Development of a method for recognizing biomedical entities in the texts of scientific articles

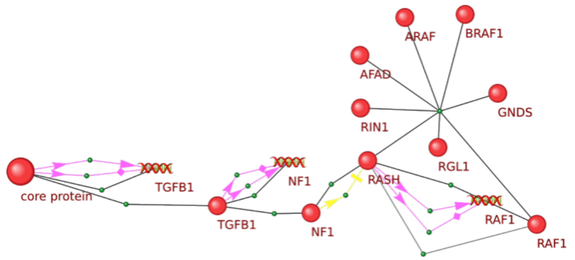
Stepan Derevyanchenko
Novosibirsk State University
Novosibirsk, Russia

Pavel Demenkov
The Federal Research Center Institute of Cytology and Genetics
The Siberian Branch of the Russian Academy of Sciences
Novosibirsk, Russia

Purpose of research

- The purpose of this work is to develop a method for recognizing the object name in the text using machine learning methods and integrating the results into the ANDSystem

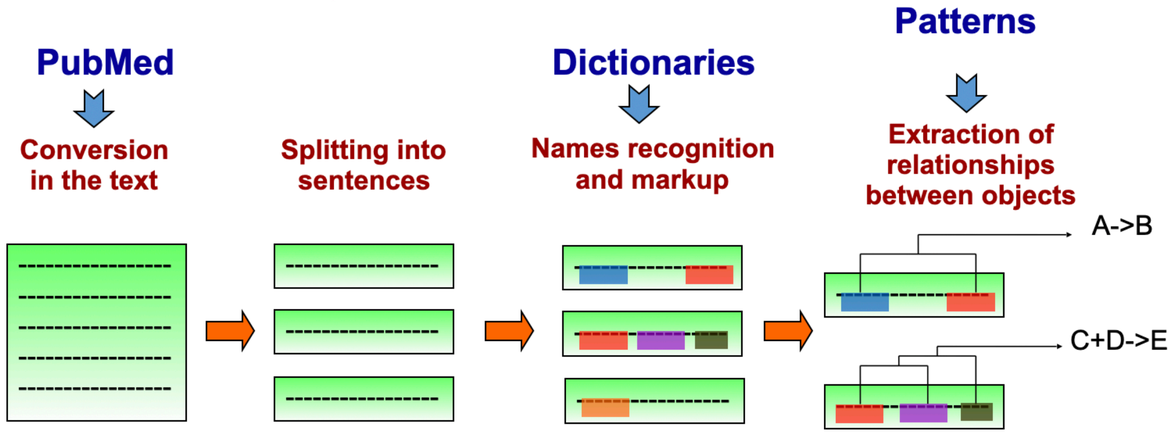
Intellectual search for knowledge in the scientific literature in the field of biology and medicine. AND System



The number of publications in the field of biology, medicine and biotechnology is growing so rapidly that the available information is fundamentally impossible to analyze for research and application purposes without automatic processing using computer tools.

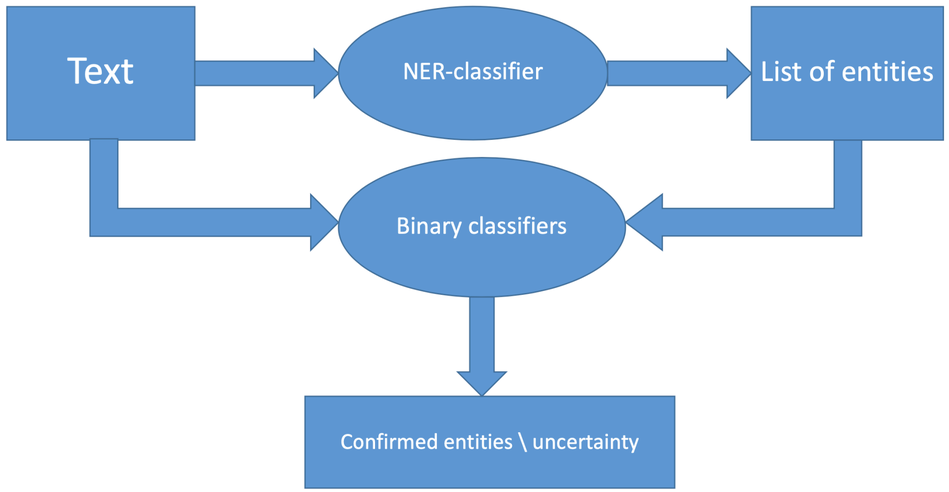
Problem statement

- Create a machine learning model for recognizing named entities in the texts of scientific articles (Named Entity Recognition)
- Perform a comparative analysis of binary classification models and select the most appropriate model for evaluating the confidence measure of the NER prediction
- Integrate the developed models and evaluate the quality of named entity recognition with the integrated machine learning model.

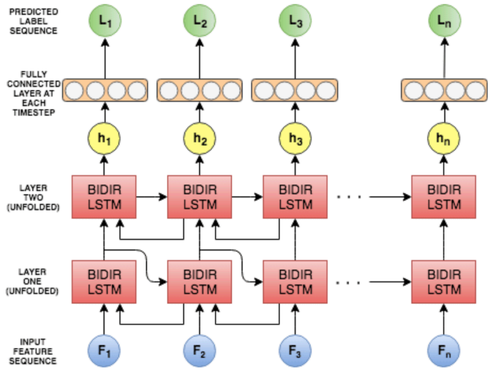


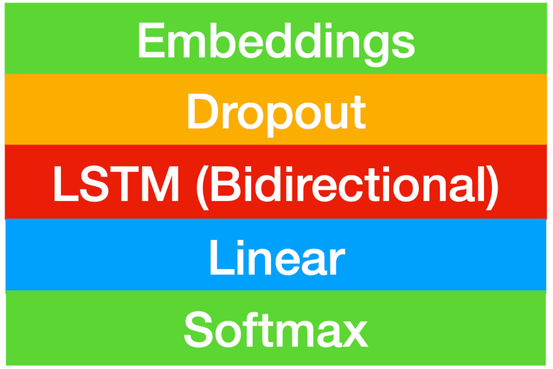
Model architecture

Pipeline

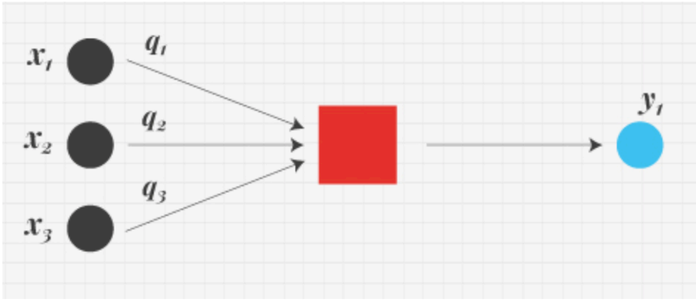


NER





Classification (Logistic Regression)

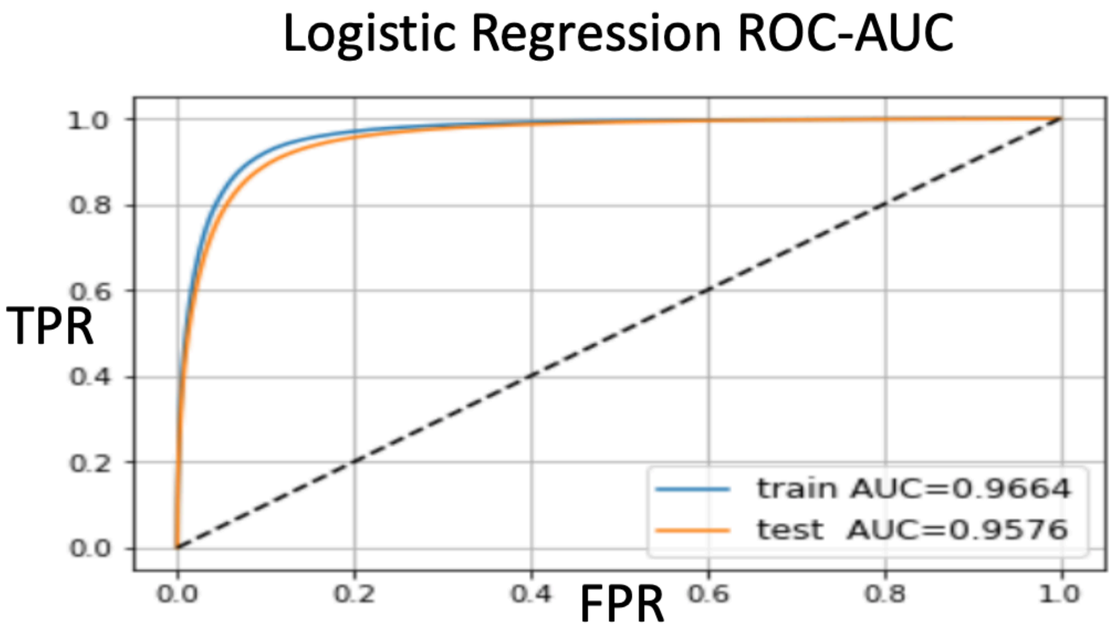


Metrics

NER

```
accuracy:  88.08%; (non-O)
accuracy:  98.17%; precision:  89.60%; recall:  87.61%; FB1:  88.59
             CELLS: precision:  90.30%; recall:  89.46%; FB1:  89.88
         COMPONENT: precision:  86.61%; recall:  80.29%; FB1:  83.33
           DISEASE: precision:  87.49%; recall:  87.73%; FB1:  87.61
              GENE: precision:  84.76%; recall:  67.04%; FB1:  74.87
        METABOLITE: precision:  87.02%; recall:  85.44%; FB1:  86.23
      MOLFUNCTION: precision:  79.53%; recall:  71.61%; FB1:  75.36
         ORGANISM: precision:  97.10%; recall:  96.57%; FB1:  96.83
          PATHWAY: precision:  83.92%; recall:  84.43%; FB1:  84.17
        PHENOTYPE: precision:  77.30%; recall:  68.64%; FB1:  72.71
          PROTEIN: precision:  78.43%; recall:  71.95%; FB1:  75.05
          SEFFECT: precision:  84.26%; recall:  75.18%; FB1:  79.46
```

Classification



Results

- The trained NER model shows the following results for metrics: Accuracy: 98.17%; Precision: 89.6%; Recall: 87.61%; F1: 88.59%
- The analysis of the models showed that Logistic Regression is the best model (by ROC-AUC) for this binary classification problem. This classification gives a measure of confidence in the prediction of NER
- Determining the confidence of a prediction using binary classification reduced the percentage of NER errors to 4.5%