Repetitorium

Gini

Im Repetitorium hat sich das Start-Up Unternehmen Gini vorgestellt. Das Moto von Gini lautet: „Fight Paperwork“. Ihr Ziel ist es Dokumente, wie zum Beispiel Banküberweisungen einfach vom Smartphone aus zu abzuhandeln. Zuerst macht man ein Foto von dem jeweiligen Dokument, dann scannt man mit Hilfe von OCR das Bild und extrahiert alle wichtigen Informationen. Am Ende kann man ein ausgefülltes Formular abschicken und muss nicht zu Fuß gehen. Es gibt aber noch Probleme, vor allem bei dem Scannen mit OCR. Dies ist unter anderem auch auf schlechte Bildqualität zurückzuführen. Gini kann bereits über 50 verschiedene Arten von Dokumenten erfassen und es sind weitere geplant.

Kolloquium

Phonologically Enhanced Character Embedding

Der erste Vortrag des Kolloquiums wurde von Tobias Ramoser gehalten. Sein Betreuer ist Martin Schmid. Viele Anwendungen der maschinellen Sprachverarbeitung sind heute nicht mehr wegdenkbar. In seiner Bachelorarbeit geht es unter anderem um die Beziehungen und Zusammenhänge zwischen Buchstaben im Vektorraum. Er will also mit Hilfe von phonologischen Features verschiedene Vektorrepräsentationen von Buchstaben erstellen. Dann verglich er die beiden Begriffe Phonetik und Phonologie. Die Phonetik beschreibt vor allem die physikalischen Eigenschaften der Lautbildung betrachtet. Die Phonologie betrachtet die Systematik der Lautbildung in einer Sprache, hier im Deutschen. Bei der Phonetik konzentriert er sich hauptsächlich auf die artikulatorische Phonetik. Dabei geht es um Artikulationsart, Ort und die Stimmhaftigkeit. Für die Automatische Vektorerstellung wird Word2Vec verwendet. Das Programm erhält Trainingsdaten, die dann von einem Modell behandelt werden. Dabei gibt es verschiedene Architekturen, die verwendet werden können. Einmal das „Continous Bag of Words“ Modell und einmal das „Skip Gram“ Modell. Bei CBoW wird aus einem gewissen Kontext ein gewisses Wort vorhergesagt. Dagegen wird beim „Skip Gram“ Modell der Kontext aus einem gewissen Wort vorhergesagt. Es gibt auch verschiedene Lernalgorithmen, wie den „Hierarchical Softmax“, das Negative Sampling und das Downsampling of frequent words. Weiterhin verwendet er das SAMPA-Alphabet. Dies ist ein auf ASCI basiertes Phonetisches Alphabet, mit dem die Aussprache der Laute dargestellt werden kann. Phoneme lassen sich so eindeutig identifizieren. Er hat insgesamt vier Experimente durchgeführt. Drei davon sollten die Vektoren erstellen und das vierte diente zur Transkription. Es wurden

phonologische Features festgelegt, also Art und Ort der Artikulation und die Stimmhaftigkeit. Dazu kamen nicht einige Unterkategorien wie zu Beispiel der Grad der Öffnung des Mundes. Sie stellen Eigenschaften der Phoneme dar. Für jedes Phonem wird dann ein entsprechender Vektor erstellet. Für viele Buchstaben gibt es mehrere mögliche Phoneme, deshalb wird immer das arithmetische Mittel dieser Phoneme ermittelt. Die Experimente werden mit vorher erstellten Zufallsvektoren verglichen. Bei dem Word2Vec Experiment stellt man phonologisch ähnliche Buchstaben nebeneinander und trainiert diese dann. Die Ergebnisse waren ja nach Einstellung sehr unterschiedlich. Er hat zur Sicherheit eine Fehlerquote berechnet, die sagt wie viele Korrekturen im Schnitt gemacht werden mussten.

Analysis of NIL Results in an Entity Linking System

Der letzte Vortrag des Tages wurde von Mai Linh Pham gehalten. Ihr Betreuer ist Yadollah Yaghoobzadeh. Es geht um „Entity Linking“, also das Verbinden von Erwähnungen aus einem Text zu Entitäten aus einer „Knowledge Base“. Zu Anfang erklärte sie einige wichtige Begriffe. Die NIL Ergebnisse zeigen Entitäten, die nicht in der KB sind. „Fine-grained“ Annotation besitzt mehr mögliche Tags als NER. In diesem Fall wurde „FIGER“ benutzt, was 112 Tags besitzt. Das Ziel der Arbeit ist es herauszufinden, ob die „Fine-grained“ Annotation von Nutzen für die „Entity Linking“ Systeme sind und ob sie bei der Analyse der NIL Ergebnisse helfen können. Zuerst wird der Output des „Entity Linking“ Systems und der Output eines „Enity annotiation“ Tools kombiniert. Die Tags werden mit einem Standard BIO-System versehen. Das verwendete „Entity Linking“ System heißt „WAt“. Dann werden die NIL-Ergebnisse extrahiert. Das bedeutet, alles, was von FIGER erkannt wurde, aber nicht von WAT verlinkt wurde. Die NIL-Ergebnisse werden daraufhin in verschieden Cluster eingeteilt. Die drei Typen sind „coarse-grained“, „fine-grained“, und top-level. Bei dem „Coarse-grained“ Clustering zeigte sie einige Bilder, wo die verschiedenen Typen abgebildet waren. Bei “Fine-grained” werden alle Typen in Multi-level und Single-level Typen eingeteilt. Die Single-Typen werden dann den Multi-Typen zugewiesen. Bei dem Top-level Clustering werden identische Tags im ersten Level gruppiert, um die Anzahl der Cluster zu reduzieren. Um die Undefinierten Typen zu reduzieren, erstellt man neue Domänen. Am Ende hat sie ihre Ergebnisse vorgestellt. “Fine-gained” Entitäts-Typen können demnach für das Clustering verwendet werden. Dabei werden auch lexikalische Eigenschaften und der Kontext der Entitäten beachtet. Außerdem sind “Fine-gained” Typen deutlich informativer als die „Coarse-grained“ Typen.