Klyueva Olga
CIS, LMU München
29.05.1917

Tomas Ebert: Corpus based identification of text segments.

Zuerst wurde über Motivation der Bachelorarbeit und den Begriff Textsegment gesprochen. Textsegment ist eine bedeutungstragende Einheit. Textsegment kann von einem Morphem über ein Wort, über eine Phrase, zu einem Satz oder zu einem Abschnitt eines Topics gehen.
Generell ist Textaufbereitung für NLP Aufgaben wortbasiert. Der häufigste preprocessing Schritt ist Tokenisierung. Das Problem darin ist, dass das Wort nicht eindeutig definiert ist, meistens intuitiv. Außerdem ist Tokenisierung fehleranfällig. Beispielweise, ob das Wort "won't" als zwei Worte oder als ein Wort betrachtet soll.

Die zentrale Frage der Bachelorarbeit ist, ob ein Wort die beste Art für eine Textsegmentierung ist oder es andere Möglichkeiten gäbe, die auf Tokenisierung verzichten.
Das Ziel der Bachelorarbeit ist einen Algorithmus zu entwickeln, der den angegebenen Satz/Text in bessere Segmente nämlich Buchstaben/N-Gramme zerlegt.

Zuerst wurden N-Gramme von 1 bis 10 aus dem Wikipedia Korpus extrahiert. Wikipedia Korpus enthält unnotierten Rohtext. Es wurde auf erste 10 000 Texte beschränkt (über 22 Millionen Zeichen), weil es genug ist, verschiedene N-Gramme zu extrahieren.
Es wurden Frequenzlisten erstellt für N-Gramme von 1 bis 10 N-Gramme. Die N-Gramme wurden mit einem Gütermaß bewertet. Gütermaß ist gleich Länge des N-Gramm multipliziert mit absoluter Häufigkeit des N-Gramm. Für den Satz soll ein möglichst hohes Gütermaß zur Verfügung stehen.
Zum Testen wird ein Satz eingegeben und der Satz soll in die N-Gramme mit hohem Gütermaß zerlegt werden. Es sind einige Probleme aufgetreten, beispielsweise, dass mit der Größe der Aufgabe die Laufzeit steigt. Um dieses Problem zu lösen, wurde die Größe des Fensters beschränkt. Innerhalb dieses Fensters werden verschiedene Möglichkeiten von N-Grammen durchgegangen und am Ende berechnet man das höchste Gütermaß von einem Teil und anschließend für die nächsten.

Generell ist die Evaluierung von Textsegmenten schwierig, weil schon gesagt wurde, dass ein Segment ein Wort oder eine Phrase sein kann. Es gibt auch keinen Goldstandard für Textsegmente. Außerdem können die Fehler, die bei der Segmentierung auftreten, für bestimmte Aufgaben relevant oder nicht relevant sein. Beispielsweise bei Information Retrieval kann die Korrektheit von Segmenten, beispielsweise innerhalb eines Satzes vernachlässigt werden. Man bekommt immer noch die korrekte Information aus dem Text. Bei der News Boundary Detektion soll man die korrekte Segmentgrenze sehen.
Die Evaluierung wird gemacht, in dem die Auswirkung auf die Anwendung überprüft wird, beispielweise Information Retrieval oder Sentiment Analyse. Das wird als Maß verwendet. Es wird festgestellt, ob das System besser als wortbasierte System ist oder nicht. Es wird Wort2Vec für Buchstabengramme verwendet. Es wurde Movie Review Datasets verwendet: Standard Datensets für Sentiment Analyse. Es wurde mit Word Embedding verglichen.

Zum Schluss werden Erkenntnisse und offene Fragen erwähnt:

- Die Buchstaben N-Gramme weisen eine Zipfische Verteilung auf.
- Die N-Gramme, die mehr als 3 Zeichen haben, sind Funktionswörter (Artikel, Prepositionen)
- Die N-Gramme, die mehr als 18 Zeichen haben, sind Inhaltswörter.

Die andere Frage ist, ob es andere Möglichkeiten gibt, die N-Gramme zu extrahieren. Es ist nicht klar, ob mit der Wahl von 1 bis 10 N-Grammen alle Varianten abgedeckt werden. Diese Wahl wurde ausgewählt, da im Englischen das durchschnittliche Wort eine Länge fünf hat.
Das Ergebnis der Evaluierung ist nicht genug aussagekräftig, wenn man das nur auf einen bestimmten Task angewendet hat. Man sollte auch auf mehrere Tasks anwenden können, um zu sagen, ob dieses System besser als ein wortbasierten System ist.