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 April 11, 2018

_____

Unterschrift

Abstract

A part-of-speech tagger is a system which automatically assigns the part of speech to words using contextual information. Potential applications for part-of-speech taggers exist in many areas of computational linguistics including speech recognition, speech synthesis, machine translation or information retrieval in general.

The part-of-speech tagging task of natural language processing is also used in the advisory artificial conversational agent called ALEX. ALEX was developed to answer questions about modules and courses at the Technische Universität Berlin. The system takes the written natural language requests from the user and tries to transform them into SQL-queries. To understand the natural language queries, the system uses a Hidden Markov Model (HMM) to assign tags to each word of the query (part-of-speech tagging). This HMM tagger is trained with manually created training templates that are filled with the data in the database to be queried. The manually created sentence-templates and the slot-filling resulted in many training data sentences with the same structure. This often led to wrong tagging results when the HMM tagger was presented with an input sentence, having a structure that doesn't occur in the training templates.

This thesis shows two different Neural Network approaches for the language modeling of the input sentences and evaluates and compares both Neural Network based tagger as well as the HMM based tagger.

Zusammenfassung

Ein Part-of-speech Tagger ist ein System, welches Wortarten anhand von Kontextinformationen automatisch den gegebenen Wörtern zuordnet. Potentielle Anwendungen solcher Tagger gibt es in vielen Bereichen der Computerlinguistik wie Spracherkennung, Sprachsynthese, maschinelle Übersetzung oder Information Retrieval im Allgemeinen.

Part-of-speech Tagging wird auch in ALEX verwendet, einem Artificial Conversational Agent. ALEX wurde entwickelt, um Fragen zu Modulen und Lehrveranstaltungen an der Technischen Universität Berlin zu beantworten. Das System nimmt die in natürlicher Sprache geschriebenen Anfragen des Benutzers und versucht diese in SQL-Abfragen umzuwandeln. Um die natürliche Sprache zu verstehen, verwendet das System ein Hidden-Markov-Model (HMM), um jedem Wort der Eingabe Wortarten zuzuweisen (Part-of-speech Tagging). Dieser HMM-Tagger wird mit manuell erstellten Trainingsvorlagen trainiert, die mit den Daten der abzufragenden Datenbank gefüllt werden. Die manuell erstellten Satzvorlagen führten zu vielen Trainingsdatensätzen mit gleicher Struktur und damit oft zu falschen Tagging-Ergebnissen, wenn der HMM-Tagger einen Eingabesatz mit einer Struktur verarbeiten sollte, die in den Trainingsvorlagen nicht vorkommt.

Diese Arbeit zeigt zwei verschiedene Ansätze für die Sprachmodellierung der Eingabesätze basierend auf neuronalen Netzwerken und bewertet und vergleicht sowohl die Neuronalen Netzwerk-basierten Tagger als auch den HMM-basierten Tagger.

Contents

List of Figures

List of Tables

Abbreviations

ACA	*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

1.1 Scope of this Thesis

The scope of this thesis is the development of a Neural Network based part-of-speech tagger for the advisory Artificial Conversational Agent (ACA) ALEX, the training of different language models and their evaluation with corresponding test sets.

In order to accomplish the new language models, two different Neural Network architectures are implemented: A Feed-forward Neural Network and Recurrent Neural Network. For the training of both Neural Network implementations, a corpus of tagged language data is generated with the help various input templates, which are created on the basis of logged user input data.

To evaluate the language models, a data set of known data[1] and unknown data[2] is created. On the basis of this evaluation, both Neural Network models and the HMM are compared to each other.

In accordance to the evaluation results, the former HMM based part-of-speech tagger is then replaced by this new tagger. To guarantee a seamless integration, the new tagger is implemented as a separate module with the same program interface the old tagger already utilizes. This way no other components of the conversational agent have to be changed and the effort of the replacement is kept minimal.

1.2 Related Work

...

1 Data, that was already used for the training of the model
2 Data, that includes words and sentence structures, that didn't occur in the training data sets

1.3 Structure of this Thesis

As introduction, this first chapter gave a short overview about the subject of natural language processing and part-of-speech tagging in general.

The second chapter describes structure and functionality of the already existing ACA ALEX with the main focus on its language model and tagging interface.

Chapter 3 explains the implementation of a part-of-speech tagging system with two different Neural Network approaches.

The training of the language models including the retrieval of the training data and tuning of the training parameter is described in Chapter 4.

Chapter 5 shows the evaluation of each language model with a generated test set and their comparison.

In conclusion the final Chapter 6 discusses and summarizes the evaluation results and gives an outlook on future work.

2 ALEX: Artificial Conversational Agent

...

2.1 System Overview

...

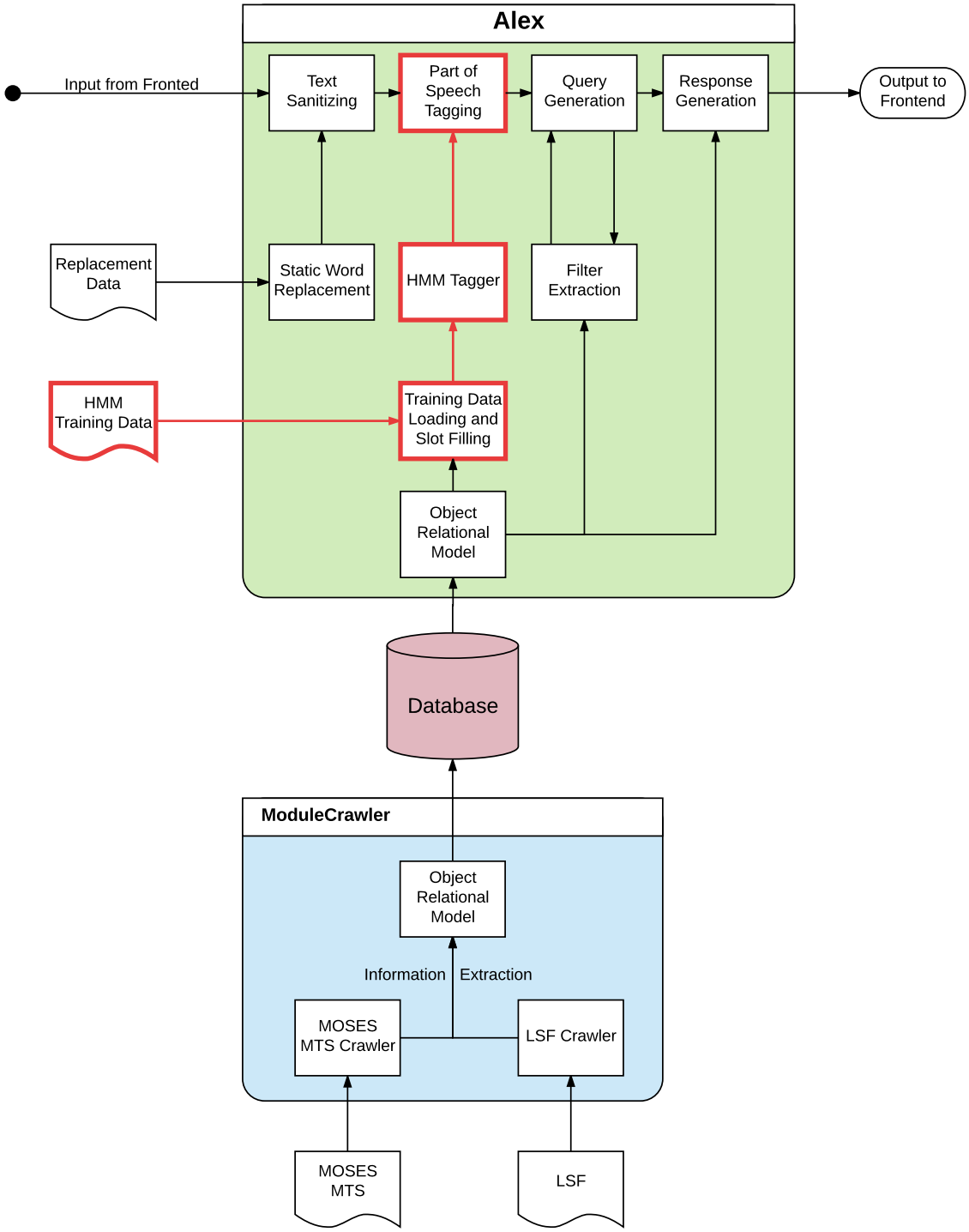

Figure 2.1: Overview of all components of ALEX. The red colored parts are components that lie within the scope of this thesis and are adapted or replaced. Original figure by Thilo Michael [1]

2.2 Hidden Markov Model

...

2.3 Tagging Interface

The modular structure of ALEX allows the separation of different functions and therefore easier replaceability of certain functionalities. Besides a web crawler for current data retrieval for the database and a frontend interface module, ALEX offers a tagging module. This module provides the training of a language model as well as the assignment of tags to the words of a given input sentence.

The implementation of this tagger utilizes a Hidden Markov Model, which is a statistical model that is particularly used for pattern recognition, speech recognition and part-of-speech tagging (see the previous section 2.2). ALEX uses 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`. These methods are used to create the language model and apply it to unknown data.

The `train` method creates a new instance of the tagger class, trains this class with the given training data and returns it. The training data itself must be a list of sentences, where a sentence is a list of tuples, containing each word of this sentence and its corresponding tag. The following exemplifies the structure of the training input data containing two sentences where each word is tagged with *TAG*:

```
[
  [ ('the', TAG), ('dog', TAG), ('is', TAG), ('running', TAG) ],
  [ ('the', TAG), ('cat', TAG), ('sleeps', TAG), ('all', TAG), ('day', TAG) ]
]
```

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

The `tag` method attaches a tag to each word of an input sentence, according to the previously trained language model. The input has to be an unknown sentence as a simple list of words:

```
[ 'an', 'unknown', 'test', 'sentence' ]
```

The output is a corresponding list of tuples containing a word and its assigned tag:

```
[ ('an', TAG), ('unknown', TAG), ('test', TAG), ('sentence', TAG) ]
```

3 Part-of-Speech Tagging

...

3.1 Feed-forward Neural Network Model

...

3.1.1 Architecture

...

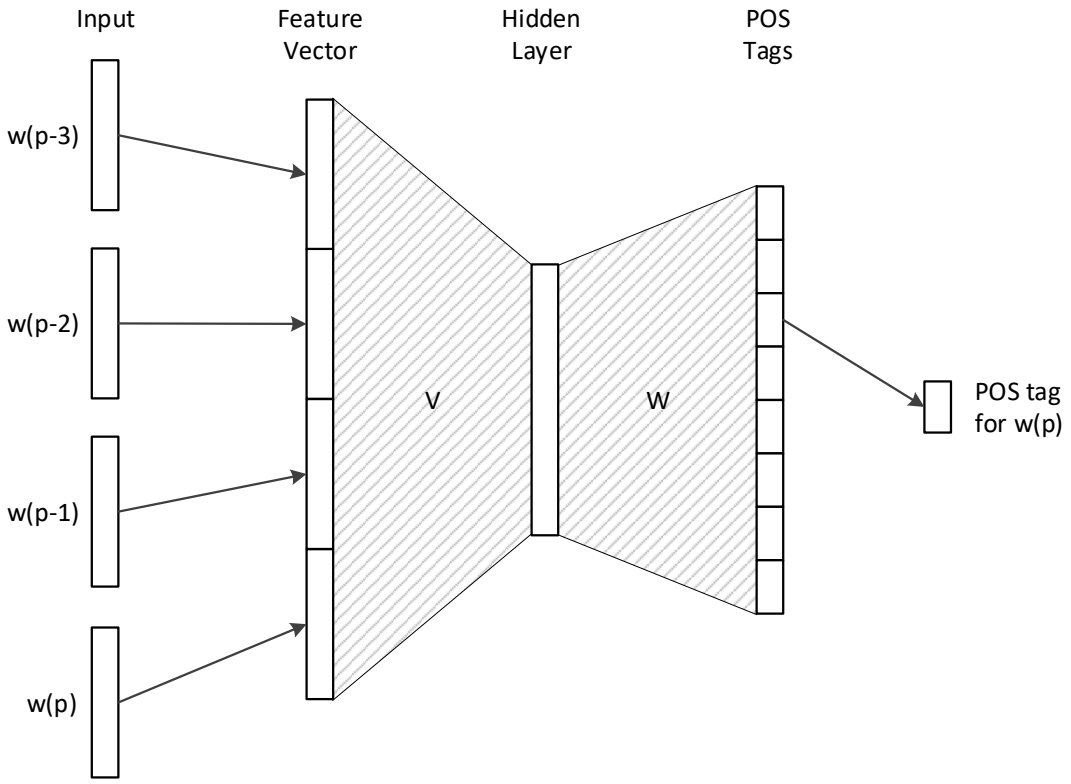

Figure 3.1: The structure of a Feed-forward Neural Network. The feature vector is built by the initial vectors of the corresponding input word on position p and by its three predecessors (here as an example).

3.1.2 Implementation

...

3.2 Recurrent Neural Network Model

...

3.2.1 Architecture

...

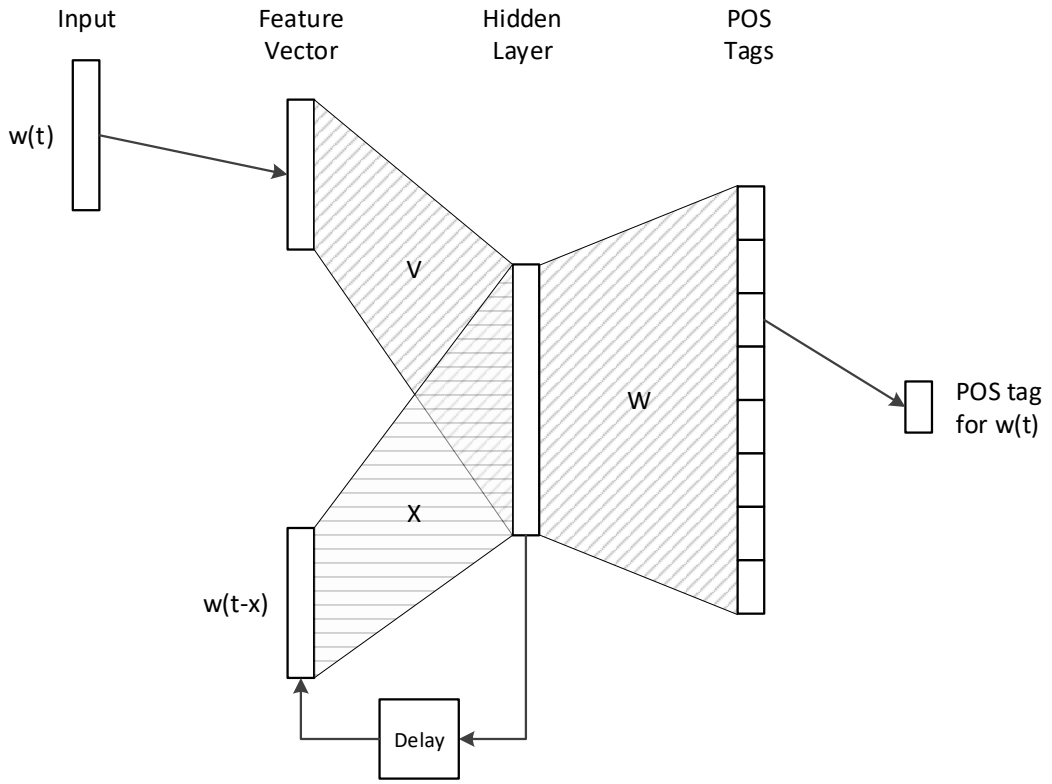

Figure 3.2: The structure of a Recurrent Neural Network. The feature vector is the initial vector of the corresponding input word at time `t`. The output of the hidden layer from previously trained words (here as an example at time `x`) is fed back into the same hidden layer for the current word.

3.2.2 Implementation

...

4 Training

...

4.1 Data Retrieval

...

4.2 Parameter Tuning

...

5 Evaluation and Comparison

...

5.1 Test Design

...

6 Discussion and Conclusion

...

6.1 Summary

...

6.2 Discussion

...

6.3 Future work

...

Bibliography

[1] Thilo Michael. Design and Implementation of an Advisory Artificial Conversational Agent, 10 2016.

A First appendix

A.1 test

...