Joseph Birkner
Mat.Nr. 11393024

Protokolle zum Colloquium (22.05.2017)

Student: Faridis Alberteris, Betreuer: Maximilian Hadersbeck

„Optimierung der linguistischen Suche in WittFind“

Im Rahmen des Projektes „Wittgenstein in Co-Text“ widmet sich das Centrum für Informations- und Sprachverarbeitung (CIS) gemeinsam mit der Uni Bergen der Transkription, Analyse und Veröffentlichung des Nachlasses, daher des Volumens der bisher unveröffentlichten Werke, des Philosophen Ludwig Josef Johann Wittgensteins (1889-1951).

Hierbei entwickelt das CIS die Software WittFind, mit der annotierte Datensätze, die in XML-Form von der Uni Bergen aus den Nachlässen eingelesen werden, durchsucht werden können. Ein wichtiger Anwendungsfall von WittFind, der leider bisher nicht optimal umgesetzt ist, ist die Auffindung von Erwähnungen von Personennamen in den Transkriptionen.

Hierfür hat die Uni Bergen im März 2017 einen neuen XML-Tag, den s.g. <*persName*>-Tag, eingeführt. So konnte Faridis die Personennamensuche in WittFind verbessern, indem sie das Suchmuster für Eigennamen änderte: Anstatt wie bisher nur das Muster `(([ADJA]|[ADJD])|[NN]) & <+EN>` zu verwenden, welches auch jegliche Eigennamen wie zum Beispiel *Venus* findet, konnte sie das Muster um den Diskriminator `& <+persName>` erweitern.

Als erstes Ergebnis konnte sie präsentieren, dass mit Hilfe des neuen Suchmusters in den neuen Daten nun 833 anstatt wie bisher nur 168 Namen gefunden werden. Aktuell ist jedoch noch unklar, wie diese Zahl durch einen zusätzlichen Diskriminator (Filter) größer geworden ist.

Als nächstes gilt es zu untersuchen, welche Arten von False Positives unter den aktuellen Suchergebnissen aus welchen Gründen vorhanden sind. Hierbei soll vor Allem differenziert werden zwischen Transkriptionsfehlern, daher solchen Fehlern im XML, die während der Faksimile-Abschrift entstanden sind, und Editionsfehlern, die bei der fälschlichen Annotierung von Eigennamen mit dem <*persName*>-Tag enstanden sind.

Student: Iulia Khobotova, Betreuer: Wenpeng Ying

„Comparing representation learning over word-level, character-level and combination of both in NLP tasks“

Mindestens seit es Mikolov 2013 gelungen ist, mit Hilfe des Word2Vec Systems semantische Embeddings für Wörter zu generieren, sind Neuronale Netze aus den Anwendungsfällen der Computerlinguistik (Maschinelle Übersetzung, Sentimentklassifikation, Textzusammenfassung, Maschinelles Sprachverständnis, POS-Tagging) nicht mehr wegzudenken.

Eine demotivierende Eigenschaft der Art von extrem großen neuronalen Netzen, wie sie aktuell eingesetzt werden, ist die große Anzahl an Parametern, die mit sehr viel „Feingefühl“ und „Trial & Error“ optimiert werden müssen. Solche Parameter beinhalten zum Beispiel die Architektur des Netzwerkes („Convolutional/Feed-Forward“ oder „Recurrent“), die Art der Eingabe (Zeichen oder Wörter), die Lernrate, die Gradient-Descent Methode (RBM/Adam/Adagrad/Adadelta), das Batch-Verfahren, die Normierung der Gewichte (L1/L2), die Wahl der Dropout-Rate, die Wahl der Aktivierungsfunktion, die Wahl der Fehlerfunktion und die Wahl der Trainingsdaten.

Iulias Ziel ist es, für die Stanford Sentiment Treebank NN-Modelle für möglichst viele Kombinationen dieser Parameter zu trainieren und zu vergleichen. Hierfür nutzt sie das Python-Framework Theano. Die Leistung der Modelle wird auf Basis ihrer Genauigkeit bei der Sentimentklassifikation evaluiert.

Student: Alexander Vordermaier, Betreuerin: Katharina Kann

"Comparison of Transfer Methods for Low Resource Morphology"

Für viele Anwendungen in der Computerlinguistik (Maschinelle Übersetzung, POS Tagging, etc.) ist es von Nutzen, über ein System zu verfügen, welches (Lemma, KNG)-Tupel bijektiv auf die entsprechende flektierte Wortform abbilden kann. Die Mengre aller flektierten Formen eines Wortes nennt man auch Paradigma. Das generelle Ziel ist also die s.g. „Paradigmenkomplettierung".

Ein generelles Problem bei dem Einsatz von Maschinellen Lernsystemen, auch beim Einsatz in der Morphologie, ist das hohe Maß an Trainingsdaten, welches zum Lernen der Zielfunktion notwendig ist. Dies ist vor Allem ein Problem bei Sprachen, für die wenig Trainingsdaten existieren. Solche Sprachen werden auch Low-Resource (LR) Sprachen genannt; das Gegenteil sind also High-Resource (HR) Sprachen. Das Thema an sich baut auf dem Paper „One-Shot Neural Cross-Lingual Transfer for Paradigm Completion" (https://arxiv.org/abs/1704.00052) (Cotterell 2017) auf, welches am 31.3.2017 von Katharina Kann, Ryan Cotterell und Hinrich Schütze veröffentlicht wurde.

Cotterell 2017 nutzen eine HR-Sprache die grammatisch einer LR-Sprache ähnlich ist, um die Genauigkeit der Paradigmenkomplettierung für die LR-Sprache zu verbessern.

Alexander beschäftigt sich mit der Anwendung des Systems von Cotterell 2017 auf die HR-Sprache Bulgarisch und die LR-Sprache Mazedonisch. Hierfür gilt es zu determinieren, welcher HR-LR Split in den Trainingsdaten am hilfreichsten für die Verbesserung der Genauigkeit der Paradigmenkomplettierung für die LR-Sprache ist. Hierfür stehen ihm für Mazedonisch Datensätze mit jeweils 50 und 200 Sätzen, und für Bulgarisch Datensätze mit jeweils 50, 100, 200, 400, 800, 1600, 2400, 4800, 9600 und 19200 Sätzen zur Verfügung.

Erste Ergenbnisse konnte er bereits präsentieren: So erzielte er für den HR-LR Split 4800-200 bereits eine Genauigkeit bei der Paradigmenkomplettierung über 80%. Als nächstes möchte er sich vor Allem der Fehleranalyse widmen.