Klyueva Olga
CIS, LMU München
12.06.1917

Tobias Ramoser: Phonologically enhanced character embeddings.

In der Bachelorarbeit geht es um Beziehungen und Zusammenhänge von Buchstaben. Es wurden phonologische Eigenschaften betrachtet, um Vektorraumpräsentationen von Buchstaben zu erschaffen. Das Ziel der Arbeit ist die Erstellung von verschiedenen Vektorraumpräsentationen von Buchstaben mit phonologischen Features. Diese Vektorraumpräsentationen wurden mit Zufallsvektoren verglichen bei der Transkription in SAMPA Alphabet.
Wichtige Begriffe sind Phonetik und Phonologie. Beide Disziplinen beschäftigen sich mit Lautbildung. Der Unterschied besteht darin, dass die Phonetik physikalische Eigenschaften von Lautbildung beschreibt und untersucht und die Phonologie beschreibt Systematik von Lautbildung, die es innerhalb einer Sprache gibt.
Die Phonetik unterteilt sich in:

1. Artikulatorische Phonetik
2. Akustische Phonetik
3. Auditive Phonetik

Das Gebiet, das für die Bachelorarbeit relevant ist, ist artikulatorische Phonetik. Dieser Bereich unterteilt sich in:

1. Artikulationsort (plosive, frikativ, nasal, literal etc)
2. Artikulationsart (bilibial)
3. Stimmhaftigkeit **(**stimmlos, stimmhaft)

In der Bachelorarbeit geht es um die deutsche Sprache. Es wurde ein Beispiel gezeigt, bei dem zwei Wörter sich in einer der phonologischen Eigenschaften unterscheiden, beispielsweise die Phonem „t“ und die Phonem „d“. Beide sind plosiv und frikativ. Der Unterschied ist in der Stimmhaftigkeit. Das Phonem „T“ ist stimmlos und das Phonem „D“ ist stimmhaft.
Als nächstes wurde über Wort2Vec gesprochen. Das ist ein Programm, das automatische Vektorerstellung von Wörtern, Dokumenten oder im Fall der Bachelorarbeit Buchstaben macht. Es wurde von Mikolov 2013 eingeführt. Das Programm erhält einen Text mit Wörtern, die mit Leerzeichen getrennt sind. Die erstellten Vektoren wurden an das Neuronetz übergeben und trainiert. Am Schluss erhält man eine Transkription des Wortes.
Es gibt continious bags of words neuronetz Modell und Skip-Gram neuronetz Modell. Das bag of words Modell sagt ein Wort aus einem Kontext vorher. Skip-Gram Modell sagt einen Kontext für ein angegebenes Wort vorher.
Es gibt verschiedene Lernalgorithmen, die ein Wort2Vec Modell trainieren: hierarchical softmax, negative samling, downsampling of frequent words.
Als nächstes wurde über SAMPA Alphabet gesprochen. Das SAMPA Alphabet ist ein ASCII-basiertes maschinenlesbares phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird.
Es wurden drei Experimente in der Bachelorarbeit durchgeführt. Während drei Experimenten wurden verschiedene Wordvektoren auf Grund von phonologischen Features erstellt.
Im vierten Experiment wurden die erstellten Vektoren mit erstellten Zufallsvektoren verglichen.
Das erste Experiment wurde Char-Vektor genannt. Es wurden phonologische Features (Artikulationsort, Artikulationsart, Stimmhaftigkeit) und Unterkategorien (plosive, nasal, frikativ, die Rundung der Lippen) festgelegt. Insgesamt gibt es fünf Eigenschaften, die man haben kann oder nicht. Danach wird die Initialisierung des Vektors mit (0,0,0,0,0) gemacht. Für jedes Phonem wird ein Vektor berechnet. Immer wenn die Eigenschaft auf ein Phonem zutrifft, wird hochgezählt. Als nächstes kommt eine Berechnung der Buchstabenvektoren mittels Durchschnittlichem aller entsprechenden Phonemevektoren.
Während dem nächsten Experiment wurden one-hot Vektoren erstellt. Der Unterschied zu Char-Vektor besteht darin, dass one-hot Vektor binär klassifiziert wurde. Die Vektorgröße wächst auf 22 Dimensionen.
Während dem dritten Experiment wurden Wort2Vec Vektoren erstellt mit der Eigenschaft Artikulationsort.
Zum Schluss wurde eine Grafik mit Ergebnissen präsentiert. Die Accuracy auf der Grafik zeigt, wie viel Prozent richtig in SAMPA Darstellung ausgegeben wurde. Am besten hat der Zufallsvektor funktioniert

mit 75 Prozent. Danach kommt der one-hot Vektor mit 70 Prozent. Char-Vektor hat 58 Prozent. Es wurde die Fehlerquote mittels Levenstein Distanz gemessen.