Technische Universität Berlin

Quality and Usability Lab

Part-of-Speech Tagging with Neural Networks for a Conversational Agent

Master Thesis

Master of Science (M.Sc.)

Author	Andreas Müller
Major	Computer Engineering
Matriculation No.	333471

Date	18th May 2018
1st supervisor	Prof. Dr.-Ing. Sebastian Möller
2nd supervisor	Prof. Dr. ???

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe. Alle Ausführungen, die anderen veröffentlichten oder nicht veröffentlichten Schriften wörtlich oder sinngemäß entnommen wurden, habe ich kenntlich gemacht.

Die Arbeit hat in gleicher oder ähnlicher Fassung noch keiner anderen Prüfungsbehörde vorgelegen.

Berlin, den December 26, 2017

_____

Unterschrift

Abstract

...

Zusammenfassung

...

Contents

List of Figures

List of Tables

Abbreviations

Alex	*Artificial Conversational Agent*
FNN	*(Feed-forward) Neural Network*
HMM	*Hidden Markov Model*
NLP	*Natural Language Processing*
NLTK	*Natural Language Toolkit*
RNN	*Recurrent Neural Network*

1 Introduction

This is just a test

2 ALEX: Artificial Conversational Agent

...

2.1 System Overview

...

2.2 Language Model

...

2.3 Tagging Interface

The modular structure of ALEXallows for easier separation of various functions and therefore easier replaceability of certain functionalities. One of these modules is the tagger, which is used to train a language model on the one hand and to assign tags to the words of a given input sentence on the other hand.

The implementation of this tagger utilizes a Hidden Markov Model (HMM), which is a statistical model that is particularly used for pattern recognition, speech recognition and part-of-speech tagging. ALEXuses an already existing implementation of the HMM Tagger from the Natural Language Toolkit (NLTK)[1], called `HiddenMarkovModelTagger`.

To replace the existing tagger, a new tagger has to provide a class with two methods: `train` and `tag`. The `train` method creates a new instance of the tagger class, trains it with the given training data and returns it. The training data itself must be a list of

1 The Natural Language Toolkit is a collection of *Python* programming libraries for natural language processing, see `http://nltk.org`

sentences, where a sentence is a list of tuples, containing each word of this sentence and its corresponding tag. Listing 2.3 exemplifies the structure of the input data containing two tagged sentences.

```
[
  [
    ('the', TAG),
    ('dog', TAG),
    ('is', TAG),
    ('running', TAG)
  ],
  [
    ('the', TAG),
    ('cat', TAG),
    ('is', TAG),
    ('sleeping', TAG)
  ]
]
```

Listing 2.1: The structure of the input data containing two tagged sentences

3 Evaluation and Comparison

...

4 Discussion and Conclusion

...

A First appendix

A.1 test

...