

Named-Entity Recognition für die Mobiliar

Studiengang: BSc in Informatik | Vertiefung: Computer Perception and Virtual Reality
Betreuer: Prof. Dr. Jürgen Vogel
Experte: Pierre-Yves Voirol (Abacus Research AG)
Industriepartner: Schweizerische Mobiliar Versicherungsgesellschaft AG, Bern

Mit der Hilfe von Künstlichen Neuronalen Netzwerken (KNN) sollen personenbezogene Daten automatisch aus Texten erkannt werden. Das finale Modell soll die Schweizerische Mobiliar unterstützen, Personen und Adressen klassifizieren zu können. Die Bachelorarbeit dringt dabei in die Bereiche Informatik, Künstliche Intelligenz sowie Computerlinguistik vor.

Die Schwierigkeit bei der Verarbeitung von natürlicher Sprache besteht vorwiegend in der Mehrdeutigkeit von Begriffen. Das Wort «Orange», das sowohl als Farbe, Frucht, französische Kleinstadt oder als ehemaliger Netz-Provider interpretiert werden kann, verdeutlicht dies. Die Rechtschreibung und Vielfältigkeit der Sprache erschweren eine Texterkennung zusätzlich.

Ausgangslage

Mit Mobi24 hat die Mobiliar Versicherung eine 24h-Anlaufstelle für die Anliegen ihrer Versicherungsnehmer geschaffen. Die dabei anfallenden Daten sollen nun genutzt werden, um ein Neuronales Netzwerk zu trainieren. Die fertige Named-Entity Recognition (NER) soll die Data Scientisten der Mobiliar bei ihrer täglichen Analysearbeit unterstützen. Es vereinfacht die Weiterverarbeitung von Texten durch Reduktion aller Named-Entities durch definierte Platzhalter. Zudem kann eine NER für die automatisierte Anonymisierung von Texten verwendet werden.

Vorgehen

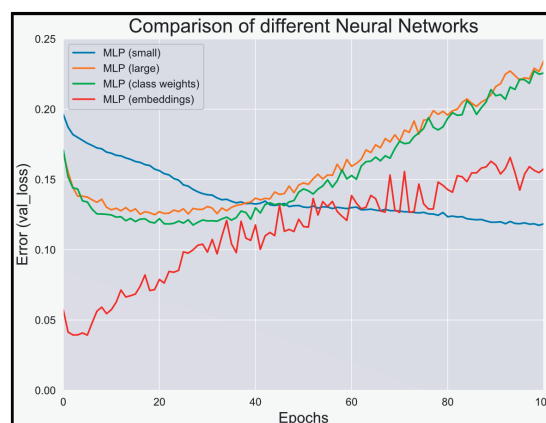
Für das Trainieren eines Neuronalen Netzwerks benötigt es gelabelte Daten. Daher markierte ich in über 1'000 Nachrichten Namen und Adressen. In der Praxis ist es häufig üblich, dass man zuerst ein Baseline-Modell erstellt, welches keine KI-Techniken verwendet. Mein Baseline basiert auf einfachen Wörterbüchern und Regular Expressions, erkennt Named Entities zu 78% und soll als unterer Schwellwert die minimalen Anforderungen stellen. Das Neuronale Netz wurde mit dem Machine Learning Framework Tensorflow erstellt und iterativ weiterentwickelt. Das finale Netz besteht aus zwei Hidden-Layern mit jeweils 32 Neuronen. Als Input wurden die Nachrichten in Textsequenzen von 9 Wörtern unterteilt und in Wortvektoren (Embeddings) umgewandelt. Embeddings repräsentieren Wörter als hoch-dimensionale Vektoren. Sie erlauben, dass Wörter mit ähnlicher Bedeutung auch im Vektorraum ähnlich dargestellt werden.

Resultate

Die Performance des Neuronalen Netzwerks wurde mit 280 Mails aus dem Mobi24 Kundendienst verifiziert. Mit einer **Trefferquote von 95.3%** beim Erkennen von Named-Entities bin ich sehr zufrieden. Dieser Wert ist jedoch mit Vorsicht zu genießen. Durch die geringe Menge an Trainingsdaten schneidet das Netz nur bei Email-Nachrichten so gut ab. Bei anders strukturierten Texten liegt die Trefferquote weit darunter. Wie an der untenstehenden Grafik zu erkennen ist, tragen die Embeddings am meisten zur Performance bei. Wenn die Fehlerrate nach einer gewissen Anzahl Durchläufen wieder zunimmt, spricht man von Overfitting. Das tritt ein, wenn sich das Netzwerk zu stark an die Trainingsdaten anpasst. Dabei lernt es die Daten auswendig, hat später aber Probleme beim Klassifizieren von neuen, noch ungesehenen Daten. Es ist erstaunlich, wie das Netzwerk aus den wenigen Daten bereits generalisierte Annahmen treffen kann. Der Baseline-Score wurde damit bei weitem übertroffen. Probleme bereiten dem Modell vor allem Mehrdeutigkeiten - in «Bäckerei Müller» wird «Müller» bspw. als Name erkannt, obwohl es streng genommen Teil einer Organisation wäre. Dieser Fehler ist jedoch nicht ganz trivial und kann auch Menschen unterlaufen.



Yves Alain Beutler
yves.beutler@hotmail.com



Fehlerrate verschiedener Neuronaler Netze inkl. finaler Version (rot)