CROSS LINGUAL TRANSFER LEARNING FOR ZERO-RESOURCE DOMAIN ADAPTATION

Alberto Abad[1,2], Peter Bell[2], Andrea Carmantini[2] and Steve Renals[2]

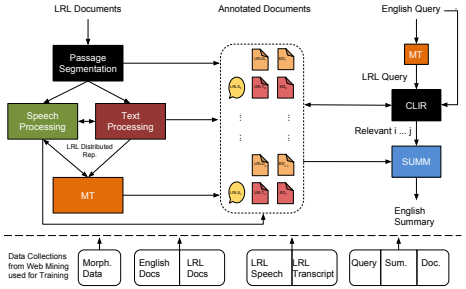
[1]INESC-ID / Instituto Superior Técnico, University of Lisbon, Portugal

[2]CSTR - Centre for Speech Technology Research, University of Edinburgh, UK

Introduction

- Building DNN acoustic models that behave robustly in different usage domains is an **open research** challenge
- Most approaches for domain adaptation relay on some (usually limited) **transcribed data** from the target domain:
  - data augmentation; auxiliary features; adaptation of selected parameters; adversarial methods; simple re-training source model; etc.
- In **zero-resource** scenarios, target domain data is not available
- This is the case of the MATERIAL program, focused on searching speech and text in **low-resource** languages using English queries:
  - ASR systems must operate on diverse multi-genre data, including **telephone conversations**, **news** and **topical broadcasts**
  - Manually annotated training data is from the telephone speech domain
  - Semi-supervised approaches are quite successful in this scenario

MATERIAL



Introduction

In this work:

- **Goal:** To improve low-resource (LR) ASR in a new target domain by using only data of a well-resource (WR) language.
- **Hypothesis:** Initial layers of a DNN encode language-independent acoustic characteristics.
- **Proposal:** An adaptation based on multi-lingual AM training to enable cross-lingual sharing of domain adaptation techniques.
- **Experiments:** Different pairs of languages; source domain is conversational telephone speech (CTS) and target domain is broadcast news (BN)

Can domain adaptation be portable across languages!?!?

Outline

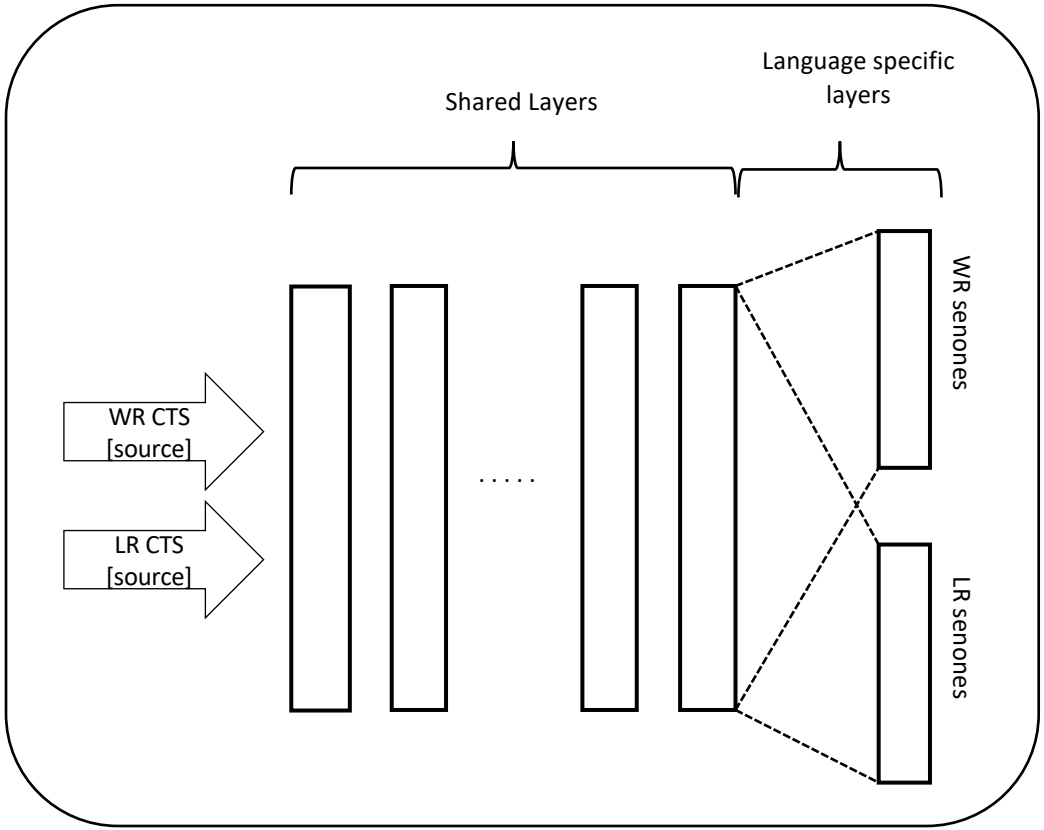
- Introduction
- Cross-lingual domain adaptation
  - Multi-lingual training
  - Domain adaptation
- Experimental set-up
- Results
- Conclusions

Cross-lingual domain adaptation

- DNN acoustic models (non-linearly) map acoustics to phonetics
  - Interpretation: Initial layers encode lower-level acoustic information (language-independent?); deeper layers codify cues closer to phonetic classes
  - Hypothesis: modifications to the initial layers to adapt to a new domain should be similar (and transferable) among different languages.
  - Proposed Solution: Use a network architecture in which parameter transforms are shared among the LR and WR languages + set of final language specific layers. This can be attained based on:
    1. Multi-lingual training
    2. Domain adaptation

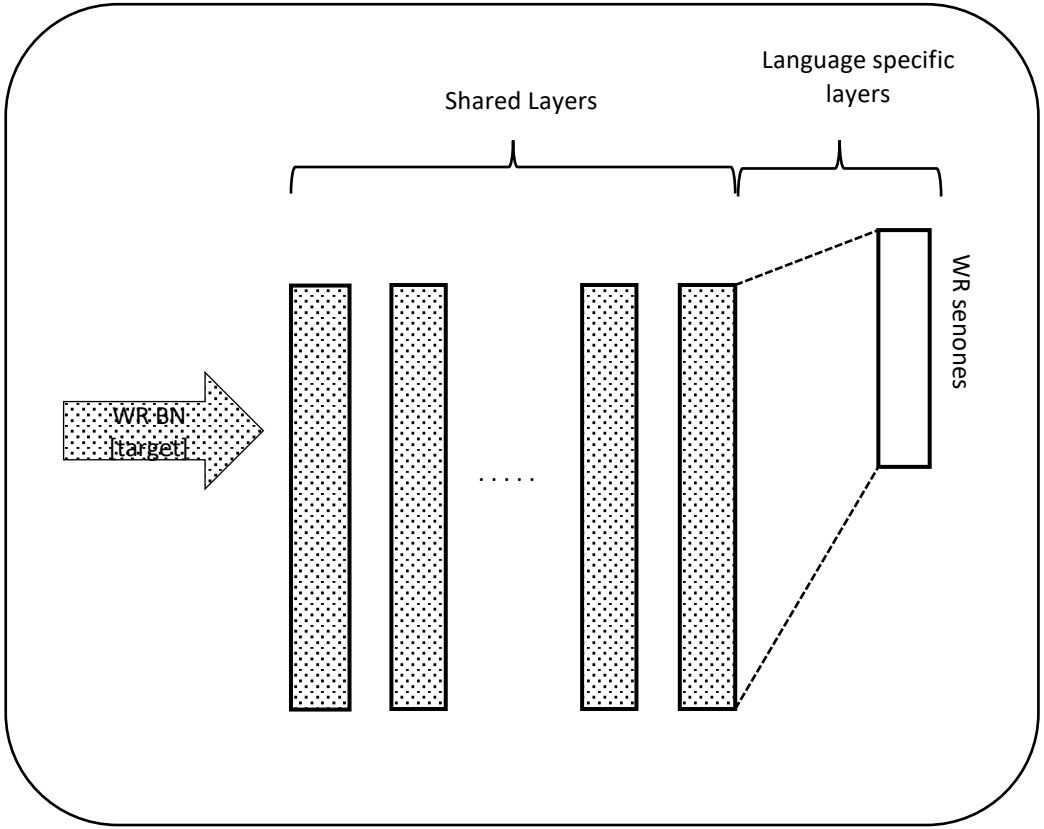
Cross-lingual domain adaptation

Multi-lingual training



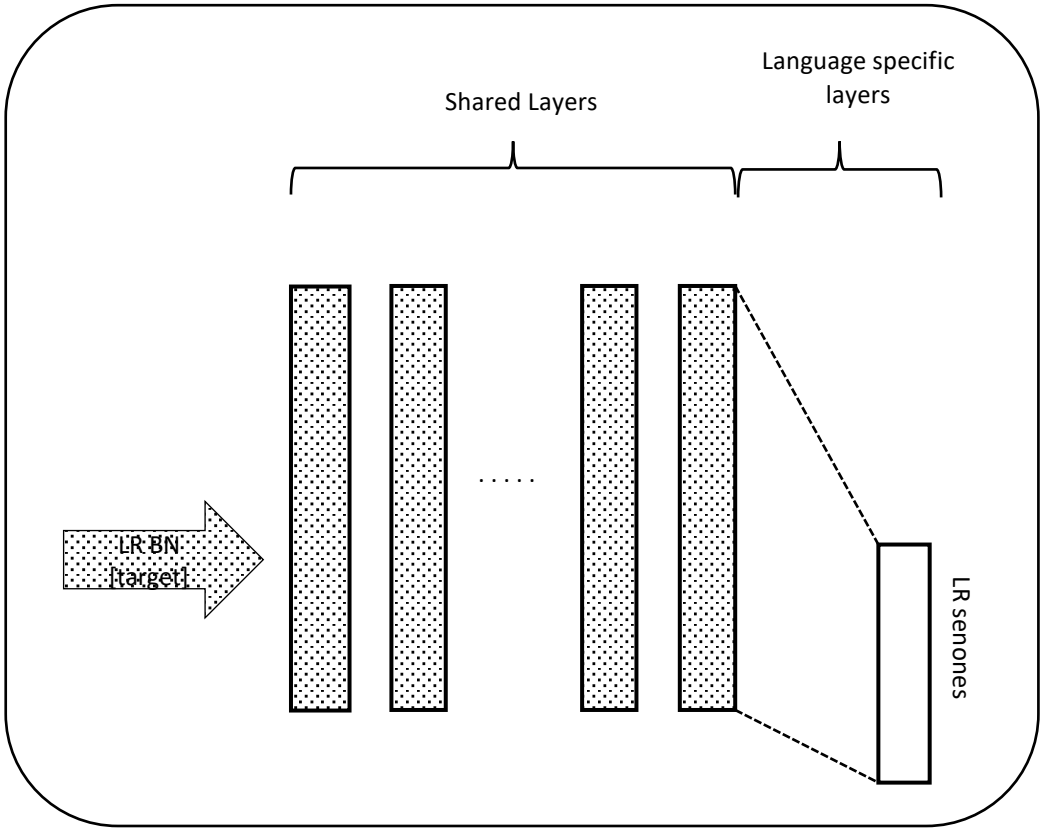
Cross-lingual domain adaptation

Domain adaptation



Cross-lingual domain adaptation

Weight transfer



Experimental set-up
Corpora

- In all experiments:
  - CTS is the source domain – transcribed data is available for all languages
  - BN is the target domain – transcribed data is only available for the WR language
- Two sets of language pairs/experiments:
  1. English as WR and Spanish as LR:
     - Source domain training data: ~200 hours from Fisher EN and ~163 hours from Fisher SP
     - Target domain BN adaptation data : ~150 hours from HUB4; (~30 hours of HUB4 Spanish, used for oracle experiments only);
     - Test sets: 1997 HUB4 English Evaluation set; Spanish HUB4 evaluation set;
  2. English as WR and MATERIAL languages (Tagalog an Lithuanian) as LR:
     - Source domain training data: ~200 hours from Fisher EN; IARPA Babel full language packs: ~80 hours for Tagalog and ~40 hours for Lithuanian
     - Target domain adaptation data : ~150 hours from HUB4;
     - Test sets: The wide-band “Analysis” test sets provided by the MATERIAL programme.

Experimental set-up
Systems description

- Kaldi used for the development of all the ASR systems:
  - Conventional recipes to obtain frame-level phonetic alignments for each language and domain.
  - DNN input features are 40 hires MFCCs + 3 pitch features; no i-vector, no speed perturbation data
  - All data downsampled to 8 kHz to match CTS source domain.
- Acoustic models are TDNN networks trained with CE loss criterion:
  - 7 TDNN hidden layers of 650 units with RELU (shared language-independent layers) + pre-final 650 units fully connected RELU and final softmax layer (language-dependent layers).
  - For training the baseline and multilingual: 3 epochs with a minibatch size of 256
  - For domain adaptation: identical configurations with learning rate of the frozen layers set to 0 and varying number of adaptation epochs.
- Use of domain matched language models in decoding:
  - CTS LMs trained on the training transcriptions
  - BN LM for Spanish trained only on the training transcriptions
  - BN LM for English trained on the transcriptions + additional 1996 CSR HUB-4 Language Model text and North American News Text Corpus
  - BN LMs for Lithuanian and Tagalog trained on around 30M words of web-crawled

Experiments
EN and SP baseline

Experiments with EN as the WR and SP mimicking LR

	Test condition			
	WR language		LR language	
	CTS source	BN target	CTS source	BN target
mono-ling BN AM	---	11.8	---	19.2*
mono-ling CTS AM	22.6	19.6	32.3	40.0▼▼
multi-ling CTS AM	23.6	19.2	32.6	**32.9**▲

- Oracle* experiment represents target upper bound performance
- Key observations:
  - Large degradation due to domain mismatch
  - Multi-ling training helps specially in mismatch LR; do not help in matched

Experiments
Cross-lingual network adaptation results

Varying number of adaptation epochs and adapted layers

		#adaptation EPOCHS					
		0.5		1		2	
		WR BN target	LR BN target	WR BN target	LR BN target	WR BN target	LR BN target
#adapted layers	1	15.7	29.1	15.6	29.0	15.5	29.1
	1-2	15.0	28.7	14.9	28.9	14.8	28.9
	1-3	14.5	28.6	14.5	**28.4**	14.4	28.4
	1-4	14.3	28.7	14.2	28.8	14.1	28.8

- Best adaptation configuration with 3 first hidden shared layers adapted for 1 epoch
- Not very sensitive for LR language → 28.4-29.1
- There seems to be a limit on the amount of transferable information

Experiments
Cross-lingual network adaptation results

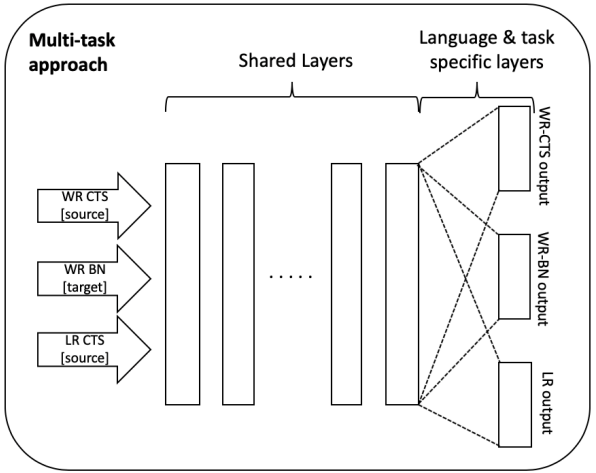
Summary of results

	WR language	LR language
	BN target	BN target
Upper bound	11.8	19.2
mono-ling CTS AM	19.6	40.0
multi-ling CTS AM	19.2	32.9
proposed CL adapt AM	14.5	**28.4**

- Using WR CTS source domain data: 40.0% → 32.9%
- Using WR BN target domain data: 32.9% → 28.4% WER.
- Overall, absolute 11.6% WER decrease and recovery of ~50% of the performance loss due to the lack of LR training data
- This is attained by using **only** WR data.

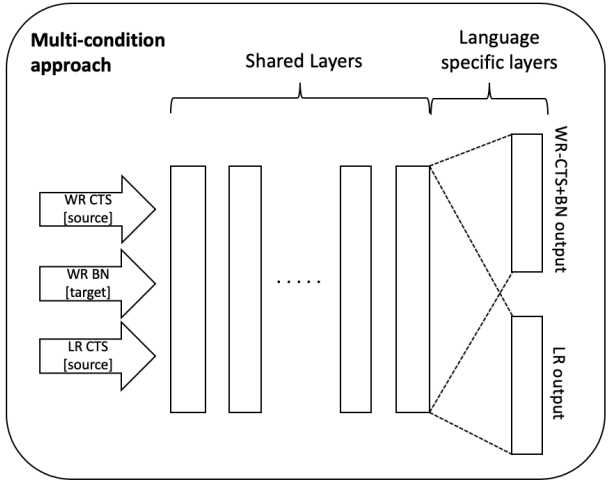
Experiments
Comparison with similar cross-lingual approaches

Multi-task approach



- Train in a multi-task way a nnet with 3 language-task pairs
- Use the LR output for decoding target BN data

Multi-condition approach



- Train in a multi-condition way a nnet with 2 lang outputs (WR data is mixed)
- Use the LR output for decoding target BN data

Experiments
Comparison with similar cross-lingual approaches

	WR language	LR language
	BN target	BN target
Upper-bound	11.8	19.2
mono-ling CTS AM	19.6	40.0
multi-ling CTS AM	19.2	32.9
proposed CL adapt AM	14.5	**28.4**
multi-task CL AM	12.4	29.1
multi-cond CL AM	12.5	29.2
multi-task CL + adapt AM	12.3	29.1
multi-cond CL + adapt AM	12.2	29.1

- Same network architecture, training and decoding recipes
- The cross-lingual proposed scheme outperforms any of the other methods
- Additional fine-tuning does not help the alternative methods:
  - Performance converges already after the initial training
  - Best adaptation configuration is attained with the minimal number of epochs and adaptation layers
- The proposed method better leverages information from the WR data for improved LR ASR

Experiments
MATERIAL languages

	Tagalog			Lithuanian		
	BN	TB	avg	BN	TB	avg
mono-ling CTS AM	53.2	58.7	57.3	45.6	43.0	44.0
multi-ling CTS AM	46.5	52.2	50.7	38.2	36.5	37.1
proposed CL adapt AM	41.9	48.5	**46.8**	31.6	32.1	**31.9**

- Same network architecture, training, decoding recipes and adaptation configuration (3 first hidden shared layers adapted for 1 epoch)
- Remarkable improvements in any of the two wide-band sub-domains:
  - BN: relative WER improvements of 21.2% for Tagalog and 30.7% for Lithuanian;
  - TB: 17.4% for Tagalog and 25.3% for Lithuanian.
- Overall, average relative WER improvement of 18.3% and 27.5% for the Tagalog and Lithuanian.

Conclusions

- We have introduced a simple, yet effective, method to transfer domain adaptation of DNNs from one language to another
- Based on a multi-lingual architecture, the method enables adaptation of a low-resourced language with absolutely no data from the target domain
- According to experimental validation, the proposed cross-lingual domain adaptation approach:
  - Outperforms other similar methods
  - Allows for remarkable improvements even in less favourable language and domain conditions
- Future work will focus on extending the method to:
  - sequence-trained models; combination with other cross-lingual information transfer methods, (e.g. multi-ling/multi-domain BNF) and SAT vector-based approaches (e.g. i-vectors).

Thank you!!

Questions and comments are welcome!!

Alberto Abad – alberto.abad@inesc-id.pt