Optimierung der linguistischen Suche beim CML-annotierten Nachlass von Ludwig Wittgenstein

Vortragende: Faridis Alberteris Azar **Betreuer:** Dr. Max Hadersbeck

Die Bachelorarbeit von Faridis entsteht im Ramen des Projekts „Wittengestein in Co-Text“ in Zusammenarbeit mit dem Wittgenstein-Archiv der Universität Bergen in Norwegen. Das Archiv hat eine Sammlung von Texten aus dem Nachlass von Ludwig Wittgenstein, die noch unveröffentlicht sind. Diese teils handschriftlichen Texte hat die Universität Bergen in XML transkribiert und an das Cis geschickt. Damit hat die Gruppe von Dr. Hadersbeck die Finder App WittFind gebaut, in der Geisteswissenschaftler in den Texten nach Begriffen suchen können. Die Aufgabe der Bachelorarbeit ist eine Optimierung der Suche mit Schwerpunkt auf Personennamen und damit eine Erweiterung des Lexikons. Die Daten des Archivs liegen in drei verschiedenen Formaten vor: NORM ist der korrekte Text, DIPLO ist wie Wittgenstein es geschrieben hat und ORG enthält alle Optionen. Die Finderapp verwendet als POS-Tagger den TreeTagger von Dr. Helmut Schmid. Die Information zu den Namen im Text ist zwar im XML enthalten, wird von TreeTagger aber ignoriert. Dies will Faridis in ihrer Arbeit beheben und hat dafür zwei Schritte:

1. Alle möglichen Fehler beim Taggen sammeln
2. Die semantische Scuhe bei WittFind verbessern und Eigennamen finden
    a. Eine neue Kategorie in Wittfind für Personennamen erzeugen
    b. Diese neue Kategorie ins Cis-Lexikon eintragen

Die vorläufigen Resultate sind, dass vor der Änderung in 13 Dateien 168 Personennamen gefunden wurden und nach der Entwicklung eines neuen Systems 833 Personennamen. Weitere Probleme sind noch Transkriptionsfehler und Probleme in der Edition. Außerdem kann das Lexikon weiterentwickelt werden, indem man die Frequenzliste mit etree statt regex erzeugt.

COMPARING REPRESENTATION LEARNING OVER WOR LEVEL: CHARACTER LEVEL and combination of both in NLP tasks buy Iuliia Khobotova supervisor Wenpeng Yin

Comparing Representation Learning over word level, character level and combination of both in NLP tasks

Vortragende: Iuliia Khobotova **Betreuer:** Wenpeng Yin

The thesis investigates how different input styles influence the accuracy of convolutional neural networks in different NLP tasks. The two questions to be answered are what the best set of parameters for the CNN is and how fast the accuracy will be calculated. The different

representations that are tried are the commonly used word embeddings, character level and a combination of both. The CNNs trained with each will be compared in NLP tasks like sentiment classification and POS tagging. There will also be experiments with changes of parameters like the embedding size, the size of the hidden layer and the batch size with the goal to find the best set of parameters for maximum accuracy. The tools used were, for sentiment analysis, the Stanford Sentiment Treebank which contains annotated data from movie reviews and for CNN implementation the Python framework Theano which consists of an input layer, a hidden layer and an output layer. Input and output layers are always the same size, but the size of the hidden layer can vary. The evaluation will be statistical and based on comparison, not accuracy.

Comparison of Transfer Methods for low resource morphology

Vortragender: Alexander Vordermaier **Betreuerin**: Katharina Kann

Das Ziel dieser Bachelorarbeit ist es, bei einer low resource Sprache, für die also zu wenige annotierte Daten vorhanden sind, annotierte Daten aus einer ähnlichen high resource Sprache zu verwenden, um das Problem zu umgehen. Die konkrete Aufgabe um die es geht, ist die Paradigmenkomplettierung von Sprachen, also die Zuordnung eines Lemmas zu seinen flektierten Formen. Das verwendete System ist das MED (morphological encoder decoder) System, dass von Katharina Kann und Hinrich Schütze für die SIGMORPHON shared task 2016 entwickelt wurde. Die Sprachen, mit denen das Vorgehen in dieser Arbeit getestet wird, ist Bulgarisch als high resource Sprache und Mazedonisch als low resource Sprache. Die drei Herangehensweisen an das Problem sind erstens, die sprachübergreifende Paradigmenkomplettierung, also das Ergänzen der Trainigsdaten mit annotierten Daten einer ähnlichen Sprache, Auto Encoding, also das Ergänzen mit unannotierten Daten der gleichen Sprache und Kombinationen aus beiden. Auto Encoding inst ein simples Verfahren, bei dem die Eingabe auch die Ausgabe ist, weil man hofft dass viele Flektionen gleich sind, wobei die Gefahr besteht, Es werden Kombinationen aus verschiedenen Datenpaketen durchprobiert. Die Größen gehen von 50 bis 12800 Wörter und alle Kombinationen werden versucht, um die beste zu finden. Beim Autoencoding werden die annotierten mit unannotierten im selben Größenformat vermischt. Weitere Aufgaben der Arbeit ist die Fehleranalyse, zum Beispiel wird oft die falsche Endung verwendet. Außerdem müssen noch andere Fehlerquellen identifiziert werden und das verwendete Modell verstanden werden.