PROTOKOLLE ZU COMPUTERLINGUISTISCHES ARBEITEN

Ines Röhrer
Centre for Information and Speech Processing, LMU
I.Roehrer@campus.lmu.de

1 Referat

Thomas' Bachelorarbeit handelt vom Thema "Corpus based Identification of Text Segments" und wird betreut von Martin Schmitt. Die Motivation für sein Thema ist, zu untersuchen, wie (gut) eine Textsegmentierung, basierend auf anderen Texteinheiten als dem „Wort" funktioniert. Bisher ist die Textufbereitung als NP-Aufgabe meist wortbasierend, dies sieht man besonders deutlich bei der Tokenisierung. Das Wort ist hierbei als Einheit nicht eindeutig definiert, sondern eher intuitiv. Die Tokenisierung ist unter anderem dher sehr fehleranfällig und lokale Anpassungen sind nötig.

Thomas will jetzt die Frage bearbeiten, ob das Wort die beste Einheit ist, um Texte digital zu segmentieren. Dazu etwickelt er einen Algorithmus, welcher einen eingegebenen Satz oder Text in seine besten Segmente zerlegt, und diesen zu Evaluieren.

Es werden N-Gramme der Längen eins bis zehn aus dem Wikipedia Korpus extrahiert, wobei er nur einen Teil des Korpus für seine Analysen verwendet. Für diese N-Gramme werden Frequenzlisten erstellt sowie ein Gütemaß berechnet, wodurch die beste N-Grammlänge für die einzelne Segmentierung bestimmt wird.

Probleme hatte er vor allem mit der Größe der Eingabe und den daraus resultierenden hohen Laufzeiten. Die gefundene ösung ist ein heuristischer Ansatz, bei welchem ein Fenster bestimmt wird, innerhalb dessen segmentiert wird. Leider ist bei diesem Ansatz die Berechnung der höchsten Güte nicht mehr garantiert.

Die Evaluierung der Textsegmente ist relativ schwierig, da es viele Uneinigeiten gibt, und Fehler je nach Anwendung mehr oder weniger relevant sein können.

Als nächstes müssen noch offene Fragen wie die der Evaluierung geklärt werden, sowie andere Möglichkeiten betrachtet werden, N-Gramme zu extrahieren.