PROTOKOLLE ZU COMPUTERLINGUISTISCHES ARBEITEN

Ines Röhrer
Centre for Information and Speech Processing, LMU
I.Roehrer@campus.lmu.de

1 Referat

In Alexander Vordermaiers Bachelorarbeit mit dem Titel ”Comparisons of Transfer Methods for low Resource Morphology”, betreut von Katharina Kann, geht es um die Paradigmen Konplettierung von Sprachen, sprich die Zuordnung aller Flexionsformen zu einem Lemma.
Für diese Paradigmen Komplettierung werden im Grunde drei verschiedene Methoden verwendet:

- sprachübergreifende Methoden
- Auto Encoding
- Kombination der ersten beiden Methoden

Für die erste Methode werden für eine sogenannte ”Low Ressource” Sprache mit wenig Quellen eine sprachlich sehr ähnliche ”High Ressource” Vergleichssprache gesucht, anschließend die annotierten Daten beider zusammengenommen und verarbeitet, wodurch man hoffentlich Brauchbare Ergebnisse bekommt.
Beim Auto Encoding werden annotierte und nicht annotierte Daten einer ”Low Ressource” Sprache vermischt und trainiert. Das große Problem ist, dass viele flektierte Formen nicht in beiden Datensätzen gleich sind.
Bei der Kombination beider Methoden werden annotierte Daten der ”High-” wie auch der ”Low Ressourche” Sprache mit nicht annotierten Daten letzterer vermischt und verarbeitet.
Verwendet werden zwei Arten von Daten, es gibt jeweils eine Datei mit der ”Source”, den Daten, die das Modell bekommt, und eine mit ”Targets”, dem Goldstandart. Die Daten in der Source Datei haben eine bestimmte Form, aufgeteilt in Informationen zu Sprache, Art des Wortes, einen Tag und das Lemma selbst.
Alex hat auch schon eingie Evaluierte Daten und eine Fehleranalyse dazu erstellt. Das Ergebnis seiner bisherigen Fehleranalyse sagt, dass zu oft die falsche Endung verwendet wird, sowie sehr viele allgemeine Fehler beim Auto Encoding auftreten. Allerdings ist so eine Analyse kompliziert, wenn man die verarbeiteten Sprachen nicht spricht.
Als nächstes möchte er seine Fehleranalyse verbessern, sowie andere, funktionierende Verfahren betrachten.