Karlsruher Institut für Technologie
Institut für Angewandte Informatik und Formale Beschreibungsverfahren (AIFB)
Tobias Weller
tobias.weller@kit.edu

29.11.2018

Abschlussarbeit

„Time-aware Word Embeddings“ zu vergeben!

Die strukturierte Repräsentation von Wörtern und Dokumenten ist im Bereich des Natural Language Processing (NLP) für die weitere maschinelle Analyse und Verarbeitung von großer Bedeutung. Word Embeddings werden u.a. im Bereich der Sentiment Analysis und Link Prediction eingesetzt. Dabei ist für die Bedeutung von Wörtern der Kontext, in dem das Wort genannt wird, relevant. Bspw. „Bank“. In dem Kontext „Die Bank erhöht die Zinsen“, ist die Bank als Geldinstitut zu verstehen. In dem Kontext „die Enten von der Bank aus füttern“ ist die Bank als Sitzmöglichkeit zu verstehen. Obwohl es sich um dieselben Wörter handelt, ist ihre Bedeutung Kontextabhängig. Bisherige Verfahren für die Erstellung von Word Embeddings wie bspw. Word2Vec oder auch die Erweiterung GloVe können diesen Zusammenhang bereits sehr gut repräsentieren. Allerdings sind Wörter nicht nur Kontextabhängig, sondern auch Zeitabhängig. Die Bedeutung von Wörtern kann sich über die Zeit hinweg verändern. So wurde das Wort „apple“ früher im Kontext eines Obsts angesehen, wohingegen es heute auch im Kontext des Unternehmens Apple Inc. steht. Ähnliches gilt für „amazon“.

Vor diesem Hintergrund ist es das Ziel dieser ausgeschriebenen Arbeit Word2Vec oder verwandte Verfahren um die Einbeziehung einer zeitlichen Komponente zu erweitern. Eine erste Arbeit in diesem Bereich stellt Dynamic Word Embeddings dar[1]. Dieses Verfahren dient als Vergleichsmodell für das eigen erstellte Verfahren. Gelernt werden soll auf verschiedenen öffentlichen Datensätzen, sowie auf einen Korpus deutscher Abiturschriften (1909 - 1976). Ebenso ist es denkbar das Verfahren Themenbezogen auszuweiten. Hierbei soll eine strukturierte Repräsentation von Themen erstellt werden, welche den zeitlichen Kontext einschließen um die zeitliche Entwicklung von Themen darzustellen.

Mögliche Aufgabenbereiche umfassen (sind aber nicht begrenzt auf)**:**

- Erstellung von Word Embeddings unter Beachtung der zeitlichen Komponente.
- Erstellung von Themenbezogenen Embeddings unter Beachtung der zeitlichen Komponente.
- Sentiment Analysis auf Themen unter Beachtung der Zeit.

Das sollten Sie mitbringen:

- Gute Kommunikationsfähigkeiten
- Strukturiertes sowie organisiertes Denken
- Interesse an semantischen Technologien und statistischen Verfahren
- Sehr gute Kenntnisse in Maschinellen Lernverfahren
- Sehr gute Python Kentnisse

Sie haben Interesse?
Dann schicken Sie eine E-Mail mit Anschreiben, Lebenslauf, und aktuellem Notenauszug.

Kontaktperson:
Tobias Weller
tobias.weller@kit.edu
Tel.: 0721/60845770

[1] https://arxiv.org/pdf/1703.00607.pdf