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

Abstract i

Zusammenfassung iii

1 Introduction 1

2 Preliminary Definitions 3
2.1 Related Work 3

3 Language Identification Tasks 5
3.0.1 Euronews 2014 5

4 Conclusion 7

List of Figures

List of Tables

1 Introduction

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.

2.1 Related Work

In this section we take a look at related work that shows different approaches of identifying Language in spoken speech and describe the differences between their work and our approach.

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. The first data corpus includes our 10 language (Arabian, German, Spanish, French, Italian, Polish, Portugese, Russian, Turkish and English) with about 20 hours of data per language provided overall. Details of this 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	
Portugese	1062	
Russian	958	
Turkish	957	
English	928	
Overall	10000	

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

4 Conclusion

Bibliography

[ABH+07] Wolfgang Ahrendt, Bernhard Beckert, Reiner Hähnle, Philipp Rümmer, and Peter H. Schmitt. Verifying object-oriented programs with KeY: A tutorial. In *5th International Symposium on Formal Methods for Components and Objects, Amsterdam, The Netherlands*, volume 4709 of *LNCS*, pages 70–101. Springer, 2007.

[AD94] Rajeev Alur and David L. Dill. A Theory of Timed Automata. *Theoretical Computer Science*, 126:183–235, 1994.

[BHS07] Bernhard Beckert, Reiner Hähnle, and Peter H. Schmitt, editors. *Verification of Object-Oriented Software: The KeY Approach.* LNCS 4334. Springer-Verlag, 2007.

[CFH+03] Edmund Clarke, Ansgar Fehnker, Zhi Han, Bruce Krogh, Olaf Stursberg, and Michael Theobald. Verification of Hybrid Systems Based on Counterexample-Guided Abstraction Refinement. In Hubert Garavel and John Hatcliff, editors, *Tools and Algorithms for the Construction and Analysis of Systems*, volume 2619 of *Lecture Notes in Computer Science*, pages 192–207. Springer Berlin Heidelberg, 2003.

[CW98] Ana Cavalcanti and Jim Woodcock. ZRC – A Refinement Calculus for Z. *Formal Aspects of Computing*, 10(3):267–289, 1998.

[DB14] John Derrick and Eerke A. Boiten. *Refinement in Z and Object-Z.* Springer London, 2014.

[GZ14] Bin Gu and Liang Zou. A refinement calculus for hybrid systems. In *Engineering of Complex Computer Systems (ICECCS), 2014 19th International Conference on*, pages 176–185, Aug 2014.

[Hen00] Thomas A. Henzinger. The theory of hybrid automata. In M.Kemal Inan and RobertP. Kurshan, editors, *Verification of Digital and Hybrid Systems*, volume 170 of *NATO ASI Series*, pages 265–292. Springer Berlin Heidelberg, 2000.

[MLP12] Stefan Mitsch, Sarah M. Loos, and André Platzer. Towards formal verification of freeway traffic control. In Chenyang Lu, editor, *ICCPS*, pages 171–180. IEEE, 2012.

[MV92] Carroll Morgan and Trevor Vickers. *On the Refinement Calculus.* Springer London, 1992.

[Pla08] André Platzer. Differential dynamic logic for hybrid systems. *Journal of Automated Reasoning*, 41(2):143–189, 2008.

[Pla10] André Platzer. *Logical Analysis of Hybrid Systems*. Springer, Pittsburgh, 2010.

[Pla15a] André Platzer. Guide for Keymaera Hybrid Systems Verification Tool. http://symbolaris.com/info/KeYmaera-guide.html, 2015.

[Pla15b] André Platzer. KeYmaera: A Hybrid Theorem Prover for Hybrid Systems. http://symbolaris.com/info/KeYmaera.html, 2015.

[RT13] Matthias Rungger and Paulo Tabuada. Abstracting and refining robustness for cyber-physical systems. *CoRR*, abs/1310.5199, 2013.

[RyS96] Jean-François Raskin and Pierre yves Schobbens. State Clock Logic: a Decidable Real-Time Logic. In *HART'97, LNCS 1201*, pages 33–47. Springer-Verlag, 1996.

[STW14] Steve Schneider, Helen Treharne, and Heike Wehrheim. The behavioural semantics of Event-B refinement. *Formal Aspects of Computing*, 26(2):251–280, 2014.

[Wol] Wolfram Research, Inc. Mathematica. https://www.wolfram.com.