Protokoll von 03.07.2017

Präsentation von Pascal Guldener

Pascal hat eine Präsentation zu dem Thema “Neural Networks for Named Entity Recognition” gemacht. Das Thema basiert sich auf dem Paper von Ronan Collobert, welcher geht um SENNA (2011). Das ist semantische oder syntaktische Extraktion mit der Hilfe von Neural Network Architecture. Die Motivation dazu ist die Lösung der verschiedenen NLP Probleme mit einem System ohne linguistische Vorkenntnisse, das wird im Named Entity recognition (NER), Part-of-Speech-tagging (POS), Chunking und Semantic Role Labeling (SRL) benutzt. Der traditionelle Ansatz nutzt das vorhandene Wissen darüber, wie die Spracharbeit eine bestimmte Aufgabe mit der Hilfe von Engineering-Task-spezifische Features lösen könnte, die Suche nach Zwischen-Darstellungen, indem sie die Ausgabe von bestehenden Systemen verwenden wurden, und diese optimieren, um die Performance zu verbessern. Generell, linguistische Features sind ziemlich teuer für die Menschen um sie zu entdecken und kodieren, auch Feature-Engineering trägt nicht dazu bei, allgemeinere Datenstrukturen zu finden, die die Bedeutung von einem Text beschreiben. Dazu gabt es zwei Abbildungen zu dem SENNA Architecture mit zwei Ansätzen-Windows.
Ein Wort wird durch einen d-dimensionalen Vektor von Reals repräsentiert, wobei d ein Hyperparameter ist, der optimiert werden soll aber bei allen Iterationen bei 50 bleiben wird. Senna trainiert diese aus einer zufälligen Initialisierung in dem ersten Schritt. In einem zweiten Schritt verbessert Senna seine Embeddings in einer unbeaufsichtigten Weise. Die erste Schicht des Netzwerks wandelt jedes Wort der Eingabe in einen Index um, der dem Lookup Table Layer zugeführt wird. Alles wird mit einer Formel berechnet und den Windows Ansatz wird verwendet.
Für das Training wurde ein Dictionary mit den häufigsten 100000 Wörtern aus Wall Street Journal (Trainingsdaten für die Benchmarks) benutzt. Für ein Preprocessing würde folgendes durchgeführt: Lowcasing aller Wörter, Hinzufügen von "caps" Features (wasupper, hadfirstupper) und Ersetzen von Sequenzen von Zahlen mit dem "NUMBER" String. Für das Training wurde es eine Stunde für NER gebracht, für POS einige Stunden und drei Stunden bei SRL. Für überwachten Benchmark wurden die Resultate für “out-of -the-box “ Neural Networks gegeben: für die NER Task 89,31%, für POS - 97% und für Chinking und SRL sind 94% und 77% entsprechend.
Die Analyse hat gezeigt, dass die Probleme bei den Embeddings liegen. Offensichtlich gibt es kaum semantische Ähnlichkeit. In Wall Street Journal 15% der Worte erscheinen etwa 90% Male, es führt zu Spärlichkeit in den verbleibenden 85%, und das führt zu eher schlechten Embeddings, deswegen sollen die Embeddings getrennt trainiert werden.
Es gibt auch nicht-überwachtes Training der Sprachmodelle. Im LM1 wurde englische Wikipedia mit den häufigsten 100000 Wörtern genutzt. LM2 Modell mit LM1 plus Reuters,da ein Wörterbuch plus 30000 häufigsten Wörter von Reuters. Folgendes Preprocessing wurde gemacht - entfernen von Wikimarkup, Tokenizing und Absätzen, die nicht-lateinische Zeichen erhalten.
Zusammenfassend kann man sagen, dass die Idee, dass Informationen über die Worte zu nutzen, die für eine Aufgabe relevant sind, ist auch relevant für damit verbundene Aufgaben. Das kann das Training viel effizienter machen.