Protokoll 3 für die Sitzung vom 19.06.2015

Ivana Daskalovska 11139620

Thema: Using morphologically-rich POS tagging to learn morphological generation
Student: Anastasiya Kryvosheya
Betreuer: Alexander M. Fraser

Morphologiereiche Sprachen stellen für viele Bereiche der Computerlinguistik, wie maschine Übersetzung, eine große Herausforderung dar. Frau Kryvosheya behandelt in ihrer Bachelorarbeit die Generierung morphologischer Formen für das Polnische und Russische mittels POS-Tagging. „Morphological Generation" ist ein Subtask von statistischer maschineller Übersetzung (SMT) und ist in den letzten Jahren sehr populär geworden. „Morphological Generation" besteht aus zwei Schritten: Übersetzung der Lemmata (die morphologisch getaggt sind) und dann die anschließende Generierung der morphologischen Form.

Ziel der Arbeit von Frau Kryvosheya ist es mit Hilfe eines getaggten Korpus ein Generierungssystem aufzubauen, welches für jedes Wort und seinen morphologischen Eigenschaften eine Form generiert. Dazu wurden ein Lemmatisierungsprogramm und ein morphologischer Tagger auf einem Korpus trainiert, um so einen noch größeren Korpus zu bekommen. Aus diesem getaggten Korpus wurde ein Wörterbuch mit Häufigkeiten erstellt, um mehrere Bedeutungen eines Wortes unterscheiden zu können. Da beim Taggingprozess Fehler auftreten können, wurde für die Generierung immer die am häufigsten vorkommende Form verwendet.

Als getaggte Korpora wurden der Russische und Polnisch National Korpus verwendet. Der Russische Nationalkorpus enthält dabei ca. 1,3 Millionen Token, wohingegen der Polnische Nationalkorpus nur knapp 130.000 Token enthält. Als ungetaggte Korpora wurde ein paralleler (Englisch, Russisch) Korpus verwendet, der ca. 23 Millionen Token enthält. Zusätzlich wurde der ungetaggte Europarl Korpus für das Polnische mit rund 7 Millionen Token als Datensatz benutzt.

Als Lemmatisierungsprogramm wurde das am CIS entwickelte Programm namens „Lemming" verwendet, welches die Lemmaform für Wörter ausgeben kann. Als morphologischer Tagger wurde das ebenfalls am CIS entwickelte Programm namens „MarMot" verwendet.

Zur Evaluation wurden die annotierten Korpora in Trainingsset (80 Prozent), Developmentset (10 Prozent) und Testdatenset (10 Prozent) aufgeteilt. Auf dem Developmentset wurde optimiert und die eigenen Regeln wurden darauf erweitert. Die anschließende Evaluierung fand auf dem Testdatenset statt.

Die Accuracy lag bei Polnisch bei 78 Prozent (ohne Regeln) und bei 89 Prozent mit Regeln. Für das Russische wurde eine Accuracy von 49 Prozent (ohne Regeln) und 53 Prozent mit Regeln erreicht.

Frau Kryvosheya konnte folgende häufige Fehler bei der Evaluierung feststellen: die Wortform war nicht korrekt, da das Lemma nicht gefunden werden konnte. Beim Russischen wurden POS Tag und morphologische Form im Wörterbuch nicht gefunden, da „Lemming" und „MarMot" falsch getaggt haben. Oft waren die morphologischen Kategorien beim Taggingprozess „umgedreht", also in falscher Reihenfolge. Auch muss die häufigste Form nicht immer die richtige sein.

Als nächstes ging Frau Kryvosheya auf die generierten Regeln ein. Für Fälle, in denen einen Lemma gefunden wurde, wurden Regeln generiert. Sofern keine Form generiert werden

konnte, wurde POS-Tag und morphologische Informationen ausgegeben und nach Häufigkeit sortiert. Für die häufigsten Fälle wurden eigene Regeln geschrieben. Z.B: S|acc|f|inan|sg, für Nomen, feminina, Singular, Akkusativ, unbelebt: wenn das Lemma auf "-a" endet, wird der letzte Buchstabe mit "y" ersetzt. Dadurch konnten 797 Formen richtig generiert werden.

Zum Schluss stellte Frau Kryvosheya fest, dass für den polnischen Teil die Ergebnisse sehr akzeptal sind. Für das Russische hingegen sind die Ergebnisse sehr schlecht, da beispielsweise der Tagger sehr oft falsche Ergebnisse zurücklieferte. Die eigenen Regeln haben beim polnischen Teil die Accuracy um fast zehn Prozent erhöht, beim russischen aber nur um drei Prozent. Außerdem spielte das Preprocessing eine große Rolle für die Qualität der Generierung morphologischer Formen.