

The ETAPE evaluation

Olivier Galibert

ETAPE final workshop

The Ester tasks

- Speech event detection
- Speaker diarization
- Speech transcription
- Named entities extraction

Overlapping voices detection

The task

- Detect the segments where voices overlap.

The metric

- Time precision and recall.
- Segment presence detection quality

Speaker diarization

The task

- Detecting speech segments and assigning them to (unnamed) speakers.

The metric

- Standard diarization error rate.
 - Map hypothesis and reference speakers.
 - Divide the time in error by the reference time.
- Two evaluation setups:
 - One mapping per show.
 - One mapping for all the shows (cross-show diarization).

Speech transcription

The task

- Transcribe all that is said, *including in overlapping parts*.
- Attribute every word to its speaker (as per the diarization).

The metrics

- Standard-ish word error rate, with words optimally distributed in multi-speaker zones.
- Speaker-attributed word error rate, where the mapping directs the word comparisons.

Named entities

The task

- Detect named entities per the Quaero guide.
- Done in manual transcription and automatic transcriptions.
- Rover of all submissions plus individual submissions.

The metric

- Slot error rate, adapted and extended. See [IJCNLP 2011] for details.

And then the problems start

- Human-made transcriptions are nowhere near precise enough to know where overlapping speech happens
- ⇒ Need to refine the boundaries of everything
- ⇒ Forced alignment of text on speech
- The complexity of the NE annotation guide is high
- ⇒ Some of the ELDA annotators had previous experience but the DGA ones didn't
- ⇒ The guide creators did their best to answer questions, but they're not officially part of the project

The data got harder

- Need to associate every word with a speaker
- Need to identify recurring speakers
- Need to correct transcriptions and named entities in parallel
- Two data producers make uniformity difficult

The metrics got harder

- Etape transcription guidelines end up with a more complex result than the Quaero ones
- What's the reference for event detection?
- NIST diarization heuristics blow up. Thankfully, we had already worked that out in Quaero
- The ASR metrics aren't really described by our friends at NIST either
- Manual EN references should be done on what, exactly?
- Aligning the EN references with the forced alignment results becomes problematic
- Projecting on the ASR outputs starts losing meaning

And the systems had a hard time too

- Incorrect speaker types in diarization
- Variable genders in diarization
- Multiple words in a line in transcription
- Meaningless confidence values
- Syntactically incorrect EN outputs
 - Changed text
 - Unbalanced tags
 - Empty entities
 - Incorrect entity types

Some lessons

- Don't do too many new things at the same time
- Really pay for the data
- Decide early who is responsible for what
- Leave a lot of time to shake the problems with the development data and the evaluation tools
- Leave a lot of time between an evaluation you know is going to be problematic and the final workshop
- Try not to lose people in the middle

The evaluation data

- Around 7 hours of audio
- Around 6:40 of actual speech
- Almost 100K words
- 7 different sources
 - 1 radio channel, France Inter
 - 2 national TV channels, with 5 different shows
 - 1 local TV channel
- 83 identified speakers
- 13K entities/components

Diarization

Laboratory	Run	Ind.	Cross.
CRIM	2	22.73	-
CRIM	X primary	24.17	28.63
Eurecom	purif_mapetape_cms	29.32	-
LIA	primary	27.27	-
LIA	X primary	27.27	37.54
LIMSI	primary	23.48	-
LIMSI	X primary	21.18	22.59
LIUM	primary bic_ilpflt2_jfa	19.01	-
LIUM	primary bic_ilpflt2_jfa_clr	19.51	20.26
Orange	4	22.45	-
ParisTech	primary	16.23	-

Transcription

		No overlap	Optimal	Speaker att.
		WER	WER	WER
LIUM+LIA	rov_bong	30.63	37.07	-
CRIM	4	24.93	31.51	54.72
LIA	2	35.63	41.43	91.91
LIUM	primary	23.60	30.16	51.31
LORIA	primary	25.87	32.18	72.57
All	rover	28.28	35.04	-

Named Entities

	Man.	Rov.	s23	s24	s25	s30
eurecom	84.78	98.82	101.45	95.03	100.72	97.28
irisa	33.81	55.51	58.35	63.40	62.53	52.71
jouve	55.63	94.24	107.71	82.67	142.96	97.19
lif	43.58	69.54	74.55	71.93	85.60	69.24
limsi	36.44	67.16	68.57	67.73	75.02	60.44
lina-lium	62.76	76.45	80.84	77.97	82.71	76.63
synapse	42.89	68.65	74.93	70.77	86.10	66.23
tours	41.01	65.97	71.01	66.89	90.32	65.37

Open questions

- What should the reference be for SES?
- Do someone want to run an alternative alignment for SRL to see whether results change?
- Speaker-attributed is too harsh. Speaker confusion error? Error weighting?
- SER for EN looks problematic. The mapping aspects are probably good, but the score may not be. Alternative scoring methods?