TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Faculty of Computer Science

IT-Based textgeneration using NLP methods

State of the art and design of a prototype

Bachelor Thesis in
Business Informatics and Management

from

Tim Löhr

Student ID 3060802

First advisor: Prof. Dr. Alfred Holl

Second advisor: Prof. Dr. Florian Gallwitz

Hinweis: Diese Erklärung ist in alle Exemplare der Abschlussarbeit fest einzubinden. (Keine Spiralbindung)

TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Prüfungsrechtliche Erklärung der/des Studierenden

Angaben des bzw. der Studierenden:

Name: Vorname: Matrikel-Nr.:

Fakultät: Studiengang:

Semester:

Titel der Abschlussarbeit:

Ich versichere, dass ich die Arbeit selbständig verfasst, nicht anderweitig für Prüfungszwecke vorgelegt, alle benutzten Quellen und Hilfsmittel angegeben sowie wörtliche und sinngemäße Zitate als solche gekennzeichnet habe.

____

Ort, Datum, Unterschrift Studierende/Studierender

Erklärung zur Veröffentlichung der vorstehend bezeichneten Abschlussarbeit

Die Entscheidung über die vollständige oder auszugsweise Veröffentlichung der Abschlussarbeit liegt grundsätzlich erst einmal allein in der Zuständigkeit der/des studentischen Verfasserin/Verfassers. Nach dem Urheberrechtsgesetz (UrhG) erwirbt die Verfasserin/der Verfasser einer Abschlussarbeit mit Anfertigung ihrer/seiner Arbeit das alleinige Urheberrecht und grundsätzlich auch die hieraus resultierenden Nutzungsrechte wie z.B. Erstveröffentlichung (§ 12 UrhG), Verbreitung (§ 17 UrhG), Vervielfältigung (§ 16 UrhG), Online-Nutzung usw., also alle Rechte, die die nicht-kommerzielle oder kommerzielle Verwertung betreffen.

Die Hochschule und deren Beschäftigte werden Abschlussarbeiten oder Teile davon nicht ohne Zustimmung der/des studentischen Verfasserin/Verfassers veröffentlichen, insbesondere nicht öffentlich zugänglich in die Bibliothek der Hochschule einstellen.

Hiermit □ genehmige ich, wenn und soweit keine entgegenstehenden Vereinbarungen mit Dritten getroffen worden sind,

□ genehmige ich nicht,

dass die oben genannte Abschlussarbeit durch die Technische Hochschule Nürnberg Georg Simon Ohm, ggf. nach Ablauf einer mittels eines auf der Abschlussarbeit aufgebrachten Sperrvermerks kenntlich gemachten Sperrfrist

von Jahren (0 - 5 Jahren ab Datum der Abgabe der Arbeit),

der Öffentlichkeit zugänglich gemacht wird. Im Falle der Genehmigung erfolgt diese unwiderruflich; hierzu wird der Abschlussarbeit ein Exemplar im digitalisierten PDF-Format auf einem Datenträger beigefügt. Bestimmungen der jeweils geltenden Studien- und Prüfungsordnung über Art und Umfang der im Rahmen der Arbeit abzugebenden Exemplare und Materialien werden hierdurch nicht berührt.

____

Ort, Datum, Unterschrift Studierende/Studierender

Preface I

The following thesis was created during my 7th and last semester at the University of Applied Science - Georg Simon OHM.

Within my last three semesters I realized, that my major interest among all IT related topics is artificial intelligence.

Together with my professor *Prof. Dr. Alfred Holl* I worked out a method-matrix for the entire structure of this paper. Without his cooperative support overseas while I was studying abroad at the City University of Hong Kong, this thesis would have not been possible for me.

Even though Natural Language Processing is just a subfield of machine learning, the current state-of-the-art research is far beyond what I can research within a bachelor thesis. In this way, I decided to write my thesis about the subfield *textgeneration* within NLP. My state-of-the-art research includes all *hot topics* within NLP and my prototyp focuses only on the textgeneration part, to dive deeper into what NLP and especially textgeneration is able to accomplish in the year 2020.

Preface Il

My interest started basically with my IT project, in which my team and I programmed an autonomously driving remote control car with a deep neural network together with a Raspberry Pi 3. From this first project on, I selected all my elective courses to be related with machine learning or data science in any possible way. I wanted to further increase my knowledge, so I searched for a website which provides courses related to AI. I found *www.udacity.com*, which offers courses in cooperation with top IT companys, such as Google, Airbnb or Microsoft. Out of curiosity, I bought the course *Natural Language Processing.* After successfully finishing it, I was encouraged to write my bachelor thesis in a subfield of *Natural Language Processing.*

For my research I encountered a lot of recently published and old papers from *https://arxiv.org/.* To read through the papers requires a lot of prior knowledge, especially in mathematics, which I learned in my abroad semester in Hong Kong.

Machine Learning and more specifically NLP is not an intuitive study. I provided the common terminologys from top researchers and tried to make the entry into this field as smooth as possible if the reader has no prior knowledge about this topic.

I still recommend some basic linear algebra and calculus knowledge to understand the formulas more easily.

Thank you very much for reading.

Abstract

– At the end , finally finished :) –

Contents

Chapter 1.

Intro

In recent months and years, neural networks have produced many *state-of-the-art* results in almost all possible disciplines of machine learning [Xie 17]. The roots of Neural Networks (NN) lie down almost 80 years ago in 1943, when **McCulloch-Pitts** [McCu 43] compared for the first time neuronal networks with the structure of the human brain. This first attempt to approach artificial neurons with neurons from the brain lead to the nowadays commonly used understanding of a simple neuron of a basic neural network (shown in figure 1.1). Those neurons connected together create an aritificial neuronal network, which can calculate any possible logical or arithmetic function.

The range in which NN's (2020) can be applied nowadays is wide. Some disciplines have only been created due of the invention of neural networks, because they solve existing- and new problems very effective and efficiently. Many frequently held conferences around the globe contribute continuous evidence of the successes of neural networks. Among those various disciplines counts for example *Pattern recognition* with Convolutional Neural Networks (CNN) [Yann 98] or the famous *CIFAR-10* dataset [Kriz], where many amateurs [Löh 19] and experts attempt annually to further increase the accuracy of predicting the 10 different image classes.

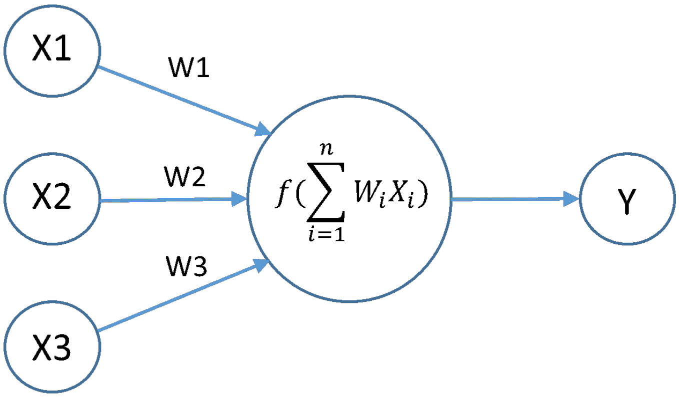

Figure 1.1.: A simple Neuron with 3 inputs and 1 output [Sing 17]

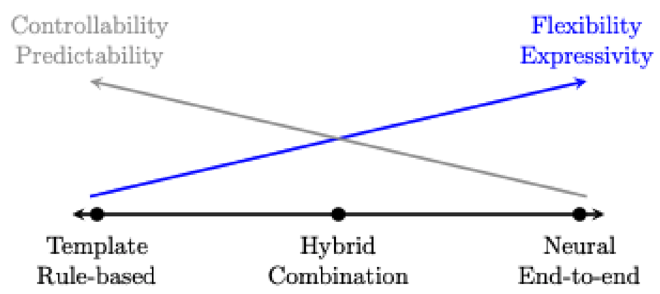

Figure 1.2.: Rule-Based vs. Neural-Text-Generations System [Xie 17]

The topic of this thesis *textgeneration* is based from the Natural Language Processing discipline, *NLP* for short. This field covers many other hot research topics, such as

- Sentiment Analysis
- Machine Translation
- Voice Recognition
- Text Generation (Neural Text Generation *NTG*)

Another term for textgeneration is denoted by *Language Modelling*, because generators use the words and grammar as input for the model. In the past five years were mainly two approaches for modelling NLP, namely the **rule-based** system and the **template-based** system (Figure 1.2) [Xie 17]. Today neural end-to-end systems are is *State-of-the-Art* [Jeka 17]. These new systems offer more flexibility and scale with proportionately better results and less data is required, because the complexity and thus the neccessary computing power has increased. However, this fact leads to a complexity problem, because it becomes very difficult to understand the decisions of the neural network. The neural network is still to a large extent basically a *black box*, although it gives surprisingly good results, especially in NLP. Nevertheless, neural network models for text processing are difficult to understand, so nowadays compromises between rule-based systems still have to be made and hybrid systems are most commonly used.

The neural text generation, also called *NTG*, has many other interesting application fields, including

- Speech recording and conversion to text
- Conversation systems e.g. chatbots
- Text summary

In order to train language models, they must be taught the probability of occurring words in relation to the preceding words. There are several approaches to achieve this goal. Language models can be trained on the level of words, whole sentences or even whole paragraphs. The granularity in which the training takes place is called *n-grams*, where *n* represents the number of preceding words.

1.1. Case study of a current NLP system

1.1.1. Case study

Image-to-Text | Captionbot Microsoft

1.1.2. Useful application areas of NLP systems

IoT, Grammerly, ok

1.1.3. Useful application areas of NTG systems

Grammerly

Chapter 2.

State of the Art

2.1. Relevante Aspekte der Mathematik

Notationen etc.

2.2. Geschichte des NLP

Zeitabfolge der geschichtlichen Hintergründe

2.3. Aktuelle Trends der Technologie

Neuronales Ende-zu-Ende

2.3.1. Einsatzgebiete von NLP-Systemen

Speech Recognition, Machine Translation

2.3.2. Einsatzgebiete von NTG-Systemen

Image-to-Text, Weatherforecast

Chapter 3.

Prototyp

In this chapter, we're actually using some code!

```
x = 1
if x == 1:
    # indented four spaces
    print("x␣is␣1.")
```

Listing 3.1: This is an example of inline listing

You can also include listings from a file directly:

```
x = 1
if x == 1:
    # indented four spaces
    print("x␣is␣1.")
```

Listing 3.2: This is an example of included listing

3.1. Zielsetzung

Image Captioning

3.2. Fachkonzept

Fachkonzept

3.2.1. Struktur

The different steps of Text Generation

- Importing Dependencies
- Loading the Data
- Creating Character/Word mappings
- Data Preprocessing
- Modelling
- Generating text

3.2.2. Neuronales Netz

LSTM https://www.analyticsvidhya.com/blog/2017/12/fundamentals-of-deep-learning-introduction-to-lstm/

Experimenting with different models

- A more trained model
- A deeper model
- A wider model
- A gigantic model

3.2.3. Prozessmodellierung

Funktionen etc.

3.2.4. Datenflussmodellierung

Diagramm

3.3. Implementierung

Code

3.4. Evaluation

Print Ergebnisse

Bild

Image Caption

Chapter 4.

Generierung von Übertragbarem Wissen

Modulare Erweiterbarkeit meines Projekts. Einordnung in Gesellschaftlichen Kontext

Appendix A.

Supplemental Information

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

List of Figures

List of Tables

List of Listings

Bibliography

[Jeka 17] O. D. Jekaterina Novikova and V. Rieser. “The E2E Dataset: New Challenges For End-to-End Generation”. 2017.

[Kriz] A. Krizhevsky, V. Nair, and G. Hinton. “CIFAR-10 (Canadian Institute for Advanced Research)”.

[Löh 19] T. Löhr and T. Bohnstedt. “Image Classification on the CIFAR10 Dataset”. 2019.

[McCu 43] W. S. McCulloch and W. Pitts. “A logical calculus of the ideas immanent in nervous activity”. *The bulletin of mathematical biophysics*, Vol. 5, No. 4, pp. 115–133, 1943.

[Sing 17] P. Singh. “Neuron explained using simple algebra”. 2017.

[Xie 17] Z. Xie. “Neural Text Generation: A Practical Guide”. 2017.

[Yann 98] L. B. Yann LeCun, Patrick Haffner and Y. Bengio. “Object Recognition with Gradient-Based Learning”. 1998.