Investigating Cross-lingual Multi-level Adaptive Networks:
The Importance of the Correlation of Source and Target Languages

Alexandros Lazaridis, Ivan Himawan, Petr Motlicek, Iosif Mporas and Philip N. Garner

Alexandros Lazaridis

8-9 Dec. 2016

13th IWSLT, Seattle

Outline

- Motivation
- Introduction
- Multitask training and MLAN schemes
- Experiments (setup & results)
- Conclusions

Motivation

- General motivation
  - Exploiting data from other languages
    - For building robust models in under-resourced languages (or dialects)
      - data may be available, in a closely related (correlated) language

Correlation: phonetic similarity between the two languages.

- Our motivation
  - SUMMA H2020 project
    - Monitoring media/news in different languages
      - NLP topic clustering
      - Multilingual automatic speech recognition (ASR) in combination with machine translation for broadcast data monitoring
        - » Good resources: Arabic, English, German, Spanish
        - » Some resources: Portuguese, Russian, Farsi
        - » Poor resources: Ukrainian, Latvian

Introduction: Knowledge Transfer

- Exploiting out-of-domain (OOD) knowledge in ASR is not new!
  - in GMM/HMM frameworks (e.g. domain or speaker adaptation)
    - Maximum a posteriori (MAP) or
    - Maximum likelihood linear regression (MLLR)
  - in DNN/HMM frameworks
    - Adaptation by adding an extra input layer with a linear activation function (speaker adaptation)
    - Exploiting OOD data to better initialize the nets
    - Multi-task training
      - The output layer (i.e., softmax layer) could vary between domains during training
      - Sharing the hidden layers across domains
    - The tandem features
      - Concatenating bottleneck (BN) features derived from the OOD network with the in-domain acoustic features

Introduction: Cross-lingual ASR (1/2)

- Cross-lingual ASR is exploiting OOD knowledge
  - Target language models are enhanced using data from a different source language
  - The target language is low-resourced:
    - Transcribed data are difficult or expensive to be acquired

- Models are able to capture common properties of the acoustics of speech which are shared across languages, despite mismatches across languages.
  - Improving the generalization of the final models to unseen speakers and conditions

  - Cross-lingual ASR techniques
    - Regularization
      - e.g. initializing nets on source data
    - Feature space
      - e.g. BN features
    - Hierarchical
      - e.g. combining nets into deep structure

Introduction: Cross-lingual ASR (2/2)

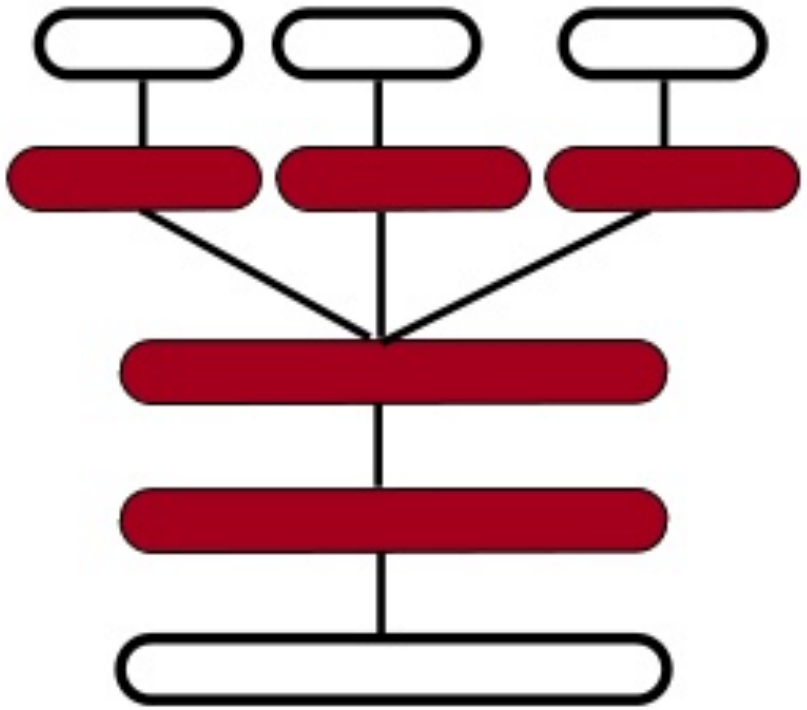
- Cross-lingual ASR is a form of adaptation
  - Major difference in respect to e.g. domain or speaker adaptation
    - Mismatch of the phonesets of source and target languages
      - Even with universal phonesets (IPA) the same phone, in practice, might differ across languages.
      - Mapping of phones manually or data-driven.
      - Separate phonesets could be used (i.e., different output layers in DNNs).

- We investigate the importance of the correlation of the source and target languages in the framework of cross-lingual adaptation.

- Multi-level adaptive networks (MLAN) scheme, aims to take advantage of BN features

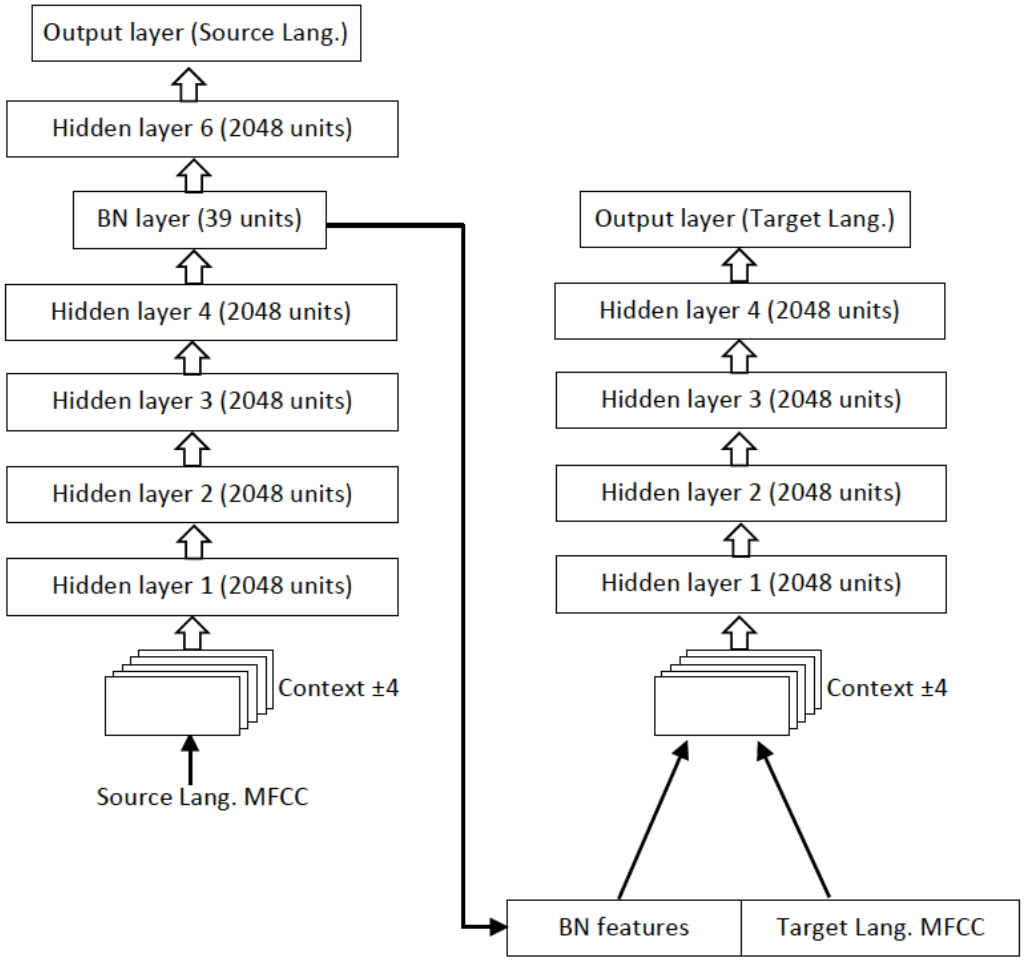
Multitask training

- The method of training a classifier to operate on two or more related tasks
  - A shared representation
  - Sharing lower hidden layers
  - Output layer (or even higher hidden layers) are task/language dependent
  - Multitask training increases the amount of data for each network
  - Aims to improve generalization of the model
  - The phoneset mismatch is dealt by swapping output layers during training



MLAN Scheme

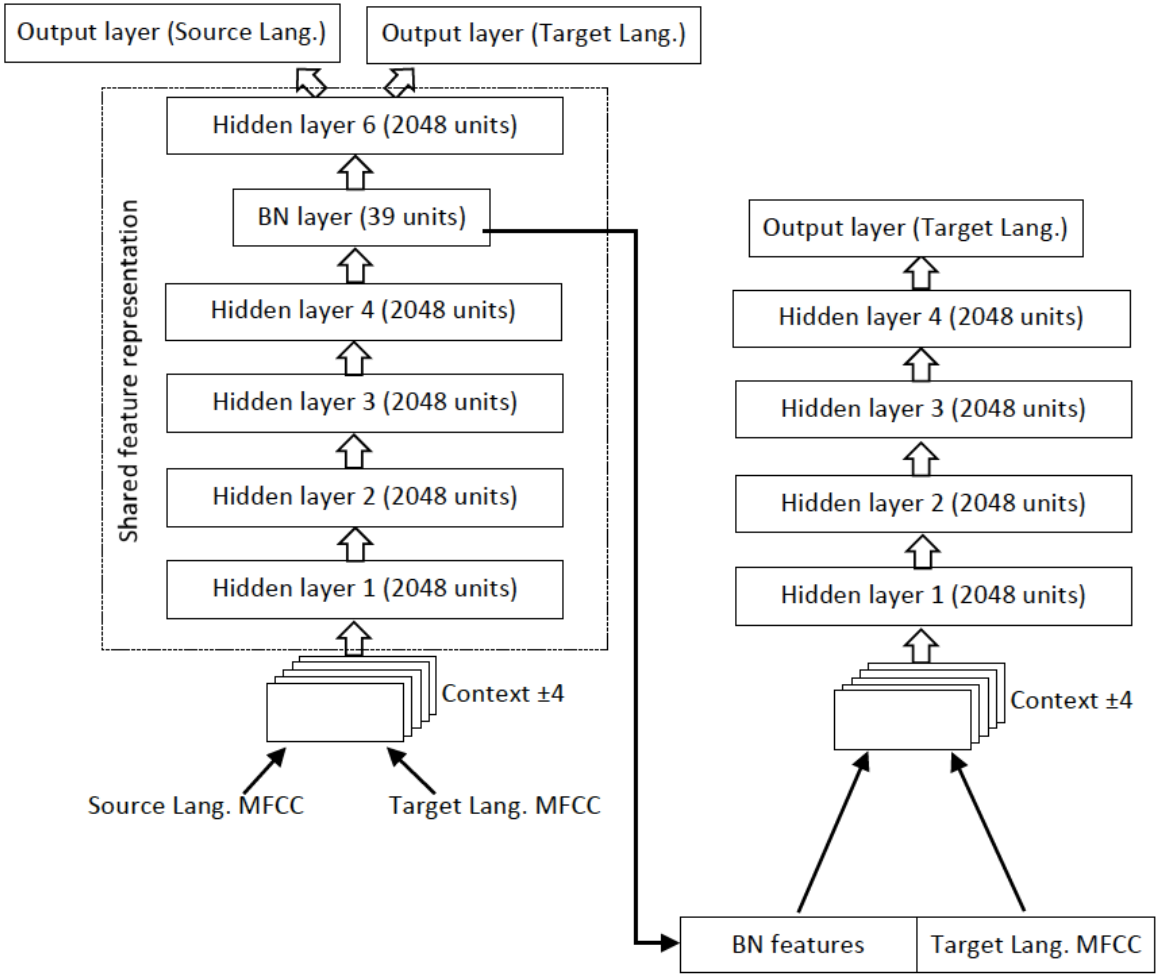
- Training a two-DNN structure
  - 1st DNN with BN layer is trained using OOD
    - Data from the source language
  - 2nd DNN trained with extracted BN features combined with in-domain (target language) spectral features



- The second-level DNN training is used to discriminatively select OOD features that are important for classification.

Multi-MLAN Scheme

- In multi-MLAN
  - Multi-task training for the 1st DNN
  - Source and target language tasks jointly
    - Using shared feature representation
    - The network effectively increases the amount of training data for each network.



- It allows a shared, language-independent speech representation to be learned.

Experimental Setup: Investigated Scenarios

- Russian used as target language (GLOBALPHONE)
  - Training set: 21 hours of speech,
    - split to 90% - 10% sets, training - cross-validation
  - Evaluation set: 1.6 hours (10 speakers)
- Source language scenarios
  - 1st: French, a language uncorrelated to the target one
    - a 47% overlap of phones in the phonesets
    - 25 hours of speech (GLOBALPHONE).
  - 2nd: Ukrainian, as correlated to the target one
    - a 79% overlap of phones in the phonesets
    - 11.5 hours of speech (GLOBALPHONE).
  - 3rd: French (size respective with the 2nd scenario)
    - 12.5 hours of speech (GLOBALPHONE).
  - 4th: English language (150 hours), as uncorrelated
    - a 45% overlap of phones in the phonesets
    - 50 hours from LIBRISPEECH
    - 50 hours from ICSIAMI
    - 50 hours from TEDLIUM

Experimental Setup: AM & LM

- Kaldi toolkit was used to build DNN/HMM system
- 39-dimensional MFCC features (including their delta and delta-delta)
- 9-frame temporal context
- 4 hidden layers of 2048 neurons/layer
- Sigmoid as activation function
- BN layer composed of 39 units.
- State alignments for training DNNs obtained from GMM/HMM system
- Dictionaries
  - CMU dictionary for English (39 phones)
  - GLOBALPHONE dictionaries
    - Ukrainian: 49 phones
    - French: 38 phones
    - Russian: 47 phones
- The decoding performed using a 3-gram LM developed based on GLOBALPHONE.

Results: General remarks

- All three cross-lingual adaptation schemes (in all scenarios) outperform the Baseline DNN
  - *Adaptation* case: Using source language net as initial net in target language
  - The *MLAN* improves the accuracy in respect to the *Adaptation*
  - The *multi-task MLAN* achieves the best performance

System	Source Language	WER(%)
Baseline	-	30.50
Adaptation	French (12.5h)	30.51
MLAN		28.86
multi-task MLAN		28.20
Adaptation	French (25h)	30.38
MLAN		28.66
multi-task MLAN		27.96
Adaptation	Ukrainian (11.5h)	30.33
MLAN		28.56
multi-task MLAN		28.00
Adaptation	English (150h)	29.83
MLAN		27.71
multi-task MLAN		27.62

Results: Importance of language correlation (1/2)

- In 1st&2nd scenarios, full French data and Ukrainian data, the performance is very similar.
  - The *Adaptation* outperform the Baseline DNN
    - by 0.4% relative improvement for French case.
    - by 0.6% relative improvement for Ukrainian case.
  - The *MLAN* outperforms the Baseline
    - by 6% relative improvement for French case.
    - by 6.4% relative improvement for Ukrainian case.
  - The *multi-task MLAN* outperforms the Baseline DNN
    - by 8.3% relative improvement for French case.
    - by 8.2% relative improvement for Ukrainian case.

System	Source Language	WER(%)
Baseline	-	30.50
Adaptation	French (12.5h)	30.51
MLAN		28.86
multi-task MLAN		28.20
Adaptation	French (25h)	30.38
MLAN		28.66
multi-task MLAN		27.96
Adaptation	Ukrainian (11.5h)	30.33
MLAN		28.56
multi-task MLAN		28.00
Adaptation	English (150h)	29.83
MLAN		27.71
multi-task MLAN		27.62

- We get similar results between correlated and uncorrelated source language cases, using double size of data for the uncorrelated source language.

Results: Importance of language correlation (2/2)

- 3rd scenario, using half of the French data
  - Matching approximately the amount of Ukrainian data
  - Accuracies of all three adaptation schemes were decreased in respect to the full French data case.
  - Using the correlated source language leads to slightly lower errors

- Validation of our hypothesis: Importance of the correlation of the source and target languages

System	Source Language	WER(%)
Baseline	-	30.50
Adaptation	French (12.5h)	30.51
MLAN		28.86
multi-task MLAN		28.20
Adaptation	French (25h)	30.38
MLAN		28.66
multi-task MLAN		27.96
Adaptation	Ukrainian (11.5h)	30.33
MLAN		28.56
multi-task MLAN		28.00
Adaptation	English (150h)	29.83
MLAN		27.71
multi-task MLAN		27.62

Results: Importance of the amount of data

- 4th scenario, using English as source language, using 150 hours of data
  - Investigating the importance of the amount of training data used from the source language.
  - In all cases outperforms the scenario using the correlated source language
  - The highest improvement, over the Baseline system,
    - by 9.2% *MLAN* relative improvement.
    - by 9.4% *multi-task MLAN* relative improvement.

System	Source Language	WER(%)
Baseline	-	30.50
Adaptation	French (12.5h)	30.51
MLAN		28.86
multi-task MLAN		28.20
Adaptation	French (25h)	30.38
MLAN		28.66
multi-task MLAN		27.96
Adaptation	Ukrainian (11.5h)	30.33
MLAN		28.56
multi-task MLAN		28.00
Adaptation	English (150h)	29.83
MLAN		27.71
multi-task MLAN		27.62

- Importance of using adequate amount of data from the source language.
- The difference between the two MLAN schemes is very small

Conclusions

- The MLAN and multi-task MLAN schemes were investigated
- French, Ukrainian and English as source languages,
- Russian as target language.
- Using French as the source language
  - the MLAN schemes needed to be trained with double the size of data of the ones used in the Ukrainian case
- Using less French data (half size),
  - the performance was decreased in respect to the scenario where the Ukrainian data were used.

- The results showed the importance of the correlation between the source and target languages.

- Using English as source language (6 times and 13 times more training data)
  - the MLAN and multi-task MLAN schemes achieved the highest relative improvement

- The results have shown the importance of the amount of source language data used.
- It is worth training using large amounts of uncorrelated data.

Thank you!

Questions?