CIS, LMU München

Studentin: Elena Atanasova

29.05.2017

Protokoll zur Sitzung vom 29.05.2017 – Computerlinguistisches Arbeiten

3. Vortrag: Thomas Ebert , "Corpus based Identification of Text Segments"

BA-Betreuer: M. Sc. Martin Schmitt

Den letzten Vortrag des Tages hält Thomas Ebert, der von Herr Schmitt betreut wird. Am Anfang der Presentation stellte er seine Ideen für die Gliederung seiner Bachelorarbeit vor und danach presentierte er die Motivation. Die Motivation ist, die Textaufbereitung für NLP-Augfaben meist wortbasierend ist und das Wort nicht eindeutig definiert ist, aber intuitiv. Viele wissen was ein Wort ist, aber wenn man es deffinieren muss, kommt man auf Schwierigkeiten. Tokenisierung ist sehr fehleranfällig, sodass lokale Anpassungen nötig sind.

Das Ziel seiner Arbeit ist es eines Algorithmus zu entwickeln, der einen eingegebenen Satz oder Text in seine 'besten' Segmente (Buchstaben N-Gramme) zerlegt. Die Überlegungen die er sich machen soll sind: Ist der nicht-symbolische Ansatz besser als der wortbasierte Ansatz und welche Chancen und Risiken bietet der nicht-symbolische Ansatz?

Herr Ebert extrahierte von N-Grammen der Länge 1 bis 10, aus dem Wikipedia Korpus, 10.000 unannotierte englische Texte, die 22.650.880 Zeichen enthalten. Als zweiter Schritt erstellte Herr Ebert die Frequenzliste für die N-Gramme, die mit einem Gütemass bewertet werden (Gütemaß = N-Gramm-Länge * log(freq)). Gütemaß ist Wert zu beurteilen wie gut das N-Gramm ist freg in dem Fall ist die absolute Häufigkeit des N-Gramms. Zum Testen wird ein Satz eingegeben, der in der N-gramme mit den höchsten Gütemaßen zerlegt wird.

Die Probleme die aufgetaucht sind, sind: mit der Größe der Eingabe steigt die Laufzeit exponentiell. Eine Lösung dafür ist die Größe des Fensters zu beschrenken (heuristischer Ansatz). Mit der Festlegung des Fensters die Berechnung der höhestens Güte ist nicht mehr garantiert, aber die Segmentierung ist ggf. noch besser, als beim symbolischen Ansatz.

Generell ist die Evaluierung von Text Segmenten schwierig, wegen die häufige Uneinigkeit über die Granularität von Segmenten. Je nach Anwendung können Fehler relevant oder irrelevant sein. z.B. bei Information Retrieval (IR) kann die Korrektheit von Segmentgrenzen teilweise vernachlässigt werden und bei "news boundary detection" nicht. Die Auswirkung auf die Endanwendung (z.B. IR, Sentiment Analysis) wird als Maß verwendet. Herr Ebert verwendete word2vec um Buchstaben N-Gramm embeddings zu erhalten, nicht nur Wörter. Sentiment Analyse wird auf Satzebene geschickt zur Evaluierung, Verwendung von Movie Review Data und Vergleich mit Word embeddings. Mögliches Modell, dass Wort embeddings ergibt, ist Cho et al. 2014 Sigmoid, das auf dem letzten Zustand eines LSTM-Encoders LSTM (Long-Short-Term Memory) angewendet wird. Die Sigmoidfunktion wird auf die Summe der gewichteten Eingabewerte angewendet um ein Ergebnis zu erhalten.

Die Evaluationsergebnisse sind leider noch nicht vorhanden. Als offene Frage bleiben noch: ob es andere Möglichkeiten gibt, die N-gramme zu extrahieren, als Beispiel wurde das Programm von Herr Schütze erwähnt.