Zusammenfassung zum Vortrag über die Arbeit „Comparison of Transfer Methods for low Ressource Methods“, gehalten von Alexander Vordermaier

Zusammenfassung von Korbinian Schmidhuber

Die Motivation der Arbeit besteht darin, dass es bei sogenannten Low Ressource Sprachen, also Sprachen, bei denen nur wenig Daten zur Verfügung stehen, schwer ist Methoden ausführlich genug zu trainieren. In der Arbeit wird dieses Problem für die Paradigmen Komplementierung von Sprachen, also die Zuordnung von Lemmata zu ihren flektierten Formen, getestet. Für sogenannte High Ressource Sprachen ist dies bereits sehr erfolgreich durchführbar.

Die Idee ist, wie auch in der vergangenen Arbeit von Kristina Smirnov, für eine Low Ressource Sprache Daten aus einer ähnlichen Sprache hinzuzufügen, wobei sich die Ähnlichkeit hierbei auf die morphologische Ähnlichkeit bezieht. Im Rahmen dieser Arbeit werden 3 Methoden getestet und miteinander verglichen:
-Sprachübergreifende Paradigmen Komplementierung, das Hinzufügen von Sprachdaten aus einer ähnlichen Sprache
-Auto Encoding, also die Abbildung jedes Lemma auf sich selbst (dies führt z.B. im Deutschen bei dem Lemma „Baum“ zu einem korrekten Ergebnis, wenn man den Akkusativ Singular bilden möchte)
-Kombination aus den beiden Methoden

Herr Schulz bemerkte, dass meht Ressourcen nicht unbedingt zu einem sehr guten Ergebnis führen und verwies dabei auf das Zipf'sche Gesetz. Es wird immer ein großes Problem bleiben, für alle Wörter alle Formen in Sprachressourcen zu finden.

In dieser Arbeit wird Mazedonisch als Low Ressource Sprache untersucht und Bulgarisch zum anreichern der Daten (in der ersten und dritten Methode) verwendet. Es werden dabei Tests mit unterschiedlichen Datensets unterschiedlicher Größe durchgeführt. Beispielsweise gibt es Datenpakete mit 50 und 200 Sätzen. All solche Datenpakete werden paarweise miteinander kombiniert und die Methode daran getestet.

Die Daten sind insgesamt in Test-, Develope- und Trainingdaten aufgeteilt, wobei es jeweils eine Source (Input) und eine Target (Output) Datei gibt. Zudem gibt es Dateien, die das Vokabular der Dateien speichert.
Jede Zeile in einer solchen Datei hat verschiedene Keywörter mit unterschiedlichen Values. Z.b. „LANG“ für die Sprache, „IN“ für die Art des Wortes, „OUT“ für einen Tag und am Ende der Zeile das Lemma.

Bei 50 Sätzen Paketen kam man mit der Auto Encoding Methode zu einer Genauigkeit von unter 20%, mit der Anreicherungs-Methode auf fast 50% und bei der Kombination beider auf etwas mehr als 50%. Bei 200 Sätzen Paketen kam man zu um einiges besseren Ergebnis. Selbst das Auto Encoding erzielte hierbei fast 60%. Trotzdem ist das Ergebnis insgesamt nicht auf einem erwünschten Genauigkeits-Niveau.

Als auftretende Fehler, die durch diese Methode zustande kommen werden genannt, dass oft eine falsche Endung verwendet wird. Bei dem Auto Encoding kommt es, wegen der Methodik an sich, die einfach für jede Form das Lemma auf sich selbst abbildet, zu sehr vielen nicht richtigen Ergebnissen.

Weitere Aufgaben, die noch im Rahmen der Arbeit angegangen werden sollen sind die Suche weiterer Fehlerquellen und das Beleuchten bereits gängiger anderer Verfahren.