Klyueva Olga
CIS, LMU München
22.05.1917

Alexander Vordermaier: Comparison of transfer methods for low resources morphology.

Die Bachelorarbeit wird von Katharina Kann betreut.
Das Ziel der Bachelorarbeit ist die Paradigmen Komplementierung von Sprachen und die Zuordnung eines Lemmas zu seinen flektierten Formen. Das System bekommt ein Lemma mit morphologischen Tags und es produziert eine korrekte Form von dem Lemma.
Es gibt Systeme, die gute Resultate erzeugen. Solche Systeme benötigen sehr viele Ressourcen, viele Daten. Es gibt low- ressource und high-ressource Sprachen. Low-Ressource Sprachen sind Sprachen, die wenige Ressourcen haben und wenig Daten. High-ressource Sprache haben viele Ressourcen, viele Daten. Es gibt wenig low- ressource Sprachen.
Um das Ziel der Bachelorarbeit zu erreichen wurde vorgeschlagen, zwei ähnliche Sprachen zu nehmen. Die Sprachen wurden vermischt und zusammen trainiert.
In Rahmen der Bachelorarbeit wurde Makedonisch als low - ressource Sprache und Bulgarisch als high-Ressource Sprache genommen. Die Ähnlichkeit der beiden Sprachen ist sehr wichtig. Unter Ähnlichkeit versteht man beispielweise gleichen Ursprung von den Sprachen, grammatikalische Ähnlichkeiten wie Kasus, Tempus etc.

Es wurden drei Methoden für die Paradigmen Komplettierung erwähnt:

1) Sprachübergreifende Paradigmen Komplettierung
2) Auto Encoding
3) Kombination auf beides

Bei der sprachübergreifenden Paradigmen Komplettierungsmethode wurden die Daten aus beiden Sprachen paarweise miteinander vermischt. Es gibt mehrere Pakete von Daten: 50/50, 200/100 etc. Es wurden annotierte Daten aus der Low-Ressource Sprache und nicht annotierte Daten aus der high Ressource Sprache genommen.
Bei der Auto Encoding Methode bekommt das System einen Input, ein Wort. Es kopiert das Wort und gibt es wieder zurück. Die Idee dahinter ist, dass Wortformen manchmal gleich sind. Es wurde das Beispiel mit dem Wort Baum gegeben: das Baum, den Baum, dem Baum etc. Das Wort Baum bleibt gleich. Unannotierte Makedonische Daten wurden dafür verwendet.
Bei der Kombination Methode wurden beide oben beschriebenen Methoden kombiniert.
Es wurden die Daten präsentiert. Es gibt Development Daten, Trainings Daten, Test Daten die in zwei Teile aufgeteilt wurden: Source Teil und Target Teil. Im Source Teil steht was das Modell übergeben bekommt. Im Target Teil stehen die Lösungen, was man bekommen soll.
Als nächstes wurden Ergebnisse in der Form von linearen Diagrammen präsentiert. Bei den 50 annotierten makedonischen mit der sprachübergreifenden Paradigmen Komplettierung Methode erreicht man 50 Prozent Genauigkeit. Bei der Auto Encoding werden nur 10 Prozent als Ergebnis erreicht. Mit Hilfe der Kombination Methode kann man 80 Prozent erreichen.
Bei der 200 unannotierten makedonischen Daten mit der Auto Encoding Methode werden 60 Prozent Genauigkeit erreicht, bei der Kombination 80 Prozent.
Als nächstes geht es um Fehleranalyse. Oft werden falsche Endungen verwendet. Bei Auto Encoding treten viele Fehler auf. Der Referent kann weder Bulgarisch noch Makedonisch.
Zum Schluss wurde angekündigt, weitere Fehlerquellen zu suchen.