NLP Course Assignment 3: Russian Nested Named Entities

Named entity extraction is one of the most popular information extraction tasks in practice – it involves searching for mentions of names, organizations, toponyms and other entities in the text. This assignment is devoted to the task of extracting nested named entities. Data partitioning allows the following cases: inside one named entity there is another named entity. For example, an entity of the Organization class "Moscow Drama Theater named after M. N. Yermolova" has a nested entity of the Person type "M. N. Yermolova".

Data

The competition is based on the NEREL corpus, collected from WikiNews news texts in Russian. The NEREL corpus contains 29 classes of different entities, and the depth of nesting of entities reaches 6 levels of markup.

Data is provided to participants in the form of marked-up documents. The markup format is BRAT.

Problem statement

The task involves extracting nested named entities. In the training set, most of the named entity types occur quite often, and some number of specially selected types occur only a few times. In the test set, all entity types are equally represented.

Thus, you have to develop extraction models for nested named entities.

How to do this assignment

1. Join the competition in CodaLab. All the further instructions are described there.

   https://codalab.lisn.upsaclay.fr/competitions/18459

2. Try to solve the problem with **at least two solutions**. Find out your best solution.
3. Make a submission to the competition **during the “Test” phase**.

   Notice that the competition has two phases: “Dev” and “Test”. During the “Dev” phase you develop your solution and improve it based on the dev.jsonl data. The next phase “Test” start two days before the competition ends. Don’t forget to submit the results when the “Test” phase starts.

   ONLY F1 SCORE OF THE TEST PHASE WILL BE CONSIDERED IN THE EVALUATION OF YOUR ASSIGNMENT! F1 FEW SHOT SCORE IS INGNORED DURING EVALUATION!
4. Create a GitHub repository and push the code of **all of your solutions** there
5. Compose a report which contains
   a. Your name and surname
   b. University email
   c. **Nickname in CodaLab**
   d. Link to the GitHub repository
   e. Detailed description of how to tried find the best solution to this problem
      - Describe each solution in details
      - Write which solution happened to be the best in terms of F1 score
6. Submit your report to Moodle

Grading Criteria

Total: 10 points

- Code quality: 3 points
  - Code is present: 1 point
  - Each solution is in its own directory in the repository: 1 point
  - Comments: 1 point
- Report quality 4 points
  - A solution is described: 2 point

- More than one solution are described: 1 point
- Best solution is reported and justified: 1 point

- Place on the leaderboard: 3 points
  - Lowest third of the leaderboard: 1 point
  - Middle third of the leaderboard: 2 points
  - First third of the leaderboard: 3 points