

Comparison of Transfer Methods for low Ressource Morphology

Das Thema wurde zuerst mit der Herangehensweise und der Motivation gestartet. Im Grunde geht es um die Paradigmen Komplettierung von Sprachen. Hier wurden Bulgarisch als High Ressource und Mazedonisch als Low Ressource verwendet, da diese beide Sprachen Gemeinsamkeiten besitzen und ähnlich sind.

Für Low Ressource Sprachen ist es schwierig im Gegensatz zu High Ressource Sprachen, da nur begrenzt Ressourcen für Low Ressource Sprachen vorhanden sind.

Es werden drei Methoden für die Paradigmen komplimentiert. Einmal das Auto Encoding, des weiteren die Sprachübergreifende Komplettierung und zuletzt die Kombination aus den beiden ersten Methoden.

Bei der Sprachübergreifende Paradigmen Komplettierung sucht man zu einer Low Ressource eine ähnliche High Ressource Sprache. Die Ähnlichkeit der Sprachen ist hier sehr wichtig.

Zunächst vermischt man die Daten aus beiden Sprachen und trainiert diese zusammen, mit der Hoffnung das man davon brauchbare Ergebnisse erzielt.

Im Verlauf wird der Ablauf erklärt, dass man eine bestimmte Anzahl von Daten von der Low Ressource und High Ressource entnimmt und diese vermischt.

Im weiterem wird hier das Auto Encoding erklärt. Es ist ein sehr einfaches Verfahren, bei dem die Eingabe auch gleichzeitig die Ausgabe ist. Die Gefahr hier jedoch ist es, dass nicht viele Flektionen der Wörter gleich sind.

Beim Kombinationen der Methoden mischt man annotierte Daten der Low Ressource Sprache mit der annotierten Daten der High Ressource Sprache und mit den nicht annotierten Daten der Low Ressource Sprache.

Um von dem bestimmten Modell verarbeitet zu werden, müssen diese Daten einen gewisse Form haben. Diese besteht aus $LANG = MC \ IN = LEMMA \ OUT = V \ OUT = PROG \ OUT = 3 \ OUT$

LANG steht hier für die Sprache. IN steht für die Art des Wortes. Beispielsweise Lemma. Das OUT steht für ein Tag und am ende steht das Lemma, das übergeben wird.

Zum Ende hin werden die annotierten Daten Ergebnisse mit einem Graphen präsentiert.

Zuletzt wird eine Fehleranalyse gemacht. Wie wir festgestellt haben, treten bei dem Auto Encoding sehr viele Fehler auf, aufgrund der Vorgehensweise. Ein anderer Grund für die Fehleranalyse ist oft die falsche Verwendung der Endung.