Referat Tobias Eder (Montag, den 15.05.2017)

Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Wir hätten gerne eine Übersetzung ohne ein Wörterbuch.
Die Übersetzung soll auf bestimmten Domain folgen.
Word-Einbettung ist der kollektive Name für eine Reihe von Sprachmodellierung und Feature-Lerntechniken in der natürlichen Sprachverarbeitung (NLP), wo Wörter oder Phrasen aus dem Vokabular auf Vektoren von reellen Zahlen abgebildet sind. Konzeptionell handelt es sich um eine mathematische Einbettung von einem Raum mit einer Dimension pro Wort zu einem kontinuierlichen Vektorraum mit viel geringerer Dimension.
Methoden zur Erzeugung dieser Abbildung sind neuronale Netze, Dimensionsreduzierung auf die Wort-Co-Vorkommensmatrix, probabilistische Modelle und explizite Repräsentation in Bezug auf den Kontext, in dem Wörter erscheinen.
Als nächste wurde hier die Word2Vec erklärt. Word2vec ist eine Gruppe von verwandten Modellen, die verwendet werden, um Wort-Embeddings zu produzieren. Diese Modelle sind flache, zweischichtige neuronale Netze, die ausgebildet sind, um sprachliche Kontexte von Wörtern zu rekonstruieren. Word2vec nimmt als Eingang ein großes Textkörpel auf und erzeugt einen Vektorraum, typischerweise von mehreren hundert Dimensionen, wobei jedem eindeutigen Wort im Korpus ein entsprechender Vektor im Raum zugeordnet ist. Wortvektoren sind in dem Vektorraum so positioniert, dass Wörter, die gemeinsame Kontexte im Korpus teilen, sich in unmittelbarer Nähe zueinander im Raum befinden.
Word2vec wurde von einem Team von Forschern unter der Leitung von Tomas Mikolov bei Google erstellt. Der Algorithmus wurde anschließend von anderen Forschern analysiert und erklärt. Einbetten von Vektoren, die mit dem Word2vec-Algorithmus erstellt wurden, haben viele Vorteile gegenüber früheren Algorithmen wie Latent Semantic Analysis.
Desweiteren gibt es die fastText, die eine Bibliothek entwickelt, um skalierbare Lösungen für die Textdarstellung und Klassifizierung zu erstellen. Unser anhaltendes Engagement für die Zusammenarbeit und den Austausch mit der Community erstreckt sich über die Bereitstellung von Code hinaus. Wir wissen, dass es wichtig ist, unsere Lernen zu teilen, um das Feld voranzutreiben. So haben wir auch unsere Forschung über FastText veröffentlicht. FastText kombiniert einige der erfolgreichsten Konzepte, die in den letzten Jahrzehnten von den Methoden der natürlichen Sprachverarbeitung und der maschinellen Lerngemeinschaften eingeführt wurden. Dazu gehören die Darstellung von Sätzen mit Beutel von Wörtern und Beutel von n-Gramm, sowie die Verwendung von Unterwort-Informationen und das Teilen von Informationen über Klassen durch eine versteckte Darstellung. Wir verwenden auch eine hierarchische Softmax, die die unausgewogene Verteilung der Klassen ausnutzt, um die Berechnung zu beschleunigen. Diese unterschiedlichen Konzepte werden für zwei verschiedene Aufgaben verwendet: effiziente Textklassifizierung und Lernwort-Vektordarstellungen.
Auch wurde etwas zu lineare Abbildungen erwähnt. Diese sind eigentlich exakte Abbzeichungen, hier werden jedoch nur Annäherungen gezeigt mit Beispiiel von verschiedenen Sprachen, wie English, Deusch, Russisch, Italienisch und weitere.
Als vorletzten Punk wurde über Korpora und Experimentaufbau gesprochen.
Es gibt vier unterschiedliche parallele Korpora. Die erste ist die General welches ca 110 Millionen Tokens besitzt. Die Medical Big welche ca 50 Millionen Tokens hat, desweiteren

die EMEA, die ca 4 Millionen Tokens besteht und als letztes die TED Talk. Diese hat ca 2 Millionen Tokens.
Zuletzt wird über die weiteren Schritte gesprochen, welche erledigt werden sollen. Einmal über die niedrigfrequente Wörter. Ob es auch bessere Abbildungen gibt? Welche anderen Methoden man noch nutzen kann.