Protokoll zur Sitzung vom 22.05.2017

Kurzvorstellung von der Bachelorarbeit von Alexander Vordermaier

BA-Betreuer: M. Sc. K. Kann

Thema: Comparison of Transfer Methods for Low-Resource Morphology

Alexander hat in der heutigen Sitzung seine Bachelorarbeit vorgestellt, deren Thema Vergleich verschiedenen Methoden für Low-Ressource-Morphologie ist. Im Grunde geht es um die Paradigmen Komplettierung von Sprachen, sprich um die Zuordnung eines Lemmas zu seinen flektierten Formen. Um die Aufgaben seiner Arbeit durchzuführen, verwendet der Student drei Methoden: sprachübergreifende Paradigmen Komplettierung, Auto Encoding und Kombination aus den beiden ersten Methoden.

Das Problem bei Low-Ressource-Sprachen ist, dass es nur begrenzt Ressourcen gibt. Für die erste Methode, sprachübergreifende Paradigmen Komplettierung, nimmt man eine ähnliche Sprache zu Hilfe. In dem Fall ist Mazedonisch die Low- und Bulgarisch die High-Ressource-Sprache. Man vermischt die annotierten Daten aus beiden Sprachen und trainiert sie zusammen. Was für ein Modell verwendet er für das Training, hat Alexander nicht erwähnt; hat aber erklärt, wie es funktioniert. Also die Paketgröße von Low-Ressource-Sprache ist 50, 200 und von High-Ressource-Sprache fängt bei 50 an und verdoppelt sich bis sie 12800 erreicht. Für das Training werden die Datenpakete paarweise vermischt. Was die zweite Methode angeht, ist Auto Encoding ein sehr simples Verfahren, bei dem die Eingabe gleichzeitig die Ausgabe ist. Man vermischt also annotierte Daten der Low-Ressource-Sprache mit ihren nicht-annotierten Daten und trainiert sie wieder paarweise. Die Paketgröße von Daten bleibt gleich wie bei der ersten Methode. Auto Encoding bringt dann gute Ergebnisse, wenn viele Flexionen der Wörter gleich sind. Bei der Kombination der beiden Methoden vermischt man annotierte Daten der Low-Ressource-Sprache mit annotierten Daten der High-Ressource-Sprache und nicht-annotierten Daten der Low-Ressource-Sprache, dabei bleibt die Paketgröße wieder gleich wie bei den ersten zwei Methoden. Um von dem Modell verarbeitet zu werden, müssen die Daten einer gewissen Form sein: Test_src, Test_trg für Test Set, analog für Development Set und Training Set, und Vokabulare zu „Source“ (was das Modell bekommt, also Input) und „Target“ (Lösung, oder Output).

Alexander hat seine Ergebnisse in Form eines Liniendiagramms präsentiert. Für beide Paketgrößen von annotierten Daten der Low-Ressource-Sprache 50 und 200 zeichnet sich die dritte Methode am besten, also die Kombination aus sprachübergreifender Paradigmen Komplettierung und Auto Encoding. Bei der Paketgröße von 50 erreicht die Accuracy ca. 50% bei der Kombination-Methode, ca. 45% bei der ersten Methode (sprachübergreifende Paradigmen Komplettierung) und nur ca. 15% beim Auto Encoding. Bei der Paketgröße 200 sind die Ergebnisse im Allgemeinen viel besser: ca. 80% bei der Kombination-Methode, ca. 75% bei der ersten Methode (sprachübergreifende Paradigmen Komplettierung) und ca. 60% bei Auto Encoding. Nach einer Fehleranalyse hat Alexander geschlossen, dass die falsche Endung oft verwendet wird und viele Fehler beim Auto Encoding auf Grund der Vorgehensweise auftreten. Da der Student keine von verwendeten Sprachen spricht, ist es für ihn nicht einfach eine ausreichende Fehleranalyse durchzuführen. Deswegen gehört zu seinen weiteren Aufgaben weitere Fehlerquellen zu identifizieren, sowie bereits gängige Verfahren zu beleuchten.