CIS, LMU München
Fach: Kolloquium Computerlinguistisches Arbeiten SS2017
Studentin: Faridis Alberteris Azar

Protokoll zur Sitzung am 12.06.17

Thema: „Phonologically-Enhanced Character Embeddings“

Student: Tobias Ramoser
Betreuer: Martin Schmitt

Das Ziel dieser Bachelorarbeit ist die Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Features. Zufallsvektoren werden auch bei der Transkription in SAMPA verglichen.

Als eine Art Einführung bzw. Erinnerung im Bezug auf die Phonologie erklärt Ramoser den Unterschied zwischen Artikulation und Stimmhaftigkeit. Wie ein Laut gebildet ist, wird Artikulationsart genannt (Plosiv, frikativ, nasal, lateral, Vibranten, Aproximanten). Der Artikulationsort ist es, wo ein Laut gebildet wird (bilabial, labiodental, alveolar, dental, velar, glottal). Wenn der Kehlkopf bei Aussprache eines Lautes vibriert, wird gesagt, dass laut der Stimmhaftigkeit, diese Laut stimmhaft ist, sonst ist es stimmlos.

Die Phonologie, wie der Student klarstellt, beschreibt die Systematik der Laute innerhalb einer Sprache (hier die Deutsche Sprache). Ein Phonem („Laut, Ton, Stimme, Sprache“) ist die abstrakte Klasse aller Laute (Phonem), die in einer gesprochenen Sprache die gleiche bedeutungsunterscheidende (distinktive) Funktion haben und seine Schreibweise ist in eckigen Klammern: […].

Tobias Ramoser zeigt die Haupteigenschften von Word2Vec, die Methode, die er für seine Zwecke benutzt. Word2Vec ist ein Programm zur automatischen Vektorerstellung von Wörtern und wurde von Mikolov im Jahr 2013 erfunden. Als Input nimmt dieses Programm Trainingsdaten in Form von Texten mit vielen Wörtern. Die Verarbeitung erfolgt durch neuronales Netz und besteht aus einer Architektur und einem Lernalgorithmus. Hier erhalten ähnliche Wörter ähnliche Vektoren, daher ist der Output von Werd2Vec die Wortvektoren und der Distanz zu einem bestimmten Wort.

In dieser Arbeit wird das SAMPA-Alphabet benutzt. Es handelt sich um ein auf ASCII-basiertes, maschinenlesbares, phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird. Dieses Alphabet wurde zwischen 1987 und 1989 entwickelt, um phonemische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft übermitteln und verarbeiten zu können.

Ramoser führt insgesamt vier Experimente durch:

(1) Zwei Implementierungen für Vektorenerstellung plus Zufallsvektoren.
(2) Word2Vec Vektoren.
(3) Transkription in SAMPA.

Seine eigene Implementierung wird „Char-Vectors“ genannt. Er verwendet die bereits definierte phonolodische Features und die Unterkategorien. Der Vektor wird mit dem 5 Features initialisiert: (0,0,0,0,0) und im Nachhinein werden die Phonemvektoren berechnet. Die Buchstabenvektoren

werden auch berechnet, indem den Durchschnitt aller entsprechenden Phonemvektoren bereechnet wird. Die zweite Implementierung heißt „One-not Vektoren“. Hier werden die gleiche Features und Unterkategorien wie im Char-Vectors verwendet. Allerdings handelt es sich hier um eine binäre Klassifizierung. Die Vektorgröße wächst somit auf 22 Dimensionen. Die Implementierung mit Zufallsvektoren wird „Randomized Vectors“ genannt. Es werden 2 Arten von Buchstabenvektoren generiert: 15 Dimensionen und 100 Dimensionen. Dafür wird das „random“-numpy-Moduls verwendet.

Im Falle vom Experiment Word2Vec Vektoren bestehen die Trainingsdaten aus Buchstaben anstatt aus ganzen Wörtern. Phonologisch ähnliche Buchstaben werden aufeinander trainiert.

Für Bewertung der Experimenten wird die Genauigkeit berücksichtigt. Die beste Ergebnisse zeigt random Experimente mit der 100-Version. Um diese Ergebnisse zu bestätigen, wird eine Fehlerquote mittels Levenshtein-Distanz berechnet.Es besagt, wie viele Korrekturen durchschnittlich gemacht werden mussten. Das bestätigt die bereits erhaltene Ergebnisse der quantitativen Auswertung.