Contribution ID: **223** Type: **Sectional reports**

Machine learning for natural language processing tasks

Thursday 13 September 2018 14:30 (15 minutes)

There are two popular algorithms for text vector extraction: bag of words and skip-gram. The intuition behind it is that a word can be predicted by context and context can be predicted from a word. The vector size of a word is the number of neurons in the hidden layer.
The task of named entity recognition can be solved by using LSTM neural networks. The features for every word can be word-embeddings (skip-gram or bag of words model), char-embeddings features, and additional features, for example, morphological.
To solve this task, we used a tagged dataset (where a human choose which words are entities like a Person, Organization, Location or Product type).
We used the softmax function in a neural network for classification. Also, is possible to use other approaches like CRF. There are many neural architectures for the problem of named entity recognition.
After that, it is possible to teach our model to predict the entities of predefined types.
There are many approaches for text classification, and for vectorization it is possible to use document-embeddings (doc2vec model) or TF-IDF. After this, it is possible to use classification algorithms like an SVM or Random Forest model. To verify the classification task, it is possible to use the most important words in class (for example 20-30 most important words can include the terms which characterize the class).

Summary

This paper explains the basics of using machine learning in natural language processing and describes a neural network architecture for named entity recognition and text classification by topic.

Authors: Mr KULNEVICH, Aleksey (Dmitrievich); Mr RADISHEVSKIY, Vladislav (Leonidovich)

Presenter: Mr KULNEVICH, Aleksey (Dmitrievich)

Session Classification: 11. Big data Analytics, Machine learning

Track Classification: 11. Big data Analytics, Machine learning