UTLN / LIS lab submission

Ricard Marxer - ricard.marxer@lis-lab.fr
Assistant Professor (Maître de conférences)
Université de Toulon, Aix Marseille Univ, CNRS, LIS
Marseille, France
Overview

- Introduction
- Data
- First (failed) attempt
- First (and final) submission
- Results
Introduction

Context

- Limited resources (human and material)
  - Recently appointed professorship
  - Paternity leave
- Focus on the low-resource-friendly task
  - Standard audio transcription
  - Callsign extraction

Disclaimer: only descriptive presentation of the submission
Language data (utterances)

- Utterance length
  - Bimodal: short ~ 12 / long ~ 80
  - Distribution dependant on callsign presence
- Words distributions
  - Few words repeated a lot (codes, numerals, etc.)
- Out of vocabulary (CMU dict)
  - Aeronames
  - Incomplete words
Language data (callsigns)

- Exchangeable words
  - aeroname, number...
- Highly structured
  - 18 patterns for 95% (w/o disfluencies)
- Clustered in time

![top-K callsign distribution](image1)
![callsign occurrences in neighbouring utterances for different window sizes](image2)
Joint ASR callsign extraction

- Proposed solution (similar to [1])
  - n-gram FST of utterances with callsign replaced with special symbol `<callsign>`
  - n-gram FST for callsign
  - combine both FST (with `fstreplace`)

- Problems
  - Disambiguation words (tags)
  - Combinatorial explosion (~230,000 `<callsign>` arcs)

---

[1] Horndasch, Axel, Caroline Kaufhold, and Elmar Nöth
"How to Add Word Classes to the Kaldi Speech Recognition Toolkit."
Standard audio transcription (ASR)

- Data characteristics
  - Noise-induced energetic masking
  - Fast speech
  - High pronunciation variability
- Plan
  - Adapt existing recent CHiME4 recipe for Kaldi
  - No speech enhancement
Standard audio transcription (ASR)

- G2P with phonetisaurus (+ some manual pronunciations)
- GMM with fMLLR
- Select clean data
- TDNN w/ i-vectors
- RNN LM rescoring
Callsign extraction

- Exploit
  - language regularities
  - word classes (NATO alphabet, numerals,...)
  - callsign sequenciality priors
Callsign extraction

- Transformer model
  - Multi-head self and cross attention
  - Positional signals

- Inputs
  - Transcription
  - Word classes
  - Context (past and future)

- Output
  - Linearised parse tree
Leaderboard and evaluation result

Development result
- WER: 10.52 %
- F1: 0.8042

Leaderboard result
- WER: 10.30 %
- F1: 0.8207

Evaluation results
- WER: 9.88 % (8\textsuperscript{th})
- F1: 0.7704 (5\textsuperscript{th})
- Score: 0.93 (6\textsuperscript{th})
Analysis of errors
Callsign errors

- Mismatches > false negatives >> false positives

Errors due to the ASR
Analysis of errors
WER vs callsign error

- Callsign error correlation
  - Callsign errors are WER dependent
  - Wrong callsigns still with small WER
- WER of individual utterances not a stable (utterances are short)
Conclusions

- Much work to be done on the ASR
  - Noise reduction
  - Cluster Adaptive Training
  - Pronunciation models
  - OOV LM
  - Joint ASR-extraction

- Much further analysis on the callsign
  - Effect of each input (transcription, context, etc.)
  - Try different input encodings
  - Data augmentation accounting for ASR errors
Questions and proposals ;-)