Referat Tobias Ramoser (Montag, den 12.06.2017)

Phonologically Enhanced Character Embeddings

Zuerst einmal wurde über die Motivation des Themas gesprochen.
Anwendungen der maschinellen Sprachverarbeitung ist heute nicht mehr aus unserem Alltag wezudenken, Beispiesweise die Spracherkennung oder eine Maschinelle Übersetzung.
Auch beschäftigen die meisten Ansätze sich mit Wörter.

Das Ziel hier ist die Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Features. Auch wollen wir ein Vergleich mit Zufallsvektoren bei der Transkription in SAMPA finden.

Als nächstes wird über die Artikulation und Stimmhaftigkeit gesprochen. Zur Artikulationsart, wie ein Laut gebildet wird, wird uns erklärt, das Plosive, Vibranten, Apporximanten und nasale gibt. Zu der Frage wo ein Laut gebildet wird, wurde uns erdeutlicht, das es in der dental, velar, glottal und labiodental gebildet wird.

Im weiteren Verlauf wird über die Phonologie gesprochen. Diese beschreibt die Systematik der Laute innerhalb einer Sprache, in diesem Fall ist das die deutsche Sprache.
Es wird ein Vergleich gemacht von „tot“ und „rot“. Wörter unterscheiden sich in genau einem Laut.

Darüber hinaus wird Word2Vec erklärt. Dieses Program wird zur automatischen Vektorerstellung von Wörtern verwendet.
Als Input werden Trainingsdaten hergenommen. Beispielsweise Texte mit vielen Wörtern.
Als Output erhält man Wortvektoren und Distanz zu einem Wort-

Im oberen Verlauft wurde SAMPA kurz erwähnt. Nun wird hier kurz verdeutlicht, was genau ein SAMPA-Alphabet ist. Es basiert auf ein ASCII Maschinen lesbares phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt werden kann.
Es wurde im Jahre 1987 entwickelt, um phonemische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft zu übermitteln und um diese zu verarbeiten.

Zuletzt wurde uns mittgeteilt, dass ein Char-Vector verwendet wird um eine eigene Implementierung zu tätigen.