Prof. Dr. Munir Georges

Technische Hochschule Ingolstadt
Fakultät Informatik

# Sprach- und Textverstehen

Wintersemester 2020/2021 – Probeklausur

**HINWEISE**

- Falls nicht anders angegeben, muss Ihre Darbietung den Lösungsweg und die eigenständige Bearbeitung erkennen lassen.
- Die Probeklausur besteht aus insgesamt 13 Seiten. Bitte überprüfen Sie die Vollständigkeit.
- Bitte beschriften Sie das Deckblatt mit Ihrem Sitzplatz und Ihrer Matrikelnummer. Weiter beschriften Sie bitte jedes Blatt mit Ihrer Matrikelnummer.
- Verwenden Sie die Aufgabenblätter um Ihre Lösung einzutragen. Sie können auch die Rückseiten für Ihre Lösung verwenden. Machen Sie dies aber bitte eindeutig kenntlich.
- Unterschleif führt zum Nichtbestehen der Prüfung. Unterlassen Sie bitte jeglichen Versuch!
- Es sind keine Hilfsmittel zugelassen.
- Schreiben Sie leserlich und nicht mit Bleistift. Verwenden Sie keine rote oder grüne Farbe.

**Studiengang:** Künstliche Intelligenz (B. Sc.)

**Datum:** Wintersemester 2020/2021

**Bearbeitungszeit:** 90 Minuten

**Raumnummer:** _____

**Platznummer:** _____

**Matrikelnummer:** _____

| Frage | 1 | 2 | 3 | 4 | 5 | 6 | Summe | Note |
|---|---|---|---|---|---|---|---|---|
| Maximum | ~6 | ~11 | ~7 | ~10 | ~12 | ~12 | 58 | |
| Erreicht | | | | | | | | |

**AUFGABE 1 Allgemeine Fragen** (~*6 Punkte*)

(a) Nennen Sie ein Beispiel für Part-Of-Speech Tagging?

Eingabe: Das ist eine Klausur -> Word sequence
Ausgabe: Art verb det Nom

(b) Was ist ein Stemmer?

-> Wort zurückgeführt auf Wortstam
↳ wird verkürzt => Wörter werden vergleichbar
-> Anzahl seltener Ereignisse wird verringert
z.B.: laufen -> lauf

(c) Wie kann der Softmax im Training effizient berechnet werden?

- hierarchische Softmax überführt das Problem in eine Baumstruktur & kann so effizienter berechnet werden
- Negative Sampling

(d) Erklären Sie was eine "Satzsegmentierung" ist und wie eine solche umgesetzt werden könnte?

Text wird in Sätze unterteilt
Satzzerlegung mit Entscheidungsbaum
Softmax v. Satz muss geg.: Umschung
z.B.: "Alles zw. Punkten"

(e) Was sagt die „Levenshteindistanz“ aus? Nennen Sie Beispiele.

min Anzahl an Charakter, die verändert werden muss um 1 Wort in 1 anderes zu verändern

HALO : Einfügen des Charakters => LD = 1
HALLO : zu viel ↳ HALLO
HOLLO : Verwechslung

(f) Zwei Sprachmodelle werden evaluiert. Auf einem Test Text erreicht ein Sprachmodell eine höhere Perplexität als das andere. Welches der beiden Sprachmodelle würden Sie bevorzugen?

niedrigere Perplexity ist besser

**AUFGABE 2 Textvorverarbeitung** (~*11 Punkte*)

(a) Markieren Sie im folgenden Text die zu normalisierenden Textbausteine so, dass sich mit dem Text ein Sprachmodell für eine automatische Spracherkennung schätzen lässt. Benennen Sie die einzelnen Schritte.

> 3 Wikipedia was launched on January 15, 2001, by Jimmy Wales and Larry Sanger. Sanger coined its name[5][6] as a portmanteau of "wiki" and "encyclopedia".<br> Overall, Wikipedia comprises more than 55 million articles,[7] attracting 1.7 billion unique visitors per month.<br>

2

- Rauschen entfernen: Zahlen, '[', ']'
- Satzzeichen '"', '.', ',' 1.7 -> 1 Punkt 7
- HTML-Tags <br>

(b) Beschreiben Sie eine gängige Methode um einen chinesischen Satz in Wörter zu segmentieren.

↳ k1 offensichtliche Worttrennung:

Max-Match-Segmentierung:

E: Vokabular & Text

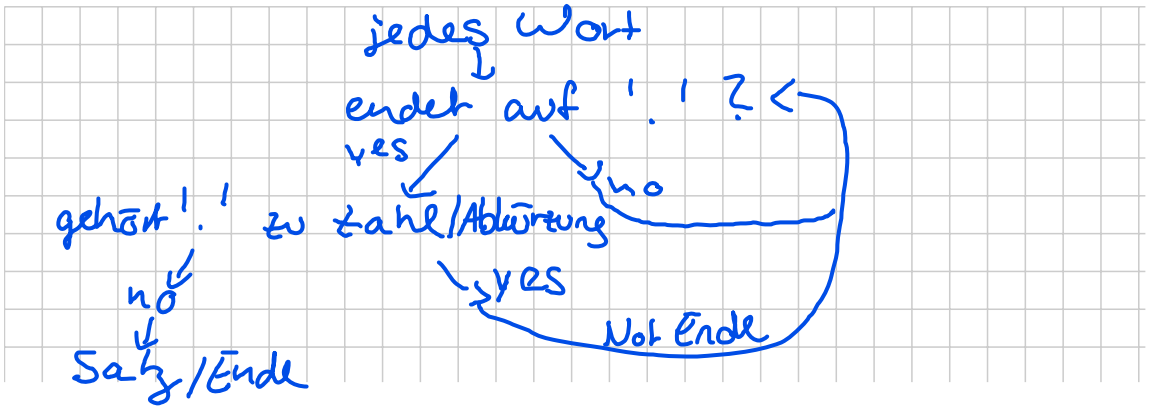
A: Text mit Segmentierung

Iteriere über Satz: Suche nach längster Zeichenkette im Vokabular

(c) Was macht eine Lemmatisierung? Nenne Beispiele.

verwandelt jedes Wort mit komplexen regelbasierten Systemen in Grundform, wie es in einem Wörterbuch vorzufinden ist z.B.: is -> be

(d) Entwerfen Sie einen Entscheidungsbaum (Decision Tree) der den oben angegebenen Text in Sätze segmentiert.




Satz: alles zw. '.' & ';'

**AUFGABE 3 Sprachmodell** (~*7 Punkte*)
Der folgende Text soll zum Trainieren verwendet werden:

| | |
|---|---|
| **Train** | THIS IS A TEXT<br>MARIE IS A RESEARCHER<br>SHE IS WORKING ON A STATISTICAL LANGUAGE MODEL |
| **Test A** | MARIE IS WORKING ON A TEXT |
| **Test B** | BEN IS A RESEARCHER |

(a) Wieviele Wörter hat ein auf "Train" geschätztes Vokabular?

|vokabular| = 12

vokabular = {THIS, IS, A, TEXT, MARIE, RE ..}

(b) Eine n-gram Wahrscheinlichkeit ist definiert als $P(w_i|w_{n-i-1}..w_{i-1})$. Berechnen Sie die Wahrscheinlichkeit für den Text "Test A". Verwenden Sie hierzu ein 1-Gram Sprachmodell welches auf dem Text "Train" geschätzt wurde:

1-gram: Marie $\frac{1}{16}$, is $\frac{3}{16}$, working $\frac{1}{16}$, on $\frac{1}{16}$, a $\frac{3}{16}$, text $\frac{1}{16}$

2-gram: {Marie|•} $\frac{1}{3}$, {is|Marie} $\frac{1}{1}$, {working|is} $\frac{1}{3}$, {on|working} $\frac{1}{1}$, {a|on} $\frac{1}{1}$, {text|a} $\frac{1}{3}$, {•|text} $\frac{1}{1}$

$= \frac{1}{27}$

(c) Wie hoch ist die Perplexität für den "Test B", wenn ein entsprechendes Sprachmodell eine Wahrscheinlichkeit von "0.003" berechnet?

$PP_B = 2^{-\frac{1}{n}\log_2 0.003}$

Anzahl Wörter = 4

(d) Welche Schwierigkeit ergibt sich, wenn ein auf "Train" geschätztes Sprachmodell auf "Test B" angewandt werden soll?

'Ben' kommt in "train" nicht vor, weshalb dafür k1 Wahrscheinlichkeit berechnet werden kann => P(Ben) = 0

Problembehebung durch Glättung

(e) Angenommen es gibt 2 Sprachmodelle. Eines mit einer Perplexität von 110 und eines mit 320. Welches würden Sie bevorzugen, und warum?

Das mit der PP von 110, da je kleiner Perplexität, desto besser passen train Daten auf Text

**AUFGABE 4 Formale Sprachen** (~*10 Punkte*)

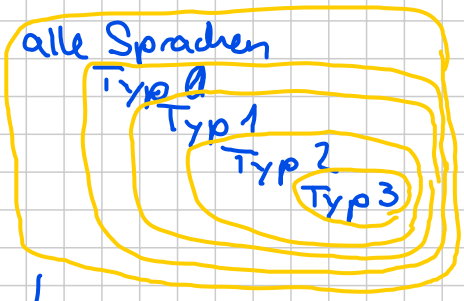
(a) Geben Sie eine kontextfreie Sprache an.

(z.B.: Programmiersprache: R, Lisp)

z.B.: $2^n 5^n$ $n \geq 0$

2,5 Symbole

(b) Welche Sprachen sind in der Chromsky Hierarchie definiert und wie hängen diese zusammen?




Typ 3: regul. Sprache

Typ 2: kontextfreie S.

" 1: kontextsensitive S.

" 0: unbeschränkte Grammatiken

=> Sprachen bauen wie in Bild aufeinander auf

(c) Welche Chomsky-Hierarchie-Sprache kann mit einem endlichen Automaten dargestellt werden? Nennen Sie ein Beispiel.

reguläre Sprache: z.B.: deutsche Sprache

$3^T p^X$ $T, X \in \mathbb{N}$ 3,p Symbole } $\rightarrow$ (3) — (p) Übergang

Startzustand Endzustand

(d) Um welche Chomsky Sprache handelt es sich bei einem Python-Quellcode? Begründen Sie Ihre Antwort.

kontextfrei, da Klammern berücksichtigt werden müssen: $(^n)^n$ $n \in \mathbb{N}$ } kontextfrei

(, ) Symbole

(e) Zeichnen Sie einen Automaten der folgende Sprache über das Vokabular "1,2" Akzeptiert: "$1^n 2^n$" mit $n = 2$?

(Keller-Automat / endlicher Automat)

$\circ \xrightarrow{1} \circ \xrightarrow{1} \circ \xrightarrow{2} \circ \xrightarrow{2} \circ$

? unterschied

**AUFGABE 5 Sprachsynthese, Text-To-Speech (TTS)** (~*12 Punkte*)

(a) Nennen Sie 3 gängige Möglichkeiten ein TTS System zu realisieren. Beschreiben Sie diese kurz.

1) konkatenativ

2) parametrisch

3)

(b) Nennen Sie 3 Qualitätskriterien für ein TTS System? (Beispiel: "Nutzer Vorliebe/Preference für eine bestimmte Stimme/Aussprache oder Wortwahl")

- Audio Qualität
- Verständlichkeit
- Natürlichkeit

(c) Welche Schritte werden in einer Text Analyse in einem TTS System unternommen? Nennen Sie Beispiele.

aus 12€ wird 'zwölf euro'

13:40 → 'Dreizehn Uhr vierzig'

z.B. → 'zum Beispiel'

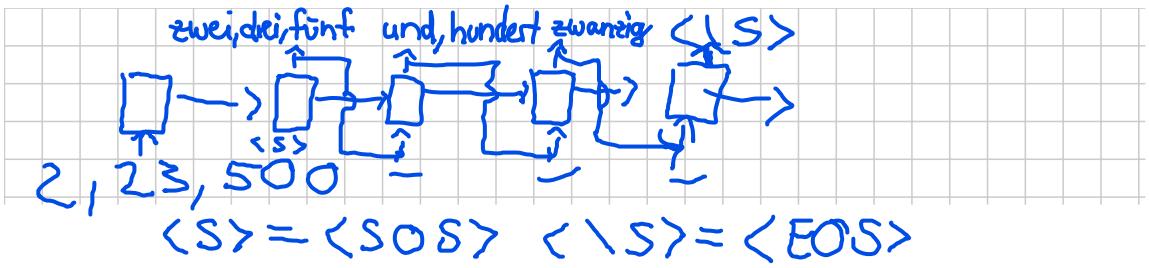
**AUFGABE 6 Rekurrente Neuronale Netze** (~*12 Punkte*)

(a) Entwerfen Sie ein rekurrentes neuronales Netz (RNN), das nach dem Trainieren für folgende Eingabe die Ausgabe berechnen könnte.

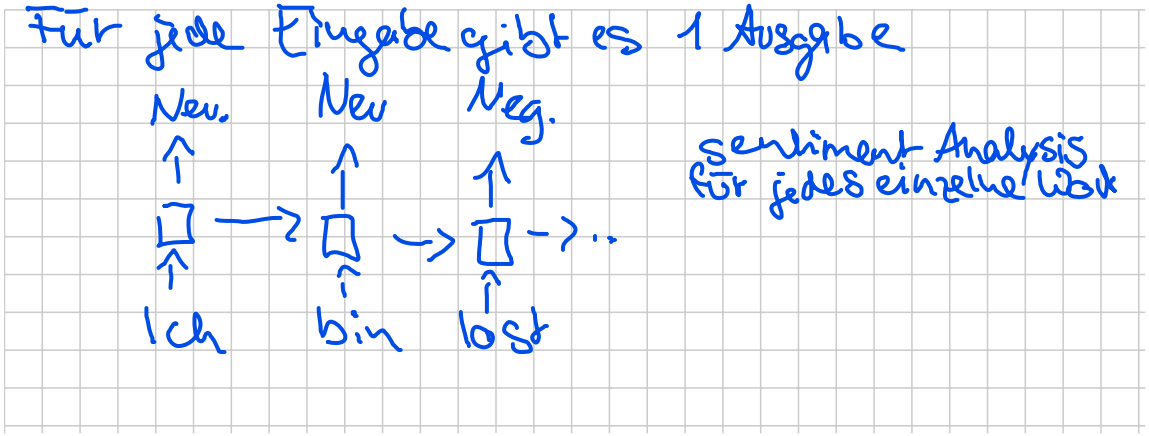
**Test 1:** **Eingabe:** 2
**Ausgabe:** zwei

**Test 2:** **Eingabe:** 23
**Ausgabe:** drei und zwanzig

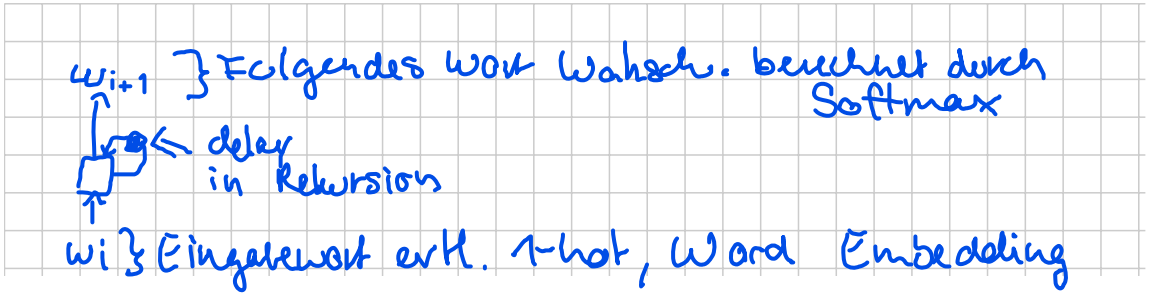
**Test 3:** **Eingabe:** 500
**Ausgabe:** fünf hundert




(b) Was wird unter einem rekurrentem Neuronalen Netz ohne Verzögerung (RNN with no delay) verstanden? Nennen Sie eine Anwendung die ein solches Netz verwendet.




(c) Wie kann ein rekurrentes neuronales Netz (RNN) zur Sprachmodellierung verwendet werden?




(d) Was ist der Unterschied zwischen "softmax" und "argmax"?