Protokolle zur Sitzung vom 29.05.2017 - Computerlinguistisches Arbeiten

31.05.2017

Drei Studenten stellten in der Sitzung am 29.05.2017 mit Hilfe einer Präsentation mit dem Beamer ihre Bachelorarbeitsthemen vor.

3. Präsentation:

Referent: Thomas Ebert

Titel: Corpus based Identification of Text Segments

Betreuer: Martin Schmitt

Thomas Ebert stellte als Letzter dieser Sitzung sein Bachelorarbeitsthema vor. In seiner Arbeit geht es um die Entwicklung eines Algortihmus', der einen eingegebenen Satz in seine “besten” Segmente zerlegt. Prinzipiell sind Textsegmente bedeutungstragende Einheiten, also Morpheme, Wörter, Phrasen, Sätze oder auch Topics (Thema eines Abschnittes). Die Textaufbereitung für diese NLP Aufgabe ist wortbasierend. Wort ist zwar nicht eindeutig definiert, jedoch sehr intuitiv, erklärt Thomas.

Wie ist die Vorgangsweise: Man extrahiert aus einem Texkorpus Buchstaben N-Gramme der Länge 1 bis 10. Dabei verwendet Thomas die ersten 10.000 Texte, mit insgesamt über 22 Mio. Zeichen eines Wikipedia-Korpus' mit unannotierten Rohtexten. Die Extraktion der Segmente (N-Gramme) demonstriert Thomas an einem Beispiel an der Tafel. Gegeben ist bspw. der Satz: “The brown fox jumps over...”. Die vordefinierte N-Gramm Länge ist 10, also startet man bei Position 0 und findet die ersten 10 Zeichen, danach geht man eine Position weiter usw.

Anschließend wird eine Frequenzliste für die N-Gramme erstellt, wobei diese mit einem Gütemaß bewertet werden, welches ein statistisches Maß zur Bewertung der N-Gramme ist und mit n*log(freq) berechnet wird, wobei “n” die N-Gramm-Länge ist und “freq” die absolute Häufigkeit des N-Gramms bezeichnet. Anschließend wird ein Satz zum Testen eingegeben. Der Satz soll in N-Gramme zerlegt werden und mittels Gütemaß wird das N-Gramm mit dem höchsten Wert zurückgegeben.

Die hier auftretenden Probleme bestehen darin, dass mit der Größe der Eingabe, die Laufzeit exponentiell ansteigt. Die Lösung hierfür wäre ein heuristischer Ansatz. Ein weiteres Problem ist die Festlegung der optimalen Fenstergröße. Außerdem ist die Berechnung der höchsten Güte nicht mehr garantiert, aber die Segmentierung ist ggf. noch besser als bei symbolischem Ansatz.

Anschließend stellt Thomas seine Evaluierung vor. Er betont die Schwierigkeit der Evaluierung von Text Segmenten: Häufig herrscht Uneinigkeit über die Granularität von Segmenten. Je nach Anwendung können Fehler relevant oder irreleveant sein, z.B. Kann die Korrektheit von Segmentgrenzen bei Information Retrieval vernachlässigt werden. Die Auswirkung auf die Endeanwendung (z.B. IR oder Sentiment Analysis) wird dann als Maß verwendet.

Thomas verwendet außerdem Word2Vec, um Buchstaben N-Gramm-Embeddings zu erhalten. Dann macht

er eine Sentiment Analyse auf Satzebene mit “Movie Review” Daten zur Evaluierung und vergleicht dies mit den Word Embeddings. Thomas schließt seinen Vortrag mit der Vorstellung eines mögliches Modells ab, welches von Cho et al. erstellt wurde. Dabei wird die Sigmoidfunktion auf die Summe der gewichteten Eingabewerte angewandt um ein Ergebnis zu erhalten.