  - Ongoing Decoder-related SU research (e.g. posterior decoding strategies).
  - A BN SU-Detection system will be constructed.
  - A CTS Disfluency Detection system will be constructed.

Diarisation Issues for RT-04

- **Scoring Reference**:
  - Are there advantages to using manually marked boundaries for diarisation? An experiment to judge the size of the effect to be carried out by CUED/MIT-LL/?ICSI.
- **Tools**:
  - What are the schedules for providing/debugging the new scoring tools for diarisation/SA-STT? Are these schedules realistic for everyone?
- **New Data**:
  - Which new episodes will be in the new RT-04 diarisation dev data?
    (same as STT dev please!).
  - How will the speaker times be marked?
    (e.g. using NIST's 'cookbook'? - if so - who will provide the forced alignment?).
  - When will the diarisation dev data be released? Can this be released independently of (and before) the SMD dev data release?

SMD Issues for RT-04

- **Tools**:
  - Which SMD scoring tool will be used during the RT-04 eval and when will it be released?
- **Data**:
  - Is the current V6 dev/training data annotation schedule realistic?
  - When will the independent annotator QC information become available?
- **Scoring Reference**:
  - Who will provide the forced alignments for SMD dev and eval reference data, and what is the schedule for creating these?