Zusammenfassung zum Vortrag zur Bachelorarbeit zum Thema „Corpus based identification of text segments“ von Thomas Ebert

von Korbinian Schmidhuber

Die Bachelorarbeit von Thomas beschäftigt sich mit dem Thema Textsegmente. „Textsegment“ ist ein recht breit gefächerter Begriff, der z.B. umfasst: Morphem, Wort, Phrase, Satz, Absatz o.Ä. Ein Textsegment bezeichnet eine bedeutungstragende Einheit.
In der Verarbeitung von Sprachdaten werden die Daten meist so aufbereitet, dass sie zunächst in Textsegmente aufgeteilt wird. Meist wird dies wortbasiert gemacht. Der Begriff „Wort“ ist jedoch sehr intuitiv und dadurch ist diese sogenannte Tokenisierung oft fehleranfällig. Die Frage, die im Rahmen dieser Arbeit deshalb untersucht werden soll ist, ob eine Segmentierung in Wörter die beste Art für die Computerlinguistik ist, Texte zu segmentieren.

Das Ziel ist also, einen Algorithmus zu finden, der einen Satz in „beste Segmente“ zerlegt. Es kommt die Frage auf, ob nicht ein nicht-symbolischer Ansatz, also ein nicht-wortbasierter Ansatz besser sei.

Um dies zu untersuchen, werden zunächst N-Gramme der Länge 1-10 aus Texten extrahiert. Die Texte, die verwendet werden, stammen aus dem Wikipedia Korpus, welches unannotierte Rohtexte enthält. Die ersten 100.000 Texte des Korpus sollen in dieser Arbeit verwendet werden.
Nachdem die N-Gramme extrahiert wurden, wird eine Frequenzliste für diese erstellt und jedes N-Gramm mit einem Gütemaß bewertet. Das Gütemaß wird mit der Formel n*log(freq) berechnet, wobei n die Länge des N-Grammes darstellt und freq die absolute Häufigkeit des N-Grammes.

Hierbei hat sich das Problem ergeben, dass mit steigender Größe der N-Gramme die Laufzeit exponential steigt. Als Lösung hierfür soll die heuristische Größe des Fensters für N-Gramme beschränkt werden. Dies führt dazu, dass die Berechnung des höchsten Gütemaßes innerhalb eines Textes nicht mehr garantiert ist, aber können die einzelnen Segmente dennoch unter Umständen besser als in einem symbolischen Segmentierungsansatz sein.

Thomas bemerkt, dass eine Evaluation von Textsegmenten generell schwierig ist. Zudem können Fehler je nach Anwendung relevant oder irrelevant sein, z.B. kann beim Information Retrieval die Korrektheit der Segmente vernachlässigt werden.
Es wird das word2vec Tool benutzt, um Buchstaben N-Gramm Embeddings zu erhalten.

Erkenntnisse, die im Rahmen der Arbeit aufkamen waren z.B., dass auch Buchstaben eine Zipf'sche Verteilung aufweisen. Bei N-Grammen der Länge 3 waren die häufigsten N-Gramme solche, die Funktionswörter umfassen.

Fragen, die noch offen sind, sind z.B., ob es noch andere Möglichkeiten gibt, N-Gramme zu extrahieren und ob das Ergebnis der Evaluierung bereits aussagekräftig ist.