CIS, LMU München

Studentin: Elena Atanasova

22.05.2017

Protokoll zur Sitzung vom 22.05.2017 – Computerlinguistisches Arbeiten

3. Vortrag: Alexander Vordermaier , "Comparison of Transfer Methods for low Ressource Morphology"

BA-Betreuer: M. Sc. Katharina Kann

Den letzten Vortrag des Tages hält Alexander Vordermaier, der von Frau Kann betreut wird. Am Anfang der Presentation stellte er seine Ideen für die Gliederung seiner Bachelorarbeit vor und danach presentierte er die Motivation. In seiner Bachelorarbeit geht es um die Paradigmen Komplettierung von Sprachen, also um die Zuordnung eines Lemmas zu seinen flektierten Formen. Es geht darum unbekannte Wörter einer Sprache, durch bekannte Wörter eine ähnliche Sprache zu finden.

Das Grundprinzip seiner Bachelorarbeit ist es, wenn es nicht genügend Trainingsdaten zur Verfügung stehen, Daten von einer anderen ähnlichen Sprache oder der gleichen Sprache mit nicht annotierten Daten zu benutzen, um das Problem unzugehen.

In seiner Bachelorarbeit verwendet er drei Methoden für die Paradigmen Komplettierung und zwar: sprachübergreifende Paradigmen Komplettierung, Auto Encoding und Kombination aus den beiden ersten Methoden.

Bei der ersten Methode versucht er zu einer „Low Ressource(LR)" Sprache, eine ähnliche „High Ressource(HR)" Sprache zu finden. Die Ähnlichkeit der beiden Sprachen ist besonderes wichtig und man vermmischt die Daten aus beiden Sprachen und trainiert sie zusammen. Die große Hoffnung ist, dass man davon brauchbare Ergebnisse enthält. In seinem Fall die LR Sprache ist die Mazedonische Sprache und die HR Sprache die Bulgarische. Man vermischt annotierte Daten der „LR" Sprache mit annotierten Daten der „HR" Sprache. Diese Daten werden dann vom Modell trainert. Die Paketgröße der LR Sprache ist 50 und 200 und die Paketgröße der HR Sprache ist 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800. Die Daten werden Paarweise vermische uns dafür hat er folgendes Beispiel erwähnt: (LR/HR) (50/400) (200/6400).

Die zweite Methode ist ein simples Verfahren, bei dem die Eingabe auch gleichzeitig die Ausgabe ist. Man hofft also, dass viele Flektionen der Wörter gleich sind. Das große Gefahr ist es, dass dies nicht der Fall ist.

Bei der dritten Methode vermisscht man die annotierte Daten der LR Sprache mit annotierten Daten der HR Sprache und mit nicht annotierten Daten der LR Sprache. Die Paketgröße der annotierte LR ist 50, 200. Die Paketgröße der HR ist 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800 und die Paketgröße der LR ist auch 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800.

Als nächstes hat er die Form der Daten vorgestellt. Um von dem Modell verarbeitet zu verden, müssen die Daten eine gewisse Form besitzen. Ein trainingsbereiter Abschnitt besteht aus folgenden Daten: Test_src, Test_trg, Dev_src, Dev_trg, Train_src, und Train_trg. Src stehet für Source und da stehen die Daten die man trainiert, trg steht für Target und da stehen die Lösungen.

Genauigkeit wird einmal für Paketgröße 50 (annotiert Mazedonisch) gemessen. Hier hat er festgestellt, dass die Ergebnisse nicht gut sind. Die zweite Methode schneidet ziemlich schlecht. Methode 1 läuft etwas besseres und kommt fast bis zu den 50 % und die Kombination verbessert das ganze nur ein bisschen. Das Ziel sind nicht unbedingt gute Ergebnisse zu bekommen, sondern festzustellen ob

diese Methode tatsächlich funktionieren. Die Genauigkeit bei der Parketgröße 200 sieht etwas besser aus. Aus Methode 2 erreicht fast 60% Genauigkeit. Die anderen beiden Methode sind gut und sogar bei der Kombination aus beiden wird die 80% der Genauigkeit erreicht. Mit diesem Ergebniss, meint Vordermaier, kann man schon anfangen, damit zu arbeiten.

Der Student hat festgestellt, dass oft die falsche Endung verwendet wird. Außerdem treten viele Fehler bei dem Auto Encoding auf Grund der Vorgehensweise auf. Für Vordermaier ist es alle noch doppelt schwer, weil er kein Mazedonisch oder Bulgarisch Muttersprachler ist.
Zuletzt wurden weitere Aufgaben aufgelistet, bei denen noch gearbeitet wird: weitere Fehlerquellen identifizieren, Modell und bereits gängige Verfahren beleuchten.