Jakob Sharab
CIS, LMU München
Matr.Nr.: 11186993

Kolloquium zu Computerlinguistisches Arbeiten -
Protokoll zur Sitzung vom 12.06.2017

Präsentation 1: Phonologically-Enhanced Character Embeddings

Der erste Vortrag wurde von Tobias Ramoser gehalten, dessen Arbeit von Martin Schmitt betreut wird. Er begann zunächst mit der Motivation hinter seinem Thema, wobei Tobias darauf aufmerksam machte, dass sich die meisten Ansätze in der Sprachanalyse mit Wörtern auseinandersetzen. In seiner Arbeit jedoch werden die phonologischen Eigenschaften von Buchstaben analysiert.
Daraufhin gab er eine Einführung in die Phonetik und die Phonologie, wobei die Phonetik generell in drei verschiedene Arten eingeteilt wird, nämlich akustische-, auditive- und artikulatorische Phonetik. Die letztere beschäftigt sich zum einem mit dem Artikulationsort, also dort wo die Laute entstehen, sowie auf welche Weise sie entstehen und ob diese stimmhaft oder stimmlos sind. Für jede dieser Eigenschaften existieren mehrere Unterkategorien, wie z.B. für den Artikulationsort, bei dem man u.a. zwischen bilabial, nasal und glottal unterscheidet. Die Artikulationsart kann beispielsweise genauer in plosiv und frikativ unterteilt werden. Die Phonologie beschreib die Systematik der Lautäußerungen einer spezifischen Sprache, wobei man hier nicht mehr von Lauten, sondern von Phonemen spricht, welche sich in mindestens einer Eigenschaft unterscheiden müssen. Tobias zeigte dies an den Phonemen [t] und [d], welche beide alveolar und plosiv gebildet werden, wobei das [t] jedoch stimmlos und das [d] stimmhaft ist.
Danach stellte er das Programm „word2vec“, welches automatisch Vektorrepräsentationen erstellt, und die Experimente die er durchgeführt hatte vor. Bei den ersten beiden Experimenten, handelte es sich um eigene, von ihm erstellte, Implementierungen. Der Unterschied dieser Experimente lag darin, dass es sich bei dem ersten Experiment um 5-dimensionale Vektoren gehandelt hat, wohingegen es bei dem zweiten Experiment um phonologische one-hot Vektoren ging, welche binär klassifiziert werden und somit 22 verschiedene Dimensionen besitzen. Bei beiden Experimenten wurden die Phoneme der jeweiligen Buchstaben identifiziert und der entsprechende Buchstabenvektor aus dem arithmetischen Mittel der Phonemvektoren berechnet. Hierfür zeigte Tobias ein Beispiel an der Tafel, für welches er die fünf Phoneme des Buchstaben „c“ aufzeigte. Das dritte Experiment bestand darin Vektorrepräsentationen mit Hilfe von „word2vec“ zu erstellen. Dabei wurde dem Modell, welches diverse Parametereinstellungen besitzt, eine von Tobias erstellte Trainingsdatei übergeben, welche in jeder Zeile jeweils phonologisch ähnliche Buchstaben enthält.
Das Ziel seiner Arbeit war es die erstellten Buchstabenvektoren mit Zufallsvektoren in der Transkription von Wörtern im SAMPA-Alphabet zu vergleichen. Die Resultate seiner Arbeit zeigen, dass die Zufallsvektoren zwar die besten Ergebnisse, mit einer Accuracy von 75%, lieferte, aber auch die Werte der one-hot Vektoren, mit 70%, und die Vektoren des ersten Experiments, mit 58% gut waren. Die Accuracy, erklärte Tobias noch nachträglich, steht dabei für die Anzahl der korrekt transkribierten Wörter. Zuletzt zeigte er noch eine Fehlerdatei, welche nicht richtig transkribierte Wörter enthielt, wobei er noch auf ein Muster hinwies, bei dem das Wort bis auf ein SAMPA-Zeichen richtig transkribiert wurde.