Проектная работа для итоговой аттестации по курсу Data Science, университет «Иннополис»

Слушатель: Константин Егоров
Наставник: Олег Дудкин

Название проекта: проект 4, извлечение именованных сущностей для русского языка.

Задача проекта: обучить и протестировать модель для извлечения именованных сущностей из текста. Провести анализ решения и альтернатив. Выбрать лучшую модель.

Источники данных:
http://bsnlp.cs.helsinki.fi/shared-task.html
https://multiconer.github.io

1. Исследование входных данных.

- Первый источник – SemEval 2022, 11 открытое задание по распознаванию именованных сущностей для 11 языков.
- Исходный набор данных содержит 15 300 размеченных предложений и 242 383 токена. Токены размечены с использованием схемы BIO (Beginning, Inside, Outside) для шести типов именованных сущностей: PER : Person, LOC : Location, GRP : Group, CORP : Corporation, PROD : Product, CW: Creative Work.

```
['# id 11b11e4f-73c6-4e3d-babd-0de83e450861\tdomain=train',
 'российская _ _ B-GRP',
 'экологическая _ _ I-GRP',
 'партия _ _ I-GRP',
 '«зелёные» _ _ I-GRP',
 '– _ _ O',
 'до _ _ O',
 'февраля _ _ O',
 '2012 _ _ O',
 'года _ _ O',
 ', _ _ O',
 'входила _ _ O',
 'в _ _ O',
 'состав _ _ O',
 'партии _ _ O',
 'ср _ _ O',
 '',
```

1. Исследование входных данных.

- Второй источник – это 3 открытое задание по распознаванию, нормализации, классификации и межъязыковому связыванию именованных сущностей в славянских языках.
- Исходный набор данных для русского языка содержит 3 191 токен в предложениях, находящихся в разных файлах.
- Разметка содержит пять типов именованных сущностей: человек, локация, организация, мероприятие, продукты (persons, locations, organizations, events, products).
- Brexit и суд над христианской девушкой Асией Биби в Пакистане, обвиняемой в богохульстве.

```
c.most_common()[:40]
```

```
[('Великобритании', 99),
 ('ЕС', 88),
 ('Brexit', 87),
 ('Дэвид Дэвис', 66),
 ('Бориса Джонсона', 63),
 ('Борис Джонсон', 60),
 ('Мэй', 58),
 ('Тереза Мэй', 58),
 ('Евросоюза', 57),
 ('Пакистане', 51),
 ('Терезы Мэй', 50),
 ('Пакистана', 46),
 ('Биби', 43),
 ('Джонсона', 42),
 ('Асии Биби', 40),
 ('Британии', 37),
 ('Джонсон', 36),
 ('Доминик Рааб', 32),
 ('Дэвис', 31),
 ('МИД Великобритании', 29),
 ('МИД', 27),
 ('Асия Биби', 26),
 ('Мухаммеда', 25),
 ('Дэвида Дэвиса', 25),
 ('Дэвиса', 25),
 ('Лондона', 24),
 ('Асию Биби', 22),
 ('Верховный суд Пакистана', 22),
 ('Евросоюзом', 22),
 ('Терезой Мэй', 21),
 ('Верховного суда', 19),
 ('Брюсселем', 18),
 ('Великобритания', 18),
 ('Бориса', 16),
 ('Рааб', 16),
 ('REGNUM', 15),
 ('США', 15),
 ('Доминика Рааба', 15),
 ('Даунинг-стрит', 15),
 ('Англии', 15)]
```

2. Исследование Python NLP библиотек.

```
ls = u'Российская экологическая партия «Зелёные» – до февраля 2012 года, входила в состав партии СР.'
```

```
ls = 'Иоанн Павел II внёс некоторые изменения в правила проведения конклавов.'
```

```
ls = 'В 1977 году окончил исторический факультет Запорожского государственного педагогического института.'
```

```
ls = 'Куно I фон Ротт – пфальцграф Баварии, граф Фобурга, граф Нижнего Изара.'
```

```
ls = '– Гаев – «Вишнёвый сад», по пьесе А. П. Чехова, реж.'
```

```
ls = 'whites off earth now!! вики'
```

```
ls = 'смерть о'брайан, рональд кларк'
```

```
ls = 'барда (река) офис шерифа округа'
```

```
ldoc = ltokenize(ls)
```

```
барда NOUN nsubj
( PUNCT punct
река NOUN appos
) PUNCT punct
офис NOUN ROOT
шерифа NOUN nmod
округа NOUN nmod
```

```
ldoc = ltokenize(ls)
```

```
смерть NOUN ROOT
о'брайан PROPN nmod
, PUNCT punct
рональд PROPN conj
кларк PROPN flat:name
```

```
ldoc = ltokenize(ls)
```

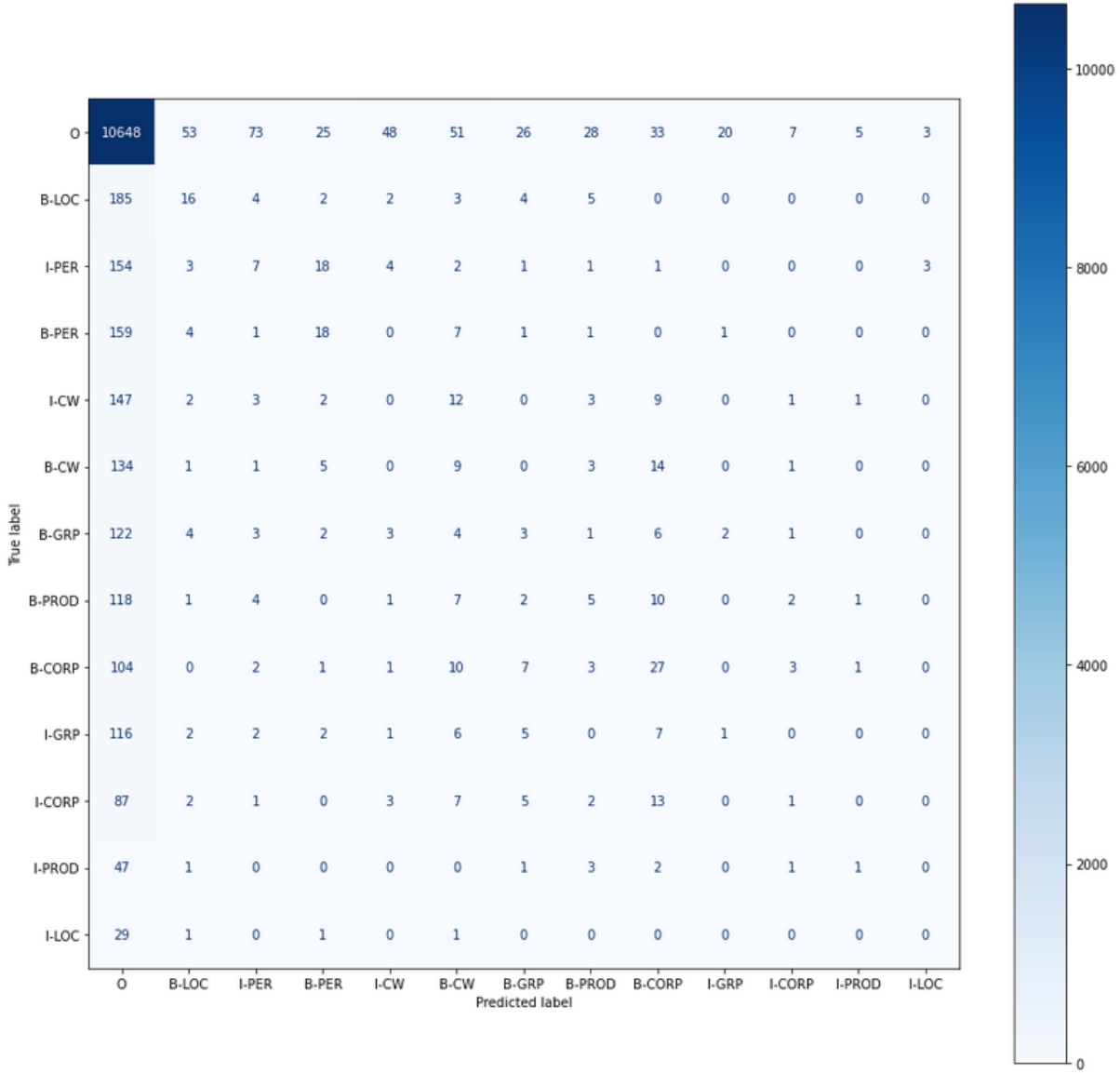
```
whites X ROOT
off X flat:foreign
earth X flat:foreign
now X flat:foreign
! PUNCT punct
! PUNCT punct
вики NOUN ROOT
```

```
ldoc = ltokenize(ls)
```

```
– PUNCT punct
Гаев PROPN nsubj
– PUNCT punct
« PUNCT punct
Вишнёвый ADJ amod
сад NOUN ROOT
» PUNCT punct
, PUNCT punct
по ADP case
пьесе NOUN conj
А. PROPN nmod
П. PROPN flat:name
Чехова PROPN flat:name
, PUNCT punct
реж NOUN conj
. PUNCT punct
```

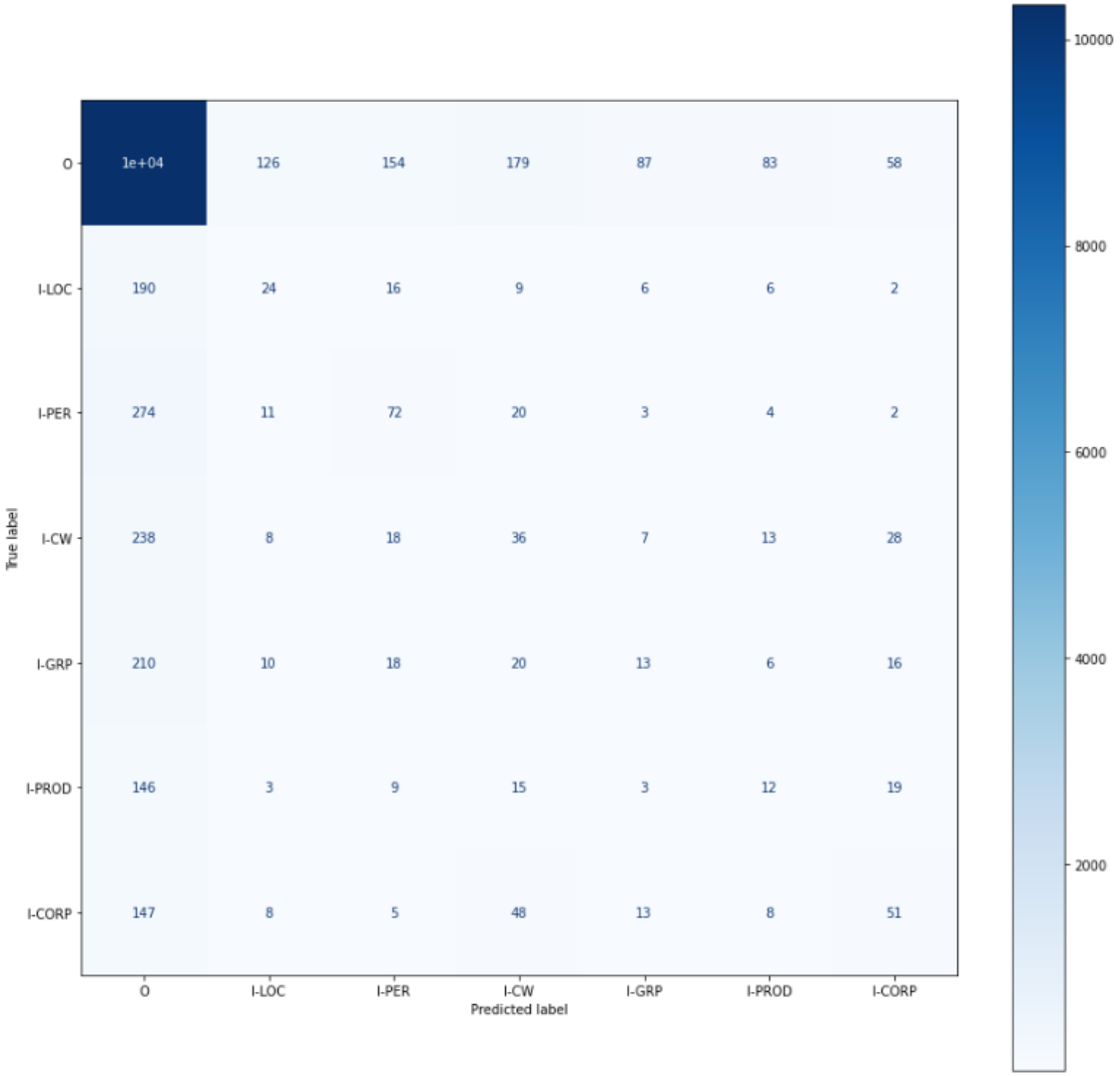
- Первый опробованный токенизатор Spacy показал хорошие результаты и было принято решение остановиться на нем.

3. Формирование эмбеддингов и обучение базовой модели.



	precision	recall	f1-score	support
O	0.90	0.93	0.91	11020
B-LOC	0.13	0.12	0.13	221
I-PER	0.05	0.05	0.05	194
B-PER	0.17	0.15	0.16	192
I-CW	0.00	0.00	0.00	180
B-CW	0.07	0.10	0.08	168
B-GRP	0.02	0.01	0.02	151
B-PROD	0.06	0.05	0.06	151
B-CORP	0.20	0.18	0.19	159
I-GRP	0.02	0.01	0.01	142
I-CORP	0.08	0.02	0.03	121
I-PROD	0.04	0.02	0.02	56
I-LOC	0.00	0.00	0.00	32
accuracy			0.81	12787
macro avg	0.13	0.13	0.13	12787
weighted avg	0.78	0.81	0.80	12787

3. Формирование эмбеддингов и обучение базовой модели.



	precision	recall	f1-score	support
O	0.90	0.94	0.92	11020
I-LOC	0.13	0.09	0.11	253
I-PER	0.25	0.19	0.21	386
I-CW	0.11	0.10	0.11	348
I-GRP	0.10	0.04	0.06	293
I-PROD	0.09	0.06	0.07	207
I-CORP	0.29	0.18	0.22	280
accuracy			0.82	12787
macro avg	0.27	0.23	0.24	12787
weighted avg	0.79	0.82	0.81	12787

4. Обучение seq2seq модели. Подготовка данных.

	words	markers	sentense_num	word_idx	marker_idx
0	российская	B-GRP	1	9681	4
1	экологическая	I-GRP	1	20293	4
2	партия	I-GRP	1	15291	4
3	«зелёные»	I-GRP	1	38551	4
4	—	O	1	43181	6

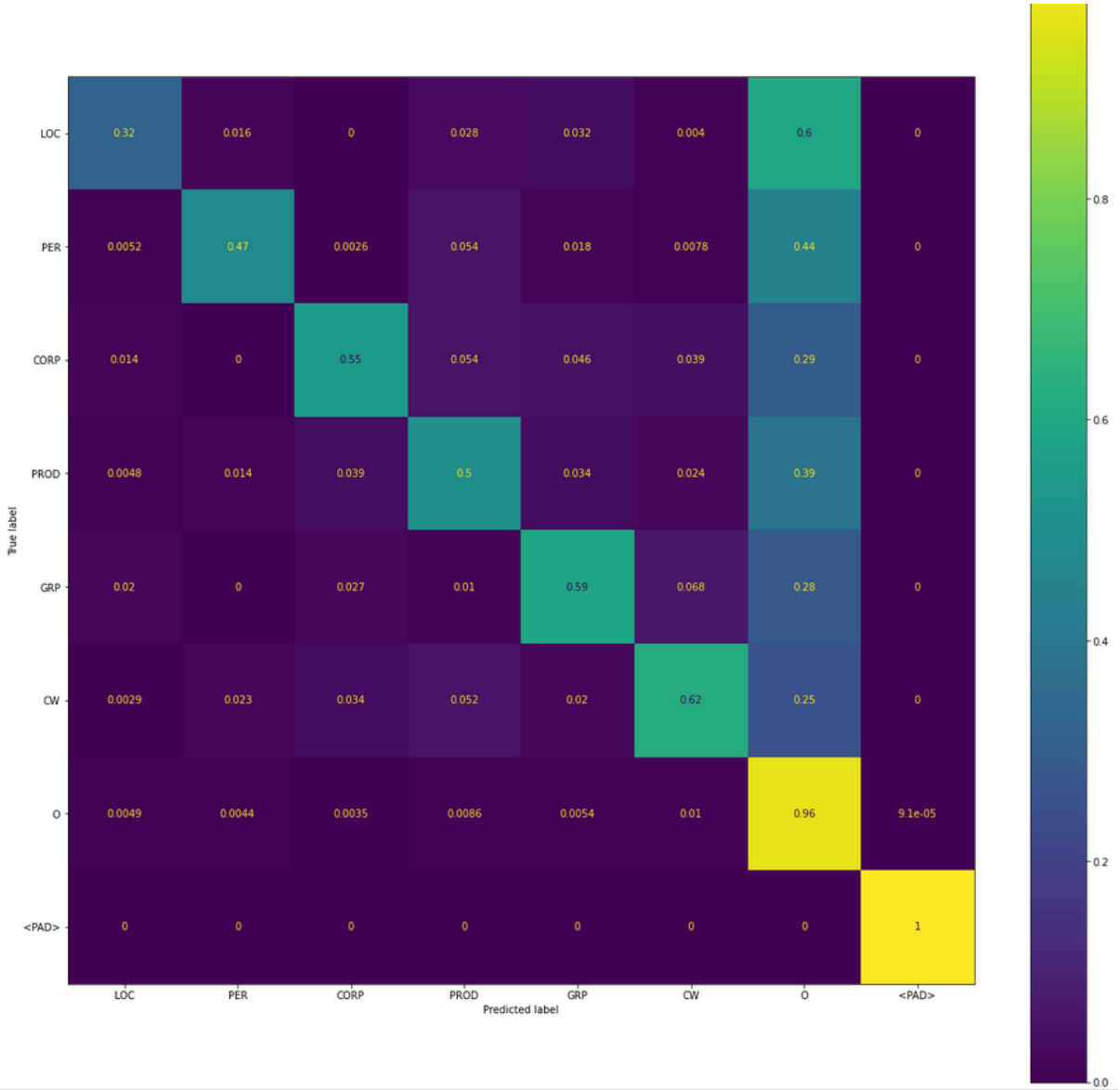
	sentense_num	words	markers	word_idx	marker_idx
0	1	[российская, экологическая, партия, «зелёные»,...	[B-GRP, I-GRP, I-GRP, I-GRP, O, O, O, O, O, O,...	[9681, 20293, 15291, 38551, 43181, 1596, 30737...	[4, 4, 4, 4, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6]
1	2	[также, посещал, два, семинара, бартольда, кёй...	[O, O, O, O, B-PER, I-PER, O, O, O, B-PER, O, ...	[24008, 47589, 42010, 40044, 31326, 34745, 239...	[6, 6, 6, 6, 1, 1, 6, 6, 6, 1, 6, 1, 6]
2	3	[в, 1999, —, 2006, играла, за, национальную, с...	[O, O, O, O, O, O, B-GRP, I-GRP, I-GRP, O]	[26415, 3416, 43181, 37240, 37335, 27712, 2989...	[6, 6, 6, 6, 6, 6, 4, 4, 4, 6]
3	4	[«, джульетта, », —, кинофильм, 2016, года, ис...	[O, B-CW, O, O, O, O, O, O, O, O, O, O, O, O]	[49441, 13541, 3797, 43181, 7570, 32408, 34363...	[6, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6]
4	5	[мякоть, спелого, плода, съедобна, в, свежем, ...	[O, O, O, O, O, O, O, O, O, O, O, O, O, O, B-P...	[43136, 34860, 13361, 5001, 26415, 29718, 1827...	[6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 3, 6]

4. Обучение seq2seq модели. Модель.

```
Model: "sequential_4"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 embedding_4 (Embedding)     (None, 59, 300)           15770400

 bidirectional_2 (Bidirectio  (None, 59, 600)          1442400
 nal)

 time_distributed_2 (TimeDis  (None, 59, 9)            5409
 tributed)

=================================================================
Total params: 17,218,209
Trainable params: 17,218,209
Non-trainable params: 0
```

	precision	recall	f1-score	support
I-CW	0.40	0.45	0.42	180
B-LOC	0.52	0.29	0.37	221
I-PER	0.55	0.38	0.45	194
I-CORP	0.58	0.21	0.30	121
_ O	0.00	0.00	0.00	0
I-PROD	0.31	0.27	0.29	56
O	0.92	0.98	0.95	11020
B-PER	0.47	0.32	0.38	192
I-GRP	0.72	0.34	0.46	142
B-PROD	0.43	0.37	0.40	151
B-CORP	0.67	0.19	0.29	159
B-CW	0.39	0.36	0.38	168
B-GRP	0.71	0.24	0.36	151
I-LOC	0.14	0.03	0.05	32
<UNKNOWN>	0.00	0.00	0.00	0
<PAD>	1.00	1.00	1.00	34413
micro avg	0.97	0.97	0.97	47200
macro avg	0.49	0.34	0.38	47200
weighted avg	0.96	0.97	0.96	47200

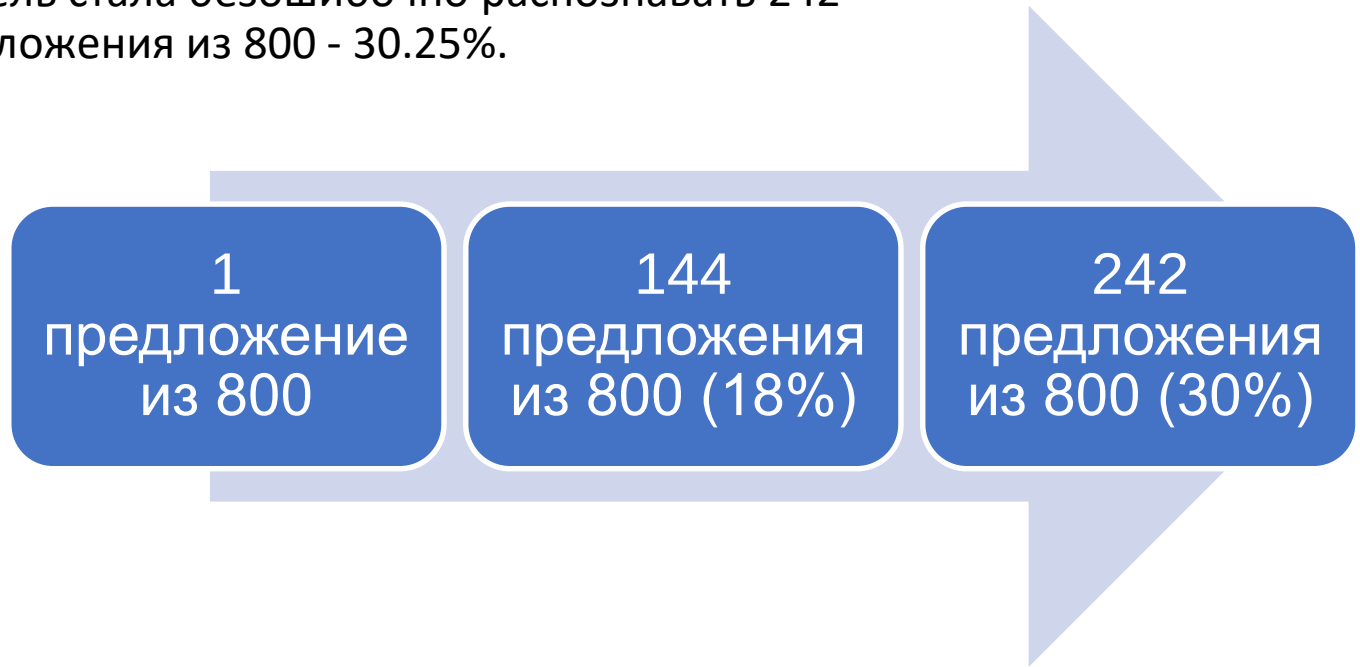
4. Обучение seq2seq модели. Доработанная модель.



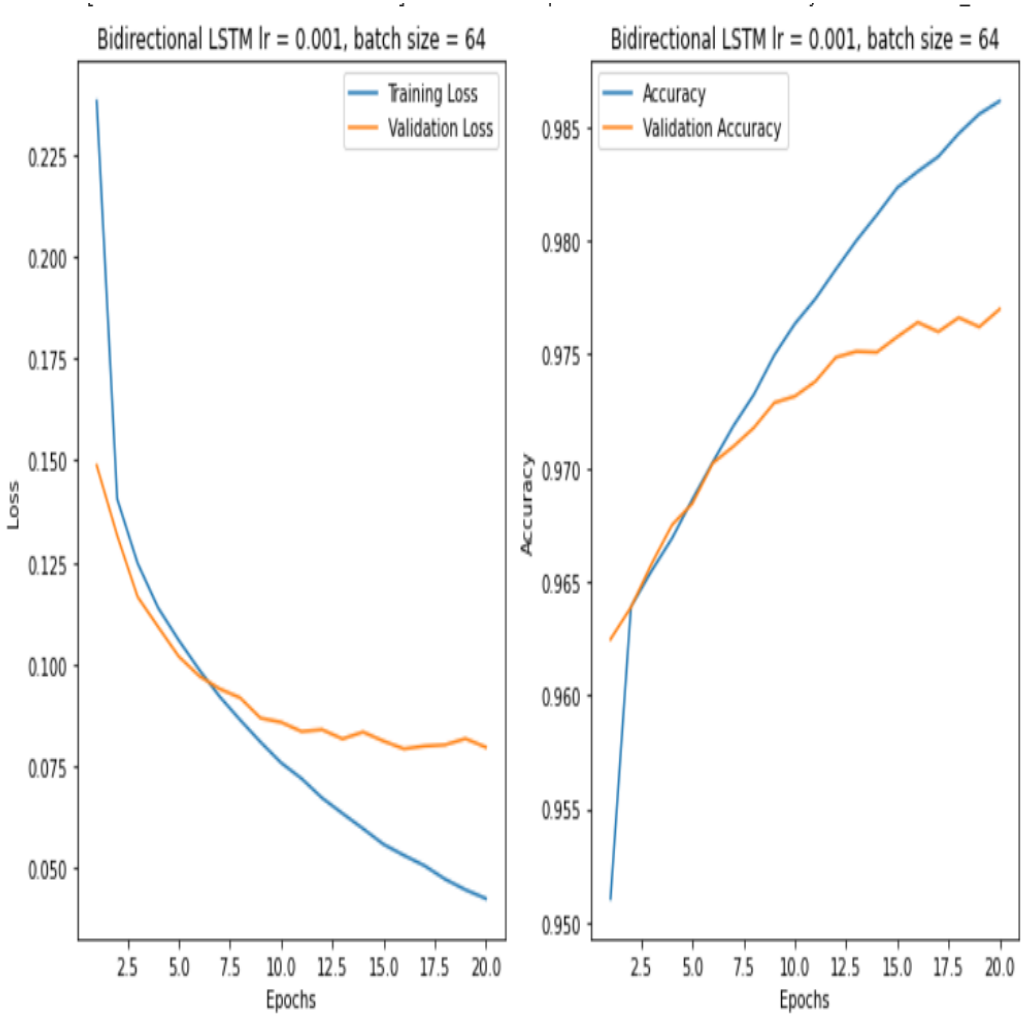
	precision	recall	f1-score	support
LOC	0.55	0.32	0.41	253
PER	0.74	0.47	0.58	386
CORP	0.70	0.55	0.62	280
PROD	0.39	0.50	0.44	207
GRP	0.63	0.59	0.61	293
CW	0.59	0.62	0.60	348
O	0.94	0.96	0.95	11020
<PAD>	1.00	1.00	1.00	34413
accuracy			0.97	47200
macro avg	0.69	0.63	0.65	47200
weighted avg	0.97	0.97	0.97	47200

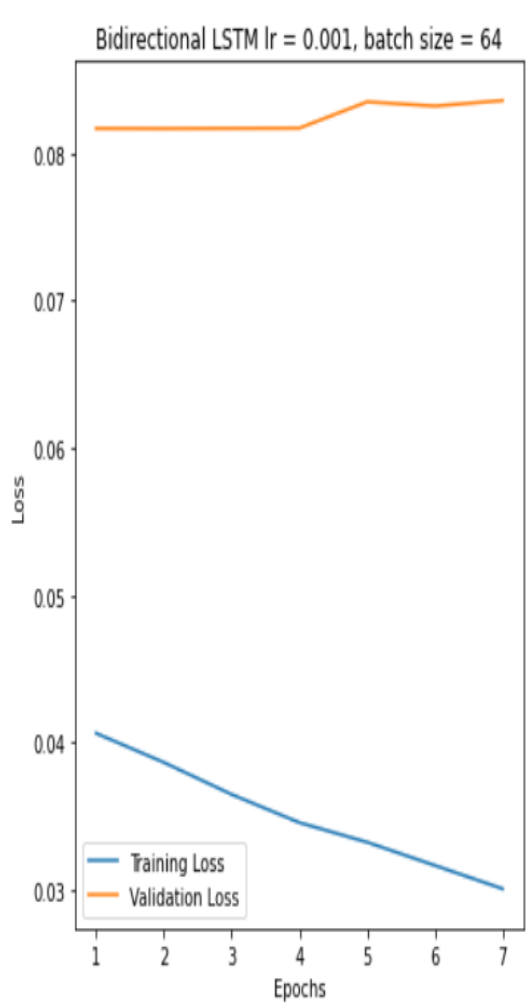
4. Обучение seq2seq модели. Динамика.

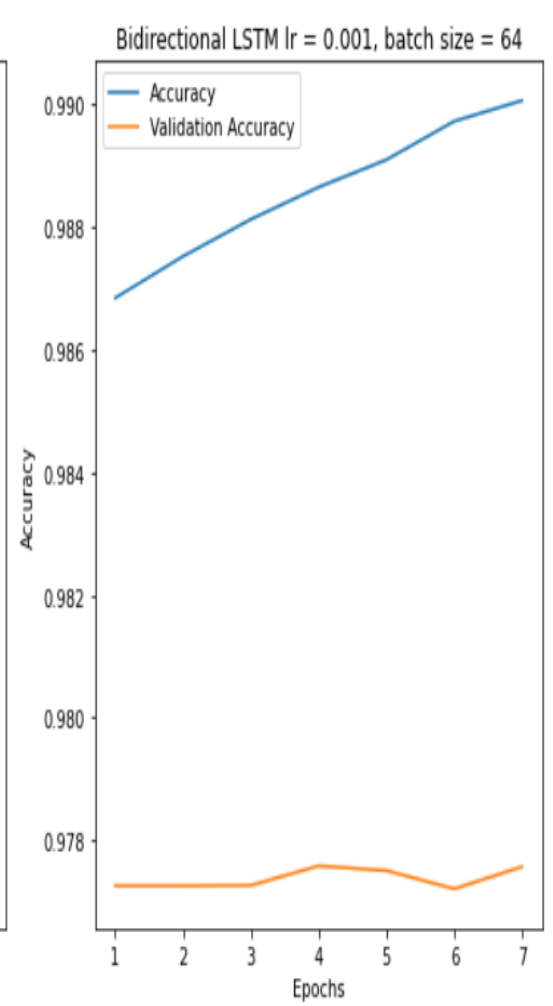
- Веденная новая метрика – процент правильно предсказанных полных предложений показала по 1 предложению на тестовой и валидационной выборках.
- Модель стала полностью правильно классифицировать 144 предложения из 800 - 18%.
- Модель стала безошибочно распознавать 242 предложения из 800 - 30.25%.



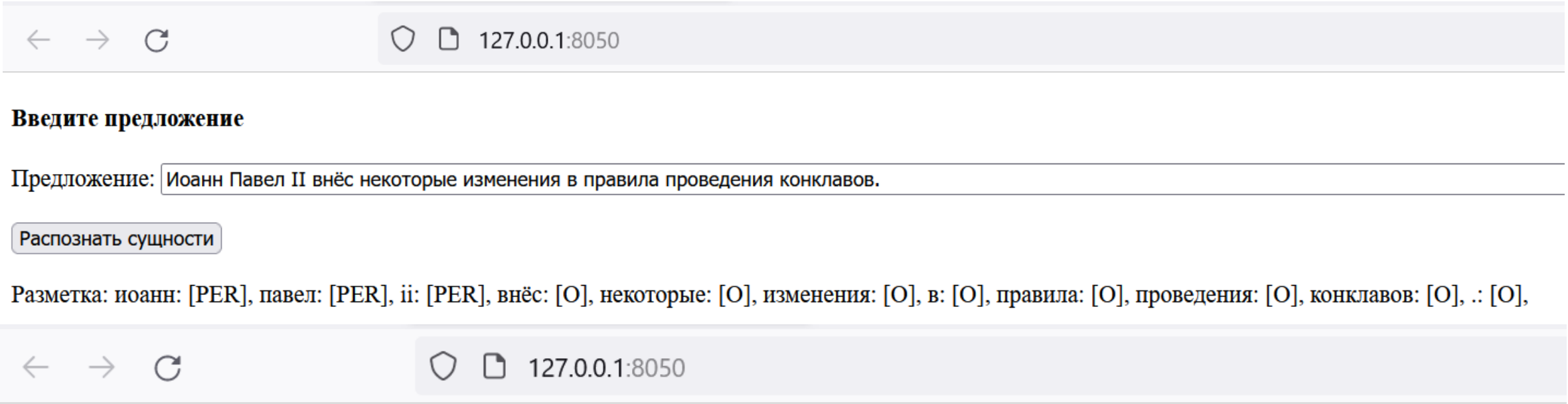
4. Обучение seq2seq модели. Кривые обучения.







5. Создание ВЕБ-приложения



Спасибо за внимание

18.12.2021

Москва