Protokoll 3 von 22.05.17

Ivana Daskalovska

Thema: *Comparation of Transfer Methods for low Ressource Morphology*

Student: *Alexander Vordermaier*

Betreuerin: *Katharina Kann*

Herr Vordermaier hat am Anfang das Thema seiner Arbeit genannt und die Motivation, die da hinter steckt erklärt. Es geht im Allgemeinen um das Vervollständigen von Paradigmen von Sprachen, also um die Zuordnung eines Lemmas zu seinen flektierten Formen. Man unterscheidet zwischen High Ressource Sprachen und Low Ressource Sprachen. Für Low Ressource Sprachen sind nur wenige Ressourcen vorhanden.

Eine mögliche Lösung wäre eine ähnliche Sprache zu verwenden, um das Problem zu umgehen. Herr Vordermaier benutzt in seiner Arbeit das Bulgarische als HR-Sprache und das Mazedonische als LR-Sprache.

Für die Vervollständigung der Paradigmen verwendet er drei Methoden:

1. Sprachübergreifende Paradigmen Komplettierung
2. Auto Encoding
3. Kombination aus beiden Methoden

Die Vorgehensweise bei der sprachübergreifenden Paradigmen Komplettierung ist folgende: man sucht zu einer Low Ressource Sprache eine ähnliche High Ressource Sprache. Dabei ist die Ähnlichkeit von großer Bedeutung. Man vermischt die beiden Sprachen und trainiert sie zusammen. Man hofft, dass dadurch nützliche Ergebnisse erzielt werden. Herr Vordermaier teilt die Daten in Source und Target auf. In Source verwendet er als Input das Lemma, die Sprache aus der das Lemma kommt und die grammatikalische Form in der das Wort vorkommt. Somit bekommt man als Output die flektierte Form des Wortes in Target.

Bei dem Auto Encoding geht man davon aus, dass viele Flexionen der Wörter gleich sind und dass die Eingabe gleich der Ausgabe ist. Bei dieser Methode ist die Gefahr sehr groß, dass dies nicht der Fall ist. Man vermischt annotierte und nicht annotierte Daten der Low Ressource Sprache in unterschiedlichen Paketgrößen:

Low Ressource(annotiert): 50, 200

Low Ressource (nicht annotiert): 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800

Im Source sind hier die annotierten Daten dieselben wie in der Oberen Methode, die nicht annotierten Daten enthalten ein „copy“ davor. Als Target bekommt man hier bei den nicht annotierten Daten dasselbe Wort wie in Source.

Bei der Kombination beider Methoden werden Daten aus alle 3 Bereichen (LR annotiert, HR annotiert und LR nicht annotiert) vermischt. Die Paketgröße beträgt dann:

Low Ressource(annotiert): 50, 200

High Ressource: 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800

Low Ressource (nicht annotiert): 50, 100, 200, 400, 800, 1600, 3200, 6400, 12800

Ergebnisse: Bei einer Paketgröße von 50 wurde bei der ersten Methode eine Accuracy von 0.45 erreicht, bei der zweiten Methode eine Accuracy von 0,1 und bei der dritten Methode eine Accuracy von 0,5.

Bei einer Paketgröße von 200 wurde bei der ersten Methode eine Accuracy von 0.7 erreicht, bei der zweiten Methode eine Accuracy von 0,6 und bei der dritten Methode eine Acuracy von 0,8.

Während der Fehleranalyse konnte Herr Vordermaier feststellen, dass oft die falsche Endung verwendet wurde. Bei dem Auto Encoding wurden wegen der Vorgehensweise viele Fehler festgestellt. Er hat Schwierigkeiten bei der Fehleranalyse, da er weder Mazedonisch noch Bulgarisch spricht.

Herr Vordrmaier hat noch vor, weitere Fehlerquellen zu identifizieren, ein Model daraus zu erstellen und bereits vorhandene Verfahren zu betrachten.