Protokoll zur Sitzung am 22.5.17 (Kolloquium)

Im Laufe dieser Sitzung wurden drei Themen in Form eines längeren Vortrages inklusive einer kleinen Präsentation auf dem Beamer von den Studenten vorgestellt.

Faridis Alberteris „Optimierung der linguistischen Suche beim XML-annorierten Nachlass von L. Wittgenstein“ Themensteller - Herr Dr. Hadersbeck

Am Anfang des Vortrages wurde kurz erklärt, was hier unter dem *Nachlass* von L. Wittgenstein verstanden wird: Das sind 5000 Seiten der nicht veröffentlichen Werke, die mithilfe von XML annoriert worden sind.

Das Ziel dieser Bachelorarbeit ist die Verbesserung der linguistischen Suche mithilfe des WittFind-Systems in diesem Nachlass. Dabei wird der Schwerpunkt auf Personennamenerkennung gelegt.

Zuerst wurde erklärt, wie genau die zu bearbeitenden XML-Dateien aussehen. Außerdem wurden kurz 3 verschiedene Typen dieser Dateien erwähnt (normale, normalisierte und diplomatische Dateien). Der Hauptunterschied zwischen den Typen liegt darin, wie die von Wittgenstein geschriebenen Texte in XML kodiert worden sind.

Es wird der *TreeTagger* (geschrieben von Hr. Dr. Schmid) auf normalisierten Dateien verwendet. Anhand eines Beispiels (*Tolstois*-Token) wurde gezeigt, wie das aktuelle System eine falsche Entscheidung trifft: der Eigennname wird nicht erkannt.

Um das zu verhindern, wurde eine neue Kategorie für *Personennamen* eingeführt. Danach wurde der Token *Tolstois* aber immer noch falsch getaggt. Also wird diese neue Kategorie gar nicht benutzt. Um sicherzustellen, dass diese Kategorie benutzt werden *muss*, um die Ergebnisse der Suche zu verbessern, wurde nach Lokalisierung aller Beispielen, wo der Tagger sich falsch verhält, ein *eTree-Parser* (Python) benutzt. Die Genauigkeit der Suche wurde dabei erhöht (168 vs. 833 Treffer).

Abschließend wurde darauf hingewiesen, dass die Arbeitsweise des verbesserten Systems im Moment noch nicht ganz klar ist: Es ist nämlich unklar, wie mit einem zusätzlichen Filter (dieser neuen Kategorie) mehr richtige Treffer gefunden werden, als ohne.

Iuliia Khobotova „Comparing representation learning over word-level, character-level and combination of both in NLP tasks“Betreuer: Hr. Dr. Wenpeng

Der zweite Vortrag wurde von Fr. Khobotova gehalten. Sie hat ihre Bachelorarbeit präsentiert.

Das Ziel ihrer Arbeit ist es, herauszufinden, wie die verschiedenen Input-Styles für eine CNN (Convolutional Neural Network) die Genauigkeit eines einfachen NLP-Systems beeinflussen können.

Zuerst wurde der Unterschied zwischen einer CNN und einer RNN (Recurrent Neural Network) geklärt. Der liegt nämlich darin, dass das Hidden Layer dieser zwei NN-Typen unterschiedlich aufgebaut wird.

In dieser Bachelorarbeit wird versucht, 2 NLP-Tasks mithilfe von einer CNN zu lösen: sentiment classification und POS-Tagging. Das Dataset besteht aus Rezensionen zu Filmen (ca. 200 000 Sätzen). Dabei werden verschiedene Parameter des CNNs geändert und die Genauigkeit des Systems nach jeder Änderung gemessen. Wie der Titel des Arbeit offenbart, wird zuerst eine traditionelle Word Embedding benutzt, dann eine character-level Repräsentation der Rezensionen und anschließend die Kombination von diesen zwei Möglichkeiten. Für die Bearbeitung dieser Aufgaben wird die Python-Bibloithek *Theano* benutzt.

Die Ergebnisse der oben genannten Messungen sollen in der Arbeit grafisch dargestellt werden.

Alexander Vordermaier „Comparison of transfer methonds for low ressource morphology “Betreuerin - Fr. Dr. Kann

Der letzte Vortrag der Sitzung wurde von Herrn Vordermaier gehalten. Er hat über seine Bachelorarbeit berichtet.

In seiner Arbeit geht es darum, alle mögliche Formen eines Lemmas festzustellen. Es gibt bereits viele effiziente Lösungen dieses Problems (neuronale Encoder-Decoder-Modelle), die brauchen allerdings relativ große Trainingsets und sind aus diesem Grund für die Sprachen, die keine große Korpora anbieten (low-resource Sprachen), nutzlos. In letzter Zeit wurde jedoch gezeigt, dass in diesem Fall zwei andere Ansätze benutzt werden können: man hat die Möglichkeit, das System auf einem anderen Dataset zu tranieren (man soll dafür aber eine ähnliche Sprache verwenden). Oder man benutzt eine *Autoencoding*-Methode in der selben Sprache.

In dieser Bachelorarbeit wird mit dem Mazedonischen (low-resource) und dem Bulgarischen (high-resource) gearbeitet. Es werden zuerst annotierte Daten für das Mazedonische (50 bzw. 200 Wörter) mit jeweils 50-12800 Wörtern aus dem Bulgarischen paarweise vermischt. Das System wird dann auf diesen vermischten Daten traniert.

Als zweite Möglichkeit wird das Autoencoding verwendet. Hier ist die Eingabe gleich der Ausgabe. Die Idee dahinter ist, dass verschiedene Wortformen manchmal gleich aussehen (z. B. Baum (Nominativ) und Baum (Akkusativ) im Deutschen). Anschließend werden diese zwei Methoden kombiniert.

Nachdem der Vortragende gezeigt hatte, wie die Daten konkret aussehen, hat er die Ergebnisse, die graphisch dargestellt wurden, präsentiert. Die Ergebnisse bestätigen die Richtigkeit des oben beschriebenen Ansatzes. Auffallend ist aber, dass die Autoencoding-Methode eine relativ hohe Genauigkeit aufweist (für ein entsprechend großes Trainingsset).

Abschließend hat der Vortragende über seine restlichen Arbeitspläne berich-

tet: er wird sich hauptsächlich mit der Fehleranalyse beschäftigen.