



Contribution ID: 155

Type: Sectional reports

Multi-instance learning for Rhetoric Structure Parsing

Friday, 9 July 2021 12:15 (15 minutes)

To accurately detect texts containing elements of hatred or enmity, it is necessary to take into account various features: syntax, semantics and discourse relations between text fragments. Unfortunately, at present, methods for identifying discourse relations in the texts of social networks are poorly developed. The paper considers the issue of classification of discourse relations between two parts of the text. The RST Discourse Treebank dataset (LDC2002T07) is used to assess the performance of the methods. The dataset is a small manually marked up corpus of texts, divided into training and test samples. Since the size of this dataset is too small for training large language models, the work uses a model-prefitting approach. Model prefitting is performed on the reddit news portal user comment dataset. Texts from this dataset are marked up automatically. Since automatic marking is less accurate than manual marking, the multiple-instance learning (MIL) method is used to train models. In the end, the resulting model will be used as part of a text analyzer for detecting elements of hatred or enmity in the texts of social networks. A distinctive feature of modern language models is a large number of parameters. Using several models at different levels of such a text analyzer requires a lot of resources. Therefore, for the analyzer to work, it is necessary to use high-performance or distributed computing. The use of grid systems from personal computers can allow attracting and combining computing resources to solve this type of problem.

This work was funded by RFBR according to the research project No. 21-011-44242

Summary

Primary authors: Mr VOLKOV, Sergey (Peoples' Friendship University of Russia (RUDN University); Federal Research Center "Computer Science and Control" RAS); Mr DEVYATKIN, Dmitry (Federal Research Center "Computer Science and Control" RAS)

Presenter: Mr VOLKOV, Sergey (Peoples' Friendship University of Russia (RUDN University); Federal Research Center "Computer Science and Control" RAS)

Session Classification: Big data Analytics and Machine learning.

Track Classification: 9. Big data Analytics and Machine learning