Protokoll zur Sitzung vom 15.05.2017 - Computerlinguistisches Arbeiten

Tobias Eder, Betreuer: Alexander Fraser, Fabienne Braun

Exploiting Bilingual Word Embeddings

Tobias startet den Vortag mit der Motivation zu seiner Arbeit. Ziel der Arbeit ist, sich an eine Übersetzung ohne Wörterbuch heranzutasten, um z.B. bessere Domain-abhängige Übersetzung zu ermöglichen. Den nächsten Schritt beginnt er mit einer Erklärung zu *word embeddings*. Während bei Audio und Grafiken eine Dichte von Daten beobachtet werden kann, tritt bei Text eine sogenannte *sparseness* auf. Das entsteht dadurch, dass man Wörter eher als atomare Einheit ansieht. Daher wird versucht, Wörter in einem Vektorraummodell zu repräsentieren. In solch einem Modell werden semantische und syntaktische Zusammenhänge mit Vektoren zu einem Kontext zusammengefasst. An dieser Stelle wirft Dr. Schulz ein, dass man das *sparseness*-Problem durch diese Vorgehensweise dennoch nicht lösen kann, sondern im Prinzip eigentlich nur verschiebt. Er führt an, dass wenn ein Wort z.B. nur ein einziges Mal im Korpus vorkommt, das statistisch nicht ausreicht und den Vektor wirkungslos macht. Außerdem warnt er davor, dass ein Vektor auch falsche Information, die in einem Korpus enthalten sein könnte, in Relation setzen würde und zu einem falschen Kontext führen würde. Das erklärt er anhand des Beispiels von Tobias, in der eine Grafik Kontexte zwischen Ländern und deren Hauptstädte herstellt. Ein falscher Satz könnte Berlin zur Hauptstadt Portugal werden lassen. Er mahnt, eine mutmaßliche Lösung zum *sparseness*-Problem kritisch zu sehen. Danach stellt Tobias einige Tools vor, die er für die Erstellung seiner Arbeit zur Hilfe genommen hat. Das erste heißt Word2Vec und erstellt einen mehrdimensionalen, kontextualen Vektorraum aus einem Korpus. Das Tool setzt sich aus mehreren neuronalen Netzen zusammen und wurde 2013 von Google entwickelt. Word2Vec bietet 2 Arbeitsweisen: CBOW (*continuous bag-of-words*), bei dem das System das Wort aus einem Kontext von Wörtern filtert, und *skip-gram*, bei dem das System das aktuelle Wort nutzt, um einen Kontext von Wörtern daraus zu bestimmen. Während CBOW schneller ist, ist *skip-gram* treffsicherer, vor allem, was seltener vorkommende Wörter betrifft. Ein weiteres Tool wurde von Facebook Research 2016 entwickelt. Es arbeitet ähnlich wie Word2Vec, jedoch zieht es bei seinen Berechnungen auch n-Gramme hinzu. Somit ist es auch bei morphologischen Differenzen möglich, kontextuale Vektoren räumlich nah beieinander anzusiedeln. Falls Wörter gar nicht in Trainingsdaten vorhanden sind, gewinnt man über n-Gramme zum Kontext. Als nächstes erklärt Tobias, dass er versucht, mit einem *predictive model*, ein Vektormodell mit 100-1000 Dimensionen, lineare Abbildungen zu erzeugen. Zur Veranschaulichung zeigt er eine Grafik, in der zwei Vektorräume für Wörter der englischen und der spanischen Sprache dargestellt waren. Man sieht, dass sich korrespondierende Wörter in einem ähnlichen Raum aufhalten. Dr. Schulz merkt an, dass der Begriff "lineare Abbildung" hier nicht greifen kann, da diese direkte Abbildungen sind, und schlägt vor, dies Annäherung an eine lineare Abbildung zu nennen. Der nächste Punkt beschreibt, welche Korpora genutzt wurden und wie das Experiment aufgebaut ist. Hier vermisst Dr. Schulz einen Evaluierungspunkt in der Gliederung, aber Tobias versichert, dass dieser auf jeden Fall vorhanden ist und behandelt wird. Bei den vier verwendeten Korpora handelt es sich um

einen generellen Korpus (z.B. aus einem Wikidump extrahiert), einem medizinischen Korpus namens Medical Big, einem pharmazeutischen Korpus mit dem Namen EMEA und Ted Talks, einem Korpus bestehend aus TED Vorträgen. Für das Sprachpaar Englisch-Deutsch (da es schließlich auch um bilinguale Embeddings geht) ist zusätzlich ein kleiner paralleler Korpus im Einsatz. Dr. Schulz erklärt mit einem Schmunzeln, dass die Medizin mit all ihren Begriffen und Synonymen eine linguistische Katastrophe darstellt. Zum Schluss fasst Tobias zusammen, womit er sich weiterhin für seine Arbeit befassen wird. Das sind zum einen Tests für seltener vorkommende Wörter und zum anderen begibt er sich auf die Suche nach besseren Abbildungsmöglichkeiten und anderen Regularisierungsmethoden. Außerdem möchte er OOV-Wörter (*out of vocabulary*) mit fastText verarbeiten.

Joseph Birkner, Betreuer: YingDing Wang

Ranking With Neural Networks Derived from Document Vectors

Zu Beginn spricht Joseph von der Motivation für seine Arbeit und dem Projekt IROM. Bei IROM handelt es sich um einen *Intelligent Recommender of MOOCS*. Das Akronym MOOCS steht für *Massive Open Online Courses*. Diese Kurse werden, wie der Name schon sagt, massiv online angeboten. Projekt IROM stellt eine intelligente Suchmaschine dar, die Ergebnisse von Suchanfragen auf die individuellen Bedürfnisse der User anpasst. Dafür stellt er das Konzept von *ubiquitous vertical search* vor. *Vertical Search* beschreibt eine Suchmaschine, die sich auf bestimme Themenbereiche und Felder spezialisiert. *Ubiquitous (*"allgegenwärtig") *vertical search* umfasst demnach viele spezielle Themenbereiche. Sobald ein User eine Suchanfrage stellt, ist ein Informationsbedarf vorhanden. Dieser soll durch eine spezialisierte *vertical search engine* gedeckt werden. Ein Suchergebnis (recommendation) wird in diesem Fall durch Information Retrieval gefunden. Kurzum kann man den Prozessablauf wie folgt beschreiben: Der Benutzer hat einen Informationsbedarf, den er in einer Suchanfrage ausdrückt, diese Anfrage wird durch domainspezifische Ranking-Algorithmen durchgegeben und anhand einer Datenbank ausgewertet. Dann wird eine rangierte Ausgabe von Recommendations zurückgegeben. Zeitgleich wird während diesem Prozess die Genauigkeit weiter durch die Meta Daten des Users zur Zeit der Suchanfrage verfeinert. Dabei handelt es sich zum Beispiel um das Geschlecht oder Alter des Users oder um dessen Suchverlauf. Der neuronale Prozess des Information Retrievals lässt sich in zwei Teile unterteilen: zum einen wird der User Input optimiert, zum anderen erfolgt eine Optimierung des Matching Algorithmus. Ersteres versucht eine effiziente Dokumentrepräsentation zu erzielen, was zu einem schnelleren und intelligenteren Ranking führt. Eines der für diese Arbeit verwendeten Tools nennt sich Doc2Vec. Es ist ähnlich zu dem im vorherigen Vortrag vorgestellten Word2Vec, mit dem Unterschied, dass es Vektorenräume für ganze Dokumente und nicht nur einzelne Wörter schafft. Joseph merkt außerdem an, dass dieser Algorithmus nicht online läuft. Im nächsten Teil berichtet er von den Tasks, die er bisher bearbeitet hat. Dabei wurde ein Prototyp erstellt, der Dokumentvektoren generiert. Dieser wird anhand von Trainingsdaten getestet, was eine Zuhörerin noch einmal hinterfragt und bestätigt bekommt. Um die Ergebnisse anschaulich zu

gestalten, macht sich Joseph ein Diagramm zunutze, das mit *plotty* erstellt wurde. Es zeigt die räumliche Nähe von semantisch korrespondierenden Dokumenten. Am Ende wird der restliche Verlauf der Arbeit erläutert. Dieser besteht aus vier Schritten. Zuerst muss das Training des LSTM Seq2Seq Modells abgeschlossen sein. Mithilfe der Ergebnisse des Trainings wird dann eine API kreiiert, die Dokument-Anfrage-Paare generiert und dafür das jeweilige rangierte Set bereitstellt. Danach möchte Joseph den Ranking-Algorithmus evaluieren und zuletzt auch die gewählten Features. Im Anschluss bittet Dr. Schulz die Studenten, sich um kürzere Vorträge zu bemühen.