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. 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. Das Thema stammte aus dem Coputerlinguisten Event Sigmorphon (2016). Das ein „Shared Task“, dessen Ziel es ist aktuelle 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 eine Kombination aus den beiden vorherigen Methoden. Die Paketkombination besteht hier ebenfalls aus Paketen der jeweiligen Verfahrensweisen. Die annotierten Pakete der Zielsprache, also die 50 oder 200 Samples großen Pakete, sind wieder dieselben. Ein Beispiel wäre also 50 (annotiert Ukrainisch), 400 (annotiert Russisch), 400 (nicht annotiert Russisch).

Sowohl das Trainings-set, als auch das Development-set und das Test-set müssen jeweils aus zwei Teilen bestehen. Einmal aus dem „Source-Teil“, darin stehen die Daten, die dem Modell übergeben werden. Das Format gibt einmal die Sprache, das übergebene Lemma und alle Tags der Zielform des Lemmas.

Im zweiten Teil, dem „Target-Teil“, steht die Lösung, also die korrekte Form des übergebenen Lemmas.

Konkrete Ergebnisse konnte Kristina uns nicht präsentieren, da sie noch auf die Vollendung des Trainings der Daten warten muss. Zusätzlich ist eine Fehleranalyse noch geplant, vor allem, da sie selbst Russisch spricht und somit effektiver auftretende Fehlerquellen identifizieren kann. Einen Teil der Daten hat die aus dem CoNLL-Sigmorphon 2017/2016 Shared Task.