ARL-IR-0000 • APR 2018

US Army Research Laboratory

A Minimal Example of MTL for ASR

by John Morgan, Stephen LaRocca and Michelle Vanni

NOTICES

Disclaimers

ARL-IR-0000 ● APR 2018

US Army Research Laboratory

A Minimal Example of MTL for ASR

by John J Morgan
Computational and Information Sciences Directorate, ARL

Stephen A LaRocca and Michelle Vanni
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE	***Form Approved*** ***OMB No. 0704-0188***

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED (From - To)**
April 2018	Internal Report	October 2016-November 2016

4. TITLE AND SUBTITLE
A Minimal Example of MTL for ASR

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
John Morgan, Stephen LaRocca and Michelle Vanni

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
US Army Research Laboratory
ATTN: RDRL-CII-T
Adelphi Laboratory Center, MD 20783-1138

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-IR-0000

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
primary author's email: <john.j.morgan50.civ@mail.mil>.

14. ABSTRACT
Multitask Learning was applied to a Large corpus of English and a small corpus of Modern Standard Arabic read speech for the purpose of improving the performance of an Automatic Speech Recognition system. An improvement in Word Error Rate over the best Singletask Learning method was observed.

15. SUBJECT TERMS
Automatic Speech Recognition of Accented Speech, Multitask Learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					John J Morgan
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER (Include area code)**
Unclassified	Unclassified	Unclassified	UU	16	301-394-1902

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

Contents

List of Tables

Acknowledgments

John Morgan wishes to sincerely thank his co-author, Dr. Stephen LaRocca.

INTENTIONALLY LEFT BLANK.

1. INTRODUCTION

For the experiments described here, we used the ANN as a framework for acoustic modeling in Automatic Speech Recognition (ASR). Recently, large ASR systems have been trained on tens of thousands of hours of speech data.[1] Artificial Neural Networks have performed well when these amounts of data are available. We assumed we are under conditions of severe training data sparcity in the Target Language (TL). We also assumed that we have access to a large corpus of speech in another language. We refer to this other language as the Background Language (BL).

Multitask Learning (MTL)[2] is a framework that enables the advantages of Deep Learning (DL) to be applied in the situation where there exists plentiful resources in the BL and scarce resources in the TL.

Our research question was formulated as follows. Can a ANN AM trained with MTL on data from two different languages improve performance of an ASR system for a TL that is a Low Resource Language (LRL)?

2. DATA

We ran an experiment to test the MTL method on two specific publically available corpora: the large Librispeech English corpus and a small corpus of Tunisian Accented MSA.

2.1 Librispeech

LibriSpeech is a corpus of read speech, based on LibriVox's public domain audio books. The corpus is available at:

`http://www.openslr.org/resources/12`.

We used the cleaned training fold of 960 hours of speech.

2.2 Tunisian MSA:

This is a corpus of ten hours of MSA collected in 2003 from a sample of 120 mixed male and female informants. The informants provided recitations and answers to questions. It can be downloaded at:

`http://www.openslr.org/resources/46`.

3. EXPERIMENT

We used the kaldi toolkit[3] to build our ASR systems. We derived our setup from the kaldi babel multilang recipe.

a ANN was built with two kinds of layers.

1. Shared Layers.

2. Language Specific Layers.

The shared layers were trained on all the training data from both the TL and BL languages.

The language specific layers were trained only on data from the TL.

We tried to write our recipe so that it only contain steps that are required to implement the MTL method. Thus, We did not use i-vectors which are standard in many kaldi recipes. however, we did include a bottleneck layer.

We built baseline Speaker Adaptive Training (SAT) Gaussian Mixture Model (GMM) Hidden Markov Model (hmm) Acoustic Models for both the Librispeech and Tunisian MSA corpora. For the Tunisian MSA baseline system we derived our pronouncing dictionary from the 2 million entries from the Qatar Computing Research Institute (QCRI) vowelized dictionary[4] available at:

`http://alt.qcri.org/resources/speech/dictionary/ar-ar_lexicon_2014-03-17.txt.bz2`

We added the Out Of Vocabulary (OOV) words from the Tunisian MSA training set and the test set. We trained our 3-gram language model with the Stanford Research Institute Language Modeling (SRILM) toolkit[5] on the transcripts from the training and test data. Our best Word Error Rate (WER) results for Tunisian MSA were obtained with online chain models.

We followed the kaldi standard recipe for the librispeech cleaned 960 hours of speech task with one exception. We extracted and trained with Perceptual Linear Prediction (PLP) pitch features instead of Mel Frequency Cepstral Coefficient (MFCC) features. We followed this path since we derived our scripts from the kaldi babel multilang recipe. In the future We plan on incorporating tonal Background Languages and we expect better results using PLP pitch instead of MFCC features.

3.1 ANN Configuration

The ANN had ten layers.

1. One input layer,
2. 6 hidden layers,
3. One Bottleneck layer,
4. One affine layer, and
5. One soft max layer.

The Rectified Linear Unit (RELU) function was used to compute activations. The dimension of the hidden layers was 1024. The dimension of the Bottleneck layer was 512. The final layer implemented a soft max function that output a probability density function over the clustered triphones. The frame Context was set to 16 frames to the left and 12 frames to the right.

3.2 AM Training

A bilingual raw feedforward deep ANN was trained on the combined set of training examples from the English Librispeech and Tunisian MSA corpora using a frame-level cross-entropy objective function. The data from the Tunisian MSA corpus was used to readjust the parameters in the last two layers of the bilingual Deep Neural Network (DNN) model to produce a new monolingual Tunisian MSA acoustic model. Similarly, a new monolingual English model was produced. These two models shared the parameters in their first eight layers, only their final 2 layers were different.

3.3 Decoding

The monolingual system with the Tunisian MSA AM was used to decode a test set of speech from four speakers, 3 Libyan males and one Tunisian female. The same Finite State Transducer (FST) decoding graph that was built for the Tunisian MSA SAT GMM hmm system was used for decoding with the MTL AM set.

4. RESULTS

The SingleTask Tunisian MSA Baseline system with chain models yielded a WER of 11.03.After MTL, the Tunisian MSA system gave a WER of 7.12. An improvement of 34.45 percent.

5. Discussion

The results obtained in the experiment described above are encouraging for the U.S. Army. The language technology needs of the U.S. Army frequently are in cases where data for training Machine Learning (ML) models are very scarce. Our results indicate that data from high resource languages can be leveraged to enable thedevelopment of language technologies for Low Resource Languages

6. References

1. Heigold G, Vanhoucke V, Senior A, Nguyen P, Ranzato M, Devin M, Dean J. Multilingual acoustic models using distributed deep neural networks,âĂİ ICASSP. 2013.

2. Caruana R. Multitask learning: A knowledge-based source of inductive bias. In: Proceedings of the Tenth International Conference on Machine Learning; 1993 ; Venue Unknown. Morgan Kaufmann; 1993. p. 41–48.

3. Povey D, Ghoshal A, Boulianne G, Goel N, Hannemann M, Qian Y, Schwarz P, Stemmer G. The kaldi speech recognition toolkit. In: In IEEE 2011 workshop; 2011 ; Venue Unknown.

4. Ali A, Zhang Y, Cardinal P, Dahak N, Vogel S, Glass J. A complete kaldi recipe for building arabic speech recognition systems. In: Spoken Language Technology Workshop (SLT), 2014 IEEE; 2014 Dec; Venue Unknown. p. 525–529.

5. Stolcke A. Srilm - an extensible language modeling toolkit. In: SRILM - An Extensible Language Modeling Toolkit; 2002 ; Interspeech. p. 901–904.

Acronyms

AM Acoustic Model. 1, 2, 4

ANN Artificial Neural Network. 1

ASR Automatic Speech Recognition. 1, 2

BL Background Language. 1–3

DL Deep Learning. 1

DNN Deep Neural Network. 3

FST Finite State Transducer. 4

GMM Gaussian Mixture Model. 2, 4

hmm Hidden Markov Model. 2, 4

LRL Low Resource Language. 1, 4

MFCC Mel Frequency Cepstral Coefficient. 3

ML Machine Learning. 4

MSA Modern Standard Arabic. iii, 1–4

MTL Multitask Learning. 1, 2, 4

NN Neural Network. 2, 3

OOV Out Of Vocabulary. 2

PLP Perceptual Linear Prediction. 3

QCRI Qatar Computing Research Institute. 2

RELU Rectified Linear Unit. 3

SAT Speaker Adaptive Training. 2, 4

SRILM Stanford Research Institute Language Modeling. 2

TL Target Language. 1, 2

WER Word Error Rate. 3, 4

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO L IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA

INTENTIONALLY LEFT BLANK.