CIS, LMU München
12.06.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 12.06.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Tobias Ramoser und Mai Linh Pham jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Phonologically – Enhanced Character Embeddings

Als erster Referent stellte Tobias Ramoser seine Bachelorarbeit vor. Betreut wurde er während seiner Arbeit von Martin Schmitt. Seinen Vortrag gliederte er in Motivation, Ziele, Theoretischer Hintergrund, Experimente und seine Auswertung. Anschließend zeigte er noch lesenswerte Quellen auf.
Als Motivation für seine Arbeit nannte er die mittlerweile nicht mehr wegzudenkenden Anwendungen unseres Alltages wie z.B. Maschinelle Übersetzung, Spracherkennung in Autonavigationsystemen oder ähnlichen. In seiner Arbeit wurden die Beziehungen und Zusammenhänge von Buchstaben untersucht. Das Ziel sollte hierbei sein, diese in einer Vektorrepräsentation mit phonologischen Features darzustellen. Nachdem die Ziele der Arbeit nun klar waren, baute er noch etwas Hintergrundwissen zur Phonetik/Phonologie auf. Hierbei ging er auf die Artikulationsart, den Artikulationsort sowie die Stimmhaftigkeit ein.
Um die Vektorraumrepräsentation zu erstellen, verwendete Tobias Word2Vec. Hierbei werden aus einem importierten Text durch ein neuronales Netz, bestehend aus Lernalgorithmen, ähnliche Wörter mit einem zueinander ähnlichen Vektor repräsentiert. Das Wort2Vec System enthält ein Continous Bag-Of-Word Modell, sowie ein Skip-Gram Modell. Beide wurden von Tobias im Laufe seiner Arbeit verwendet. Außerdem kann das Modell mit verschiedenen Lernalgorithmen verwendet werden. Hier können Hierarchical Softmax, Negative Sampling und Downsampling of frequent words verwendet werden.
Um das Training der neuronalen Netze durchzuführen, wurde das SAMPA- Alphabet verwendet. Dies ist ein auf ASCII basiertes, von Maschinen lesbares Alphabet mit dem die Aussprache von Lauten dargestellt wird.
Insgesamt wurden so 4 Experimente durchgeführt. Darunter zwei Implementierungen für Vektorenerstellung inklusive Zufallsvektoren, ein Word2Vec Modell sowie die Transkription in SAMPA. Als Vektorenmodelle werden neben eigen implementierten Char-Vektoren auch One-hot und zufallsbasierte Vektoren verwendet.
Die Evaluation seiner Experimente zeigt, dass das Zufallsbasierte Vektormodell am besten performt, dicht gefolgt von One-Hot und Char-Vektoren. Am schlechtesten funktionierte dagegen das word2vec Modell. Die für die Auswertung verwedeten Werte wurden durch die Levenshtein-Distanz berechnet.