Protokoll zum Referat von Kristina Smirnov zur Arbeit zum Thema „Comparison of transfer methods for low-resource morphology“ am 15.05.17

Die Arbeit beschäftigt sich mit dem Problem, dass man in manchen Sprachen nicht genug anotierte Daten zur Verfügung hat um eine Software gut genug zu trainieren, morphologische Tasks korrekt durchzuführen, zum Beispiel die Generierung von morphologisch korrekten Formen in dieser Sprache. Der Ansatz der Arbeit besteht darin zu testen, ob das Hinzufügen von Sprachdaten einer ähnlichen Sprache, oder das Hinzufügen von nicht anotierten Daten dieser Sprache zu einer Verbesserung führen. Auch eine Kombination dieser beiden Methoden soll getestet werden.

Das Modell, für das Kristina in Ihrer Arbeit die oben genannten Methodiken zu einer Verbesserung testet ist das MED, ein System entwickelt von ihrer Betreuerin Katharina Kann und Hinrich Schütze, das für den SIGMORPHON 2016 Shared Task entwickelt wurde. MED steht für Morphological Encoder-Decoder. Dieses Modell schafft es selbst für eine geringe Menge an vorhandene Daten ein sehr gutes Ergebnis zu erzielen.

Die Methodiken zur Verbesserung des Modelles bei wenig Sprachdaten testet Kristina für Russisch und Ukrainisch. Dabei soll angenommen werden, dass für Russisch nicht genügend anotierte Sprachdaten verfügbar sind. Es sollen drei Tasks durchgeführt werden:
- Anotierte russische Sprachdaten sollen mit anotierten ukrainischen Sprachdaten kombiniert werden
-Anotierte russische Sprachdaten sollen mit nicht anotierten russischen Sprachdaten erweitert werden
-Beide Methoden kombiniert, also Sprachdaten aus allen drei Bereichen

Die Daten, die für die Arbeit verwendet werden sind von dem CoNLL-SIGMORPHON 2016/17 Shared Task und haben das Format:
Lemma – Zielform – Morphosyntaktische Beschreibung
Bei diesem Format kann das Problem der Mehrdeutigkeit auftreten, weshalb es zu Problemen mit Daten in einem solchen Format kommen kann.

Für die Evaluation ihrer Arbeit wird Kristina einerseits statistische Methoden verwenden und andererseits selbst – als russische Muttersprachlerin – die Ergebnisse analysieren und versuchen, ein Muster zu erkennen und daraus resultierend Verbesserungsvorschlägen zu entwickeln