Zürcher Hochschule
für Angewandte Wissenschaften

School of Engineering

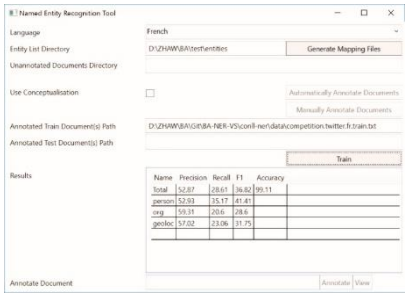
InIT Institut für angewandte Informationstechnologie

Named Entity Recognition mit Deep Learning mit wenig Daten

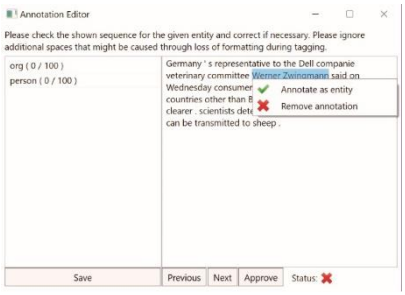
Diplomierende
Stefano Dolce
Nicole Falkner

Dozent
Mark Cieliebak

In dieser Arbeit wurde ein Named Entity Recognition Tool erstellt, welches anhand einer Liste von Entitätsnamen Dokumente automatisch annotiert und ein neuronales Netzwerk trainiert. Dazu wurden verschiedene Ansätze wie Transfer Learning, ein Netzwerk pro Entität zu trainieren und das Verwenden von teilannotierten Daten überprüft. Als Basis für diese Arbeit diente eine Kombination aus einem Bidirectional Long Short Term Memory Network und einem Convolutional Neural Network. Dabei wurde das Long Short Term Memory Network für die Wort-Ebene und das Convolutional Neural Network für die Zeichen-Ebene verwendet.
Ein weiterer Teil dieser Arbeit war die Teilnahme am CAp 2017 Wettbewerb für Named Entity Recognition auf französischen Tweets. Das finale System war für den Wettbewerb noch nicht verfügbar, aber mit dem Transfer-Learning-Ansatz wurde der 5. Platz erreicht mit einem F1-Score von 50.05.
Transfer Learning kann das Resultat eines neuronalen Netzwerks verbessern und ist insbesondere in Kombination von teilannotierten Daten interessant. Auch das Trainieren eines eigenen Netzwerks pro Entität verbessert das Resultat. Eine Kombination mit Transfer Learning bildet die Basis für das vorgeschlagene Named Entity Recognition Tool. Dabei dienen teilannotierte Daten als Quelle für das Transfer Learning auf die eigentliche Zieldomäne, wobei jede Entität einzeln trainiert wird. Da aber auf jeden Fall annotierte Daten notwendig sind, bietet das Tool eine Oberfläche, um die automatisch annotierten Daten zu korrigieren.
Das beste System basiert darauf, dass von mehreren Quellen in teils unterschiedlichen Sprachen ein Transfer Learning pro Entität angewandt wird. Um gute Resultate zu erhalten, wurden pro Datensatz insgesamt 3000 manuell annotierte Sequenzen verwendet. Auf den französischen CAp-2017-Daten wird mit Transfer Learning von dem CoNLL-2003-Datensatz und teilannotierten französischen Tweets ein F1-Score von 60.78 auf den Entitäten «geoloc», «person» und «organisation» erreicht. Auf englischen Newstexten mit Transfer Learning von teilannotierten englischen Newstexten wird für dieselben Entitäten ein F1-Score von 85.75 erreicht.



Named Entity Recognition-Netzwerke über eine einfache Benutzeroberfläche trainieren



Automatisch annotierte Texte ansehen und falls notwendig korrigieren