Tobias Ramoser

Phonologically Enhanced Character Embeddings

Betreuer: Martin Schmidt

Die Arbeit behandelt einen Wortrepräsentationsansatz der Wörter nicht als atomare Einheiten auffasst, sondern buchstabenbasierte Repräsentation mit einbezieht. Hierzu soll die Repräsentation nicht orthographisch, sondern phonologisch motiviert sein, also phonologische Features einen Einfluss auf die Embeddings nehmen. Zum Test der phonologischen Embeddings wurde der SAMPA task zur Transkription von Wörtern in ein phonologisches Schriftsystem verwendet. Das SAMPA Alphabet ist ein ASCII-basiertes phonetisches Alphabet zur Transkription von Wörtern und soll universell maschinell lesbar und verarbeitbar sein.

Zur Grundlage der Arbeit hat Tobias nochmals die Grundlagen von Phonologie und Phonetik erläutert. Phonologie als Untersuchung der Laute einer gegebenen Sprache stehen hierbei im Gegensatz zur Phonetik als allgemeiner Lehre der Lautbildung und Klassifikation. Spricht man von einem Phonem ist es ein Laut der sich in einer eindeutig phonetischen Eigenschaft von anderen Lauten der Sprache unterscheidet. Eine Sprache besitzt ein konkretes Inventar an Phonemen.

Die Arbeit vergleicht vier unterschiedliche phonologisch-basierte Vektorräume miteinander. Der erste ist ein selbst-definierter Buchstabenvektor nach eigener Implementation, welcher phonologische Informationen eines Buchstabens in reelle Zahlen codiert. Dabei ist der entstehende Wortvektor der Durchschnitt der konstituierenden Phonem-Vektoren. Die Methode arbeitet mit Vektoren der Dimensionalität 5x1.

Der zweite Ansatz ist ein one-hot Vektor, bei welchem eine Stelle innerhalb des Vektors eine phonologische Eigenschaft codiert. Die Werte dieser Vektoren sind binär und die Dimensionalität ist auf 22x1 festgelegt.

Im dritten Ansatz werden Vektoren mit zufälligen Zahlenwerten befüllt. Die Methode zählt als Baseline für den Versuch. Diese wurden in den Dimensionen 15x1 und 100x1 getestet.

Der vierte Ansatz nutzt das Toolkit Word2Vec, welches durch das trainieren eines Sprachmodells eine gewichtete Input-Matrix für Wörter erhält. Im Gegensatz zum wortbasierten Training, für das Word2Vec normalerweise eingesetzt wird, sind diese Vektoren ebenfalls buchstabenbasiert. Dimensionen der Vektoren sind ebenfalls 15x1 und 100x1.

Diese Vektormodelle wurden als Input Layer für ein Neural Network für den SAMPA task genutzt. Die Experimente fanden extern statt. In der Praxis erwiesen sich hierbei die randomisierten

Vektoren als am effektivsten. Hinzu kam eine qualitative Analyse, welche die Fehleranfälligkeit an konkreten Wörtern mit Hilfe von Editierdistanzen bestimmt. Auch diese Analyse bestätigt die gute Performance der zufälligen Vektoren im Task.