Jakob Sharab
CIS, LMU München
Matr. Nr: 11186993

Protokoll zur Sitzung vom 15.05.2017

Präsentation 1: Exploiting Bilingual Word Embeddings to Establish Translational Equivalence

Die erste Präsentation wurde von Tobias Eder gehalten, der seine Bachelorarbeit bei Fabienne Braune und Alexander Fraser schreibt. Zunächst erläuterte er uns, weshalb sein Thema aktuell interessant ist und die Motivation dahinter. Das Ziel ist es Übersetzungen auch ohne ein entsprechendes Wörterbuch zu erhalten, da es beispielsweise gerade auf bestimmten Domänen unbekannte Wörter gibt, die nicht im Wörterbuch vorhanden sind. Aber auch um dem generellen Problem entgegenzuwirken, dass wenn man für eine bestimmte Sprache kein Wörterbuch hat, man auch keine Übersetzung erhält.
Danach ging er auf die Theorie hinter Word Embeddings ein, bei der es besonders bei der Textverarbeitung oft zu *sparse data* Problemen kommt. Als Lösungsansatz dafür nannte Tobias, die Darstellung der Wörter in einem Vektorraummodell, woraufhin Herr Schulz einwarf, dass auch dies keine perfekte Lösung sei. Weiter erklärte Tobias, dass in einem solchen Modell die syntaktischen und semantischen Beziehungen festgehalten werden können und ähnliche Wörter sich somit im Modell clustern sollten. Jedoch sind Vektorraummodelle sehr hochdimensional, weshalb man zur Veranschaulichung die Zahl der Dimensionen reduziert.
Im Anschluss daran wurden zwei Programme zur Vektordarstellung vorgestellt, welche Tobias für seine Arbeit benutzt. Das erste Programm „Word2Vec“ wurde 2013 von Google entwickelt und ist in der Lage zwei verschiedene Arten von Modellen zur Vektordarstellung zu verwenden, nämlich entweder mit Hilfe eines CBOW- oder eine Skipgrammodells. Das zweite Programm, welches von Facebook entwickelt wurde, war „fastText“, ist in der Lage auch für OOV-Wörter, d.h. unbekannte Wörter, Vektordarstellung zu erstellen. Eine weitere Funktion des Programmes ist zudem die Textklassifikation mit Hilfe eines linearen Modelles, was Tobias für seine Arbeit jedoch nicht verwendet. Daraufhin erklärte er, dass die linearen Abbildungen der beiden Programme durch lineare Regression geschehen, aber für seine verwendeten Texte nicht unbedingt geeignet sind, da es leicht zu *overfitting* kommt, weshalb man dem mit Regularisierung, z.B. durch das „bestrafen“ großer Gewichte, entgegenwirken muss.
Anschließend wurden die verwendeten Korpora und der Experimentaufbau erläutert. Die 4 Korpora die zum Einsatz kamen, waren der „General“-, „Medical“-, „EMEA“- und der „TED Talks“- Korpus. Tobias erklärte, dass er unterschiedliche Embeddings verwenden wolle, nämlich das CBOW- und das Skipgrammmodell, und es um die Übersetzung vom Englischen ins Deutsche geht. Dazu verwendet er einen kleinen parallelen Korpus, der aus ca. 5000 Wörtern besteht, und eine Auswahl an Wörtern, ca. 1000 hochfrequente, aus den verschiedenen Korpora, welche nicht im parallelen Korpus auftauchen. Die Erwartung sei nach Tobias, dass die Übersetzung bei diesen Wörtern gut funktioniert. Danach, erklärte Tobias, habe er domänspezifische Testsets verwendet, bei denen eine unterschiedliche Performance bei den beiden verschiedenen Modellen (CBOW und Skipgram) zu erwarten sind. Das CBOW Modell sei z.B. besser für größere Korpora geeignet.
Abschließend gab Tobias noch einen Ausblick auf die Schritte, welche er noch durchführen will. Zum einen will er sich mit der Frage beschäftigen, wie gut die Übersetzung bei niedrigfrequenten Wörtern funktioniert und ob es evtl. bessere Abbildungen als die, mit dem Regressionsmodell gibt. Aber auch, ob es noch andere Regularisierungsmethoden gibt. Zudem will er noch evaluieren, wie gut die Übersetzung von OOV-Wörtern in „fastText“ funktioniert.

Präsentation 2: Ranking With Neural Network Derived Document Vectors

Der zweite Vortrag wurde von Joseph Birkner gehalten. Er begann, damit „Ubiquitous Vertical Search“ im Zusammenhang mit dem Projekt „IROM“ (Intelligence Recommendation of Massive Open Online Courses) vorzustellen. Dabei definierte er „Vertical Search“ als Suche innerhalb einer bestimmten Domäne. Im Vergleich dazu soll es mit der „Ubiquitous Vertical Search“ möglich werden die Lösung, die vorher nur auf einer bestimmten Domäne funktionierte, auch auf andere Domänen anzuwenden. Daraufhin veranschaulichte er uns die „Ubiquitous Vertical Search“ noch an einem Schaubild, an dessen Anfang ein Benutzer, mit einem „Information Need“, stand, welcher diese in einer Suchanfrage formuliert, woraufhin diese Anfrage, zusammen mit verschiedenen Dokumenten, einem Ranking-Algorithmus übergeben wird, der diese Dokumente nach Relevanz sortiert wieder ausgibt.
In Josephs Arbeit geht es im Allgemeinen um Information Retrieval, wobei der Fokus zwar auf „Representation Optimization“ liegt, er sich aber auch mit „Deep Relevance Matching Models“ beschäftigt. Das Ziel dabei ist es ein System zu entwickeln, welches eine gute „Document Representation“ erstellt, die ein effizientes Bewerten der Relevanz von Dokumenten ermöglicht. Er erklärte, dass es traditionell so sei, dass die „Document Representation“ durch ein „TF-IDF“-Modell geschieht. Dieses Modell erzeugt eine Matrix mit der Zahl aller möglichen Terme (TF) und der inversen Dokumenthäufigkeit (IDF). Das Problem hierbei ist jedoch u.a., dass die Reihenfolge der Wörter ignoriert wird.
Als nächsten Punkt stellte Joseph das Ziel der Darstellung semantischer Räume für Dokumente, ähnlich zu der Vektorraumdarstellung von Wörtern, vor, wofür er kurz das Prinzip der „Latent Semantic Spaces“ (LSS) erläuterte. Als Beispiel zeigte Joseph, wie bei „Word2Vec“ traditionell die Darstellung der semantischen Ähnlichkeiten zwischen Wörtern funktioniert und erwähnte dabei, ähnlich wie Tobias, dass es dabei oft zu *sparse data* Problemen kommt. Daraufhin nannte er als Ziel, nun einen Autencoder für Dokumente und nicht für Wörter zu entwickeln, also nicht „Word2Vec“, sondern „Doc2Vec“. Dafür müsse man zunächst Vektoren für Dokumente, mit Hilfe eines „Recurrent Neural Network“ erstellen und dann ein großes Korpus nehmen, um einen „seq2seq“ Autoencoder zu trainieren. Anschließend nutzt man den vorher trainierten Autoencoder, um *feature*-Vektoren aus den Dokumenten zu extrahieren.
Daraufhin zeigte er uns ein Vektorraummodell von Dokumenten, welches ein Anderer, der sich mit einem ähnlichen Thema beschäftigt hatte, erstellt hatte. Dort waren die Ähnlichkeiten zwischen verschiedenen Kursen, aufgrund ihrer Kursbeschreibung, dargestellt worden.
Als letzten Punkt nannte er die verschiedenen Aufgaben, die er bis jetzt bewältigt hatte und die er noch plant durchzuführen. Dazu gehörten z.B. ein LSTM (long short term memory) „seq2seq“ Modell zu trainieren, oder auch die extrahierten *features* mit Hilfe einer Heatmap zu evaluieren.

Präsentation 3: Comparison of Transfer methods for LOW Resource Morphology

Die dritte Präsentation wurde von Kristina Smirnov gehalten, deren Arbeit von Katharina Kann betreut wird. Kristina begann zunächst die Forschungsfragen zu nennen, mit welchen sie sich in ihrer Arbeit beschäftigt. Die Hauptfrage dabei war, wovon man Gebrauch machen kann, wenn kaum Daten einer Sprache vorhanden sind und ob es helfen würde annotierte Daten, aber auch nicht annotierte, von einer ähnlichen Sprache hinzuzufügen und ob es sinnvoll wäre diese beiden miteinander zu kombinieren. Die Effizienz dieser verschiedenen Szenarien will sie anhand der korrekten Erzeugung der morphologischen Formen bewerten.
Danach stellte Kristina kurz das Modell vor, welches sie für ihre Arbeit verwendet, welches von Hinrich Schütze und Katharina Kann an der LMU entwickelt wurde. Dabei handelt es sich um einen morphologischen Encoder und Decoder, der auch mit kleinen Datensätzen gute Ergebnisse erzielt. Anschließend stellte sie uns ihre Vorgehensweise vor, um die oben genannten Fragen zu beantworten.

In einem ersten Szenario, geht es darum annotierte russische Daten mit annotierten ukrainischen Daten zu kombinieren. Als zweite Kombination will sie annotierte russische Daten und nicht annotierte russische Daten verwenden und als letztes Daten aus allen drei Kategorien. Dafür wählt sie immer eine Kombination aus Wörtern der Sprache, für die es kaum Daten gibt, immer mit einer anderen Anzahl von Wörtern, für die es viele Daten gibt.
Als nächstes erläuterte sie, wie die Daten aussehen, welche ihr zur Verfügung stehen. Diese stammen aus dem Shared Task, aus welchem das Modell von Herrn Schütze und Frau Kann entstand und enthalten einmal das Lemma, dann die Zielform und die morphosyntaktische Beschreibung.
Abschließend kam Kristina zum Punkt „Ergebnisse und Evaluation“, bei dem sie erklärte, dass sie noch auf die Ergebnisse warte, da das Modell diese erst verarbeiten muss, sie aber plane diese am Schluss graphisch aufzuarbeiten. Des Weiteren wolle sie noch genauer darauf eingehen, was für Fehler in einzelnen Fällen aufgetreten sind, und ob dabei ein Muster zu erkennen sei, welches dabei hilft das System zu verbessern.