12.06.17

Protokoll zur Sitzung vom 12.06.2017 – Computerlinguistisches Arbeiten

1. Referat: Tobias Ramoser, Phonologically-Enhanced Character Embeddings (Martin Schmitt)

Der Vortrag beginnt mit einem kurzen Überblick über die Bachelor-Thesis. Die Motivation für die Arbeit ist, dass Anwendungen der maschinellen Sprachverarbeitung nicht mehr aus dem Alltag wegzudenken sind. Beispiele dafür sind zum einen die maschinelle Übersetzung und zum anderen die Spracherkennung.

Die bisherigen Ansätze beschäftigen sich meistens mit Wörter. In dieser Arbeit liegt der Fokus auf die Beziehungen und Zusammenhänge zwischen Buchstaben im Vektorraum, im Speziellen also den phonologischen Vektorraumpräsentationen von Buchstaben. Das Ziel der Arbeit ist die Erstellung von verschiedenen Vektorrepräsentationen von Buchstaben mit phonologischen Features und der Vergleich mit Zufallsvektoren bei der Transkription in SAMPA.

Als nächstes wiederholt Tobias einige Aspekte aus der Phonologie, wie z.B. die Artikulationsart, der Artikulationsort und die Stimmhaftigkeit. Er definiert Phonologie als die Beschreibung der Systematik der Laute innerhalb einer Sprache. In diesem Fall wird die deutsche Sprache benutzt.

Das Programm, das für die automatische Vektorerstellung von Wörtern benutzt wurde, ist Word2Vec. Es bekommt Trainingsdaten, also einen Text mit vielen Wörtern, als Input, welche durch ein neuronales Netz verarbeitet werden. Es besteht aus verschiedenen Architekturen Lernalgorithmen, wobei ähnliche Wörter ähnliche Vektoren erhalten. Das Output besteht aus Wortvektoren und die Distanz zu einem Wort. Eine Architektur ist das Continuos Bag-of-Words Modell, welches ein zweischichtiges neuronales Netz ist. Es sagt aus einem Kontext ein gewisses Wort voraus. Die zweite Architektur ist das Skip-Gram Modell, welches ein zweischichtiges neuronales Netz ist. Es sagt den Kontext für ein gegebenes Wort voraus. Die verwendeten Lernalgorithmen sind das Hierarchical Softmax, das Negative Sampling und das Downsampling of frequent words. Anschließend erläutert Tobias das SAMPA-Alphabet. Es ist ein auf ASCII-basiertes, machinen-lesbares phonetisches Alphabet.

Für die Arbeit werden insgesamt 4 Experimente, 2 Implementierungen für Vektorenerstellung und Zufallsvektoren durchgeführt bzw. erstellt. Weitere Bestandteile des Experiments sind Word2Vec Vektoren und die Transkription in SAMPA. Tobias erläutert, dass Char-Vectors eine eigene Implementierung ist. Weitere Implementierungen sind One-Hot Vektoren und randomized Vectors.

Zum Schluss präsentiert Tobias seine Ergebnisse und eine quantitative Analyse. Die Berechnung einer Fehlerquote wird mittels Levenshtein-Distanz berechnet. Diese sagt aus, wie viele Korrekturen durchschnittliche gemacht werden mussten und bestätigt die Ergebnisse der quantitativen Auswertung.