Oksana Konovalova

o.konovalova@innoppolis.university

Nickname CodaLab: ksko02

Link to the GitHub repository: https://github.com/ksko02/NLP_course/tree/main/assignment3

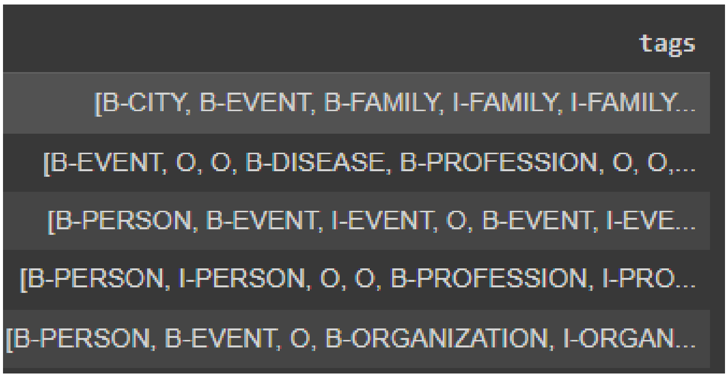
Solution 1

While searching for a solution, I found the Natasha library that can define named entities. Before that I had a chance to try spaCy (baseline.ipynb presents only work with Natasha), I was very interested to see how the Natasha library would handle this task. Оказалось, что библиотека Natasha так же, как и spaCy определяет только три entity types: PERSON, LOCATION, ORGANIZATION. I expected more entity types from this library, as it was originally made for Russian language. For example, spaCy for English has many more entity types (MONEY, DATA, etc.). In general, they work about the same and are not particularly suited to the task at hand, as more entity types need to be defined.

Solution 2

While searching for a solution, I saw that it is possible to make fine-tuning BERT models to solve the NER problem. On huggingface I found the rubert-tiny model (https://huggingface.co/cointegrated/rubert-tiny). It is a small model trained for Russian text. It was poorly trained and consequently had poor metrics. After that I tried the rubert-tiny2 model (https://huggingface.co/cointegrated/rubert-tiny2), an improved version of rubert-tiny. The result improved by 0.14 F1 metric, so it was decided to use the second version.

1. For this solution, I moved entity markup to the word level, using IOB notation to separate multiple entities of the same type in a string.



tags
[B-CITY, B-EVENT, B-FAMILY, I-FAMILY, I-FAMILY...
[B-EVENT, O, O, B-DISEASE, B-PROFESSION, O, O,...
[B-PERSON, B-EVENT, I-EVENT, O, B-EVENT, I-EVE...
[B-PERSON, I-PERSON, O, O, B-PROFESSION, I-PRO...
[B-PERSON, B-EVENT, O, B-ORGANIZATION, I-ORGAN...

IOB notation

2. Used the Dataset library to facilitate the process of tokenization and align labels, which I did through AutoTokenizer.
3. Set up the model and metrics.
4. Made fine-tuning models on my data.
5. Made predicted test data with the help of pipeline.

Best Solution

	F1 metric
Natasha	0.06
Fine-tuning BERT (rubert-tiny2) model	0.3

The best solution is the fine-tuning BERT (rubert-tiny2) model, because compared to Natasha it takes into account all necessary entity types and is trained on training data.