Kolloquium, 22. Mai 2017

Dayyan Smith

Comparison of Transfer Methods for Low Resource Morphology

Vortragender: Alexander Vordermaier

Betreuerin: Katharina Kann

Alex versucht in seiner Bachelorarbeit Paradigmen Komplettierung, die Generierung einer bestimmten flektierten Form aus einem Lemma, für "low resource" Sprachen mit Daten aus einer "high resource" Sprache zu verbessern. In manchen Sprachen gibt es nur wenig Trainingsdaten und vielleicht ist es möglich Daten einer verwandten Sprache zu nehmen um die Trainingsdaten zu vermehren. Alex erforscht dies für Mazedonisch (low resource) und Bulgarisch (high resource). Drei Methoden werden untersucht. In der ersten, der sprachübergreifenden Paradigmen Komplettierung, werden mazedonische und bulgarische Trainingsdaten in verschiedenen Anteilen gemischt. Die zweite Methode, Auto Encoding, ist ein sehr simples Verfahren, bei dem die Eingabe gleichzeitig auch die Ausgabe ist. Was genau hier passiert konnte während des Vortrags nicht geklärt werden. Aber: Es werden annotierte mazedonische Daten mit nicht annotierten bulgarischen Daten gemischt, wieder in unterschiedlichen Zusammensetzungen. Für die dritte Methode, eine Kombination der vorigen zwei, werden annotierte mazedonische Daten mit annotierten bulgarischen und nicht annotierten mazedonischen Daten gemischt. Auch hier werden wieder unterschiedliche Kombinationen getestet. Bei 50 annotierten mazedonischen Datensätzen steigt die Accuracy vorerst bei allen drei Methoden, wobei Auto Encoding nach etwa 1000 high resource sample bei einer Accuracy von 0.15 stagniert. Die anderen beiden Methoden erreichen hier eine Accuracy von etwa 0.5. Bei 200 annotierten mazedonischen Datensätzen erreicht man mit Auto Encoding eine Accuracy von 0.6 und die beiden anderen Methoden erreichen knapp 0.8.