Protokoll zur Sitzung vom 15.05.2017

Kurzvorstellung von drei Bachelorarbeit-Themen

(1) Tobias Eder, BA-Betreuer: Dr. A. Fraser

Exploiting Bilingual Word Embeddings to Establish Translation Equivalence

Tobias hat in der heutigen Stunde seine Bachelorarbeit vorgestellt, der Grundgedanke deren eine zweisprachige domain-abhängige Übersetzung ohne Wörterbuch ist. Es wird mit so einer Methode, wie Word Embedding erreicht.

Die Information wird verschieden dargestellt, und zwar als ein Audio, ein Bild oder ein Text. Erste zwei Formate präsentieren die Information ziemlich dicht, das letzte dafür sehr dünn. Als Word Embeddings wird eine Darstellung bekannt, bei der ein Wort mit einem Vektor versehen ist. Ist der Kontext eines Wortes ähnlich zum Kontext eines anderen, dann haben diese Wörter semantische und syntaktische Gemeinsamkeiten und dann befinden sich die Wort-Vektoren sehr nah zueinander im so genannten Vektor-Raum-Modell. Das Vektor-Raum-Modell ist sehr hochdimensional, es gibt aber viele Möglichkeiten, die Anzahl der Dimensionen zu reduzieren (z.B.: Clustering).

Es wurden zwei Algorithmen erwähnt: Word2Vec und fastText. Word2Vec wurde vom Google im Jahr 2013 veröffentlicht und stützt sich auf CBOW- und Skipgram-Modelle. Das CBOW-Modell betrachtet den ganzen Kontext und schaut, welche Wörter in dem Kontext aufgetreten sind. Das Skipgram-Modell verwendet das einzelne Wort und analysiert, in welchem Kontext es aufgetreten wurde, um den Kontext vorherzusagen. Der andere erwähnte Algorithmus, fastText, wurde vom Facebook Research im Jahr 2016 veröffentlicht. Es ist ein Word-Representation-Learning, das Subword-Information verwendet. Tobias hat auch nicht-dimensionale Predictive Models präsentiert, indem er demonstriert hat, wo sich die gleichen Wörter aus verschiedenen Sprachen auf einer linearen Abbildung befinden. Das Ganze wurde durch lineare Regression gemacht.

Das Experiment wurde für Deutsch und Englisch durchgeführt. Dafür hat Tobias vier verschiedene parallele Korpora verwendet: General (ca. 50M Tokens), Medical Big (ca. 50M Tokens), EMEA (ca. 4M Tokens) und TED Talks (ca. 2M Tokens) und unterschiedliche Embeddings (CBOW, Skipgram). Zu jedem Korpus wurde noch ein kleiner paralleler Korpus vorbereitet mit ca. 5000 Wörtern jeweils. Für diese Wörter hat man eine Übersetzung, die nicht manuell gemacht wurde. Als erster Schritt werden für ca. 1000 hochfrequente Wörter (die nicht im parallelen Korpus sind und für die es keine Übersetzung gibt) die Abbildungen gesucht anhand Regressionsmodells. Weitere Schritte, die von Tobias noch nicht durchgeführt wurden, sind folgendes:

- Suche nach Abbildungen für niedrigfrequente Wörter;
- Suche nach besseren Abbildungen;
- Verwendung von anderen Regularisierungsmethoden;
- Evaluation auf Wörtern, die es nicht in Korpora gibt.