

Bachelorarbeit Proposal v1

Clemens Damke

8. Dezember 2016

1 Problemstellung

Im Rahmen menschlicher Kommunikation entstehen zahlreiche Informationen, deren Auswertung nicht immer trivial möglich ist. Ein manuelles Auswerten ist aufgrund der Datenmenge in der Regel nicht möglich. Maschinelle Analysen hingegen werden durch die Komplexität natürlicher Sprachen erschwert. Oft wird das Potential bereits existenter Daten daher nicht voll ausgeschöpft.

In diesem Kontext soll untersucht werden, wie aus einer Menge an gegebenen Kommunikationsdaten Relationen zwischen Entitäten extrahiert werden können.

Entitäten sind z. B. Personen, Organisationen oder Orte. Das gesuchte Verfahren soll allerdings flexibel genug sein, auch eine beliebige Menge anderer Entitätsklassen erkennen zu können.

Eine Relation kann z. B. die Freundschaft zwischen zwei Personen oder die Anwesenheit einer Person an einem Ort sein. Auch hier gilt, dass das gesuchte Verfahren flexibel genug sein soll eine beliebige Menge anderer Relationsklassen erkennen zu können.

Gefundene Entitäten und Relationen, die die selben realen Entitäten bzw. Relationen repräsentieren, sollten, so gut wie möglich, als Duplikate erkannt und unifiziert werden.

Abhängig vom Kommunikationspartner kann z. B. die Person „Edward F.“ sowohl durch das Wort „Eddi“ als auch durch das Wort „Papa“ referenziert werden.

2 Anforderungen

Gegeben ist ein Stream von Nachrichten, denen ein Absender, ein textueller Inhalt, eine Menge von Empfängern und ggf. weitere Metadaten, wie z. B. Absendeort, Absendezeit oder benutztes Endgerät, zugeordnet sind.

Ziel ist es aus diesem Nachrichtenstrom einen probabilistischen Wissensgraphen zu konstruieren, der alle Nachrichten und die darin enthaltenen Entitäten als Knoten enthält. Das Verfahren muss inkrementelle Updates des Graphen unterstützen, da laufend neue Nachrichten hinzukommen.

Bei jeder eintreffenden Nachricht muss also ein NLP (natural language processing), insbesondere eine NER (named entity recognition), der Nachricht erfolgen. Relationen zwischen Entitäten innerhalb der Nachricht können ebenfalls bereits erkannt werden. Hierfür kann und soll eine bestehende Lösung NLP verwendet werden.

Die so gefundenen lokalen Entitäten und Relationen müssen anschließend in den bestehenden Wissensgraphen eingefügt und mit im Graph bestehenden Entitäten in Relation gebracht werden. Den so gefundenen Relationen, soll dabei ein Konfidenzwert zugeordnet werden. Ein Verfahren hierfür soll gefunden werden. Zu prüfen ist konkret, ob Distanzen und Richtungen zwischen Wortvektoren in einem mittels word embedding (z. B. word2vec) konstruierten Vektorraums geeignet sind.

Das gefundene Verfahren soll implementiert und mit realen Datensamples getestet werden. Idealerweise wird die Lösung zudem auch im Hinblick auf ihre Skalierbarkeit entworfen.

3 Verwandte Arbeiten