

## **Comparison of Transfer methods for low Resource Morphology**

Der letzte Vortrag des Tages war von Kristina Smirnov. Die Betreuerin von Kristina ist Frau Katharina Kann. Dieser Vortrag war etwas kürzer als, die anderen, da auch nicht mehr viel Zeit vorhanden war. In ihrem Thema ging es darum, das Problem der Ressourcenknappheit in manchen Sprachen zu umgehen und deren Paradigmen zu vervollständigen, zum Beispiel mit der Vermischung einer ähnlichen Sprache mit vielen Ressourcen. Konkret ging es um die Sprachen Russisch und Ukrainisch. Am Anfang kam sie auf den Ausgangspunkt des Themas zu sprechen. Nämlich kam das Thema aus dem Computerlinguisten Event Sigmorphon (2016). Dort gibt es regelmäßig „Shared Tasks“ um Problemstellungen in der Computerlinguistik zu lösen.

Um dies zu erreichen, wendet sie drei Verfahren an. Zum einen werden annotierte Daten aus einer Zielsprache, der Sprache mit wenigen Ressourcen (Ukrainisch), mit annotierten Daten aus einer Quellsprache, der Sprache mit vielen Ressourcen (Russisch), vermischt. Diese Methode nennt sich „Cross Lingual Paradigm Completion“. Die Daten der Zielsprache bestehen aus zwei Paketen mit 50 oder 200 annotierten Samples. Die Daten der Quellsprache bestehen aus 50, 100, 200, ..., 12800 großen Paketen. Die Größe verdoppelt sich also immer. Dann werden die Pakete paarweise vermischt und das Modell mit den Paketpaaren trainiert. Die zweite Methode heißt „Auto Encoding“. Dabei werden wieder zwei annotierte Datenpakete aus der Zielsprache, sprich 50 und 200 große Pakete, mit 50,100,200,...12800 großen unnotierten Paketen, ebenfalls aus der Zielsprache, vermischt. Kristina hat dies jedoch nicht mit ukrainisch getan, sondern mit russisch, da sie selbst russisch spricht.

Die dritte Methode ist einfach eine Kombination aus den beiden vorherigen Methoden.

Konkrete Ergebnisse konnte Kristina uns nicht präsentieren, da sie noch auf die Vollendung des Trainings der Daten warten muss. Zusätzlich ist eine Fehleranalyse noch durchaus angebracht, vor allem, da sie selbst russisch spricht. Einen Teil der Daten hat die aus dem CoNLL-Sigmorphon 2017/2016 Shared Task.