ZUSAMMENFASSUNGEN COLLOQUIUM, REPETITORIUM

Pascal Guldener

17.7.2017

1 Colloquium

1.1 Sprachmodelle und Smoothing

Der Referent Sohail Malih srpach über die Bedeutung von Sprachmodellen und wofür Smoothing eingesetzt wird. Zunächst grenzte er stichpunktartig lingusitische von statistischen Sprachmodellen ab, wobei er recht abstrakt blieb und ohne Vorwissen die Unterschiede kaum klar geworden sind. Genannt wurden, Verwendung grosser Datenmengen bei statistischen Verfahren und morphologische Analysen bei den linguistischen. Sprachmodelle sollen einen Teil der Struktur von Sprache abbilden. Da natürliche Sprache viele Strukturebenen wie Phonologie, Morphologie, Syntax, Semantik, Pragmatik hat hängt die Güte eines Modells auch davon ab wieviel Struktur aus den jeweiligen Ebenen sich im Modell wiederfinden lässt. Sodann erwähnte er N-Gramm Modelle welche Corpora als Mengen von N-Grammen betrachtet, wobei das N natürlich ein Hyperparamter darstellt der zur optimierung des Modells verändert werden kann, sowie Klassenbasierte Modelle welche als Verbesserung von N-Garm-Modellen bezeichent wurden ohne das näher zu erleutern und HMM-Modelle, welche Übergangwarscheinlichkeiten zwischen Kontexten und Token benutzen. Die Arithmetik hinter der Berechung von Wahrscheinlichkeitsketten liefert problematische Resultate bei ungesehenen Tokens da diese im Modell eine Wahrscheinlichkeit von 0 haben. Eine Lösung für diese Problem stellen Smoothingverfahren dar welche Wahrscheinlichkeitsmasse aus gesehenen Token nach bestimmten Regeln auf ungesehene Wörter verteilen. Bei der Interpolation wird die Wahrscheinlicheit der *n-1*-Gramme für das Smoothing verwendet um die Schätzung über Wahrscheinlichkeiten ungesehener N-Grammer zu verbessern. Dann kam er zur Bewertung von Sprachmodellen und erwähnter kurz die Schlüsselbegriffe "Entropie" und "Perplexität", wobei das eine den Informationsgehalt eines Ereignisses meint, also zB. die größe der Änderung in der Warscheinlichkeit eines Ereignisses und das andere den "Verzeigungsfaktor" vor einem Ereignis.

1.2 Wortähnlichkeit: Levenshtein-Abstand, Wagner-Fischer-Algorithmus und Brill-Moore: Verbessertes Fehlermodell für Noisy Channel Rechtschreibkorrektur

Im Vortrag von Ivana Daskalovska ging es um den Levensthein-Abstand, der Vorstellung eine Algorithmus zu dessen Berechnung und um ein Verfahren zur Korrektur von Schreibfehlern basierend auf der Warscheinlichkeit von Editiersequenzen. Zunächst ging Sie auf die Eigenschaften der Levenstheinschen Distanzfunktion ein und führte die zugrundeliegenden Editieroperationen, Einfügen, Löschen und Substitution und abgeleitete Editieroperationen an. Da Wladimir Levenshtein selbst kein Verfahren zur Berechnung angegeben hat stellte Sie das Verfahren von Wagner-Fischer vor, welches auf einer Matrix operiert. Diese 2-Dimensionale Matrix mit Spalten/Zeilenzahl in der Länge der betrachteten Wörter. Man berechnet nun jeden für jeden Pfad von Editieroperationssequenzen und die Kosten und bekommt dann das Minimum durch den Pfad mit den kleinsten Teilkosten. Die Zeitkomomplexität beläuft sich dadurch

auf $\mathcal{O}(nm)$ was aber duchr Verbesserungen auf $\mathcal{O}(n/logn)$ gebracht werden kann. Der Algorithmus wird in der Rechtschreibprüfung, OCR Korrektur, Duplikatenerkennung und Sequenz-Alignierung verwendet. Ein weiterer Ansatz von Brill-Moore ist maschinelles Lernen der Editierwarscheinlichkeit von einem String zum anderen. Dabei verwendet man zwei probabilistische Modelle, das Source-Model (Wahrscheinlichkeitsverteilung über Wöerter) und das Channel-Model (Wahrscheinlichkeitsverteilung über ein falsches Wort gegeben die höchste Warscheinlichkeit für das richtige Wort). Das Channel-Model wird mit Wortpaaren aus falsche geschriebenen Varianten mit der richtiggeschriebenen Variante trainiert und anschliessend werden die einzelnen Buchstaben aligniert basierend auf der Editierdistanz zwischen diesen. Für die Ersetzungwarscheinlichkeit werden noch Kontextzeichen miteinbezogen. Eine Verbesserung dieses Verfahren stellt noch die Einbeziehung von Positionen im Wort dar, da Rechtschreibfehler beispielsweise häufiger am Wortende als am Anfang passieren.

2 Repetitorium

2.1 Firmenvorttrag - TrustYou

Die Vertreter von TrustYou stellten Vorgehensweisen, Tool-Pipelines und Ausblicke auf ihr Geschäftsfeld vor. Insbesondere gingen sie auf den für das Sammeln der Daten (Hotelbewertungen) benutzen Crawler und Technologien (Python, Rechencluster, Hadoop,Word2Vec) ein und stellten die Toolpipline für die Aufbereitung der Daten und Verfügbarmachung per API vor. Dabei machen Sie Gebrauch von nicht näher gannanten ML-Systemen und beabsichtigen auch das deployen von DeepLearning Frameworks. Auch gaben Sie einige Codebeispiele für die Erstellung von Wortrepresäntationen in Form von Vektoren durch Word2Vec um diese für ML verwertbar zu machen.