Referat Anastasiya Kryvosheva (Montag, den 19.06.2017)

Using morphologically-rich POS tagging to learn morpholocial generation

Sprachen, die eine reiche Morphologie aufweisen, stellen eine Herausforderung für viele Bereiche der Computerlinguistik dar. Charakteristisch dafür ist eine komplexe Flektierungssystem. In dieser Arbeit wurden anhand Russisch und Polnisch das Ganze behandelt.

Wie wir erkennen, ist SMT (statisical machine translation) in den letzten Jahren populär geworden. SMT besteht aus zwei Schritte. Zu mal aus der Übersetztung von Lemmas, welche morphologisch getagged sind und die Generierung der korrekten morphologischen Form. Somit ist Morpholocial generation ein Subtask von SMT und wird m zweiten Schritt angewendet.

Ziel dieser Arbeit ist es, mit Hilfe eines getaggten Korpus ein Generierungssystem aufzubauen, das für jedes Wort und seine morphologische Eigenschaften ein Form generiert. Anhand eines Beispiels auf der Folie. wird dieses verdeutlicht.

Im Verlauf wird gezeigt, das Lematizer und morpholigscher Tagger auf ein annotierten Korpus trainiert wurden, um einen größeren annotierten Korpus am Ende zu erhalten.
Auch wurde aus dem getaggten Korpus ein Wörterbuch mit Häufigkeiten dargestellt, um zwischen den Formen zu disambiguieren zu können.
Durch das Taggen wird gezeigt, das sehr einige Fehler entstehen können und somit oft mehrere Möglichkeiten dadurch aufkommen.
Um diese Möglichkeiten abzudecken wurden einige Regeln geschrieben.

Zuletzt wurde über die Evaluation und Probleme des Themes gesprochen.
In Polnischen Sprache erhalten wir eine Accuracy von 0,78% ohne irgendwelche Regeln anzuwenden. Mit bestimmten Regeln haben wir eine deutlich höhrere Accuracy von 0,89%. Vergleichen wir diese mit der russischen Sprache erhalten wir ohne Regeln eine Accuracy von 0,49% und nur ein leicht erhöhtes Accuracy von 0,53% mit angwandten Regeln.

Zum Fazit des Themas wurde erwähnt, dass für den polnische Teil die Ergebnisse akzeptabel. Im Vergleich waren die Ergebnisse für Russisch sehr schlecht. Der Grund hierfür zeigt, dass der Tagger falsch getaggt hat. Diese Regeln haben beim polnischen Teil die Accuracy um fast 10% verbessert, jedoch nur um ca 3% im russischen.