Klyueva Olga
CIS, LMU München
15.05.1917

Den ersten Vortrag des Tages hielt Tobias Eder über das Thema “Exploiting bilingual word embeddings to establish translational equivalence“. Die Bachelorarbeit wird von Herrn Dr. Fraser unterstützt.
Die Motivation dafür ist eine Übersetzung eines Wortes ohne Wörterbuch in eine andere Sprache.

Zuerst wurde es über die Begriffe word embeddings und bilingual embeddings gesprochen.
Unter word embeddings versteht man die Repräsentation von Wörtern aus Texten in einem Vektorraummodell. Pro Wort wird ein Vektor gefunden.
Unter bilinguale embeddings versteht man eine Annäherung an eine Lineare Abbildung von einem Vektorraum zu dem anderen, von einem Wort einer Sprache zu dem Wort in der anderen Sprache, wie es bei einer Übersetzung gemacht wird.

Es wurde die wichtigste Annahme für einen Vektorraum erwähnt: eine Abbildung von einem Wort in einem bestimmten Vektorraum hält zu anderen Vektoren syntaktisch und auch semantisch Zusammenhänge aufrecht. Das heißt, wenn der Kontext von einem bestimmten Wort ähnlich zu dem anderen Wort ist, dann gibt es syntaktisch oder semantisch Ähnlichkeit zueinander.

Als nächstes wurde über Vektormodelle gesprochen. Es ist ein hochdimensionaler Vektorraum. Es wurde gezeigt mit einem Beispiel, dass wenn es zweidimensional reduziert wird, kann man sehen, dass semantisch ähnliche Wörter oder verwandte Wörtern sich clustern lassen. Durch dieses Clustern kann man anhand geringer Distanz von einem Wort zu dem anderen feststellen, welche Ähnlichkeiten vorliegen.

Im Rahmen der Arbeit wurden zwei Vektormodelle betrachtet: Wort2Vec (von Google) und FastText.
Es wurde genauer über diese Modelle gesprochen. Zu Wort2Vec Vektormodelle gehört Word – Embedding Toolkit, die zwei verschiedene Modelle hat, CBOW (continuous bag of words Modell) und Skipgram Modell. Das erste nimmt einen bestimmten Kontext und versucht vorherzusagen, welches Wort an der Stelle rein gehört. Skipgram Modell nimmt ein bestimmtes Wort und sagt im Voraus in welchem Kontext gerade dieses Wort aufgetreten ist.

Zweites Tool ist FastText, das identisch mit CBOW ist. Trotz morphologischer Differenzen bei den Wörtern kann man Wörter mit selbem Stamm finden.

Dann kommt es zu der Frage, wie das Korpus aussieht und zum Aufbau von einem Experiment und Evaluation.
Es gibt vier parallele Kopula: General (110M Token), Medical Big (50M Token), EMEA (4M Token), TED Talks (2 M). Es handelt sich um eine Englisch-Deutsch Übersetzung.

Der Experimentaufbau ist folgender: man nimmt unterschiedliche embeddings von CBOW und Skipgram Modelle, mit ähnlichen Parametern. Es wurden pro Korpus etwa 5000 übersetzte Wörter betrachtet. Es wurde festgestellt, dass die Übersetzung mit General zu 95 Prozent genau ist, mit EMEA zu 78 Prozent übereinstimmt.

Zunächst wird eine Auswahl von 1000 hochfrequenten Wörtern aus einem Korpus, die nicht im parallelen Korpus auftreten, betrachtet. Hiervon wird die Lineare Abbildung von einem Vektorraum zum anderen gesucht. Die Lineare Abbildung wird durch eine Lineare Regression gebildet. Das heißt, anhand der Matrix wird ein richtiges Wort an der Stelle gefunden.

Für jede Domain wird ein Testset mit 1000 hochfrequenten Wörtern erzeugt und manuell überprüft, ob die Übersetzungen stimmen.

Es werden die unterschiedlichen Vektormodelle verglichen und Abbildungen in Vektorräume untersucht. Es wird eine Antwort auf die Frage gesucht, wo ist der Unterschied zwischen CBOW und Skipgram Modellen, welches Modell je nach Domain welche Resultate liefert.
Es wäre zu erwarten, dass das Skipgram Modell bessere Resultate für eine große Anzahl von Tokens beispielsweise für General liefert, während das CBOW Modell sich besser für Medical Big eignet.

Am Ende des Vortrags wurden einige nächste Schritte in der Arbeit erwähnt: als erstes wird mit niederfrequenten Wörtern gearbeitet. Man versucht mit den gleichen Abbildungen ins Deutsche zu übersetzen und zu untersuchen, wie gut es funktioniert.

Es wurden auch andere Regularisierungsmethoden gesucht, weil beispielsweise Jahreszahlen nicht richtig übersetzt wurden (Jahreszahl 1980 wird mit Jahreszahl 1933 übersetzt). Außerdem wird versucht, die Wörter, die nicht im Kopula auftreten, übersetzen zu lassen.