CIS, LMU München

Studentin: Elena Atanasova

15.05.2017

Protokoll zur Sitzung vom 15.05.2017 – Computerlinguistisches Arbeiten

3. Vortrag: Kristina Smirnov, "Comparison of transfer methods for low-resource morphology"

BA-Betreuer: Katharina Kann

Den letzten Vortrag des heutigen Tages hält Kristina Smirnov. Sie begann mit einem Überblick über das Thema und stellte ihre Ideen für die Gliederung ihrer Bachelorarbeit vor.

Die Motivation der Arbeit ist, dass man in der Computerlinguistik oft das Problem hat passende annotierte Daten zu finden. Die Überlegungen, die sie sich am Anfang gemacht hat waren: Was kann man machen, wenn es nicht genug Sprachdaten gibt? Würde es helfen, Daten einer ähnlichen Sprache hinzuzufügen? Würde es helfen, nicht-annotierten Daten der gleichen Sprache hinzuzufügen?

Das Ziel Ihrer Bachelorarbeit ist es anhand bestimmter Kriterien flektierte Wortformen im Russischen zu finden. Da die Anzahl der annotierten Daten in diesem Gebiet für das Russische sehr gering ist, versucht die Frau Smirnov das gesamte über das Ukrainische und über nicht annotierten Daten im Russischen zu machen. Das Model mit dem sie arbeitet wurde vom Herrn Hinrich Schütze und Frau Katharina Kann entwickelt. Es handelt sich um einen Morphologischen Encoder-Decoder:

The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection.

Die nicht annotierte Daten, die benutzt werden, kommen aus Wikipedia und annotierte aus dem SIGMORPHON 2016 Shared Task on Morphological Reflection.

Es werden verschieden Kombinationen von LR und HR benutzt, wobei LR = 50, 200 und HR: 50, 200, 400, 800, 1600, 3200, 6400 und 12800. Die Struktur der Implementierung wird anhand eines Beispiels an der Tafel gezeigt.

Man gibt die Sprache und das Lemma ein und bestimmt in welcher grammatikalischen Form das Wort vorkommen soll. Das System soll anhand von diesen Informationen die flektierte Form des Lemma ausgeben.

Resultate und Evaluation: Das System wurde bereits angewendet, aber die Ergebnisse hat sie noch nicht, sie folgen in den nächsten Tagen. Am Ende werden sowohl die statistische Evaluierung, als auch linguistische gemacht, um zu sehen, wo genau die Fehler auftretten.