Protokoll 3 für die Sitzung vom 29.05.2017
Ivana Daskalovska 11139260

Thema: Corpus based Identifikation of Text Segments
Student: Thomas Ebert
Betreuer: Msc. Martin Schmitt

Am Anfang der Präsentation hat Herr Ebert den Titel seiner Bachelorarbeit und eine Übersicht über die Gliederung vorgestellt. Dann hat er erläutert was man alles unter einem Textsegment verstehen kann. Ein Textsegment kann ein Morphem, ein Wort, eine Phase, ein Satz oder ein *Topic* sein. Die Textaufbereitung für eine NLP-Aufgabe ist in den meisten Fällen wortbasierend. Also wird meist eine Tokenisierung durchgeführt. Wörter sind aber nicht eindeutig definiert, sie sind intuitiv. Außerdem ist die Tokenisierung sehr fehleranfällig und benötigt lokale Anpassungen, je nachdem welche Sprache benutzt wird. Die zentrale Frage der Bachelorarbeit von Herrn Ebert ist es herauszufinden, ob das intuitive Konzept von Wort, wirklich die beste Art ist, um mit Hilfe von einem Rechner einen Text zu segmentieren. Eventuell gibt es auch eine andere Möglichkeit, die auf Tokenisierung komplett verzichtet und bessere Ergebnisse liefert.
Das Ziel seiner Bachelorarbeit ist die Entwicklung eines Algorithmus mit dessen Hilfe ein eingegebener Satz oder Text in seine „besten" Segmente, in diesem Fall Buchstaben N-Gramme, zerlegt werden kann. Er versucht damit herauszufinden, ob der nicht-symbolische (nicht wortbasierende) Ansatz besser als der wortbasierte Ansatz ist und welche Möglichkeiten dieser Ansatz bietet und was die Risiken sind, die durch die Anwendung entstehen.
Um sein Ziel zu realisieren geht Herr Ebert wie folgt vor: Er extrahiert N-Gramme der Länge 1 bis 10 aus dem Englischen Wikipedia Korpus. Dieser Korpus enthält nur unannotierte Rohtexte, keine POS-Tags. Zum Extrahieren werden die ersten 10.000 Texte des Korpus verwendet, die 22.650.880 Zeichen enthalten. Davon wird eine Frequenzliste für die N-Gramme der Länge von 1 bis 10 erstellt. Alle N-Gramme werden mit einem Gütemaß bewertet. Das Gütemaß kann berechnet werden indem man die N-Gramm Länge mit dem Logarithmus der absoluten Häufigkeit des N-Grammes multipliziert. (Gütemaß = n*log(freg)). Insgesamt soll ein möglichst hohes Gütemaß für jeden Satz zur Verfügung stehen. Herr Ebert testet sein System indem er einen Satz eingibt, der dann in N-Gramme mit den höchsten Gütemaßen zerlegt wird. Dabei entsteht ein Laufzeit Problem, die Laufzeit steigt exponentiell mit der Größe der Eingabe. Herr Ebert probiert das Problem mit Hilfe eines heuristischen Ansatzes zu lösen. Hierbei wird die Größe des Fensters z.B. auf 10 beschränkt und für jedes Fenster werden alle verschiedenen Möglichkeiten durchgegangen. Das Gütemaß wird für einen Teil berechnet und dann für den nächsten Teil usw. Mit der Festlegung der Fenstergröße wird die Berechnung des höchsten Gütemaßes nicht mehr garantiert. Man kann also nicht mehr sicher sein was das höchste Gütemaß für den gesamten Satz ist, da es passieren kann, dass ein N-Gramm mit einem sehr hohen Gütemaß durch die Trennung verloren geht. Die Segmentierung ist aber ggf. besser als bei dem symbolischen Ansatz.

Evaluation:
Herr Ebert meinte die Evaluierung von Text-Segmenten sei schwierig da es häufig Uneinigkeit über die Granularität von Segmenten gibt. Abhängig von der Anwendung seien Fehler relevant bzw. nicht relevant. Er hat hier ein Beispiel genannt:
Beim Information Retrieval könne die Korrektheit von Segmentgrenzen vernachlässigt werden und bei „news boundary detection" nicht.
Herr Ebert überprüft in seiner Bachelorarbeit die Auswirkung auf die Endanwendung (IR, Sentiment Analyse usw.). Dort verwendet er als sein Verfahren und überprüft, ob dieses besser als wortbasierende Verfahren ist. Er verwendet word2vec um Buchstaben N-Gramm Embeddings zu trainieren. Die Segment Analyse zur Evaluation geschieht auf Satzebene. Es werden Movie Review Daten verwendet, die mit Word Embeddings verglichen werden.

Ein mögliches Modell (Cho et al. 2014) wendet die Sigmoid-Funktion auf den letzten Zustand eines LSTM-Encoders (Long-Short- Term-Memory) an. Die Sigmoidfunktion wird auf die Summe der gewichteten Eingabewerte angewendet um das Ergebnis zu erhalten.

Seine bisherigen Erkenntnisse zeigen, dass auch die Buchstaben N-Gramme eine Zipf‘sche Verteilung aufweisen. Die häufigste N-Gramme, die größer als 3 Zeichen sind enthalten oft Funktionswörter und die häufigste N-Gramme größer als 8 Zeichen enthalten tendenziell Inhaltswörter.
Es sind noch keine Ergebnisse für die Evaluierung vorhanden. Daher ist die Frage, ob das Ergebnis der Evaluierung aussagekräftigt ist, noch offen. Ebenso weiß man noch nicht, welche andere Möglichkeiten es noch gibt um N-Gramme zu extrahieren, da die Aufteilung in N-Gramme von 1 bis 10 sehr willkürlich ist. Man müsste das System auch auf andere Tasks anwenden, um zu wissen, ob das nicht symbolische System besser ist als wortbasierte Systeme.

Bei verbleibender Zeit wird Herr Ebert das Verfahren („multiple random segmentation“) von Herrn Schütze (Aus „Nonsymblic text representation“), welches eine zufällige Auswahl an Buchstaben N-Grammen verwendet. Dabei wird eine minimale und maximale Segementlänge angegeben und Subsegmente durch Durchlaufen über den Text extrahiert.