Kurzfassung

Für das Textverstehen domänenspezifischer Texte ist die Named-Entity-Recognition (NER) von zentraler Bedeutung. Für die deutsche Sprache existiert jedoch erst seit kurzem (2018) ein domänspezifischer Entitätendatensatz, der über eine größere Anzahl spezifischer Entitäten verfügt. Auf diesem wurden zwei State-of-the-Art neuronale Netze (BiLSTM und BERT) miteinander verglichen. Die besseren Ergebnisse wurden dabei mit der BiLSTM-Architektur erzielt. Bei der Erkennung einiger Eigennamen-Entitätenklassen war hingegen das BERT-Modell überlegen. Durch einen groß angelegten Hyperparametervergleich wurden Dropout, Variational Dropout, sowie die verwendeten Wort-Embeddings als die wichtigsten Hyperparameter des BiLSTM-Netzes identifiziert. Es wurde gezeigt, dass das Kombinieren von verschiedenen Wort-Embeddings zu einer verbesserten Entitätenerkennung führt. Da die Modellqualität eines für die Entitätenerkennung trainierten Netzes wesentlich von der Verfügbarkeit seiner Trainingsdaten abhängt, wird nach Wegen gesucht, die Entitätenerkennung zu verbessern, ohne dafür mehr annotierte Daten zu benötigen. In dieser Arbeit wurde gezeigt, dass das Wissen von einem neuronalen Netz, welches für die Erkennung bestimmter Entitätenklassen trainiert wurde, das Lernen einer neuen Entitätenklasse vereinfachen kann. Die damit erzielten Verbesserungen waren bei einer kleinen Trainingsdatengröße am größten. Jedoch ergaben die Untersuchungen, dass nicht jede Entitätenklasse gleich stark von dem übertragenen Wissen profitiert. Die größten positiven Effekte wurden dabei bei Eigennamen-Entitäten festgestellt. Die Ergebnisse aller Untersuchungen wurden mit den Roh-Vorhersagen der neuronalen Netze veröffentlicht.

Abstract

Named Entity Recognition (NER) plays a major role in improving text comprehension of domain-specific texts. However, only recently (2018) a German domain-specific entity dataset with a larger number of specific entities was published. Using this dataset two state-of-the-art neural networks (BiLSTM and BERT) were compared. The usage of the BiLSTM-architecture led to overall better results. However, the BERT-model achieved a better recognition of some named entity categories. A large-scale hyperparameter comparison identified Dropout, Variational Dropout and the used word embeddings as the most important hyperparameters of the BiLSTM network. Furthermore, it has been shown that combining different word embeddings results in improved entity recognition. Since the model quality of a network trained for entity recognition depends essentially on the availability of its training data, ways are sought to improve entity recognition without requiring more annotated data. In this work, it has been shown that the knowledge of a neural network trained for the recognition of certain entity classes can simplify the learning of a novel entity class. The greatest improvement was achieved with a small training data size. However, research has shown that not every entity class benefits equally from the transferred knowledge. The greatest positive effects were observed with named entities. Finally, the results of all investigations were published with the raw predictions of the neural networks.