Karlsruhe Institute of Technology

Online Neural Network-based Language Identification

Master's Thesis of

Daniel H. Draper

at the Department of Informatics
Institute for Anthropomatics and Robotics

Reviewer:	Dr.-Ing. Sebastian Stüker
Second reviewer:	
Advisor:	M.Sc. Markus Müller

12. December 2016 – 11. May 2017

Karlsruher Institut für Technologie
Fakultät für Informatik
Postfach 6980
76128 Karlsruhe

I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 12th of May, 2017

. . . . . .

(Daniel H. Draper)

Abstract

Zusammenfassung

Contents

List of Figures

List of Tables

1 Introduction

Language Identification describes the task of differentiating between spoken speech in different languages and being able to correctly identify which speech-segment consists of which language. Neural Networks refer to Artificial Neural Network's, a Machine Learning approach to classification tasks employed greatly throughout all sciences and especially in computer science and tasks concerned with the processing on spoken speech. This thesis tries to find a low-latency, fast, or "online", approach to Language Identification.

The following chapter gives an introductory view of the applications of Language Identification, and introduces the tasks this thesis tries to solve. It also presents related work and how this thesis can be put into perspective to those works. Afterwards we give preliminary theoretical explanations and definitions, including an introduction to Neural Networks in Sec. **??**. The chapter afterwards introduces the language identification tasks this thesis deals with, and the different data corpusses used, followed by our solution split into two chapters: the preprocessing and actual neural networks trained with its training results. Evaluation results are then presented followed by the conclusion and an outlook.

1.0.1 Applications

Automatic Speech Recognition (ASR) is used in many applications and devices today, especially in the rise of handheld mobile devices like smart-phones and tablets. It has progressed quickly in the last five years and has found commercial success. Famous examples include Google[1]'s "Ok, Google" and Apple[2]'s Siri. Which both include voice search[FHBM08], a form of voice control, that even is extensible in the case of Google and Android e.g[bAO14]. Many other applications have emerged, including spoken language translation[3], especially for this thesis in the realm of Lecture Translation[MNN+16] .

The task of Language Identification can be applied in all of those fields, as Automatic Speech Recognition is trained on one language and therefore always requires manual changing of the language as to use the correct language for the speech recognizer. Robust and low-latency language identification would eliminate the need for this.

Spoken language translation, as used for example in the European Parliament where already components of ASR and Machine Translation are employed and are being actively developed[4][VMH+05].

[1] Google: `www.google.com`
[2] Apple: `www.apple.com`
[3] IWSLT: `iwslt.org`
[4] TC-STAR: `tcstar.org`

This thesis will focus mostly on the KIT's lecture Translator[5] as the system trained was implemented for it. We believe our results are general enough to be transferable to other applications with implementation-specific changes.

1.0.2 Language Identification in History

1.1 Related Work

This section takes a look at related work that shows different modern approaches of identifying Language in spoken speech and describe the differences between their work and our approach.

[5]Lecture Translator: `https://lecture-translator.kit.edu`

2 Preliminary Definitions

In the following chapter we want to define and explain terms and concepts used throughout this thesis as well as give an outlook to related work and the general language identification approaches.

3 Language Identification Tasks

This chapter introduces the datasets used to train the networks employed in this approach. While Language Identification is applicable in many different scenarios, in this thesis the focus lies on trying to establish a low-latency online approach for recognizing the spoken language in a university-lecture environment. Because finding a suitable test setup for online data retrieval is hard the data used was cut to short lengths to make an evaluation as to correctness of the recognition possible in an ”online-like” scenario.

This means that the output of the net is evaluated after short samples of speech and therefore can be seen as indicative of online performance of the neural net.

3.0.1 Euronews 2014

Our first data set we retrieved from Euronews [1] 2014. Euronews is a TV channel that is broadcast in 13 different languages simultaneously both on TV and over the Web and is semi-automatically transcribed. The data corpus includes our 10 language (Arabian, German, Spanish, French, Italian, Polish, Portuguese, Russian, Turkish and English) with about 20 hours of data per language provided overall. Details of this breakdown can be seen in table 3.0.1.

Language	**Number of Speakers**	**Length overall**
Arabian	1055	
German	928	
Spanish	932	
French	1016	
Italian	935	
Polish	1229	
Portuguese	1062	
Russian	958	
Turkish	957	
English	928	
Overall	10000	

Table 3.1: The Euronews corpus speaker breakdown with total utterances length

The speaker list was then split into three smaller datasets: the train set, development set and test set using the common Simple Random Sampling. The sizes were 80% train set, and 10% for both the development and test set. Table **??** shows the split data for the three sets.

[1] Euronews: http://www.euronews.com/

3.0.2 Lecture Data

3.0.3 European Parliament

4 Feature Preprocessing

This chapter deals with the feature preprocessing used to form normal speech into feature vectors to be understood by neural networks. The setup we used is based on the standard capabilities of the Janus Recognition Toolkit[1]. It is then run through a six layer Automatic Speech Recognition network that was pre-trained on 10 languages. The 2^{nd} last layer of the ASR net is a Bottleneck feature layer, where the feature vectors are extracted and then used as input for the trained Language Identification Network. The following sections describe this Feature Preprocessing for data as well as the first ASR network used to create the BNF features the LID net requires.

4.0.1 Feature Access

4.0.2 Feature Description

The extracted ADC features from the audio files are then used for further preprocessing. We first use a standard Mel filter bank to extract only the necessary coefficients from the ADC0 feature.

First a spectrum is applied to the ADC0 Feature, therefore calculating the Fast Fourier Transformation of the Digitalized Signal.

[1]Janus Recognition Toolkit(JRTk): http://isl.anthropomatik.kit.edu/cmu-kit/english/1406.php

5 LID Network

This chapter describes the actual Language Identification Neural Network trained as well as the results of different network/data setups used. Most network experiments in the Network Geometry, meaning the number of hidden layers as well as the layout of the neurons in these layers were tried using the Euronews corpus. Results from this corpus were then transferred over to the other corpusses, meaning that the network layout that worked best for Euronews was then adjusted for the lecture data but otherwise the geometry was kept intact.

5.1 Basic Setup

Many different tools exist for deeplearning, the most acclaimed being Tensorflow [AAB+16], DL4J/ND4j[1] and Theano [BBB+11]. Pre-existing work on Language Identification using ASR BNFs wused a python wrapper around Theano for training, that was developed by Jonas Gehring [Geh12]. This thesis continues the use of this wrapper. The basic layout of the network used and improved upon by this thesis were input vectors from the ASR net 4.0.1 with 966 coefficients. The net setup were 5 layers of denoising auto-encoders with each 1000 nodes and a tanh activation function using the mean squared error as the loss function.

The neural net was then trained using mini batches of size 2m and a learning Rate of 0.01. The pretrained net was then retrained with a 1000 neuron to 10 coefficient output to get to our 10 Languages as classes to classify against. In the basic setup this was trained using a learning rate of 1 and the exit condition of a minimum change of 0.005 / 0.0001 for the training/validation data respectively.

The beginning benchmark to improve upon was then set to the frame-based validation/-train error of 0.23 / 0.27 respectively.

The following sections describe different network layouts and changes we made to the training of the neural network and the improvements we managed to make upon our initial result.

5.2 Improving Network Layout

Different experiments were undertaken with the network layout. This includes a 6 hidden layer-pretraining as well as a change in the geometry. The differences in the frame-based errors can be seen in Table **??**

[1] DL4J: `https://deeplearning4j.org/index.html`

6 Conclusion

Bibliography

[AAB+16] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Gregory S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian J. Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Józefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Gordon Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul A. Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda B. Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *CoRR*, abs/1603.04467, 2016.

[bAO14] N. bt Aripin and M. B. Othman. Voice control of home appliances using android. In *2014 Electrical Power, Electronics, Communicatons, Control and Informatics Seminar (EECCIS)*, pages 142–146, Aug 2014.

[BBB+11] James Bergstra, Frédéric Bastien, Olivier Breuleux, Pascal Lamblin, Razvan Pascanu, Olivier Delalleau, Guillaume Desjardins, David Warde-Farley, Ian Goodfellow, Arnaud Bergeron, et al. Theano: Deep learning on gpus with python. In *NIPS 2011, BigLearning Workshop, Granada, Spain*, volume 3. Citeseer, 2011.

[FHBM08] A.M. Franz, M.H. Henzinger, S. Brin, and B.C. Milch. Voice interface for a search engine, April 29 2008. US Patent 7,366,668.

[Geh12] Jonas Gehring. *Training deep neural networks for bottleneck feature extraction.* 2012. Karlsruhe, KIT, Pittsburgh, Carnegie Mellon Univ., Interactive Systems Laboratories, Masterarbeit, 2012.

[MNN+16] Markus Müller, Thai Son Nguyen, Jan Niehues, Eunah Cho, Bastian Krüger, Thanh-Le Ha, Kevin Kilgour, Matthias Sperber, Mohammed Mediani, Sebastian Stüker, and Alex Waibel. Lecture translator - speech translation framework for simultaneous lecture translation. In *Proceedings of the Demonstrations Session, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 82–86, 2016.

[VMH+05] David Vilar, Evgeny Matusov, Saša Hasan, Richard Zens, and Hermann Ney. Statistical machine translation of european parliamentary speeches. In *Proceedings of MT Summit X*, pages 259–266, 2005.