Katja Bertholdt
CIS, LMU München
12.06.17

Protokoll zur Sitzung vom 12.06.17 – Computerlinguistik Kolloquium

1. Vortrag: Tobias Ramoser, „Phonologically-Enhanced Character Embeddings“

Den ersten Vortrag des Tages hält Tobias Ramoser. Das Thema seiner Bachelorarbeit ist „Phonologically-Enhanced Character Embeddings“ und wird von Martin Schmitt betreut. Die Motivation seiner Arbeit liegt darin, sich mit Beziehungen und Ähnlichkeiten zwischen einzelnen Buchstaben zu beschäftigen, wohingegen sich die meisten alltäglichen Anwendungen der maschinellen Sprachverarbeitung mit Wörtern und deren Beziehungen beschäftigen. Für dieses Vorhaben werden Buchstaben mit phonologische Features in verschiedenen Vektorraumpräsentationen erstellt und diese mit Zufallsvektoren bei der Transkription in das SAMPA-Alphabet verglichen.
Zunächst gibt Ramoser einen Überblick über die Disziplinen „Phonetik“ und „Phonologie“.
Die Phonetik betrachtet die physikalischen Aspekte der Lautbildung und lässt sich in drei Teilbereiche untergliedern: die artikulatorische, die akustische und die auditive Phonetik. Ramoser beschäftigt sich in seiner Arbeit mit der artikulatorischen Phonetik. Diese lässt sich wiederum untergliedern in Artikulationsart (durch welche Organe wird der Laut gebildet?), Artikulationsort (an welcher Stelle wird der Laut gebildet?) und Stimmhaftigkeit (Unterscheidung zwischen stimmhaft und stimmlos). Jeder dieser Punkte kann unterschiedliche Werte annehmen.
Die Phonologie beschreibt die Systematik der Laute innerhalb einer spezifischen Sprache. Hier führt Ramoser ein Beispiel für die deutsche Sprache an, in dem er die Worte „rot“ und „tot“ gegenüberstellt. Diese unterscheiden sich in einem Phonem. Das bedeutet, dass sie sich in mindestens einer phonetischen Eigenschaft unterscheiden.
Nach den vorangegangenen Begriffserklärungen stellt Ramoser das Word2Vec Programm vor. Damit können Vektoren von Wörtern, Dokumenten oder Buchstaben erstellt werden. Dazu werden Trainingsdaten einem neuronalen Netz übergeben und dieses liefert in Ramosers Fall Buchstabenvektoren und die Distanz zu einem Buchstabenvektor als Output. Für das neuronale Netz stehen zwei Architekturen zur Verfügung. Zum einen das Skip-gram Model und das Continuous bag-of-words Model (CBOW). Beim Skip-gram Model wird zu einem gegebenen Wort der Kontext vorausgesagt, bei CBOW wird zu einem gegebenen Kontext ein Wort vorausgesagt.
Danach stellt Ramoser das SAMPA Alphabet vor. Dabei handelt es sich um ein in den Jahren zwischen 1987 und 1989 entwickeltes, ASCII-basiertes, maschinenlesbares, phonetisches Alphabet, mit welchem die Aussprache der Laute dargestellt wird.
In seiner Arbeit hat Ramoser vier Experimente durchgeführt. Durch drei Experimente wurden Vektoren erstellt und das vierte Experiment besteht aus einer Transkription, bei dem die Vektoren aus den ersten drei Experimenten mit den Zufallsvektoren verglichen werden.
Das erste Experiment hat Ramoser „Char-Vectors“ genannt. Dazu hat er fünf phonologische Features (Artikulationsort, Artikulationsart, Stimmhaftigkeit, Grad der Öffnung des Mundraums bei Vokalen und die Rundung der Lippen) definiert, die jeweils unterschiedliche Werte (Unterkategorie des jeweiligen Features) annehmen können. Da jedes Phonem über Werte dieser fünf Features verfügen kann, werden Vektoren mit fünf Einträgen initialisiert. Für jedes Phonem wird ein Vektor mit den entsprechenden Werten für das jeweilige Feature berechnet. Da Ramoser sich in seiner Arbeit jedoch mit Buchstaben beschäftigt, wurde aus den Werten der Phoneme das arithmetische Mittel ermittelt, da ein Buchstabe mehrere Phoneme besitzen kann.
Im zweiten Experiment wurden „One-hot“ Vektoren erstellt. Dabei wurden die gleichen Features und Unterkategorien verwendet, wie bei den Char-Vectors. Allerdings erfolgt hier eine binäre Klassifizierung, weshalb die Dimension des Vektors auf 22 Stellen anwächst. Für jede

Unterkategorie erhält der Vektor des Phonems einen Eintrag mit dem Wert 0 oder 1.
Außerdem hat Ramoser „Randomized Vectors“ mithilfe des numpy-Moduls random erstellt. Dabei hat er zwei Arten von Buchstabenvektoren erstellt, zum einen mit 15 und zum anderen mit 100 Dimensionen.
Zuletzt wurden „Word2Vec“ Vektoren erstellt. Die Trainingsdaten bestehen hierbei aus Buchstaben. Daraus hat Ramoser eine Datei erstellt, in der phonologisch ähnliche Buchstaben nebeneinander gestellt wurden und diese aufeinander trainiert.
Die erstellten Wortvektoren wurden dem neuronalen Netz übergeben und dieses hat die Wörter in SAMPA-Darstellung ausgegeben.
Seine Ergebnisse stellt Ramoser in einem Balkendiagramm dar, in dem zu sehen ist, zu wie viel Prozent der Input mit dem SAMPA-Output übereinstimmt. Im Vergleich zwischen den Vektoren haben dabei die „Randomized Vectors“ mit ca 75% accuracy am Besten abgeschnitten, während die Word2Vec Vektoren mit 15 Dimensionen mit fast 0% die schlechteste accuracy liefern.
Zur Analyse hat Ramoser eine Fehlerquote mittels der Levenshtein-Distanz berechnet, die aussagt, wie viele Korrekturen durchschnittlich gemacht werden mussten. Bei den Fehlern in seinem Output hat Ramoser nur schwer ein Muster erkennen können.