Protokolle zur Sitzung vom 15.05.2017 - Computerlinguistisches Arbeiten

18.05.2017

Drei Studenten stellten in der Sitzung am 15.05.2017 mit Hilfe einer Präsentation mit dem Beamer ihre Bachelorarbeitsthemen vor.

1. Präsentation:

Referent: Tobias Eder
Titel: Exploiting bilingual word embeddings to establish translational equivalence
Betreuer: Alexander Fraser, Fabienne Braune

Im Vortrag von Tobias Eder geht es darum, zu unteruschen ob ein Wort auch in anderen Sprachen ähnlich im Vektorraum dargestellt wird. Tobias begann mit dem der Motivation und dem Ziel seiner Arbeit: Die Übersetzung ohne Wörterbuch. Zum allgemeinen Verständnis seiner Arbeit gab Tobias eine Einführung in das Thema Word Embeddings. Durch Word Embeddings kann man semantische und syntaktische Ähnlichkeit von Wörtern darstellen.
Die technische Umsetzung seines Themas besteht darin, zu untersuchen, ob dasselbe Wort in verschiedenen Sprachen im Vektorraum ähnlich positioniert ist. Dafür verwendet er zwei verschieden Programme, welche Word Embeddings berechnen. Das erste Programm mit dem er arbeitet heißt 'Word2Vec'. Dieses Programm arbeitet mit zwei verschiedenen Möglichkeiten das Wort darzustellen. Zum einen mit dem Skipgram-Modell, welches den Kontext gegeben einem Wort vorhersagt, zum anderen das CBOW (Continous Bag-of-words Modell), welches ein Wort gegeben dem Kontext vorhersagt. Das zweite Programm heißt 'fasttext', welches mit Buchstaben n-Grammen arbeitet und versucht unbekannte Wörter mit Hilfe ähnlicher n-Gramme zu berechnen.

Tobias stellt anschließend die ihm zu Verfügung stehenden zweisprachigen Korpora und den Experimentenaufbau vor. Er verwendet vier verschieden große Korpora: Einen Wikipediatext, ein Medizinisches Korpus, ein Korpus mit phamazeutischem Inhalt und ein kleineres Korpus, welches gesprochene Sprache beinhaltet. Von Tobias aktuell ausgeführte Stichproben des Wikipediatexts ergaben dabei eine korrekte Übersetzung von 95%.
Der Aufbau seines Experiments ist folgendermaßen gegliedert:
Er verwendet einen kleinen Korpus, ca. 5000 Wörter, und eine Auswahl von ca. 1000 hochfrequenten Wörtern, die nicht im Korpus stehen. Dann bildet er die Wörter mit einem Regressions-Modell ab. Seine Auswertung ergibt unterschiedliche Perfomance der Modelle.

Anschließend stellt Tobias seine nächsten Schritte seiner Arbeit vor. Diese bestehen darin bessere Regularisierungen und Abbildungen zu finden und die Wörter mit 'fasttext' zu evaluieren.

2. Präsentation:

Referent: Joseph Birkner
Titel: Ranking With Neural Network Derived Document Vectors
Institut Informatik

Der zweite Vortrag von Joseph Birkner in Zusammenarbeit mit dem Institut für Informatik handelt von der Darstellung von Dokumenten im Vektorraum.

Joseph vermittelte hierfür Hintergrundwissen zu Information Retrieval. Grunsätzlich geht es um einen Nutzer, welcher einen Informationsbedarf hat. An das System wird dann eine Anfrage gestellt, welches mittels Ranking Algorithmus und Zugriff auf eine Datenbank die Information zurückliefert. Dabei stellte er das Projekt 'IROM' , welches versucht Online-Kurs-Dokumente zu erkennen und dem Nutzer zu liefern und den Begriff 'Vertical Search' vor, welcher die Suche innerhalb einer spezifischen Domäne definiert. Als Motivation für die Arbeit nennt er die Optimierung von Repräsentationen im Vektorraum und die Kodierung von Dokumenten. Laut Joseph geschehen traditionelle Dokumentenrepräsentationen durch ein „TF-IDF“-Modell. Bei diesem Modell wird durch eine Matrix eine Zahl aller möglichen Terme (TF) und die inverse Dokumenthäufigkeit erstellt (IDF).

Für die Darstellung von Wörtern im Vektorraum stellt Joseph das Programm 'Word2Vec' vor. Word2Vec ist ein neuronales Netzwerk, auch Autoencoder genannt, welches das nächste Wort gegeben eines Kontextes vorhersagt. Das Ziel seiner Arbeit ist jedoch die Darstellung von Wörtern, sondern von Dokumenten im Vektorraum, also 'Doc2Vec'. Dabei verwendet er ein sehr großes Dokumenten Korpus um einen 'sequence2sequence' Autoencoder zu trainieren, mit welchem er wiederum Feature Vektoren von Dokumenten extrahiert - das Konzept hierbei: Das Dokument wird dadurch charakterisiert, in wie weit es vom Durchschnitt abweicht.
Zur Veranschaulichung seiner bereits erledigten Arbeit zeigt Joseph die aktuelle Darstellung einiger domänenspezifischer Dokumente in einem 2-dimensionalen Vektorraum. Er greift dabei einige interessante Punkte heraus und erklärt Auffälligkeiten.

Seine weiteren Schritte sind das Trainieren des Autoencoders, mittels unterschiedlicher Kombinationen der Trainingsdaten ; die Erzeugung eines APIs (Programmierschnittstelle) und die Evaluierung der Ranking Performance und der verwendeten Features.

3. Präsentation

Referent: Kristina Smirnov
Titel: Comparison of transfer methods for low-resource morphology
Betreuer: Katharina Kann

Im letzten Vortrag dieser Sitzung von Kristina Smirnov geht es darum, was man machen kann, wenn es zu wenig Sprachdaten gibt. Kristina zeigt zu Beginn mögliche Antworten auf diese Problemstellung, und zwar die Verwendung einer ähnlichen Sprache und die Verwendung von nicht annotierten Daten der gleichen Sprache. Die Evaluierung dieser Lösungsansätze erfolgt auf Basis der korrekt erzeugten morphologischen Formen.

Kristinas Aufgabe besteht darin, gewisse Datensätze zu kombinieren um oben genanntes Problem zu lindern. Ihre Aufgaben dabei sind die Kombination von annotierten russischen Daten mit annotierten ukrainischen Daten, die Kombination von annotierten russischen Daten mit nicht annotierten russischen Daten und die Kombination aller drei genannten Datensätze.

Katharina Kann & Hinrich Schütze sind die Entwickler des hier verwendeten Modells: 'MED: The LMU System for the SIGMORPHON 2016 Shared Task on Morphological Reinflection'. Hierbei handelt es sich um einen morphologischen Encoder-Decoder, welcher laut Kristina relativ gut arbeitet, selbst wenn es nur wenig Datenmengen gibt.

Als nächstes erklärt Kristina die Daten mit denen sie arbeitet genauer. Es handelt sich hierbei um 'CoNLL-SIGMORPHON 2017/2016 Shared Task'. Das Format der Daten ist wie folgt aufgebaut: Lemma - Zielform - morphosyntaktische Beschreibung. Sie verwendet 50 'low ressource' Datensätze für russisch, kombiniert mit 50 'high ressource' Datensätze für ukrainisch. Danach werden 200 'HR' Datensätze verwendet usw. bis zu 12.800. 'LR' Datensätze bis 200. Für jede Kombination erhält man: Training Set für 'source' und 'target', Development Set, Test Set und ein Vokabular für jedes Training Set.

Als letzten Punkt stellt Kristina Ergebnisse und Evaluation vor. Ihre Ergebnisse möchte sie in einer graphischen Präsentation vorstellen. Die Evaluation soll Fragen wie: welche Art von Fehlern treten auf? Gibt es ein gewisses Muster, welches dabei hilft das System zu verbessern?