Mitschrift: Sohail Malih

Referat Thomas Ebert Betreuer: Msc. Martin Schmitt (Montag, den 29.05.2017)

Corpus based identification of text segments

Textaufbereitung für NLP-Aufgabe sind meist wortbasierend (Tokenisierung)
Wort ist nicht eindeutig definiert aber intuitiv.
Tokenisierung set sehr fehleranfällig, somit sind lokale Anpassungen nötig
Ist das intuitive Konzept 'Wort' die beste Art für einen Computer einen Text zu segmentieren.

Das Ziel hier ist es, die Entwicklung eines Algorithmus der einen eingegebenen Satz oder Test in seine 'besten' Segmente (Buchstabe N—Gramme) zu zerlegen.

Wie gehen wir hier vor?
Wie extrahieren von N-Grammen der Länge 1-10 aus dem Wikipedia Korpus(englisch).
Der Korpus enthält unantorierte Rohtexte.
Erste 10000 Texte (22.650880 Zeichen) des Korpus werden zum extrahieren verwendet.

Nun wird ein Frequenzlist für N-gramme wird erstellt.
N-Gramme werden gemutmaßt bewertet.

Zum testen wird ein Satz eingegeben.
Dieser Satz soll in N-gramme mit höchstem Gütemaß zerlegt werden.
Welche Probleme entstehen hier?
Mit Größe der Eingabe steigt sowohl die Laufzeit exponentiell.
Die Lösung hier wäre ein heuristischer Ansatz.
Man legt die Größe des Fenster fest und berechnet die höchste Güte, welche nicht mehr garantiert, aber jedoch eine Segmentierung ist.

Zur Evaluierung ist es recht schwierig bei Textsegmenten.
Häufig Uneinigkeit über die Granularität von Segmenten.
Je nach Anwendung können Fehler relevant oder irrelevant sein z.b. bei IR kann die Korrektheit von Segmentgrenzen vernachlässigt werden.
Auswirkung auf die Endanwendung (z.b. IR)

Verwendung von wor2vec um Buchstaben n-gramm embeddings zu erhalten.
Sentiment analyse auf Satzeben zur evaluation.
Es werden Vergleich mit word embeddings gemacht.

Zuletzt werden Erkenntnisse und offene Fragen gestellt und erwähnt.
Auch Buchstaben weisen ein zipfsche Verteilung auf.
Häufigste n-gramme welche größer als 3 enthalten sind Funktionswörter und häufigste n-gramme die größer als 8 enthalten...sind ebenfalls Funktionswörter.