Kolloquium, 12. Juni 2017

Dayyan Smith

Phonologically-Enhanced Character Embeddings

Vortragender: Tobias Ramoser

Betreuer: Martin Schmitt

In seiner Bachelorarbeit möchte Tobias Buchstaben in einer phonologischen Vektorraumpräsentation darstellen und Repräsentationen mit verschiedenen phonologischen Features mit Zufallsvektoren vergleichen. Dabei wird mit dem SAMPA Alphabet gearbeitet. Es is ein auf ASCII basierendes, maschinen-lesbares, phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird. Entwickelt wurde es Ende der 1980er für phonemische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft. Laute lassen sich durch Artikulationsart, Artikulationsort und Stimmhaftigkeit charakterisieren. Ein Laut kann beispielsweise als plosiv, frikativ oder nasal beschrieben werden, je nachdem wie der laut entsteht, und beispielsweise als bilabial oder labiodental, wenn er mit beiden Lippen gemacht wird bzw.““‘ wenn für den Laut Lippe und Zähne notwendig sind. Bei der Stimmhaftigkeit gibt es nur zwei Möglichkeiten: stimmhaft oder stimmlos.

In insgesamt sechs Experimenten werden zufällige Vektoren mit 15 sowie 100 Dimensionen, mit word2vec erstellte Vektoren mit 15 sowie 100 Dimensionen, char vectors und one-hot vectors (hier werden die gleichen Features wie bei char vectors benutzt, aber mit binärer Klassifizierung).

Die Ergebnisse werden quantitativ und qualitativ analysiert. Die quantitative Analyse zeigt, dass die Accuracy der zufälligen Vektoren mit einer Dimension von 100 am besten ist, gefolgt von den one-hot Vektoren. Von der qualitativen Analyse werden diese Ergebnisse bestätigt. Für diese Analyse wird die Fehlerquote für die verschiedenen Arten der Repräsentationen mittels Levenshtein-Distanz berechnet. Sie gibt an wie viele Korrekturen gemacht werden müssten, um vom berechneten Wort zum richtigen Wort zu kommen.ram