Universität Stuttgart

Context-aware Neural-based Dialog Act Classification On Automatically Generated Transcriptions

IMS
D. Ortega
Chia-Yu Li
G. Vallejo
P. Desinov
Thang Vu

Motivation

Explore the effect of training and testing a context-aware neural-based dialog act (DA) classifier on transcriptions generated from two different automatic speech recognition (ASR) systems, so that the DA classification is taken into a more realistic scenario.

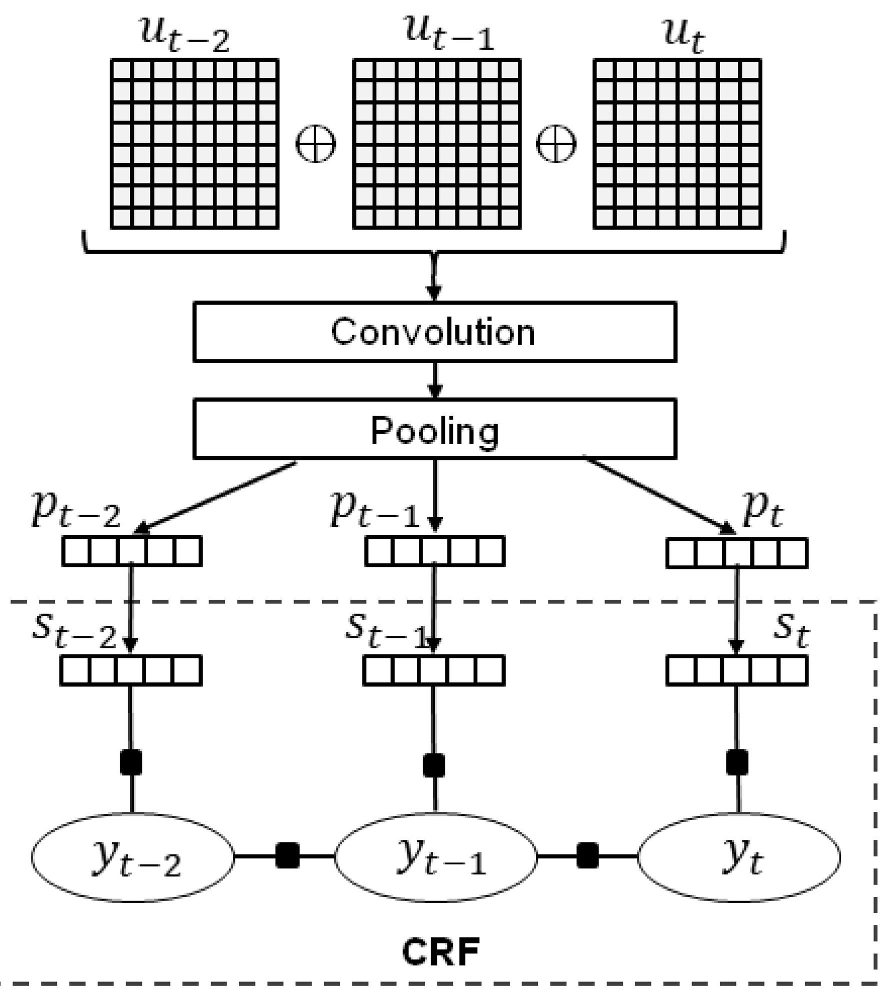
Utterance	Dialog Act
A: *Are you a musician yourself?*	Yes-no-question
B: *Uh, well, I sing.*	Affirmative non-yes answer
A: *Uh-huh.*	Acknowledge (Backchannel)
B: *I don't play an instrument.*	Statement-non-opinion

Manual transcription (MT) extract from Switchboard [1]

Dialog Act Classification Model

Our two-fold model consist of:

- Convolutional neural networks (CNNs) for utterance representation.
- Conditional random fields (CRFs) for sequence labeling.



Model architecture. ⊕ stands for concatenation.

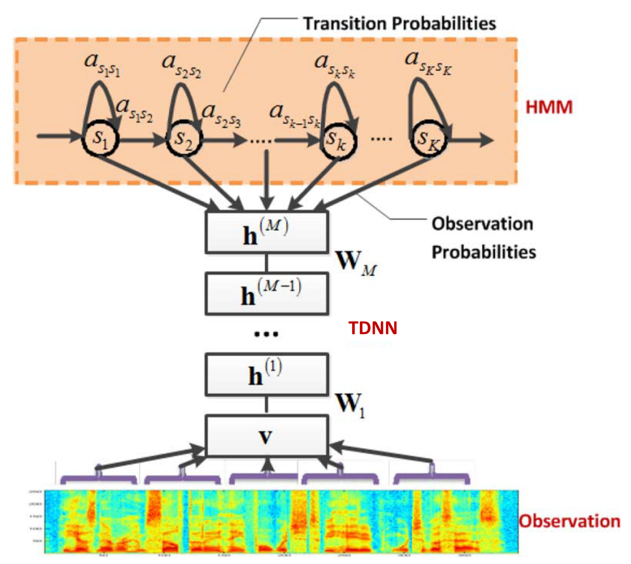
- The model takes the current and n previous utterances (context) as input in a grid-like representation [2].
- For evaluation, only the DA predicted for the current utterance is taken into account.

Automatic Speech Recognition

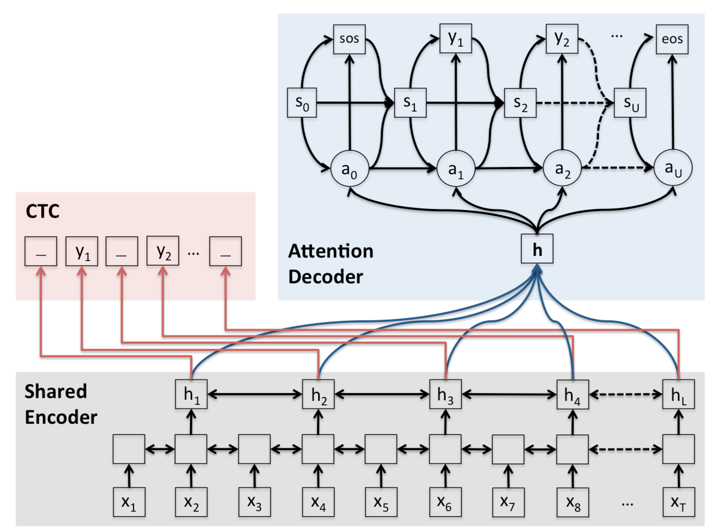
Two types of ASR architectures:

- Hybrid Time Delay Neural Network and Hidden Markov Model (TDNN/HMM) trained with lattice-free maximum mutual information.
- Joint CTC-Attention End-to-End (E2E): shared-encoder representation trained by both Connectionist Temporal Classification (CTC) and attention model using the following combined trainig loss:

$$\mathcal{L} = \alpha\mathcal{L}^{ctc} + (1-\alpha)\mathcal{L}^{att}$$



TDNN/HMM from [3, 4]



CTC-Attention E2E from [5]

Hyperparameters:

- TDNN: 6 layers with default settings for spliced indices in Kaldi recipe; using MFCC and iVector features with LDA.
- CTC-Attention E2E: five layers of 1024 BLSTM units for Encoder and a layer of 1024 LSTM units for Decoder; using 80-bin logMel filter banks and pitch as suggested in Espnet recipe [6].

Experimental Setup

Datasets:
MRDA: ICSI Meeting Recorder Corpus [7]
SwDA: Switchboard DA Corpus

Dataset	C	|V|	Train	Val	Test
MRDA	5	12k	78k	16k	15k
SwDA	42	20k	193k	23k	5k

C: # of classes, |V|: Vocabulary size, Train/Val/Test: # of utts.

Hyperparameter	Value
Activation function	ReLU
Filter width	3, 4, 5
Filters per width	100
Pooling size	Utterance-wise
Embeddings	Word2vec [8]

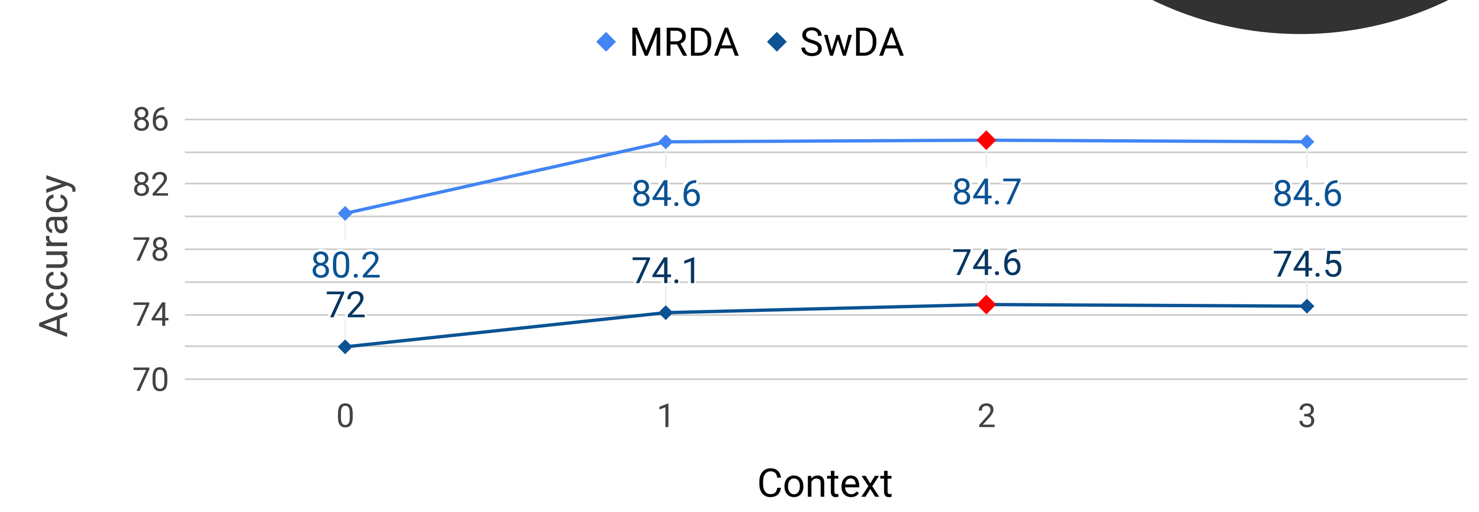
CNN hyperparameters

Dataset	ASR System	Train (WER)	Val (WER)	Test (WER)
MRDA	TDNN/HMM	9.89	19.28	21.48
	CTC-Attention E2E	2.30	16.80	18.80
SwDA	TDNN/HMM	13.8	14.28	18.02
	CTC-Attention E2E	29.0	8.90	18.80

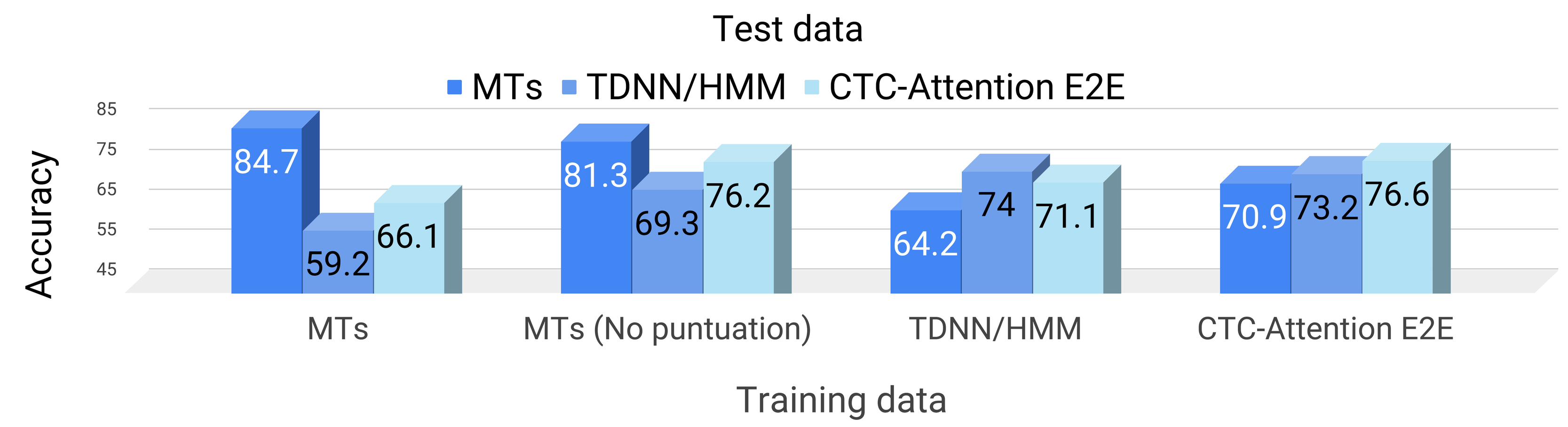
Best ASR performance in terms of WER (%)

Experimental Results

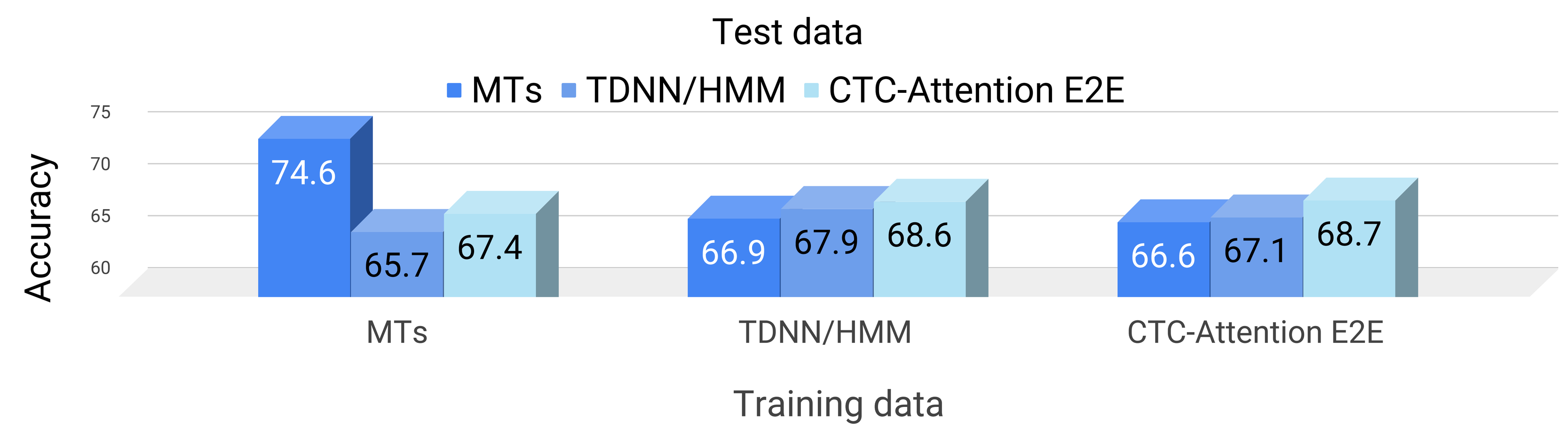
- Experiments varing the context



- Experiments on MRDA



- Experiments on SwDA



Conclusion

- We explored dialog act classification on automatic transcriptions by means of CNNs and CRFs.
- Although the WERs from both ASR syetems are comparable, the End-to-End ASR system might be more suitable for dialog act classification.
- Punctuation yields central cues for the task. Therefore, it should be integrated into the ASR output in future works.

[1] S. Calhoun et al. The NXT-format Switchboard Corpus: A Rich Resource for Investigating the Syntax, Semantics, Pragmatics and Prosody of Dialogue. In *LREC*, 2010.

[2] D. Ortega and N. T. Vu. Neural-based Context Representation Learning for Dialog Act Classification. In *SIGDIAL*, 2017.

[3] G. E. Dahl et al. Context-Dependent Pre-trained Deep Neural Networks for Large-Vocabulary Speech Recognition. In *IEEE Trans. Audio, Speech, Language Process*, 2012.

[4] V. Peddinti et al. A time delay neural network architecture for efficient modeling of long temporal contexts In *INTERSPEECH*, 2015.

[5] S. Kim et al. Joint CTC-Attention based End-to-End Speech Recognition using Multi-task Learning. In *ICASSP*, 2016.

[6] S. Watanabe et al. ESPnet: End-to-End Speech Processing Toolkit In *INTERSPEECH*, 2018.

[7] E. Shriberg et al. The ICSI meeting recorder dialog act (MRDA) corpus In *SIGdial Workshop on Discourse and Dialogue at HLT-NAACL*, 2004.

[8] T. Mikolov et al. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

www.ims.uni-stuttgart.de

{daniel.ortega, chiayu.li, thang.vu}@ims.uni-stuttgart.de