CIS, LMU München
15.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 15.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Tobias Eder, Joseph Birkner und Kristina Smirnov jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Exploiting Bilingual Word Embeddings to Establisch Translational Equivalence

Tobias Eder hält das erste Referat der Sitzung. Betreut und unterstützt wird er hierbei von Fabienne Braune und Alexander Fraser. Das Ziel der Arbeit soll es sein, Übersetzungen ohne Wörterbücher zu ermöglichen. Dies ist seiner Meinung nach vor allem in speziellen Bereichen sinnvoll, in der es einen sehr großen fachbezogenen Wortschatz gibt wie zum Beispiel der Medizin. Diese, auf Wortfamilien basierte Methode, soll auch unbekannte Wörter übersetzbar machen.
Zu Beginn, stellt Tobias das Kernprinzip von Word Embeddings vor und wie sie helfen können, das Sparse Problem etwas einzugrenzen.
Anders als bei sehr dichten Stukturen wie Audio und Videodateien hat man bei Text sehr wenige Daten. Dies ist darauf zurückzuführen, dass Wörter als atomare Einheiten auftreten und somit bis auf den Kontext nahezu keine weiteren Daten erheben kann. Deshalb soll versucht werden, ähnliche Wörter in Vektorraummodelle zu clustern. Somit können syntaktische und semantische Beziehungen zu ähnlichen und anderen Wörtern dargestellt werden. Der Abstand zwischen zwei verglichenen Wörtern ist hierbei ein Maß für ihre Ähnlichkeit. Auf Grund der Zipschen Verteilung ist dies nicht mit Embeddings lösbar.
Anschließend stellt Tobias seine verwendeten Tools für die Umsetzung vor. Dazu zählen Word2Vec (entwickelt von Google 2013). Dieses Tool bringt ein Bag of Words Modell (welches ein Wort anhand eines Kontextes sucht) und ein Skipgram Modell (welches Kontext anhand eines Wortes sucht). Außerdem verwendet er FastText Tool (entwickelt von Facebook Research 2016).
Sein Trainingskorpus besteht im wesentlichen aus vier Korpora. General (110M Tokens), Medical Big (50M Tokens), EMEA (4M Tokens), TED Talks (2M Tokens). Aus diesem Korpus wurde ein 5000 Token großer Parallelkorpus mit dem Moses Toolkit erstellt. Das Toolkit performt hierauf sehr gut, und generiert zu 93% die korrekten Wörter. Zur Regularisierung verwendet Tobias Glättung indem er hohe Gewichte bestraft. Schlussendlich verliert er noch ein paar Worte zur laufenden Evaluation und gibt einen Ausblick auf seine weiteren geplanten Schritte. Um seine Ergebnisse bewerten zu können, wählt er 1000 hochfrequente Wörter aus und bildet sie in einem Regressionsmodell (Lineare Regression) ab. Diese wurden von ihm manuell überprüft.
Demnächst möchte er noch einen besonderen Blick auf niedrigfrequente Wörter werfen, an seinen Abbildungen arbeiten, da hier manchmal noch Jahreszahlen „übersetzt“ werden. Außerdem will er noch andere Regularisierungsmethoden überprüfen. Abschließend zeigt Tobias noch seine Referenzen weiterführende Quellen.

CIS, LMU München
15.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 15.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Tobias Eder, Joseph Birkner und Kristina Smirnov jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Ranking With Neural Network Derived Document Vectors

Das zweite Referat hielt Joseph Birkner zu seiner Bachelorarbeit mit dem Titel „Ranking with neural network derived document vectors“.
Joseph geht sehr genau auf die Funktionsweise von „Neural Networks“ ein und gliederte seinen Vortrag in Vision, Motivation, Objective und Task. Das Ziel der Arbeit ist es, ein System zu generieren, welches Dokumente klassifizieren bzw. encoden kann. Dafür verwendet Joseph einen Korpus, welcher bereits von einer vorhergehenden Bachelorarbeit erstellt wurde. Das Themengebiet ist noch sehr neu, da es sich erst vor etwa 3 Jahren aufgetan hat.
Grundsätzlich beschreibt der Referent die Funktionalität in einem rudimentären Neural Information Retrieval System. Der Benutzer muss eine Anfrage erstellen, diese wird anschließend auf die entsprechende Datenbank ausgeführt. Das System gibt dann der Wahrscheinlichkeit nach sortierte Vorschläge aus, durch die der Benutzer „klicken“ kann. Anschließend kann er die Suchanfrage bearbeiten um so bessere und genauere Ergebnisse zu erziehen.
Für das Ranking benutzt er eine Art Vorkommenshäufigkeit, welche jedoch auch mit Problemen verbunden ist. Er führt als Beispiel den Film „Star Wars“ und die Token „Mann“ und „Lichtschwert“ an. Obwohl das Wort „Mann“ vermutlich viel häufiger im Korpus vorkommt als das Wort „Lichtschwert“, assoziiert man „Lichtschwert“ eher mit dem Filmtitel. Hier muss also eine Gewichtung oder Anpassung in irgendeiner Form verwendet werden.
Momentan arbeitet er mit Hilfe von Doc2Vec daran, einen Kontext mit einem Fehlenden Wort in das Tool Einzupflegen, damit Dokumentenvektoren zu generieren und dann das Originale Wort vom Kontext möglichst wieder zu bekommen. Diese Vorgehen trainiert und evaluiert er momentan auf dem selben Korpus, weshalb noch keine richtige Evaluation erstellt werden kann. Seine Ergebnisse zeigt er anschließend mit dem python Modul „plotty“ an.
Schlussendlich zeigte er noch seine Referenzen und Quellenangaben.

CIS, LMU München
15.05.2017
Michael Strohmayer
11137111
michael.strohmayer@campus.lmu.de

Protokoll zur Sitzung 15.05.2017 – Computerlinguistisches Arbeiten

Im Laufe der Sitzung wurden von den Studenten Tobias Eder, Joseph Birkner und Kristina Smirnov jeweils das Thema ihrer Bachelorarbeit vorgestellt.

Coparison of transfer Methods for low-ressource Morphology

Kristina Smirnov hält das letzte Referat des Tages. Auch sie stellt ihre Bachelorarbeitsthema vor. Die Aufgabenstellung ist hier, zu einem gegebenen Lemma und dessen Form die konkreten Vorkommen in einem Korpus zu finden. Dafür arbeitet sie mit einem De- und Encoder. Dieses Modell macht es möglich, Experimente auf Korpora anderer Sprachen zu testen. Auf dieses Modell wird zurückgegriffen, weil für die gegebene Sprache russisch sehr wenige Ressourcen zur Verfügung stehen (low-ressource). Deshalb sind Ergebnisse auf diesen verfügbaren Trainingsdaten nicht aussagekräftig genug. Um bessere Ergebnisse zu erziehen, soll einerseits die Zielsprache, andererseits eine ähnliche Sprache zur Hilfe genommen werden. Wegen ihrer Ähnlichkeit wurde hier Russisch und Ukrainisch gewählt. Hierbei wird auf verschiedenen Ressourcen trainiert und evaluiert. Im Allgemeinen wurde das Vorgehen in drei Tasks aufgeteilt.

Task 1: Russisch annotiert und ukrainisch annotiert
Task 2: Russisch annotiert und russisch nicht annotiert
Task 3: Russisch annotiert, russisch nicht annotiert und ukrainisch

Auf diese Korpora werden verschieden große Anfragensets ausgeführt. Ihre wesentliche Aufgabe besteht darin, die Ergebnisse, welches das Modell liefert, zu vergleichen.
Für die Auswertung berechnet sie Precision, Recall und F-Score. Nachdem sich die Fehleranalyse mit diesen rein statistischen Werten jedoch als schwierig gestaltet, verwendet sie noch ihre eigene Fehleranalyse. Diese basiert auf ihrem bilingualen Sprachkenntnis von Ukrainisch und Russisch. Eines Ihrer Probleme äußert sie außerdem in der Dauer der Evaluationsverfahren. Diese brauchen sehr lange, da für jeden Korpus und jeden Task in den verschiedenen Größen ausgeführt werden muss. Kristina beschrieb Kommunikation mit ihrer Betreuerin, Ungenauigkeit in der Aufgabenstellung und Flüchtigkeitsfehler als weitere Probleme.
Zum Schluss gab die Referentin noch einen Ausblick über eine mögliche Verbesserung des Systems und präsentierte ihre Referenzen.