TECHNISCHE HOCHSCHULE NÜRNBERG
GEORG SIMON OHM

Fakultät Informatik

IT-Basierte Textgeneration mit NLP methoden

Bachelorarbeit im Studiengang Wirtschaftsinformatik

vorgelegt von

Tim Löhr

Matrikelnummer 3060802

Erstgutachter:	Prof. Dr. Alfred Holl
Zweitgutachter:	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

Kurzdarstellung

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.

Abstract

ToDo12s

Contents

Chapter 1.

Einleitung

In den letzten Monaten und Jahren bringen Neuronale Netze viele *State-of-the-Art* Ergebnisse in beinahe allen möglichen Disziplinen des Maschinellen Lernens hervor. Eine der Disziplinen ist das Natural Language Processing, kurz *NLP*. Unter diesem Begriff verbergen sich noch viele weitere weitverbreitete Disziplinen, unter Anderem sentiment analysis, machine translation, voice recognition und auch text generation. Da die Textgeneration auch *Sprachmodellierung* genannt, ein Kernelement einiger NLP-Disziplinen ist, gibt es viele Vorgängerversionen. Vor einigen Jahren wurden hauptsächlich die beiden Ansätze des Regelbasierten-Sytems und des Templatebasierten-Systems (Figure 1.1) verwendet, wohingegen *State-of-the-Art* die Neuronalen Ende-zu-Ende Systemen sind. Diese neuen Systeme bieten wesentlich mehr Flexibilität und skalieren weit bessere Ergebnisse mit wegen benötigten Daten, da die Komplexität und somit die benötigte Rechenleistung drastisch gestiegen ist. Aus dieser Tatsache heraus ergibt allerdings ein Komplexitätsproblem, da es sehr schwer wird die Entscheidungen des Neuronalen Netzes nachzuvollziehen. Das Neuronale Netz ist weitestgehend immer noch eine *Black Box*, obwohl es erstaunlich gute Ergebnisse liefert, vor allem im NLP. Nichtsdestotrotz sind Neuronale Netz Modelle um Text zu verarbeiten schlecht zu verstehen, deswegen müssen heutzutage immer noch Kompromisse zwischen den Regelbasierten System geschlossen werden und Hybride Systeme verwendet werden.

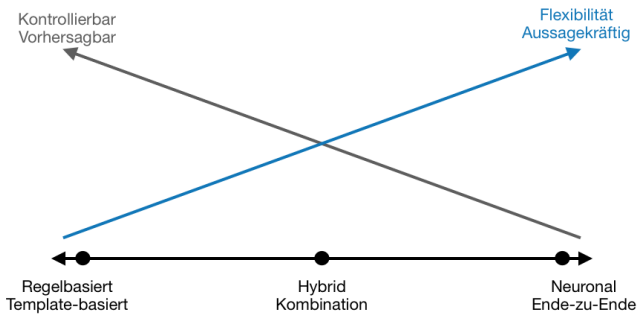

Figure 1.1.: Regelbasiertes- vs. Neuronales Text generations System

Die natürliche Text Generation, auch *NTG* genannt, hat wiederum viele nützliche und interessante Einsatzgebiete. Darunter zählen

- Spracherfassung und Umwandlung in Text
- Konversationssysteme z.B. Chatbots
- Textzusammenfassung

Um Sprachmodelle zu trainieren muss ihnen die Wahrscheinlich von auftretenden Wörtern in Abhängigkeit der vorangegangen Wörter beigebracht werden. Es gibt mehrere Ansätze um dieses Ziel zu erreichen. Sprachmodelle können auf Ebene der Wörter, ganzer Sätze oder sogar ganze Paragraphen trainiert werden. Die Granularität in welcher das Training stattfindet wird als *n-grams* bezeichnet, wovon n die Anzahl der vorangegangen Wörter repräsentiert.

1.1. Fallbeispiel eines aktuellen NLP-Systems

1.1.1. Fallbeispiel

Image-to-Text | Captionbot Microsoft

1.1.2. Nützliche Einsatzgebiete von NLP-Systemen

IoT, Grammerly

1.1.3. Nützliche Einsatzgebiete von NTG-Systemen

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!

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

Listing 3.1: This is an example of inline listing

You can also include listings from a file directly:

```
1 x = 1
2 if x == 1:
3     # indented four spaces
4     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

Wie ist mein Programm Strukturiert

3.2.2. Neuronales Netz

LSTM

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