Protokol 3: Alexander Vordermaier „Comparison of Transfer Morphology for low Resource Morphology“. Betreuer: Katharina Kann

Die Motivation für die Arbeit ist Paradigmenkomplettierung für die Sprachen möglich zu machen, für die es wenig Daten zur Verfügung stehen. Hier geht es um die Zuordnung eines Lemmas zu seinen flektierten Formen.

Man muss bei den Sprachen zwischen High Ressource und Low Ressource unterscheiden. Bei den ersten handelt es sich um die Sprachen, die viele Daten zur Verfügung haben und bei den anderen, wo es eher weniger gibt. Als HR gilt in dieser Bachelorarbeit Bulgarisch und als LR Mazedonisch.

Da es für Mazedonische Sprache wenig Daten gibt und sie verwandt mit Bulgarischen ist, die mehr Daten hat, will man versuchen durch Kombination der Daten der beiden Sprachen davon zu profitieren.

Dafür nutzt man drei folgende Methoden: Sprachübergreifende Paradigmen Komplettierung, Auto Encoding und Kombination von den beiden Methoden.
Beim Autoencoding kopiert man den Input und benutzt als Output. Hier hofft man, dass viele Flektionen der Wörter gleich sind.

Das Verfahren sieht folgend aus: Zu einer LR Sprache sucht man eine HR Sprache, wobei die Ähnlichkeit eine große Rolle spielt und kombiniert die Daten so, dass LR weniger ist. Folgende Größen wurden getestet: LR: 50 und 200 Tokens gemischt mit HR 50,100,200,400,800,1600,3200,6400,12800. Man benutzt annotierte Daten für Mazedonisch und unnanottierte für Bulgarisch. Annotierte stammen aus Shared Task 2016 und unannotierte aus dem Opus Corpus.

Die Evaluation wurde anhand von Grafen dargestellt, wobei die Kombinationen von 50 LR im Vergleich zu 200 viel schlechter abschneidet. Bei 200 erreicht die Accuracy 80%, was ein ziemlich guten Ergebnis liefert. Bei der Fehleranalyse wurde festgestellt, dass oft die falsche Endung verwendet wird und bei Auto Encoding viele Fehler auftreten.