LISTEN Workshop / Summer School

Bonn, Germany, July 17-19, 2018

Speaker-Adapted Confidence Measures for ASR using Deep Bidirectional Recurrent Neural Networks

Contents

1 Introduction

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

- Confidence Estimation (CE) aims at providing Confidence Measures (CM) for an Automatic Speech Recognition output (ASR)

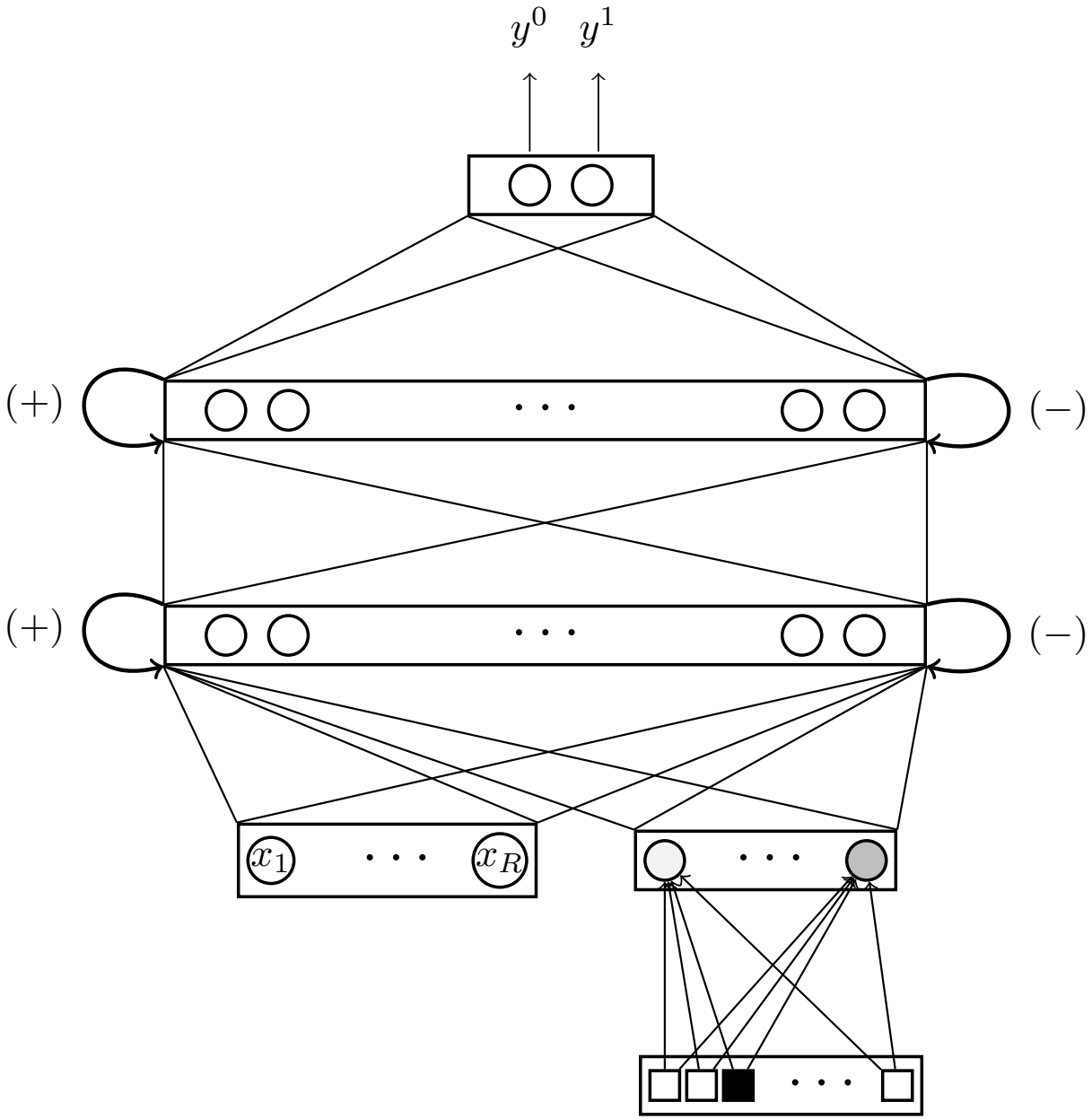
- CM reflect the reliability of any recognition output
  - Usually a score between 0 and 1
  - Can be applied at different levels of granularity: **word**, sentence, etc.

- CE has been largely addressed following three main approaches:
  1. Utterance Verification
  2. Word posterior probabilities (lattices, n-best,etc.)
  3. As a two-class classification problem

Introduction

- CE classifiers has benefit from advances in deep learning
  - Classifiers based on: DNNs, CRFs, etc.
- More recent improvements to CE include RNNs
- CE models can also benefit from speaker adaptation
  - For both CRF and RNN models
- In this work:
  - We extend our previous results to a multi-task empirical evaluation
  - Introduced word-embeddings into the CE RNN architecture
  - Proposed a new a novel CE-based unsupervised adaptation method for acoustic BLSTM

2 DBRNN architecture for CE



3 Speaker-Adapted DBRNN-CMs

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Input:

- Θ: speaker independent DBLSTM/DBRNN
- $\{\mathcal{X}, \mathcal{Z}\}_1^M$: speaker supervised data

Output:

- Θ': speaker-adapted DBLSTM/DBRNN
- τ^*: confidence measure decision threshold

Procedure:

1. Split $\{\mathcal{X}, \mathcal{Z}\}_1^M$ into training ($\mathcal{T}$) and validation ($\mathcal{V}$)
2. Use Θ, $\mathcal{T}$ and $\mathcal{V}$ for metaparameters tuning:
   - Learning rate, number of epochs and τ^*
3. Estimate Θ' from Θ using $\{\mathcal{X}, \mathcal{Z}\}_1^M$

4 Tasks and ASR Systems

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

LibriSpeech: 2-pass (fMLLR) BLSTM-HMM ASR-**En** system

Set	Duration (h)	Speakers	Words	Vocab	WER
Train	961	1210	9.4M	89K	4.7
Dev-other	5.3	33	51K	7.4K	12.5
Test-other	5.1	33	52K	7.6K	13.5

poliMedia: 2-pass (fMLLR) BLSTM-HMM ASR-**Es** system

Set	Duration (h)	Videos	Speakers	Words	Vocab	WER
Train	813	9.5K	>205	8.3M	36.6K	14.5
Dev	3.4	26	5	35K	2.6K	11.3
Test	3.2	23	5	30K	2.4K	12.5

5 Experiments

- 20 word-level predictor features:
  - 8 based on ASR models: decoding score, acoustic log-score, etc.
  - 12 based on lattices: forward, backward and posterior probabilities, etc.
- Same training data was used for ASR and CE models.
- Three sets of experiments were carried out:
  1. Experiments on CE
  2. Experiments on speaker-adapted CM
     - Experiments carried out on the LibriSpeech system (ASR-En)
     - 8 speakers extracted from TED-LIUM corpus for evaluation
     - 4 videos per speaker with WER between 10% and 30%
  3. Experiments on improving ASR performance

Experiments on CE

Task	CM	AUC	CER(τ^*)	95%-CI CER
LibriSpeech	WP	85.3	10.71	[10.44, 10.97]
	CRF	89.6	9.29	[9.04, 9.54]
	BRNN	91.1	8.82	[8.58, 9.07]
	BLSTM	91.0	8.85	[8.60, 9.09]
	BRNN+BLSTM	91.5	8.65	[8.41, 8.89]
poliMedia	WP	83.6	9.67	[9.33, 10.00]
	CRF	90.0	7.69	[7.39, 7.99]
	BRNN	91.6	7.00	[6.71, 7.29]
	BLSTM	92.0	6.77	[6.48, 7.05]
	BRNN+BLSTM	92.1	6.75	[6.47, 7.04]

Experiments on Speaker-Adapted CM

	CER(τ^*)			
Speaker	CER(0)	¬Adapt	Adapt	R.I. [%]
1	18.61	13.86	13.21	4.7
2	16.00	12.23	11.53	5.7
3	19.74	14.42	14.16	1.8
4	19.03	13.21	12.81	3.0
5	12.07	9.03	8.29	8.2
6	12.06	9.04	8.79	2.8
7	20.19	14.09	14.06	0.2
8	22.03	15.87	15.48	2.5
All	17.35	12.66	12.21	3.6

Experiments on Improving ASR

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

- Confidence measures are introduced in CE criterion as:

$$\mathcal{C}(\mathcal{X}, \mathcal{S}) = -\frac{1}{T}\sum_{t=1}^{T} \log p(s_t \mid \mathbf{x}_t) \cdot cm(s_t) \quad (1)$$

- Results:

	WER%		
Recognition setting	LibriSpeech	poliMedia	TED-LIUM
2-pass	13.50	12.53	20.78
3-pass	13.06	12.37	20.02
3-pass+CM	13.05	12.06	19.63

6 Conclusions

- New study has confirmed our previous results:
  - CE BRNNs outperform CRF and WP (Best result BRNN+BLSTM)
  - Speaker adaptation of CMs improves CE

- A novel unsupervised speaker-adaptation technique for DBLSTM has been proposed
  - Relative reductions in WER in the range of $3\% - 5.5\%$