Präsentation von Anastasiya Kryvosheya

In der Präsentation hat Anastasiya über das Thema ihrer Bachelorarbeit “Using morphologically rich POS tagging to learn morphological generation“ gesprochen, die Arbeit wurde von Dr. Alexander M. Fraser betreut. Zuerst hat Anastasiya einen Überblick zur Präsentation gegeben, die Hauptpunkte davon sind die Taskbeschreibung, Verlauf, benutzte Korpora und Externe Tools, sowie Evaluation und Probleme, am Ende hat sie ein Fazit gemacht. Generell, Sprachen, die eine reiche Morphologie aufweisen, stellen eine Herausforderung für viele Bereiche der Computerlinguistik dar, besonderes dafür ist ihre komplexe Flektierungssystem. In ihrer Arbeit hat sie den Vergleich der polnischen und russischen Sprachen gemacht.
Allgemein Statistical Machine Translation ist in den letzten Jahren sehr populär geworden. Sie besteht aus zwei Schritten: erste ist eine Übersetzung von Lemmas, die morphologisch getagged sind und die zweite ist eine Generierung der korrekten (morphologischen) Form. Somit morphological generation ist ein Subtask von SMT und wurde im zweiten Schritt angewendet.
Also das Ziel der Bachelorarbeit ist mit Hilfe eines getaggten Korpus ein Generierungssystem aufzubauen, das für jedes Wort und seine morphologische Eigenschaften eine Form generiert.
Der Verlauf wurde nach folgende Weise aufgebaut: Lemmatizer und morphologischer Tagger wurden auf annotierten Korpus trainiert um einen größeren annotierten Korpus zu bekommen. Dann aus dem getaggten Korpus wurde ein Wörterbuch mit Häufigkeiten erstellen, um zwischen den Formen zu disambiguieren. Da die Fehler beim Taggen vorkommen können, entstehen oft mehrere Möglichkeiten, bei der Generierung wird die häufigste Form genommen und um mehrere Fälle abzudecken wurden Regeln geschrieben.
Für die Arbeit wurde getaggte und Korpora: Russian National Corpus mit 1.291.448 Tokens, Polish National Corpus - 126.182 Tokens, getaggt - Yandex English-Russian Parallel Corpus mit 23.271.021 Tokens und Europarl Parallel Corpus (Polish) mit 7.087.016 Tokens.
Für Tools wurden folgenden externe Tools verwendet: Lemming - es ist ein Lemmatizerprogramm, das im CIS geschrieben wurde, um die Lemmas für die Wörter auszugeben und MarMOT - ein Programm, dass die Wörter mit POS tag und morphologische Eigenschaften taggt Getaggte Korpora wurden auf drei Teilen geteilt: 80% - Trainset. Wird Lemming und MarMOT übergeben, um den ungetaggten Korpus zu taggen, 10% - Developmentset.
Bei der Evaluation sind folgende Probleme aufgetreten: im Polnisch - Accuracy 0.78% ohne Regeln; 0.89% mit Regeln. Im Russisch Accuracy: 0.49% ohne Regeln; 0.53% mit Regeln.
Die Fehleranalyse hat gezeigt daß meistens die Form war nicht korrekt, da das Lemma nicht gefunden wird. POS und morph wurde in Dictionary nicht gefunden (für Russisch sehr viele Fälle, da Lemming und MarMOT falsch getaggt haben: falsche Lemma und beim POS+morph taggen sind die Kategorien oft umgedreht, in falscher Reihenfolge) und die häufigste Form ist oft nicht die richtigste. Was die Regeln betrifft - für die Fälle, wo ein Lemma gefunden wurde, wurden die Regeln generiert. Es wurden die POS+morphs ausgegeben, für die keine Form generiert werden konnte und nach Häufigkeit sortiert. Für die häufigsten Fälle wurden Regeln geschrieben - 797 Form wurden dadurch richtig generiert.

Für polnisches Teil waren Ergebnisse akzeptabel, für Russisch war sehr schlecht, weil den Tagger haben falsch getaggt. Regeln haben beim polnischen Teil die Accuracy um fast 10% verbessert, beim russischen nur um ca. 3%.