Universität Regensburg

Prof. Dr. U. Kruschwitz 22.2.2022

# Exam
# "Natural Language Engineering 1"
# WS 2021/22

You have 120 minutes to work on this take-home-exam (open-book). You should answer all questions. The total number of marks is 100. Please submit a PDF document containing your answers using the GRIPS module „Einführung in die Informationslinguistik 1 / Natural Language Engineering 1 (WS 2021/22)“ no later than 5:00pm today. Make sure your submission includes your **name and registration number**.

# *Prüfung*
# *"Einführung in die Informationslinguistik 1"*
# *WS 2021/22*

*Die Prüfung wird als Take-Home-Prüfung (‚Remote Open-Book-Prüfung‘) organisiert. Sie haben 120 Minuten Zeit. Beantworten Sie bitte alle Fragen. In der Klausur sind insgesamt 100 Punkte zu erreichen. Bitte geben Sie über den GRIPS-Kurs „Einführung in die Informationslinguistik 1 / Natural Language Engineering 1 (WS 2021/22)“ Ihre Lösung als PDF-Dokument ab, und zwar heute bis spätestens 17:00 Uhr. Bitte geben Sie in Ihrer Arbeit* ***Namen und Matrikelnummer*** *mit an.*

**Question 1 / Aufgabe 1:** (40 marks / 40 Punkte)

**Basics /** ***Grundlagen*****.**

**Question 1.1 / Aufgabe 1.1:** (10 marks / 10 Punkte)

A major problem in Natural Language Engineering (NLE) is ambiguity. Illustrate this problem using two sample sentences, one demonstrating structural ambiguity and one demonstrating semantic ambiguity.

*Mehrdeutigkeit ist ein wesentliches Problem der automatischen Sprachverarbeitung. Belegen Sie diese Aussage mit zwei Beispielsätzen, jeweils einem zur Veranschaulichung von struktureller und semantischer Mehrdeutigkeit.*

**Question 1.2 / Aufgabe 1.2:** (10 marks / 10 Punkte)

This is a simple regular expression (RE) using Perl syntax:

```
/^(\(?\+?[0-1]*\)?)?[0-1\- \(\)]*$/
```

Write down one string that matches this RE and one that does not. Justify your examples.

*Oben sehen Sie einen regulären Ausdruck in Perl-Notation. Geben Sie je ein Beispiel für Texteingaben, die von diesem Ausdruck erkannt bzw. nicht erkannt werden und erläutern Sie kurz Ihre Antwort.*

**Question 1.3 / Aufgabe 1.3:** (10 marks / 10 Punkte)

Briefly discuss the problems a tokenizer might encounter when processing texts which contain apostrophes.

*Erklären Sie kurz, welche Probleme ein 'Tokenizer' bei Texten haben kann, die Apostrophe enthalten.*

**Question 1.4 / Aufgabe 1.4:** (10 marks / 10 Punkte)

Provide three reasons as to why dense word embeddings might have become so popular.

*Benennen Sie drei Gründe dafür, warum 'dense word embeddings' so populär geworden sind.*

**Question 2 / Aufgabe 2:** (30 marks / 30 Punkte)

**Grammars and Parsing /** ***Grammatiken und Syntaxanalyse*****.**

**Question 2.1 / Aufgabe 2.1:** (10 marks / 10 Punkte)

Digital assistants have become ubiquitous in recent years. Imagine you want to develop a grammar that defines the commands your own personal assistant should be able to accept (and eventually act upon). Develop a simple context-free grammar (CFG) that accepts (at least) these inputs:

*"book a dentist's appointment for tomorrow"*

*"stop the alarm"*

*"restart"*

Ensure that ungrammatical inputs such as the following should be rejected by your grammar:

** "book"*

** "book for tomorrow"*

** "stop alarm"*

*Digitale Assistenten sind seit einiger Zeit allgegenwärtig. Stellen Sie sich vor, Sie wollen eine Grammatik entwerfen, die Kommandos für Ihren eigenen Assistenten erkennt (um diese später entsprechend auszuführen). Entwickeln Sie eine einfache kontextfreie Grammatik, die die oben genannten korrekten Eingaben akzeptiert (die ersten drei Bespiele), inkorrekte Eingaben (wie die anderen drei) jedoch nicht erkennt.*

**Question 2.2 / Aufgabe 2.2:** (10 marks / 10 Punkte)

Briefly discuss whether a right-linear grammar can be developed that defines the same language that your grammar accepts.

*Diskutieren Sie kurz, ob die von Ihrer Grammatik definierte Sprache auch mit Hilfe einer rechtslinearen Grammatik beschrieben werden kann.*

**Question 2.3 / Aufgabe 2.3:** (10 marks / 10 Punkte)

Outline the steps that a top-down chart parser performs when applying your grammar to the following input:

*"restart the alarm"*

*Skizzieren Sie die Schritte eines 'Top-Down-Chart-Parsers', der auf Ihre Grammatik zugreift bei der Texteingabe "restart the alarm".*

**Question 3 / Aufgabe 3:** (30 marks / 30 Punkte)

**Applications /** ***Anwendungen*****.**

**Question 3.1 / Aufgabe 3.1:** (10 marks / 10 Punkte)

Imagine you want to build a word processor that is able to predict the next word you will type, based on what you have typed so far. Such a system could be trained on the user's specific writing style. Outline how you could train such a system on n-gram probabilities using maximum likelihood estimation. Use the mini-corpus below (starting with "*Welcome to ...*") as an example to illustrate what the system would suggest to follow after the last two tokens ("*WE WILL*") using different types of n-grams.

*Stellen Sie sich vor, Sie sollen eine Anwendung programmieren, die bei einer Texteingabe in der Lage ist, auf Basis der bereits eingebenen Wörter das nächste Wort vorherzusagen. Man kann sich vorstellen, dass dieses System auf den jeweiligen Schreibstil eines einzelnen Autors trainiert wird. Skizzieren Sie, wie Sie unter Anwendung von N-Gram-Wahrscheinlichkeiten und Maximum-Likelihood-Schätzungen solch ein System trainieren würden. Zeigen Sie am folgenden Minitext, welches Wort so ein System am Ende des Textes (nach "WE WILL") vorhersagen würde für unterschiedliche Typen von N-Grams.*

"Welcome to Women in Data Science (WiDS) Regensburg! The goal of this conference is to showcase the exciting work that is conducted by women in the field of data science. Accomplished specialists from academia and industry will give talks on relevant topics. We are sure that the event will be a great opportunity to exchange ideas and inspire each other!

This year's edition of WiDS Regensburg will be even more fun than the last iteration, because we can actually meet in person! We are planning to hold the event at the wonderful Degginger in the historic city center of Regensburg. Be prepared for a great location, engaging discussion and tasty food! You were happy with the virtual setting? Don't worry, we will make this a hybrid event and stream most of the talks, so that people from all over the world can join us!

Relevant information about our program, the invited speakers, the organizing team and the registration procedure will appear on this website over the course of the next months. Stay tuned!

NOTE: WE WILL"

**Question 3.2 / Aufgabe 3.2:** (10 marks / 10 Punkte)

What is the motivation for *text pre-processing* in statistical language processing? Use the text in Question 3.1 to illustrate your answer.

*Was ist der Grund für die Anwendung des 'Pre-Processing'-Schritts in statistischer Sprachverarbeitung? Beziehen Sie sich in Ihrer Antwort auf den Beispieltext in Aufgabe 3.1.*

**Question 3.3 / Aufgabe 3.3:** (10 marks / 10 Punkte)

Outline how you might go about intrinsically and extrinsically evaluating the word processor described in Question 3.1.

*Skizzieren Sie, wie Sie das in Aufgabe 3.1 beschriebene System intrinsisch und extrinsisch evaluieren könnten.*