ZUSAMMENFASSUNGEN COLLOQUIUM, REPETITORIUM

Pascal Guldener

3.7.2017

1 Colloquium

1.1 Learning String Edit Distance

In Anton Serjogins Vortrag stellte dieser Ansätze zum maschinellen Lernen von String Abständen dessen Motivation und Anwendungmöglichkeiten vor.

Zunächst erleuterte er die Modifizierungsoperationen mit welchen man einen String x in einen String y umwandeln kann, anhand von Beispielen. Diese sind Einfügen, Löschen und Ersetzen. Dann erklärte er die gängigste Metrik dafür, den Levenstheinabstand, nämlich die minimalen Anzahl an Modifizierungoperationen welche nötig sind um von einem String x in einen String y zmzuwandeln. Er beschrieb dann den ansatz der Transduktion, bei welcher der Prozess sei ein String in einen anderen umzuwandlen. Wenn man dem die Betrachtung der Operationen als Zufallsvariablen ansieht gelangt man zur Ähnlichkeit zwischen diesen Strings als *stochastischer Abstand*. Eine Variante ist die Viterbi-Distance, welche den warscheinlichste der möglichen Sequenzen von Operationen zugrundelegt, eine weitere die Stochastic-edit-distance, bei der die Warscheinlichkeit des Stringpaares ausschlaggebend ist. Die Herausforderung bestand nun darin die Paramter für einen *memoryless Transducer* mit Hilfe von Expectation Maximization zu schätzen. Dafür wurde ein gelabaltes Korpus verwendet und die gemeinsame Warscheinlichkeit von Wörtern nach Bayes berechnet. Deser Classifier wurde für mehrere Experimente verwendet und angewandt auf verschiedene Teile eines Aussprache-Lexikons bzw auf eine Mischung aus dem Testkorpus und dem Aussprachelexikon und auf ein aus dem Testkorpus generiertes Aussprachelexikon. Dabei konnte ein klarer Performanzgewinn über den traditionellen Levenstheinabstand festgestellt werden.

1.2 Das Zipfsche Gesetz

wird in der nächsten Sitzung abgeschlossen

2 Repetitorium

2.1 Firmenvortrag Bayrische Staatsbibliothek

Vertreter der bayerischen Staatsbibliothek stellten Hintergründe zur technischen Umsetzung der Aufgabe ihre Bestände online druchsuchbar und verfügbar zu machen vor. Anhand von Pipelines stellten sie Tools und Datenflüsse. Dabei gaben sie Ausblick auf zukünftige Anwendungen und Problemstellungen insbesondere aus dem Bereich des Information Retreival. Konkret ging es dabei um die Suche in Notentexten, welche natürlich ganz anders beschaffen und strukturiert sind als Textcorpora. Auch die Form der Query einer solchen Suche wurde als Demonstration des aktuallen Enwicklungsstands demonstriert.