Protokoll für die Sitzung vom 12.06.17
Ivana Daskalovska 11139260

Thema: Phonologically-Enhanced Character Embeddings
Student: Tobias Ramoser
Betreuer: M.Sc. Martin Schmitt

Am Anfang der Präsentation stellte Herr Ramoser den Titel seiner Bachelorarbeit und eine Übersicht über die Gliederung vor.
Der Fokus seiner Arbeit liegt auf Anwendungen der maschinellen Sprachverarbeitung (Maschinelle Übersetzung, Spracherkennung usw.), die immer mehr im Alltag verwendet werden. Die meisten Ansätze beschäftigen sich dabei mit Beziehungen und Zusammenhängen von Wörtern im Vektorraum. Herr Ramoser versucht im Rahmen seiner Bachelorarbeit herauszufinden, ob die phonologische Vektorraumpräsentation von Buchstaben besser wäre. Das Ziel seiner Arbeit ist die Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Merkmalen (Features) und der Vergleich mit Zufallsvektoren bei der Transkription in SAMPA.

Der thematische Hintergrund seiner Bachelorarbeit setzt sich dabei aus
Phonetik und Phonologie, word2Vec und SAMPA-Alphabet zusammen.

Herr Ramoser gab zu diesen Teilen jeweils eine kurze Erklärung. Er begann mit einer Einführung in die Grundlagen der Phonetik bzw. ging auf Artikulation und Stimmhaftigkeit ein. Bei der Artikulation werden die Artikulationsart und der Artikulationsort betrachtet. Die Artikulationsart zeigt wie ein Laut gebildet wird. Dabei teilt man die Laute in Plosive, Frikative, Nasale, Laterale, Vibranten und Aproximenten ein. Der Artikulationsort zeigt wo ein Laut gebildet wird. Dabei werden die Laute in Bilabiale, Labiodentale, Alveolare, Dentale, Velare und Glottale eingeteilt.
Was die Stimmhaftigkeit betrifft, so unterscheidet man stimmhafte und stimmlose Laute. Wenn der Kehlkopf bei der Aussprache vibriert, dann ist der Laut stimmhaft, andernfalls stimmlos.

Die Phonologie beschreibt die Systematik der Laute innerhalb einer Sprache. Im konkreten Fall wird die Deutsche Sprache betrachtet. Bestandteile der Phonologie sind die Phoneme. Phoneme müssen sich in einer phonetischen Eigenschaft unterscheiden, wie z.B. „rot“ und „tot“. Diese Wörter unterscheiden sich genau in einem Laut.

Word2Vec ist ein Programm zu automatischen Erstellung von Vektorrepräsentationen von Wörtern (auch Word Embeddings genannt). Als Eingabe für Word2Vec sollte ein großer Textkorpus hergenommen werden. Die Daten werden durch ein neuronales Netz verarbeitet. Dabei enthalten ähnliche Wörter einen ähnlichen Vektor. Als Output bekommt man dann einen Wortvektor. Auch die Distanz zu einem anderen Wort kann ausgegeben werden.

In seiner Bachelorarbeit vergleicht Herr Ramoser zwei Ansätze, welche in Word2Vec verwendet werden können, um Vektorrepräsentationen zu erstellen:

1. Continuous Bag-of-Words Modell: zweischichtiges neuronales Netz, welches versucht das aktuelle Wort auf Basis von Wörtern um das Wort herum vorrauszusagen
2. Skip-Gramm Modell: Zweischichtiges neuronales Netz, das das aktuelle Wort verwendet um den Kontext um die Wörter herum vorrauszusagen

Die Lernalgorithmen, die Herr Ramoser in seiner Bachelorarbeit verwendet sind:

1. Hierarchical Softmax
2. Negative Sampling
3. Downsampling of frequent words

SAMPA-Alphabet ist ein auf ASCII-basiertes, maschinen-lesbares, phonetisches Alphabet, mit dem die Aussprache der Laute dargestellt wird. Das SAMPA- Alphabet wurde zwischen 1987- 1989 entwickelt, um

phonemische Transkriptionen der offiziellen Sprachen der damaligen Europäischen Gemeinschaft zu erstellen und verarbeiten zu können.

Herr Ramoser führt im Ramen seiner Bachelorarbeit insgesamt vier Experimente durch: zwei Implementierungen für Vektorenerstellung und Zufallsvektoren, Word2Vec Vektoren sowie Transkription in SAMPA.

Bei den Buchstaben-Vektoren handelt es sich um eigene Implementierungen. Dabei verwendet Herr Ramoser phonologische Merkmale und Unterkategorien wie vorher definiert. Der Vektor wird initialisiert mit (0,0,0,0,0). Zur Berechnung der Phonemvektoren: Immer wenn eine Eigenschaft auf ein Phonem zutrifft erhält es den dementsprechenden Vektor. Die Buchstabenvektoren werden mittels Durchschnitt aller entsprechenden Phonemvektoren berechnet.

Bei den One-Hot Vektoren werden diegleichen Merkmale und Unterkategorien wie bei Buchstaben-Vektoren verwendet. Allerdings wird hier eine binäre Klassifizierung durchgeführt. Die Vektorgröße „wächst“ somit auf 22 Dimensionen.

Bei den Zufallsvektoren (Randomized Vectors) handelt es sich um zufällig generierte Vektoren. Es werden zwei Arten von Buchstabenvektoren generiert:

1. Vektoren mit 15 Dimensionen
2. Vektoren mit 100 Dimensionen

Bei Beiden wird das „random“ Modul von NumPy verwendet.

Ergebnisse und Analyse:

Nach der quantitativen Analyse werden die besten Ergebnisse von den Zufallsvektoren geliefert. Als zweitbestes schneiden die One-Hot Vektoren ab. Bei der qualitativen Analyse wird die Fehlerquote mittels Levenstein-Distanz berechnet. Die qualitative Analyse sagt aus, wie viele Korrekturen durchschnittlich gemacht werden müssen. Sie bestätigt die Ergebnisse der quantitativen Auswertung. Auch hier schneiden die Zufallsvektoren am besten ab.