Protokoll zur Sitzung vom 19.06.2017

Kurzvorstellung von der Bachelorarbeit von Anastasiya Kryvosheya

BA-Betreuer: Hr. Dr. Alexander M. Fraser

Thema: Using morphologically-rich POS tagging to learn morphological generation

Anastasiya hat in der heutigen Sitzung ihre Bachelorarbeit vorgestellt, in der es um die morphologische Generierung mittels Verwendung vom morphologisch reichen POS Tagging geht. Sie hat mit Russischem und Polnischem gearbeitet, Sprachen mit komplexen Flektiersystemen. Im Russischen gibt es zum Beispiel 6 Kasus, im Polnischen sind 7 Kasus, die sich durch die Endungen des Substantivs und Adjektivs unterscheiden. Sprachen mit so einer reichen Morphologie stellen eine Herausforderung für viele Bereiche der Computerlinguistik dar. Statistical machine translation (SMT) ist in den letzten Jahren weit verbreitet geworden und besteht aus 2 Schritten: (1) Übersetzung von morphologisch getaggten Lemmata und (2) Generierung von korrekten Formen. Morphologische Generierung ist somit ein Subtask von SMT und wurde im zweiten Schritt angewendet.

Ziel der Bachelorarbeit war, mithilfe eines getaggten Korpus ein Generierungssystem aufzubauen, das ein Wort und die nötige morphologische Eigenschaft als Input bekommt und die entsprechende korrekte Form als Output ausgibt, z.B.: bajka subst|pl|f => bajki. Die getaggten Korpora wurden erst einmal auf 3 Sets aufgeteilt: Train- (80%), Delelopment- (10%) und Testsets (10%). Trainset wurde für Lemming und POS Tagging verwendet. Developmentset wurde benutzt, um Regeln zu schreiben. Auf dem Testset wurde letztendlich Accuracymessung dürchgeführt.

Um das Ziel der Bachelorarbeit zu erreichen, hat Anastasiya erst einmal Lemmatizer und morphologischen Tagger auf dem annotierten Korpus trainiert, um einen größeren annotierten Korpus zu kriegen. Es wurden für den Zweck einige externe Tools benutzt. Für Lemming wurde ein Lemmatizer verwendet, das am CIS geschrieben wurde. Um Wörter dann mit POS+morph zu taggen, wurde MarMOT angewendet. Danach wurde aus dem getaggten Korpus eine Dictionary mit Häufigkeiten erstellt nach dem Muster: lemma { pos + morph { surfaceform : freq } }, um zwischen den Wörtern zu desambiguieren. Da beim Taggen Fehler vorkommen können (Klein- / Großschreibung u.a.), sollte man oft mehrere Varianten aufschreiben. Um möglichst mehr Fälle abzudecken, wurden Regeln geschrieben. Für das Experiment hat Anastasiya sowie getaggte (Russian und Polnisch National Corpora), als auch ungetaggte Korpora (Yandex English-Russian Parallel Corpus und Europarl Parallel Corpus für Polnisch) verwendet. Herrn Schulz Meinung nach, ist es nicht sehr schlau, dass Anastasiya nur Korpora verwendet hat, um ein Lexikon zu bauen. Die Sache ist die, dass in einem Korpus sehr oft nur eine Form des Wortes auftaucht und es nicht genug für ein gutes Lexikon ist.

Nach der Evaluation hat Anastasiya folgende Ergebnisse gekriegt: Accuracy für Polnisch 0,78% ohne Regeln / 0,89% mit Regeln und für Russisch 0,49% ohne Regeln / 0,53% mit Regeln. Die Studentin hat 2 Kreisdiagramme vorgestellt, für die beiden Sprachen. Wie man aus dem Kreisdiagramm für Russisch sehen kann, liegen so ziemlich niedrige Ergebnisse daran, dass POS+morph (also die morphologische Eigenschaft, die Anastasiyas Generierungssystem als Input kriegen sollte) nicht gefunden wurde (etwa 35%). Jeweils etwa 15% für die beiden Sprachen hatten restliche nicht-erwähnte Diagrammenanteile: „surface not found“ und „lemma not found“. Anastasiya hat gemeint, dass Lemming und MarMOT für Russisch oft falsch getaggt haben, weil beim POS+morph Tagging die Kategorien oft

eine falsche Reihenfolge hatten. Das andere Problem für die beiden Sprachen war, dass die häufigste Form nicht immer die Richtige war. Für die Fälle, wo ein Lemma gefunden wurde, wurden die Regeln generiert, die die häufigsten Fälle abgedeckt haben. Dadurch wurden 797 Formen richtig geliefert, also wurde die Accuracy für Polnisch um 10% verbessert und für Russisch nur um 3%. Preprocessing spielt eine große Rolle, hat Anastasiya erwähnt.