Joseph Birkner

Ranking with Neural Network Derived Document Vectors

Betreuer: Francois Bry und Yingding Wang

Die Bachelorarbeit von Joseph Birkner findet im Rahmen des Projekt IROM statt, der Intelligent Recommendation of Massive Open Online Courses und nutzt Methoden des Neural Information Retrieval, einer sehr jungen Disziplin im Rahmen von IR, um aus einem Angebot an Onlinekursen bestimmte Useranfragen effizient und qualitativ besser zu beantworten. Dabei stützt sich die Arbeit auf mehrere generelle Grundannahmen des IR, wie den Information Need eines Users und der Beantwortung durch Ranking von Dokumenten, welche diesem Information Need nach Möglichkeit entsprechen sollen. Da bisherige Modelle des IR oft von einer term frequency und inverse-document frequency ausgehen, werden syntaktische Strukturen im Bag of Words Modell der relevanten Terme vernachlässigt.

Die Arbeit setzt thematisch auf dem Punkt des Deep-structured semantic matching, oder kurz DSSM an und soll bei der Repräsentationsoptimierung helfen. Dabei geht es um die bessere Darstellung von Dokumentvektoren in einem latent semantischen Vektorraum, ein ähnlicher Ansatz, der auch bei Modellen wie Word2Vec zum Einsatz kommt. Durch semantisierte Vektorräume, in denen thematische Relationen abbildbar werden und die nicht nur naiv den tf-idf Werten entsprechen könnte somit eine Verbesserung der Ranking Methoden und des Dokument-Matchings gefunden werden. Zu diesem Zweck nutzt die Arbeit Autoencoder Neural Networks, mit deren Hilfe ein unbeaufsichtigtes Lernen stattfinden kann. Hierbei lernt das neuronale Netz aus einem gegebenen Dokumentinhalt das Dokument vorherzusagen. Dies soll im weiteren Schritt dabei helfen semantisch ähnliche Nutzeranfragen nah an die passenden Dokumente zu platzieren. In einem Recurrent Neural Network werden zu diesem Zweck die Sequenzen des betreffenden Dokuments als Eingabe verarbeitet.

Bisher ist der erste Prototyp für Dokumentvektoren erstellt worden. Hierzu wurde das RWTH-language model als Ausgangsbasis verwendet. Der Prototyp verwendet bisher sowohl Trainings- als auch Testdaten aus dem IROM Task. In den folgenden Schritten sollen mit Hilfe von Tensorflow weitere Modelle trainiert werden und APIs für Vektoren aus der LSTM (long short-term memory) Architektur und einer Suchquery angegliedert werden.

Zur weiteren Evaluation des Modells soll unter anderem das TREC (Text REtrieval Conference) Dataset verwendet werden. Hinzu kommt in einem zweiten Schritt eine genauere Auswertung einzelner Features des RNN. Da die RNN Architektur mit einem Hidden Layer arbeitet ist die Gewichtung einzelner Features im Prozess nicht absehbar, weswegen Joseph auf die Methodik aus Karpathy 2015 zurückgreifen will und einzelne Features als Heatmap auf den Dokumenten abbilden lassen will, um dadurch die Entscheidungsprozesse des RNN transparenter zu machen.

Joseph hat unter anderem während seines Vortrags auf die Python Bibliothek plotly zur Visualisierung von Datensätzen hingewiesen.

Kristina Smirnov

Comparison of Transfer Methods for Low-Resource Morphology

Betreuer: Katharina Kann

Die Arbeit von Kristina Smirnov entsteht in der Folge des Sigmorphon 2016 Shared Task über morphologische Flexion. Insbesondere beschäftigt sich die Arbeit mit Verfahren zur Adaption von externen Daten, falls in einer Sprache nicht genug annotierte Daten zum Training von Modellen zur Verfügung stehen. Für diesen Fall werden zwei mögliche Lösungsansätze vorgestellt: Die Verwendung von fremden annotierten Daten aus einer anderen, morphologisch verwandten Sprache, oder die Verwendung von nicht-annotierten Daten der ursprünglichen Sprache.

Zu diesem Zweck sollen in der Arbeit mehrere Encoder-Decoder Netzwerke auf unterschiedlichen Datensätzen trainiert werden. Die Arbeit nimmt sich eine Bearbeitung des Sigmorphon 2016 Shared Task durch Katharina Kann und Hinrich Schütze zur Grundlage. Dabei gilt zunächst die Unterteilung in drei verschiedene Ansätze zur morphologischen Flexion im russischen:

1. Das Mischen von russischen und ukrainischen Trainingsdaten
2. Netzwerk mit rein russischen Trainingsdaten, die nur zum Teil annotiert sind
3. Die Verknüpfung beider vorheriger Datensätze

Die Trainingsdaten sind dabei wie folgt organisiert: In den annotierten Daten ist ein Kürzel für die vorliegende Sprache, samt dem Grundlemma und der morphologischen Form gegeben. Das flektierte Wort selbst dient in den Trainingsdaten als Label für den Datensatz.

Für den direkten Vergleich miteinander werden eine Reihe an neuronalen Netzwerken trainiert, die in jeweils unterschiedlicher Anzahl die Trainingsdaten mischen und auf den jeweils gleichen Test und Development-Daten evaluiert werden sollen. Durch diese Gegenüberstellung sollen nicht nur die Unterschiede in der Leistungsfähigkeit der Netzwerke nach den obigen 3 Paradigmen miteinander verglichen werden, sondern auch ein Maß für die Steigerung der Performanz der Netzwerke nach unterschiedlicher Größe der Trainingsdatensätze einsehbar sein.

Als weiteren Schritt in der Evaluation soll nach der statistischen Analyse eine linguistisch motivierte Fehlersuche in den falschen Ausgaben der Netzwerke vorgenommen werden, in welcher bestimmte Fehlertypen identifiziert und nach Möglichkeit im Training der Netzwerke adressiert werden sollen, um dadurch den Morphologie-Task weiter zu verbessern.

Repetitorium Einheit Latex 2

In der Dokumentstruktur von Latex bietet es sich unter anderem an, pro Sektion oder Unterkapitel eine eigene Datei zu verwenden, und diese im Hauptdokument mit den Stylevorgaben zusammenzufügen. Dies erlaubt eine leichtere Zusammenarbeit bei Teamartikeln, oder eine schnellere Übersicht über verschiedene Teile der Arbeit.

Latex erleichtert das Referenzieren innerhalb einer Arbeit. Mit dem Befehl \label kann man einen Querverweis auf eine andere Stelle im Dokument, z.B. ein Unterkapitel, anfügen. Durch \eqref lassen sich Formeln indexieren und referenzieren, \pageref bietet die Möglichkeit bestimmte Seiten direkt zu referenzieren.

Mit Bibtex lassen sich Quellen und Zitationen im Text einfach verwalten. Hierzu wird eine eigene bib Datei benötigt, in welcher Angaben über die Quelle festgehalten werden. Zur Verwaltung von Literatur können auch Tools wie Jabref verwendet werden. Zitieren im Text funktioniert über die Kommandos \citet als eingebundenes Zitat im Text, oder \citep in Parenthese.