

Named Entity Recognition

Распознавание именованных сущностей

Цели вебинара

1

Что такое NER

2


Подходы к решению задачи

3

Метрики качества

4

Практика на Python



Распознавание именованных сущностей

Распознавание именованных сущностей

Выделение именованных сущностей - Named Entity Recognition (NER)

- Имена
- Локации
- Организации
- Даты, время
- Суммы, проценты
- Названия и др.

News NER example

Kofi Atta Annan is a Ghanaian diplomat who served as the seventh Secretary General of the United Nations from January 1, 1997, to January 1, 2007, serving two five-year terms. Annan was the co-recipient of the Nobel Peace Prize in October 2001.

Kofi Annan was born on April 8, 1938, to Victoria and Henry Reginald Annan in Kumasi, Ghana. He is a twin, an occurrence that is regarded as special in Ghanaian culture. Efua Atta, his twin sister, shares the same middle name, which means 'twin'. As with most Akan names, his first name indicates the day of the week he was born: 'Kofi' denotes a boy born on a Friday. The name Annan can indicate that a child was the fourth in the family, but in his family it was simply a name which Annan inherited from his parents.

In 1962, Annan started working as a Budget Officer for the World Health Organization, an agency of the United Nations. From 1974 to 1976, he was the Director of Tourism in Ghana. Annan then returned to work for the United Nations as an Assistant Secretary General in three consecutive positions.

Person

Location

Organization

Date

Nationality

Title

Зачем это нужно?



Зачем это нужно?

- Понимание текста
- Поиск информации
- Для предварительной обработки
- Для перевода
- Разрешение анафоры
“Прискакал Иванушка на белом коне. Принцесса выбежала ему навстречу и поцеловала его.”
- Вопросно-ответные системы, чат-боты
- *Поиск значимых последовательностей в нуклеотидных и аминокислотных последовательностях*

Выделение именованных сущностей

Выделение именованных сущностей
– это задача с учителем или без?



Сложности

1. Многозначные слова и омонимы

Вашингтон – столица США.

Джордж *Вашингтон* – первый президент США.

Сложности

1. Многозначные слова и омонимы

Вашингтон – столица США.

Джордж *Вашингтон* – первый президент США.

2. Вариативность сущности

Вас приветствует *Магазин Профессиональных Металлоискателей*.

Вас приветствует *Волхонка Престиж* – ваш любимый магазин брендов по доступным ценам.

Вам пишет *магазин зоотоваров Немо*.

Сложности

1. Многозначные слова и омонимы

Вашингтон – столица США.

Джордж *Вашингтон* – первый президент США.

2. Вариативность сущности

Вас приветствует *Магазин Профессиональных Металлоискателей*.

Вас приветствует *Волхонка Престиж* – ваш любимый магазин брендов по доступным ценам.

Вам пишет *магазин зоотоваров Немо*.

3. Нужна разметка

The background of the slide is an aerial photograph of a city skyline, likely New York City, with numerous skyscrapers. The image is filtered with a blue color scheme. A horizontal band across the middle of the image features a network of white lines and dots, resembling a digital or data network, set against a gradient from teal on the left to dark blue on the right.

Способы решения

Основанный на правилах

С использованием регулярных выражений

Хорошо подходит только для простых типизированных случаев:

- e-mail
- телефоны
- IP-адреса
- даты

Основанный на правилах

С использованием регулярных выражений

Хорошо подходит только для простых типизированных случаев:

- e-mail
- телефоны
- IP-адреса
- даты

С использованием словарей: имена, фамилии, локации.

Основанный на правилах

С использованием регулярных выражений

Хорошо подходит только для простых типизированных случаев:

- e-mail
- телефоны
- IP-адреса
- даты

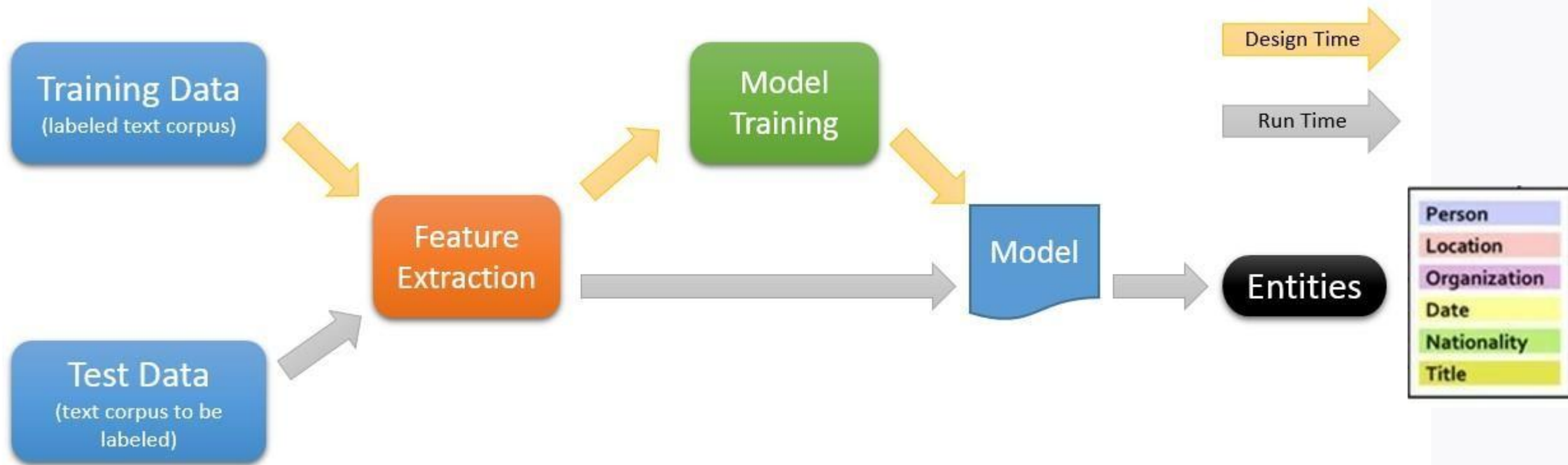
С использованием словарей: имена, фамилии, локации.

Поиск по капитализации

- первая буква большая, остальные маленькие
- все буквы маленькие
- все буквы большие
- нестандартная капитализация (iPhone)

Классификация

Каждый токен относится к одному из нескольких возможных классов.



Можно применять классические алгоритмы классификации:
логрег, SVM, деревья, бустинг

BIOES схема

К метке сущности (например, PER для персон или ORG для организаций) добавляется префикс, который обозначает позицию токена в сущности.

Карл Фридрих Иероним фон Мюнхгаузен родился в Боденвердере

B-PER I-PER I-PER I-PER E-PER OUT OUT S-LOC



BIOES схема

К метке сущности (например, PER для персон или ORG для организаций) добавляется префикс, который обозначает позицию токена в сущности.

Карл Фридрих Иероним фон Мюнхгаузен родился в Боденвердере

B-PER I-PER I-PER I-PER E-PER OUT OUT S-LOC

B (beginning) – первый токен в сущности, которая состоит больше чем из 1 слова

I (inside) – токен находится в середине сущности, которая состоит больше чем из 1 слова

O/OUT – токен не относится ни к какой сущности

E (ending) – последний токен сущности, которая состоит больше чем из 1 слова

S (single) – сущность состоит из одного слова

Проблема:

«Уральский федеральный университет имени первого Президента России Б. Н. Ельцина»



CoNLL-U формат

Таблица, где строка – один токен, а колонки — конкретный тип признаков токена.

1	В	в	PR	PR=	OUT
2	январе	январь	S	S, муж, неод=пр, ед	OUT
3	1995	1995	UNDEF	UNDEF	OUT
4	года	год	S	S, муж, неод= (вин, мн род, ед им, мн)	OUT
5	в	в	PR	PR=	OUT
6	Подмосковье	подмосковье	S	S, гео, сред, неод= (пр, ед вин, ед им, ед)	S-LOC
7	состоялся	состояться	V	V, сов, нп=прош, ед, изъяв, муж	OUT
8	учредительный	учредительный	A	A= (вин, ед, полн, муж, неод им, ед, полн, муж)	OUT
9	съезд	съезд	S	S, муж, неод= (вин, ед им, ед)	OUT
10	Общероссийского	общероссийский	A	A= (вин, ед, полн, муж, од род, ед, полн, муж род, ед, полн, сред)	B-ORG
11	общественного	общественный	A	A= (вин, ед, полн, муж, од род, ед, полн, муж род, ед, полн, сред)	I-ORG
12	движения	движение	S	S, сред, неод= (вин, мн род, ед им, мн)	I-ORG
13	"	"	UNDEF	UNDEF	I-ORG
14	Яблоко	яблоко	S	S, сред, неод= (вин, ед им, ед)	I-ORG
15	"	"	UNDEF	UNDEF	E-ORG
16	.	.	UNDEF	UNDEF	E-ORG
17					

В январе 1995 года в Подмосковье состоялся учредительный съезд Общероссийского движения «Яблоко».

CoNLL-U формат

Таблица, где строка – один токен, а колонки — конкретный тип признаков токена.

Всего 10 признаков, в примере ниже 6 типов: номер слова в тексте, словоформа (т. е. само слово), лемма (начальная форма слова), POS-таг (часть речи), морфологические характеристики слова и, наконец, метка сущности, выделяемой на данном токене.

1	В	в	PR	PR=	OUT
2	январе	январь	S	S, муж, неод=пр, ед	OUT
3	1995	1995	UNDEF	UNDEF	OUT
4	года	год	S	S, муж, неод= (вин, мн род, ед им, мн)	OUT
5	в	в	PR	PR=	OUT
6	Подмосковье	подмосковье	S	S, гео, сред, неод= (пр, ед вин, ед им, ед)	S-LOC
7	состоялся	состояться	V	V, сов, нп=прош, ед, изъяв, муж	OUT
8	учредительный	учредительный	A	A= (вин, ед, полн, муж, неод им, ед, полн, муж)	OUT
9	съезд	съезд	S	S, муж, неод= (вин, ед им, ед)	OUT
10	Общероссийского	общероссийский	A	A= (вин, ед, полн, муж, од род, ед, полн, муж род, ед, полн, сред)	B-ORG
11	общественного	общественный	A	A= (вин, ед, полн, муж, од род, ед, полн, муж род, ед, полн, сред)	I-ORG
12	движения	движение	S	S, сред, неод= (вин, мн род, ед им, мн)	I-ORG
13	"	"	UNDEF	UNDEF	I-ORG
14	Яблоко	яблоко	S	S, сред, неод= (вин, ед им, ед)	I-ORG
15	"	"	UNDEF	UNDEF	E-ORG
16	.	.	UNDEF	UNDEF	
17					

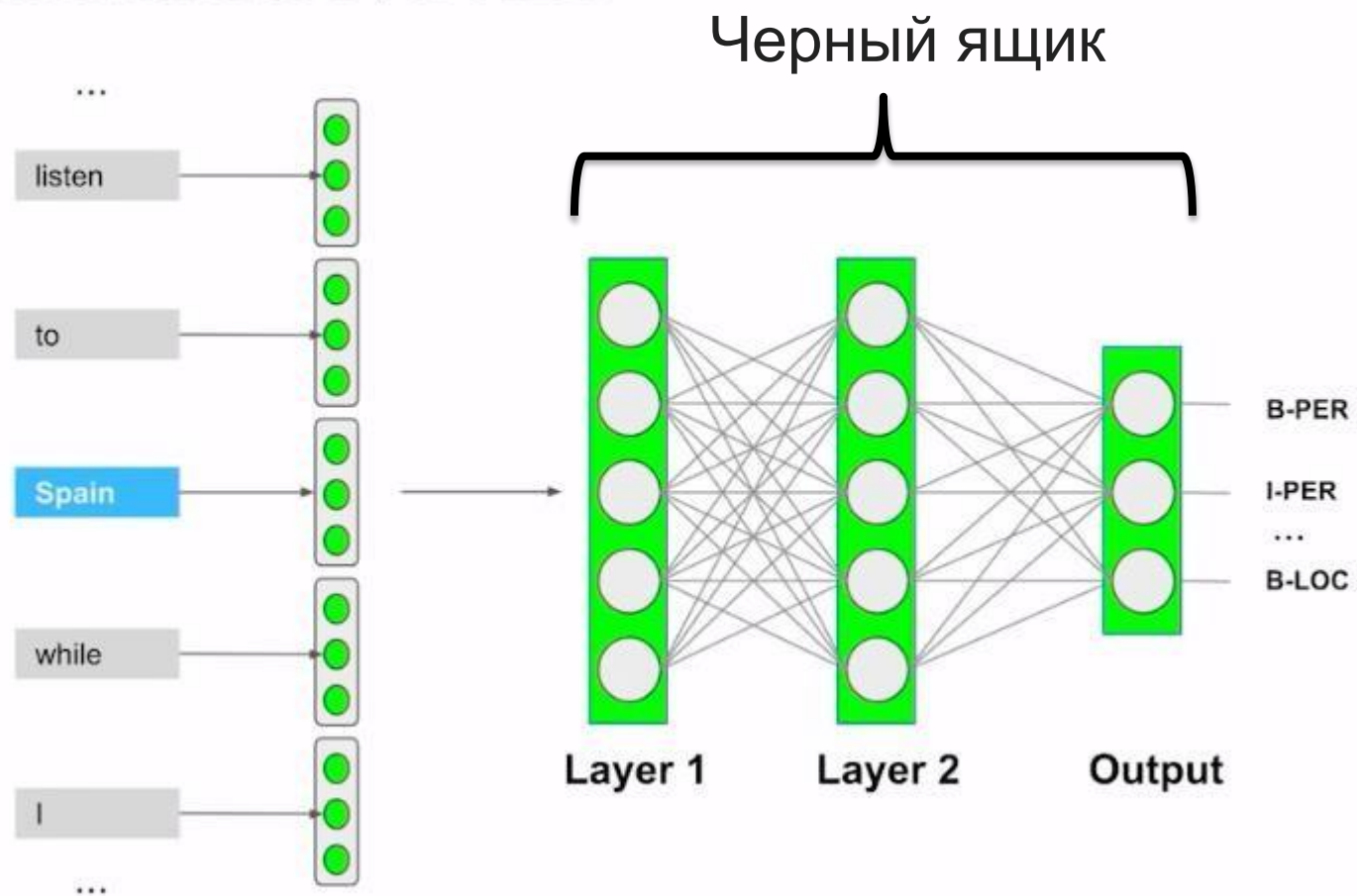
В январе 1995 года в Подмосковье состоялся учредительный съезд Общероссийского движения «Яблоко».



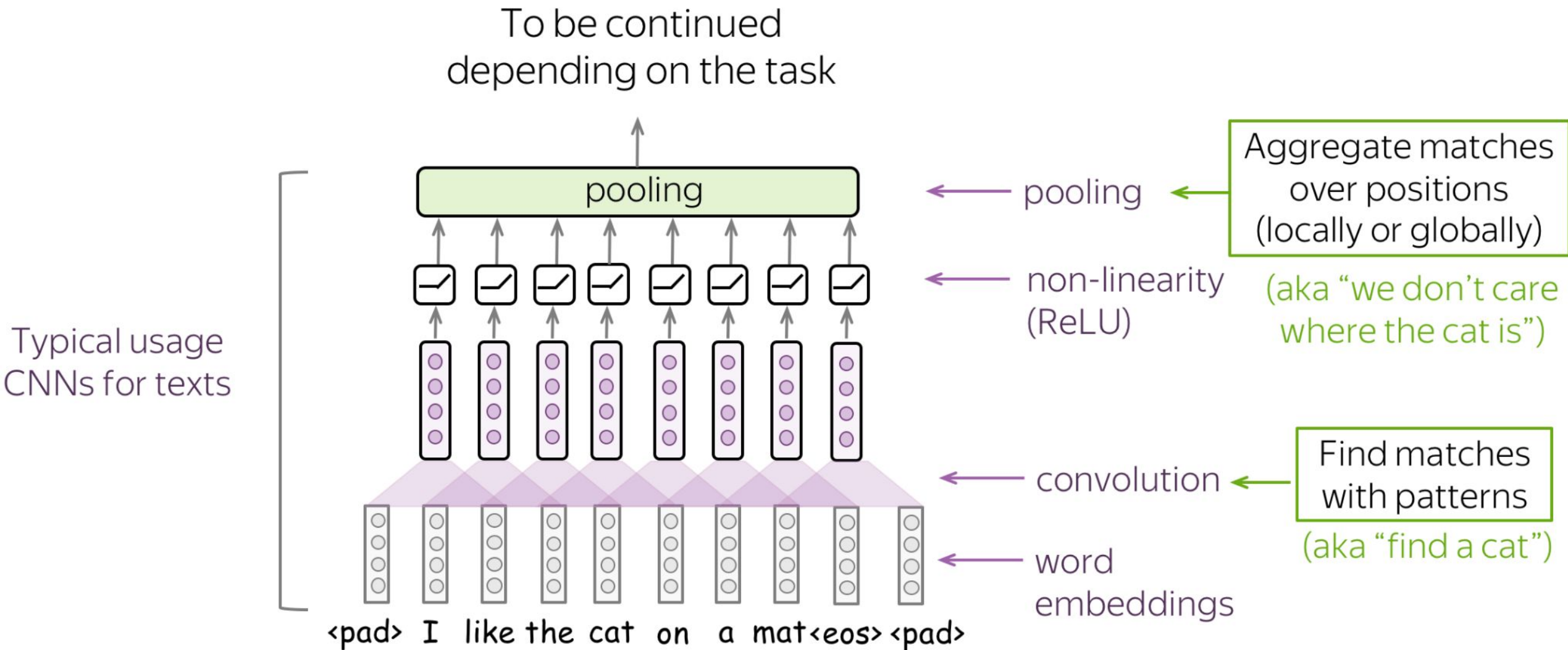
Нейронные сети

Feed Forward Network

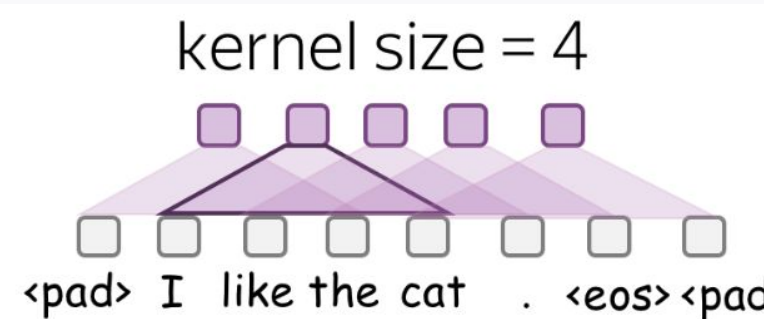
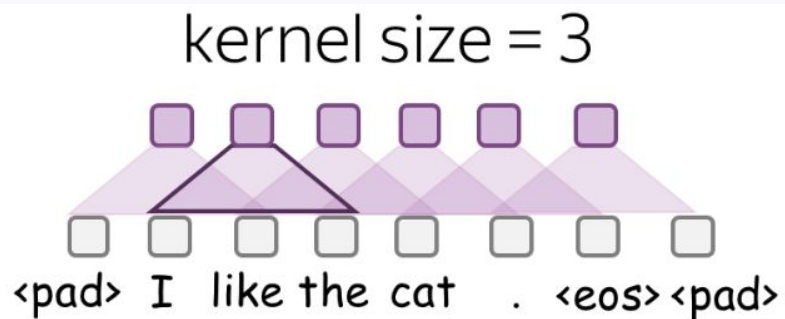
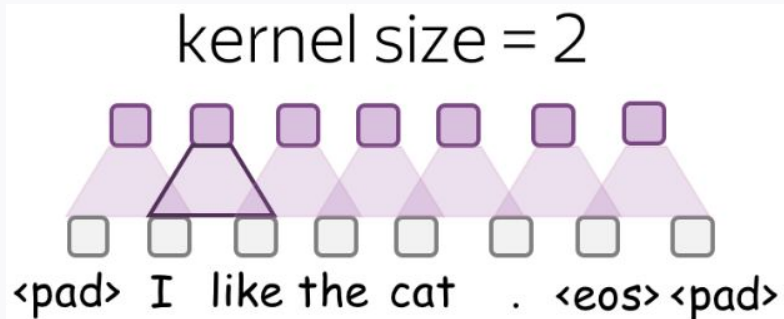
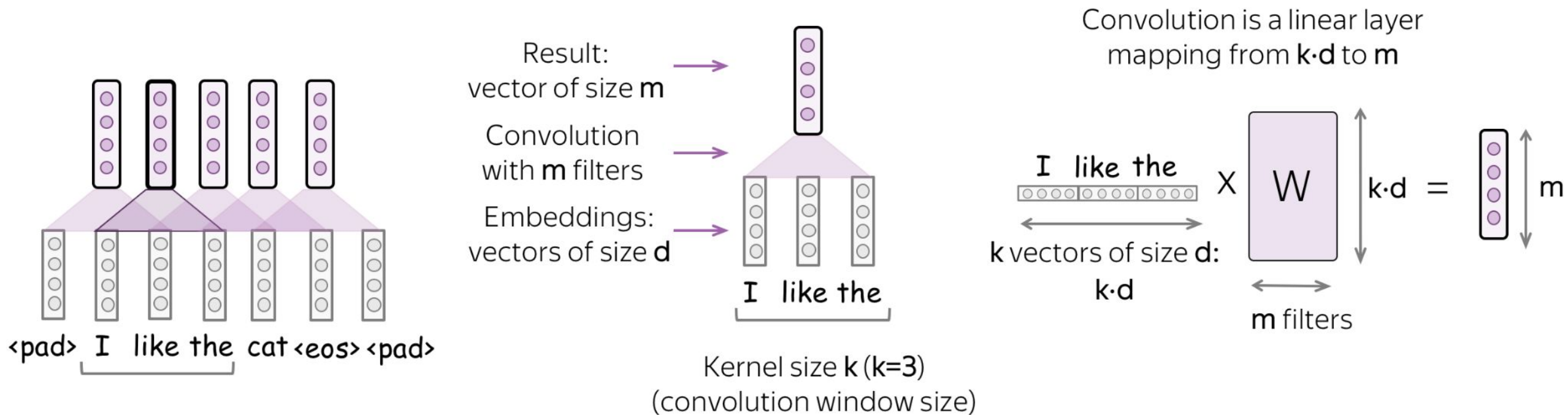
Feed forward network for NER



Сверточные сети для текста



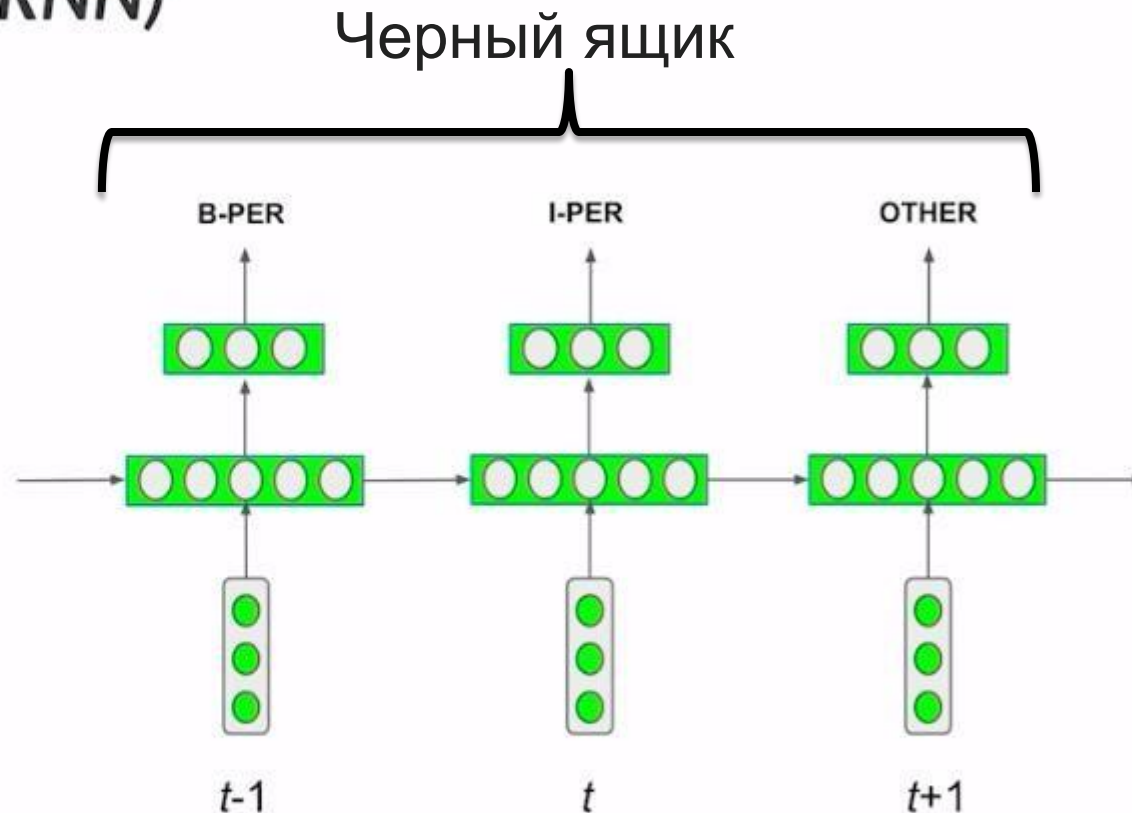
Сверточные сети для текста



RNN

Recurrent neural network (RNN)

- At each time step we process one word concatenated with the output from previous time steps
- It **remembers** information for many time steps



More advanced

- LSTM
- BiLSTM
- Multilayer LSTM
- **TRANSFORMERS!**



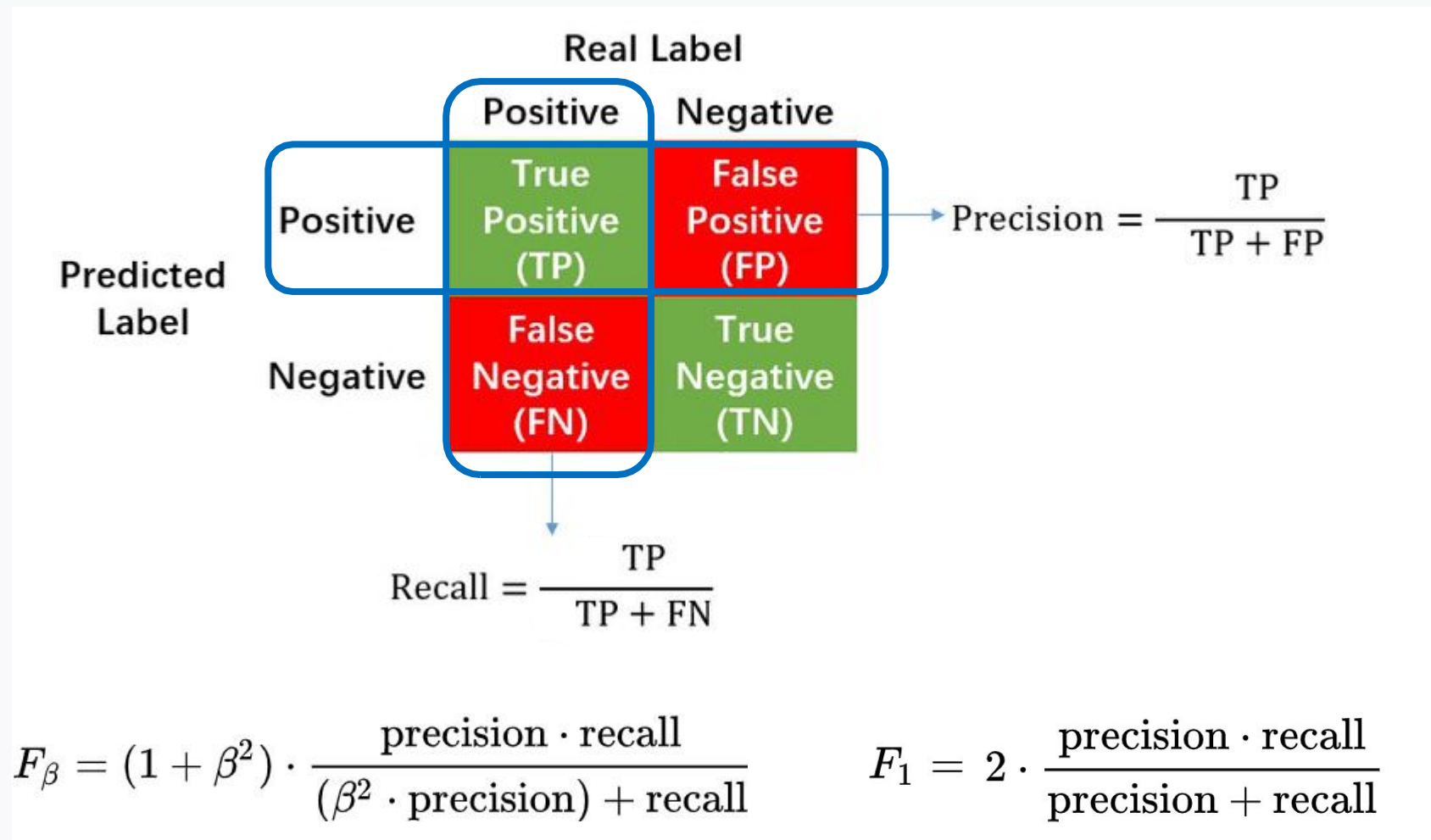
Additional features

- **Позиция слова в предложении**
- **POS-тег (positional of speech) – часть речи:** *существительное, глагол, предлог...*
- **Морфологические признаки:** *род, падеж, склонение...*
- **Синтаксические признаки:** *подлежащее, сказуемое, дополнение...*
- **Капитализация:** *ООО Рога и копыта, МГУ, H₂O*

The background of the slide is an aerial photograph of a city skyline, likely New York City, with numerous skyscrapers. The image is overlaid with a semi-transparent blue gradient. A network of thin, light blue lines connects various points across the blue area, creating a web-like pattern. The text "Метрики качества" is centered in the middle of the slide in a white, sans-serif font.

Метрики качества

Метрики качества



Метрики качества

Micro-average

$$micro - Pr = \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FP_i}$$

$$micro - R = \frac{\sum_i^n TP_i}{\sum_i^n TP_i + \sum_i^n FN_i}$$

Confusion Matrix				
Actual				
P r e d i c t e d		Class 1	Class 2	Class 3
	Class 1	TP1	FP12	FP13
	Class 2	FN21	TP2	FP23
	Class 3	FN31	FN32	TP3

Метрики качества

Macro-average

$$macro - Pr = \frac{\sum_i^n Pr_i}{n}$$

$$macro - R = \frac{\sum_i^n R_i}{n}$$

Confusion Matrix				
Actual				
P r e d i c t e d		Class 1	Class 2	Class 3
	Class 1	TP1	FP12	FP13
	Class 2	FN21	TP2	FP23
	Class 3	FN31	FN32	TP3



Библиотеки



Библиотеки NER

Английский язык

- Spacy <https://spacy.io/usage/rule-based-matching#entityruler>
- OpenNLP https://www.tutorialspoint.com/opennlp/opennlp_named_entity_recognition.htm
- StanfordNLP
- <https://stanfordnlp.github.io/CoreNLP/ner.html>
- Google Language API <https://cloud.google.com/natural-language/docs/analyzing-entities>

Библиотеки NER

Английский язык

- Spacy <https://spacy.io/usage/rule-based-matching#entityruler>
- OpenNLP https://www.tutorialspoint.com/opennlp/opennlp_named_entity_recognition.htm
- StanfordNLP
- <https://stanfordnlp.github.io/CoreNLP/ner.html>
- Google Language API <https://cloud.google.com/natural-language/docs/analyzing-entities>

Русский язык

- Yargy <https://github.com/natasha/yargy>
- Natasha <https://github.com/natasha>
- DeepPavlov <http://docs.deeppavlov.ai/en/master/features/models/ner.html>



natasha

Утилиты создания разметки

<https://www.isahit.com/blog/the-best-free-text-labeling-tools-for-text-annotation-and-categorization-in-natural-language-processing>

1. BRAT

BRAT comes with detailed instructions on how to install it. If you just want to install and run brat on your local machine, then the standalone server is what you want. Firstly, you must place the data section of the instructions to learn how to set up the annotation files. As BRAT is not compatible with Python, you would have to modify the command `python standalone.py` to `python2 standalone.py`. BRAT is noted to work exceptionally well with Google Chrome.

2. DOCCANO

DOCCANO is easier to use. When installing DOCCANO, you don't necessarily need to understand what Docker is. This can be done provided Docker is installed. To get abreast with its functionality, try out doccano's live demos.

3. INCEpTION

INCEpTION provides a comprehensive user guide that describes at length how to install and run it. Running INCEpTION is especially easy, because you can execute the downloaded JAR file without installing it.

