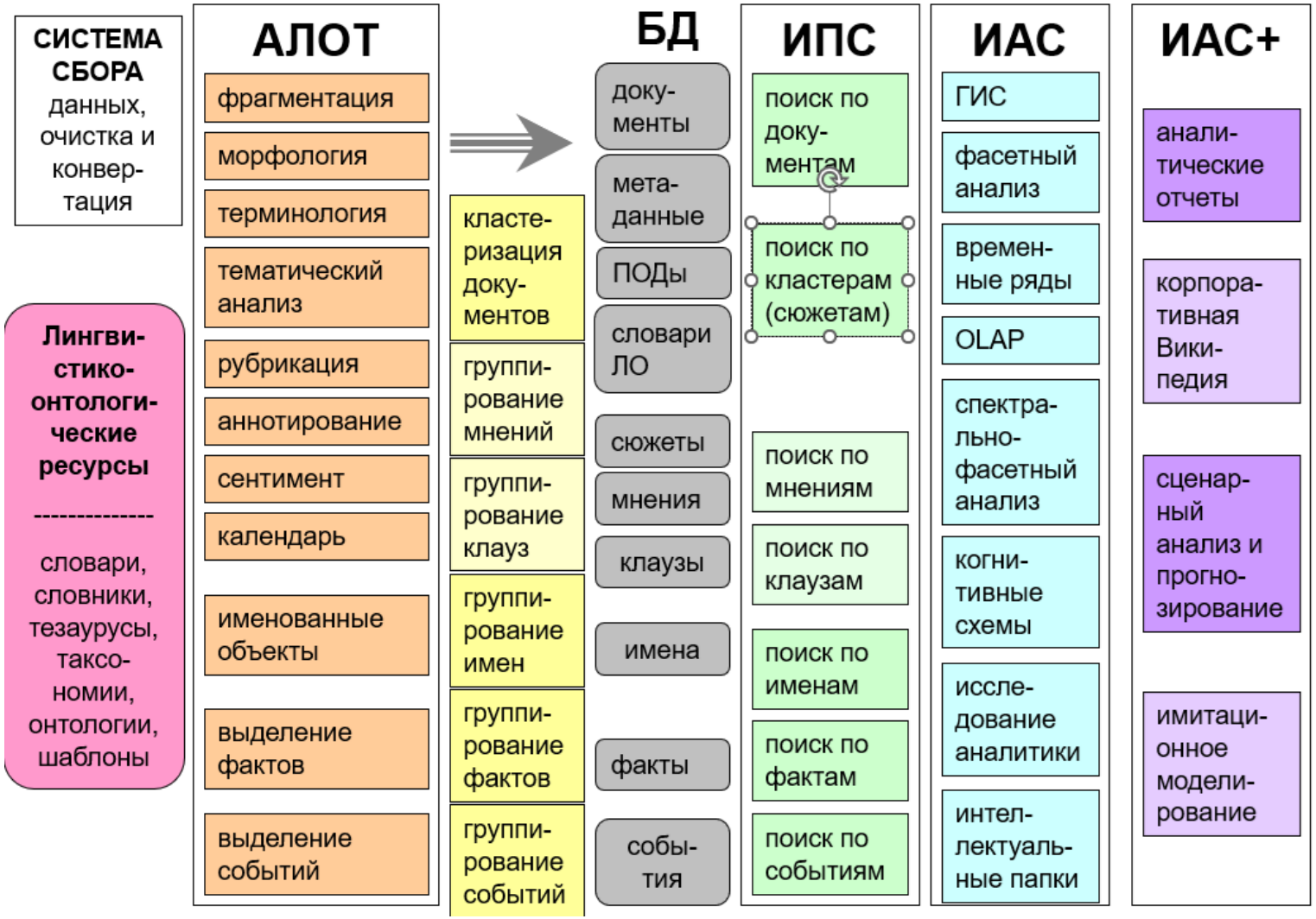
Извлечение информации из текстов, графы знаний и большие языковые модели

Лукашевич Н.В.

доктор технических наук
ведущий научный сотрудник НИВЦ МГУ имени М.В. Ломоносова,
профессор, зав. кафедрой ВМК МГУ,
профессор филологического факультета МГУ

Наша специализация: информационно-аналитические системы



Извлечение информации из текстов

Задача преобразования неструктурированных данных в структурированные

- Именованные (конкретные) сущности

  NE - Named Entities

  - персоны, компании, адреса, даты
  - упоминания генов и белков и пр.

- Отношения выделенных сущностей:
  - Место работы, должность
  - Взаимодействие белков

- Связанные с ними события и факты

  Events

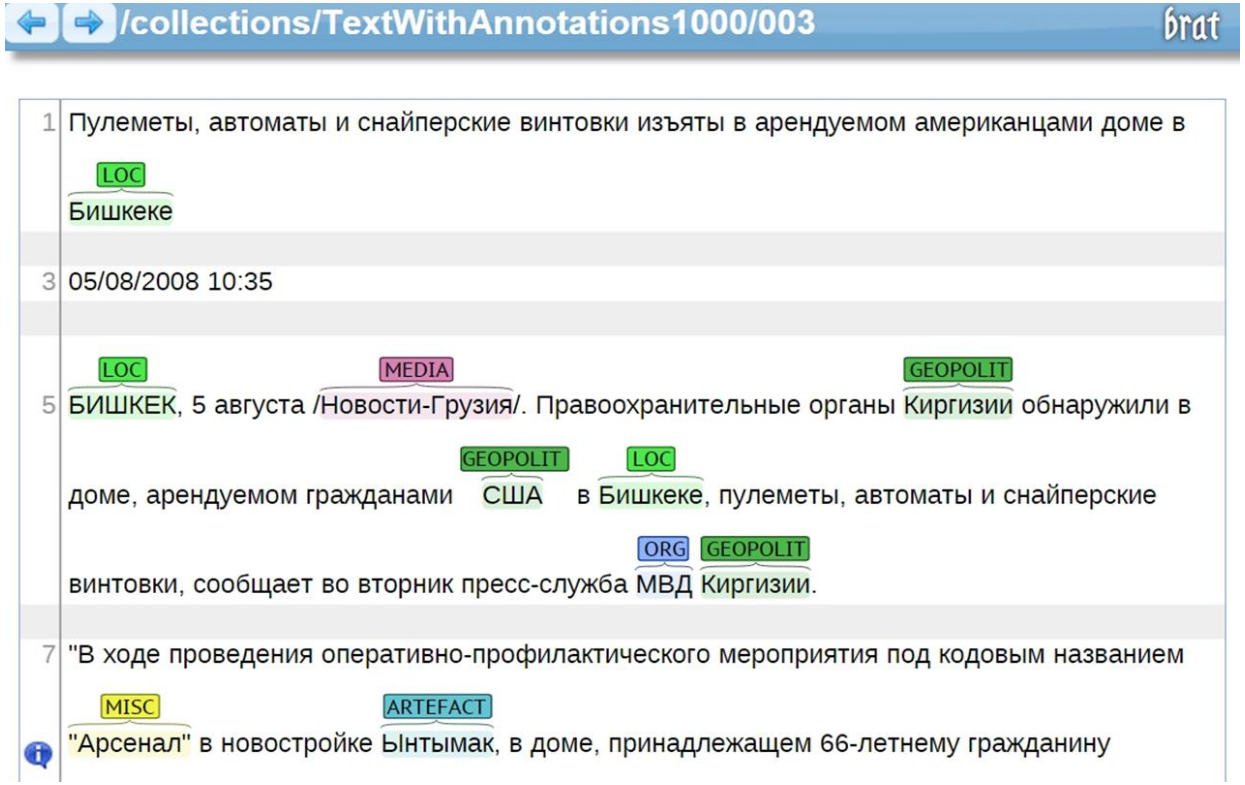
  слияние/поглощение компаний…

  приобретение контрольного пакета акций

План

- Новые аспекты извлечения информации из текстов
  - Извлечение вложенных именованных сущностей и отношений между ними
  - Извлечение событий
  - Связь задач извлечения информации с графами знаний
- Нейросетевые подходы для извлечения информации:
  - Извлечение сущностей
  - Извлечение отношений
  - Изменения при извлечении вложенных сущностей и отношений
  - Сквозное (end-to-end) извлечение информации
  - Извлечение информации большими языковыми моделями

Извлечение Плоские (flat) – задача разметки последовательностей: нужно приписать нужный тег каждому токену



- Кодирование BIO

- Пресс-служба O
- МВД B-ORG
- Киргизии B-GEOPOLIT

Размеченные данные Collection5 (Mozharova, Loukachevitch, 2016)

Текущая базовая технология – кодировщики трансформеров . Модель BERT (позднее RoBERTa)

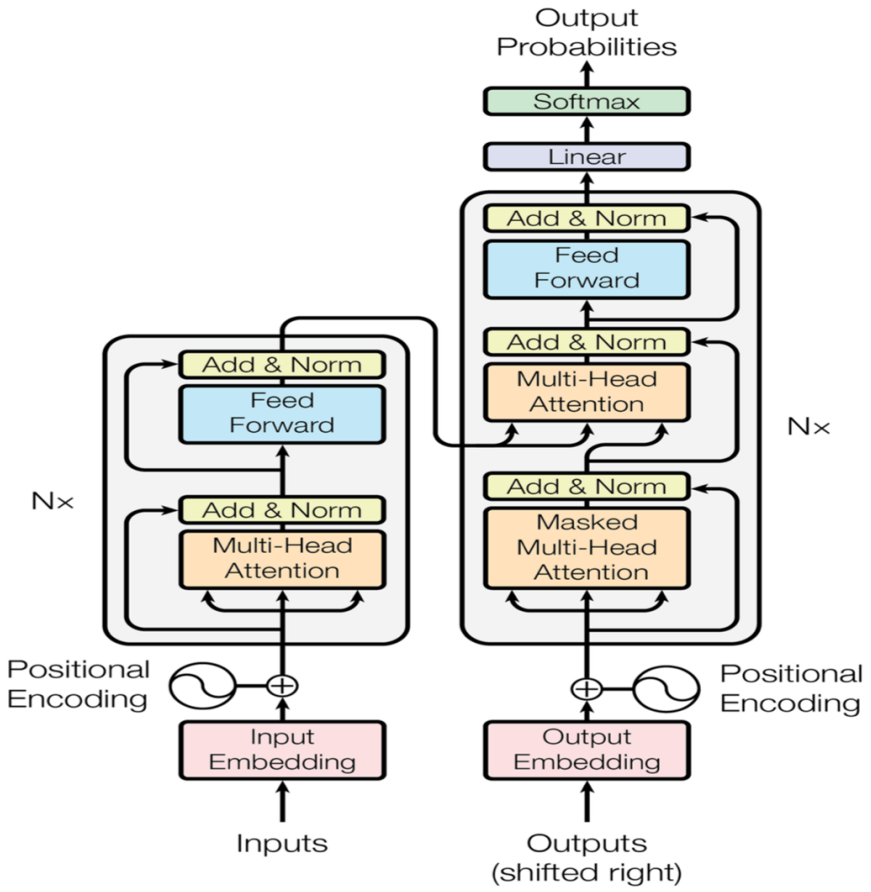
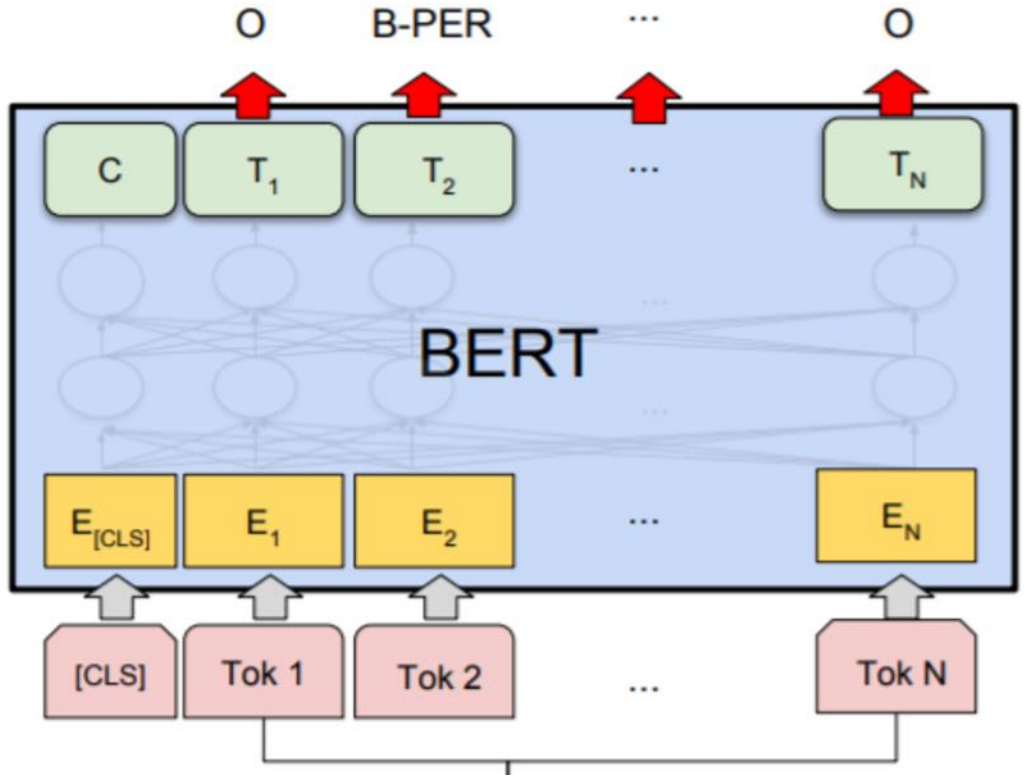
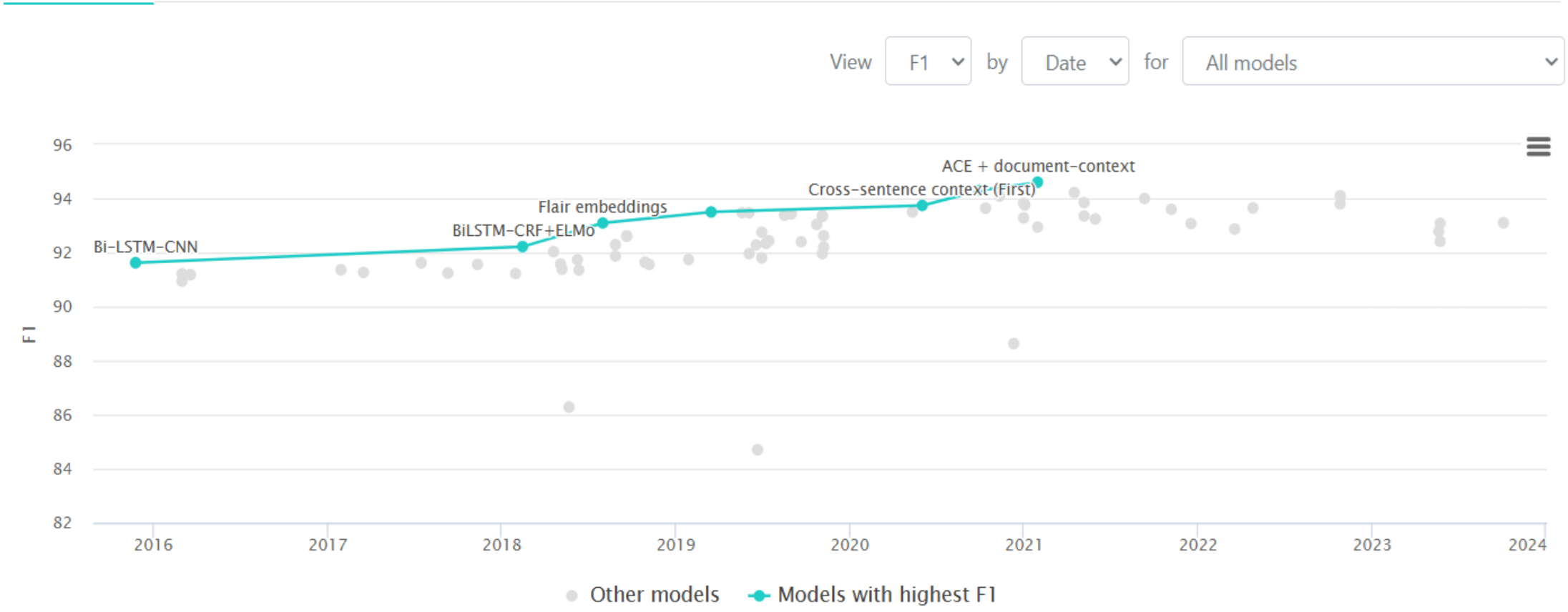


Figure 1: The Transformer - model architecture.



BERT – предобучен на текстовых коллекциях предсказывать маскированную сущность

Изменение результатов на датасете CONLL-2003: 4 типа сущностей (paperwithcode.com)



BERT можно дообучать на целевых коллекциях и улучшать качество изменения именованных сущностей

Распределение сущностей в Sec_col

Тип сущности	Описание	Количество
ORG	организации (для организаций, исключая группы хакеров)	3791
PROGRAM	программы (программные продукты и их составляющие: коды, процедуры)	3497
TECH	технологии (именованные методы и подходы)	2962
LOC	локации (географические локации)	1376
PER	персоны (имена людей, которые не являются хакерами)	1015
DEVICE	устойства (различные электронные и компьютерные устройства)	539
VIRUS	вирусы (вредоносное ПО и уязвимости)	480
EVENT	события	301
HACKER	хакеры (отдельные хакеры и хакерские группы)	60

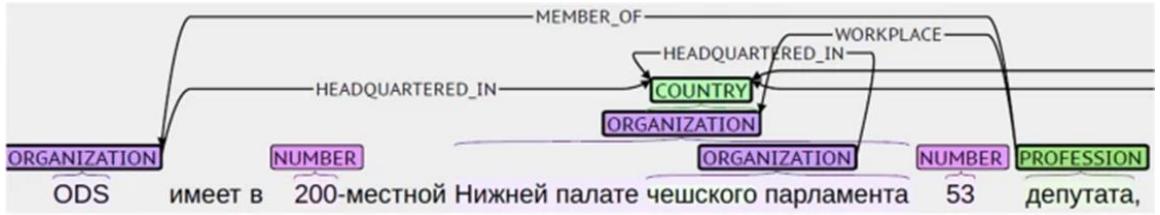
	CRF	BERT	RuBERT	RuCyBERT
DEVICE	31.78	34.04	43.13	**46.77**
EVENT	42.70	60.38	64.49	**67.86**
HACKER	26.58	42.69	52.43	**61.03**
LOC	82.30	90.00	**91.28**	90.01
ORG	68.15	76.10	**78.95**	78.58
PER	67.10	80.99	84.32	**84.56**
PROGRAM	62.15	63.15	64.77	**66.57**
TECH	60.65	67.08	67.60	**69.24**
VIRUS	40.90	40.21	46.92	**54.72**
F-micro	63.95	69.37	71.61	**72.74**
F-macro	53.59	61.63	65.99	**68.82**
F-macro std	–	1.52	0.93	**0.86**

Tikhomirov M. M., Loukachevitch N. V., Dobrov B. V. Recognizing named entities in specific domain //Lobachevskii Journal of Mathematics. – 2020. – Т. 41. – С. 1591-1602.

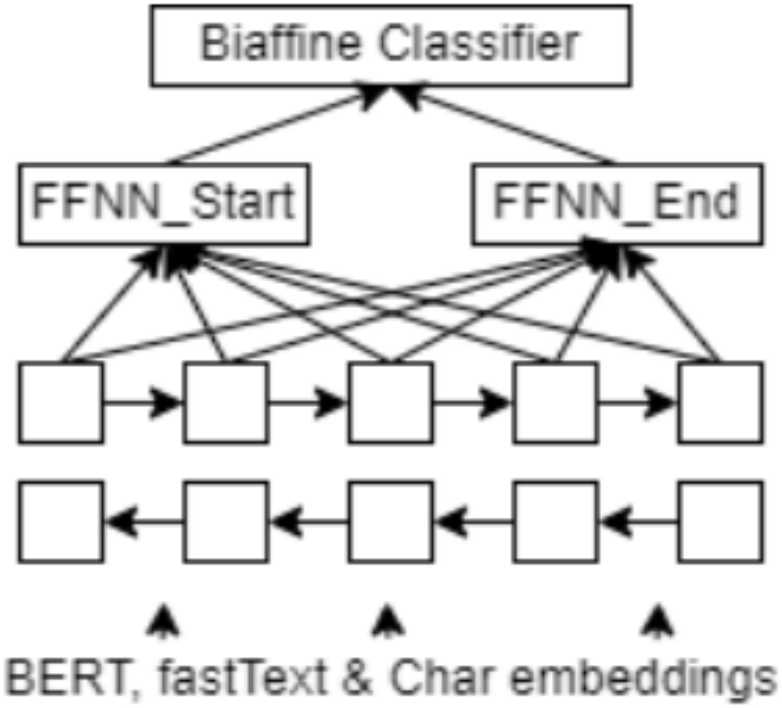
Вложенные именованные сущности

- Сущности могут содержать другие вложенные сущности
  - Примеры: датасет NNE (2019), русскоязычный датасет NEREL
- Можно описывать и извлекать отношения между вложенными сущностями
- Ставить задачу извлечения вложенных сущностей как задачу разметки токенов нецелесообразно
  - Нужно извлекать начало и конец сущности

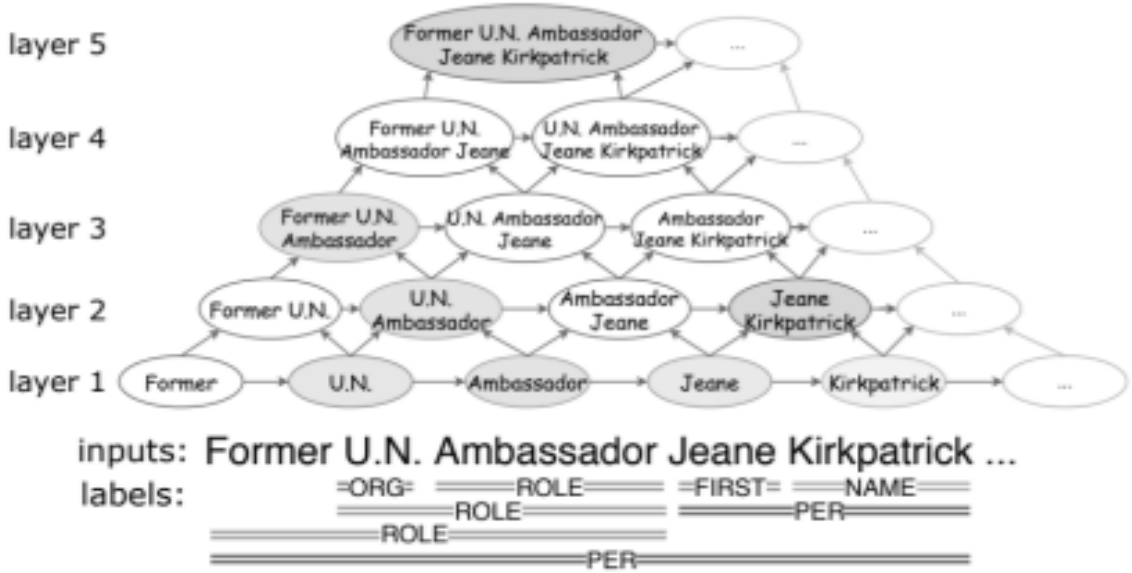
!! До появления нейросетевых моделей задача извлечения вложенных сущностей практически не решалась



Методы извлечения вложенных именованных сущностей



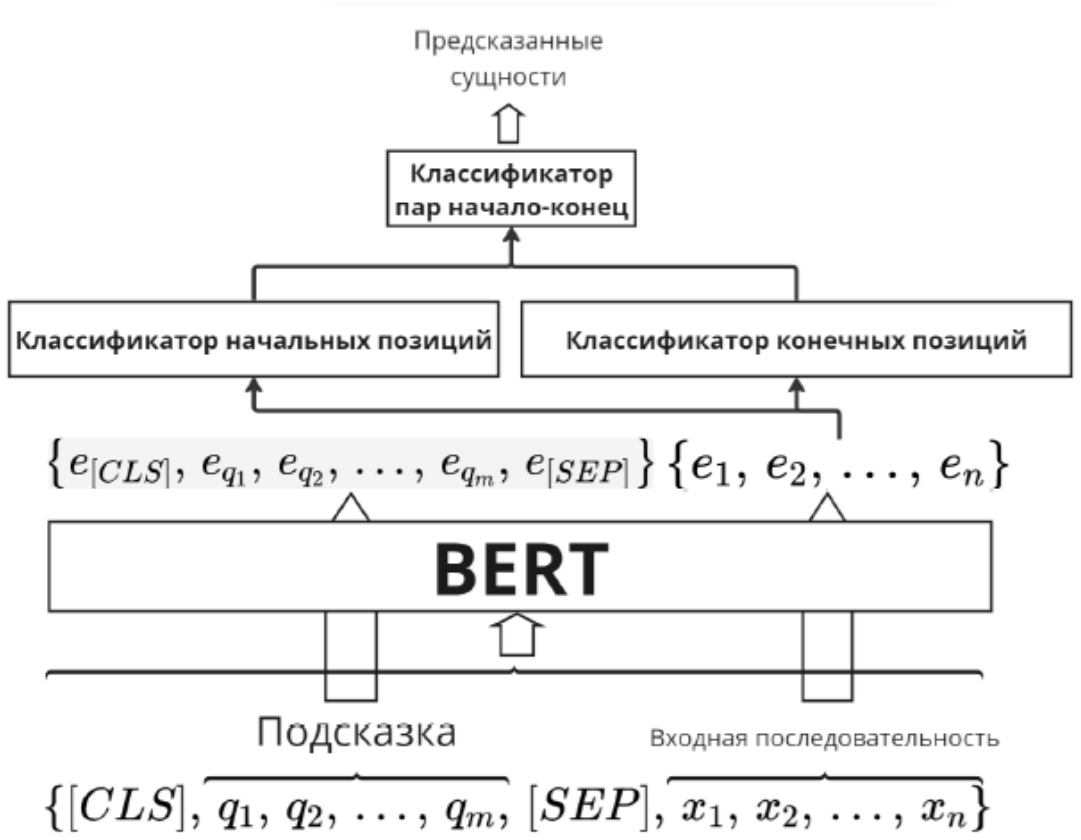
Архитектура Biaffine NER



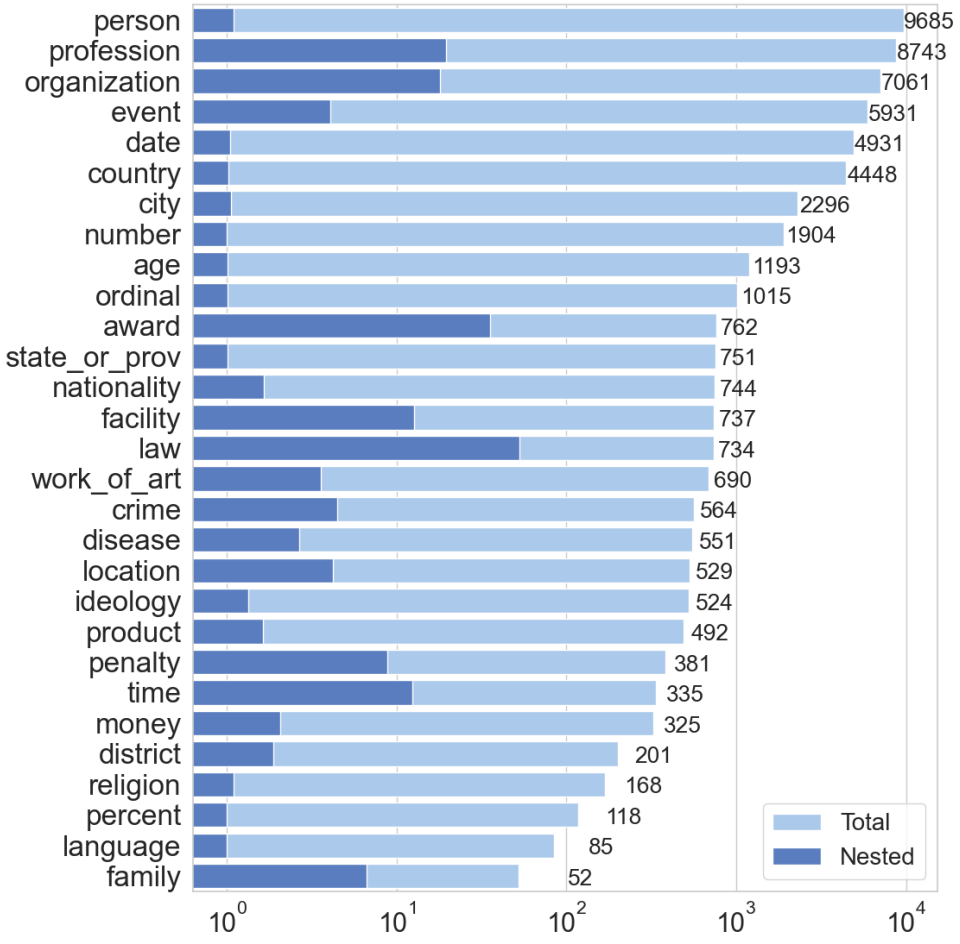
Архитектура Pyramid NER

Метод Machine Reading Comprehension в задаче извлечения именованных сущностей

- Модель MRC даёт ответ A на вопрос Q, выбирая его из контекста C. Таким образом, ответом является подпоследовательность слов исходного предложения
- Запросы (промпты, подсказки):
  - Толкования сущностей
  - Ключевое слово типа
  - Наиболее частотные сущности в обучающих данных



Датасет NEREL – датасет для русского языка с вложенными сущностями



	Dataset	Lang	#NE inst. (Types)	Max Depth	#Rel inst. (Types)
1	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	en	34.5K (4)	1	–
	Ontonotes (Hovy et al., 2006)	en	104K (19)	1	–
2	ACE2005 (Walker et al., 2006)	en	30K (7)	6	8.3K(6)
	NNE (Ringland et al., 2019)	en	279K (114)	6	–
	No-Sta-D (Benikova et al., 2014)	de	41K (12)	2	–
	Digitoday (Ruokolainen et al., 2019)	fi	19K (6)	2	–
	DAN+ (Plank et al., 2020)	da	6.4K (4)	2	–
3	TACRED (Zhang et al., 2017)	en	(3)	1	22.8K (42)
	DocRED (Yao et al., 2019)	en	132K (6)	1	56K (96)
4	Gareev (Gareev et al., 2013)	ru	44K (2)	1	–
	Collection3 (Mozharova and Loukachevitch, 2016)	ru	26.4K(3)	1	–
	FactRuEval (Starostin et al., 2016)	ru	12K (3)	2	1K (4)
	BSNLP (Piskorski et al., 2019)	ru	9K (5)	1	–
	RuREBUS (Ivanin et al., 2020)	ru	121K (5)	1	14.6K (8)
	RURED (Gordeev et al., 2020)	ru	22.6K (28)	1	5.3K(34)
	NEREL -	ru	100K (29)	6	49 (73K)

Loukachevitch N. et al. NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links // Language Resources and Evaluation. – 2024. – Т. 58. – №. 2. – С. 547-583.

Результаты методов извлечения вложенных именованных сущностей на датасете NEREL

Method	P	R	F1
Biaffine, fT	81.64	77.69	79.62
Biaffine, RuBERT, fT	80.71	77.84	79.25
Pyramid, fT	75.87	72.40	74.09
Pyramid, RuBERT, fT	79.54	79.91	79.73
Second best, fT	78.48	63.65	70.29
Second best, RuBERT	82.53	84.41	83.46
SpERT, RuBERT	82.90	82.14	82.52
MRC	**85.04**	**84.95**	**84.99**

Проблема: MRC работает долго

MRC – разные подсказки (промпты)

- **Определения**
- **Контекстный**
- **Лексический** (“премия” -> AWARD)
- **Структурный** (“премия” -> [“премия” | AWARD])
- **Полный лексический**
- **Полный структурный**
- **N самых частотных примеров сущностей** (DATE - это сущность, такие как в 2011 году, в понедельник, завтра.)
- **N самых частотных компонентов сущностей** (DATE - это сущность, такие как год, завтра, месяц.)

Вид подсказки	Общая постановка		
	Точность	Полнота	F1
Определения	78.76	72.44	74.31
2 самых част. сущ.	78.59	72.19	74.17
5 самых част. сущ.	79.23	71.58	73.89
10 самых част. сущ.	78.13	70.64	73.09
2 самых част. комп.	78.65	**73.05**	**74.63**
5 самых част. комп.	78.54	72.77	74.62
10 самых част. комп.	78.04	71.82	73.76
Опр. и 2 комп.	78.37	71.74	73.96
Опр. и 5 комп.	77.83	72.62	74.26
Опр. и 10 комп.	77.60	71.36	73.22
Контекстный	75.97	67.05	69.39
Структурный	80.69	71.47	74.43
Лексический	79.94	71.69	74.30
ПЛ, внутр.	**80.80**	71.37	74.37
ПЛ, внешн.	80.18	72.11	74.54
ПС, внутр.	80.35	72.02	74.40
ПС, внешн.	80.16	71.60	74.07

Rozhkov I. S., Loukachevitch N. V. Prompts in few-shot named entity recognition //Pattern Recognition and Image Analysis. – 2023. – Т. 33. – №. 2. – С. 122-131.

Разметка вложенных именованных сущностей в датасете NEREL-BIO

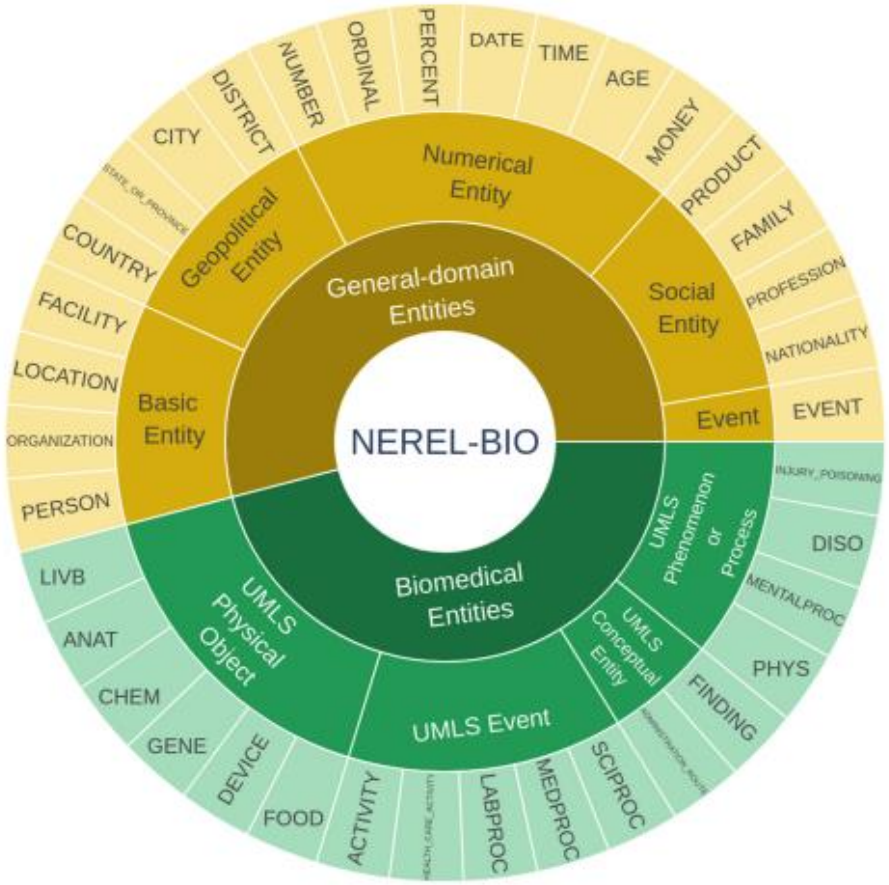
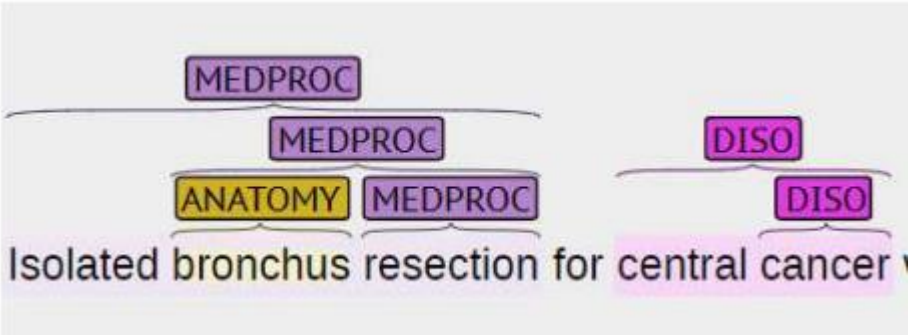


Table 1 Statistics of NEREL-BIO.

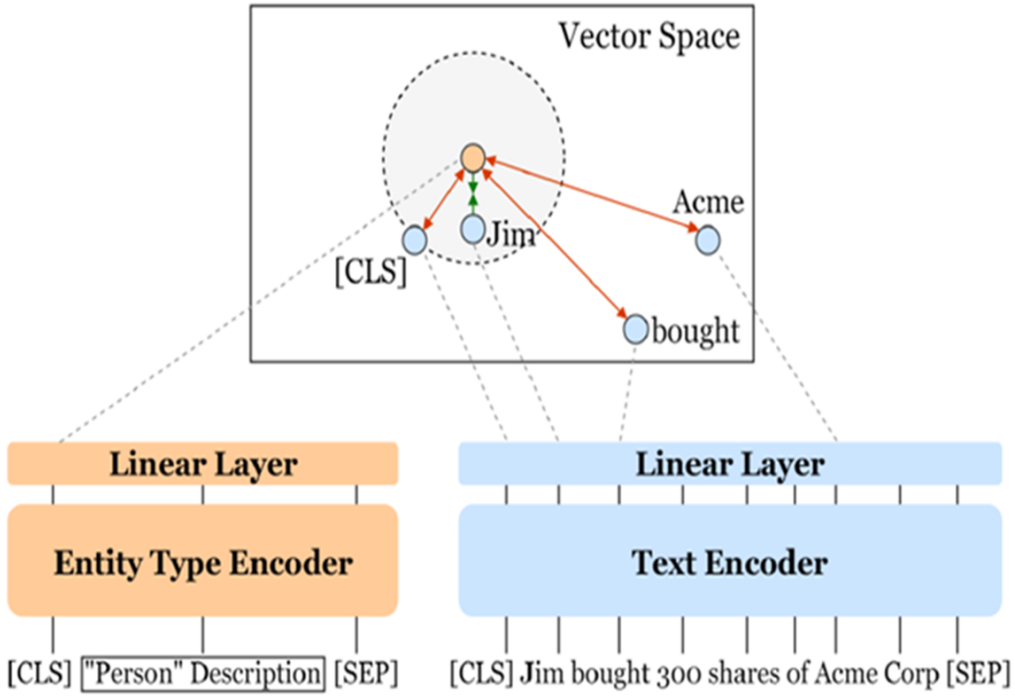
Collection	No. of doc	No. of entities	No. of nonzero entity types
Abstracts in Russian	766	66 888	37
Abstracts in English	105	10 651	32



Loukachevitch N. et al. NEREL-BIO: a dataset of biomedical abstracts annotated with nested named entities // Bioinformatics. – 2023. – Т. 39. – №. 4. – С. btad161.

Метод извлечения именованных сущностей Binder (2023)

- Извлекает вложенные сущности
- Использует контрастивное обучение
- У каждого типа сущности должно быть описание Ei
- Основная идея контрастивного обучения:
  - Нужно описания сущностей и входные данные преобразовать в некоторое векторное пространство, чтобы сущности оказались ближе к своему правильному типу



$$\ell_{\text{span}} = -\log \frac{\exp(\text{sim}(\mathbf{s}_{i,j}, \mathbf{e}_k))}{\sum_{\mathbf{s}' \in \mathcal{S}_k^- \cup \mathbf{s}_{i,j}} \exp(\text{sim}(\mathbf{s}', \mathbf{e}_k))},$$

Заключение к разделу NER

- Плоские vs. Вложенные сущности
- Метод для использования на основе контрастивного обучения Binder

- !! Именованные сущности могут извлекаться и на правилах
  - Расположены в одном предложении и компактно
  - Требует развитая инструментальная система записи правил

Тестирование RuNNE-2022 на основе NEREL

User	**Team**	**# of runs**	**F1**$_{full\_set}$	**F1**$_{few-shot}$	System Summary
Baseline		-	0.674	0.447	RuBERT
			Participating Teams		
ksmith	Pullenti	44	**0.811**	**0.710**	Rule-based
abrosimov_kirill	Saldon	20	0.741	0.644	Sodner model, labelling
fulstock	MSU-RCC	7	0.749	0.604	MRC model
svetlan		23	0.607	0.572	n/a
LIORI		6	0.653	0.433	n/a
botbot		8	0.460	0.414	n/a
bond005	SibNN	20	0.743	0.404	Siemese network, Viterbi alg.
Stud2022		24	0.477	0.395	n/a
mojesty		2	0.619	0.172	n/a

Artemova E. et al. Runne-2022 shared task: Recognizing nested named entities //arXiv preprint arXiv:2205.11159. – 2022.

Методы извлечения отношений

- Сложная задача для методов на правилах и классических методов машинного обучения
- Ситуация изменилась с приходом нейросетевых подходов, еще лучшие результаты дали кодировщики трансформеров (BERT)

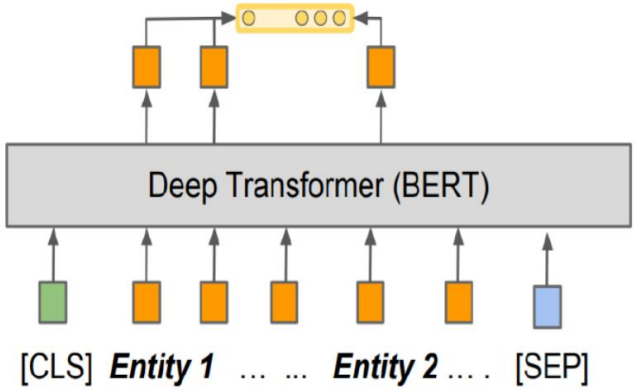
Влияние признаков на качество извлечения отношений (Zhou et al. 2005)

Features	P	R	F
Words	69.2	23.7	35.3
+Entity Type	67.1	32.1	43.4
+Mention Level	67.1	33.0	44.2
+Overlap	57.4	40.9	47.8
+Chunking	61.5	46.5	53.0
+Dependency Tree	62.1	47.2	53.6
+Parse Tree	62.3	47.6	54.0
+Semantic Resources	63.1	49.5	55.5

BERT в задаче извлечения отношений (Soares et al., 2019)

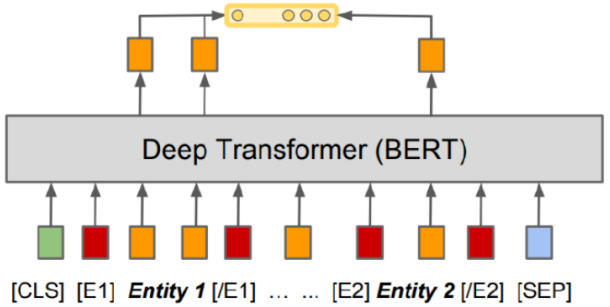
1) Standard – [CLS]
2) **Standard – mention pooling**

- **Вход:** стандартный
- **Выход:** max pooling векторов между сущностями



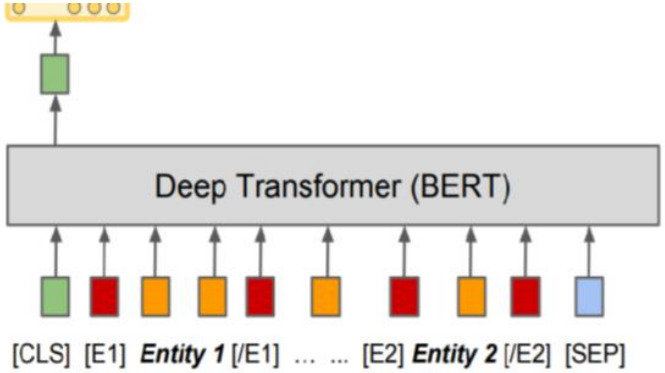
Подходы сведения к задаче извлечения отношений

1) Standard – [CLS]
2) Standard – mention pooling
3) Positional emb. – mention pool
4) Entity markers – [CLS]
5) **Entity markers – mention pool.**

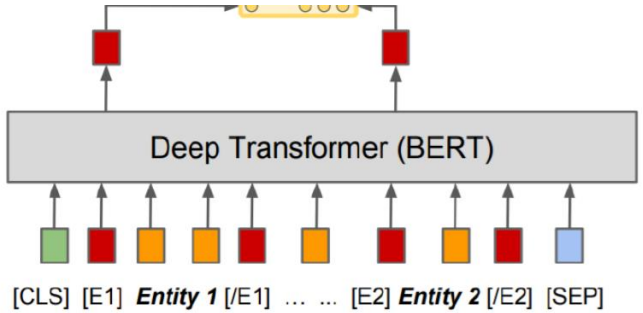


1) Standard – [CLS]
2) Standard – mention pooling
3) Positional emb. – mention pool
4) **Entity markers – [CLS]**

- **Вход:** добавляются спец токены на границах сущностей
- **Выход:** эмбединг [CLS] токена

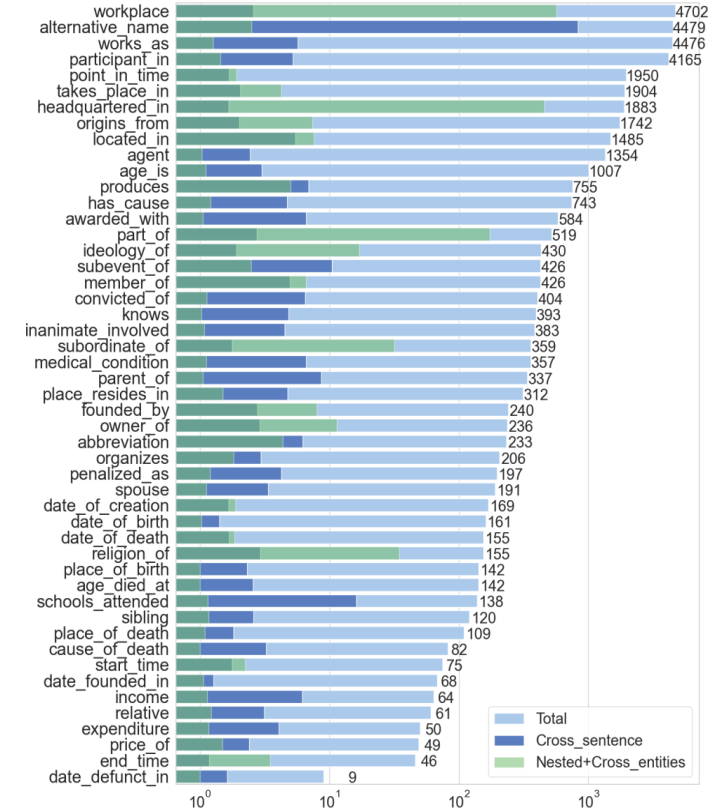


1) Standard – [CLS]
2) Standard – mention pooling
3) Positional emb. – mention pool
4) Entity markers – [CLS]
5) Entity markers – mention pool.
6) **Entity markers – entity start**



NEREL: 49 типов отношений

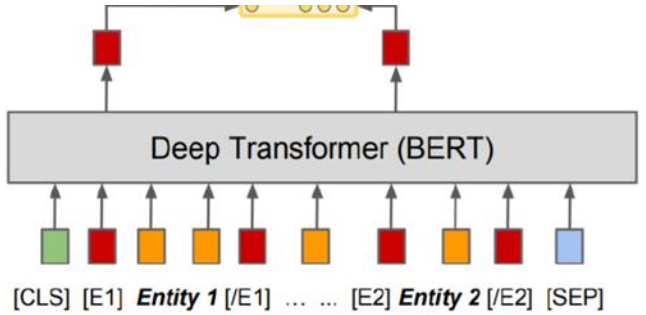
- Отношения людей:
  - Workplace, place_of_birth..
- Отношения организаций
  - Workplace, headquartered..
- Отношения событий
  - Roles
  - Temporal relations
  - Place of event
  - Causal relations
- Синонимичные отношения
  - Abbreviation
  - Alternative_name



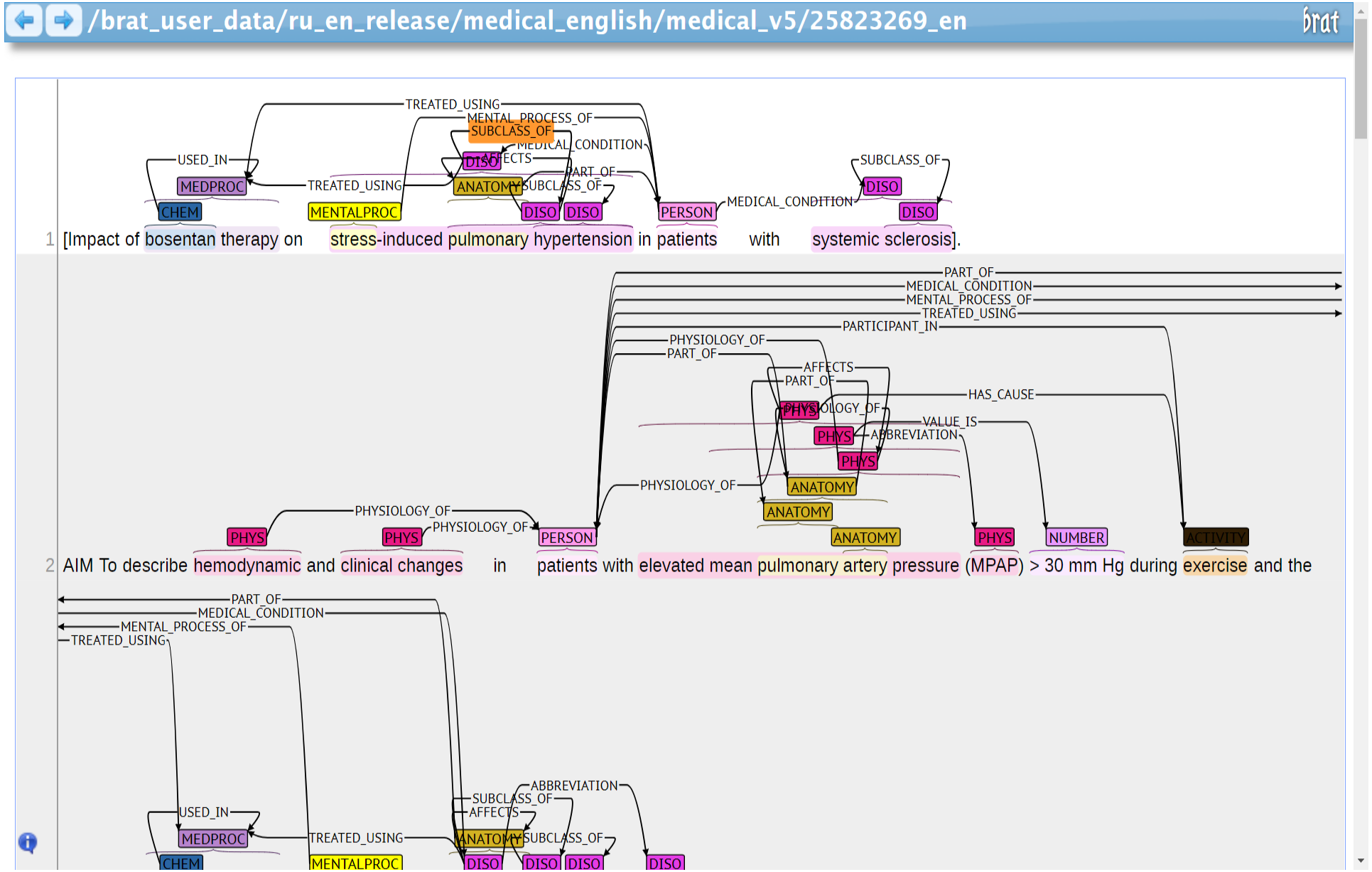
Результаты извлечения отношений для датасета NEREL (Пакет OpenNRE)

Метод	F-score	
Cls-pooling	51.0	
Entity-pooling	65.0	
ent.marker-ent	**80.1**	
ent.marker-cls	75.9	

1) Standard – [CLS]
2) Standard – mention pooling
3) Positional emb. – mention pool
4) Entity markers – [CLS]
5) Entity markers – mention pool.
6) **Entity markers – entity start**



Разметка отношений NEREL-BIOeng



Извлечение информации из текстов и графы знаний

- Извлеченные сущности и отношения нужно загружать в базу знаний
- Для этого нужно проверять, есть ли такая сущность и такое отношение в базе

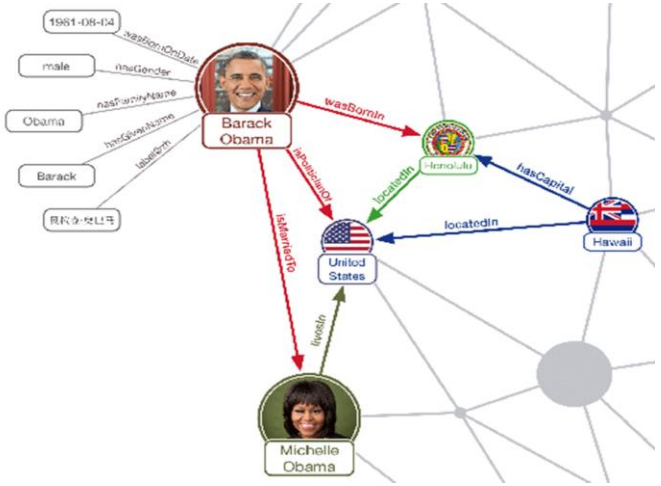
- Связывание сущностей (Entity Linking)
  - Извлеченные сущности нужно связать с какой-то базой знаний.
  - А если нет такой сущности в базе, то поставить пустой концепт

- Нужны большие графы знаний, чтобы использовать их в современных задачах
  - Графы знаний никогда не полны, нужно пополнять

- Пополнение графов знаний из текстов
  - Извлечение именованных сущностей
  - Извлечение отношений между именованными сущностями
  - Entity linking (связывание сущностей с графом)

- Примеры графов знаний: Викиданные, UMLS

Графы знаний



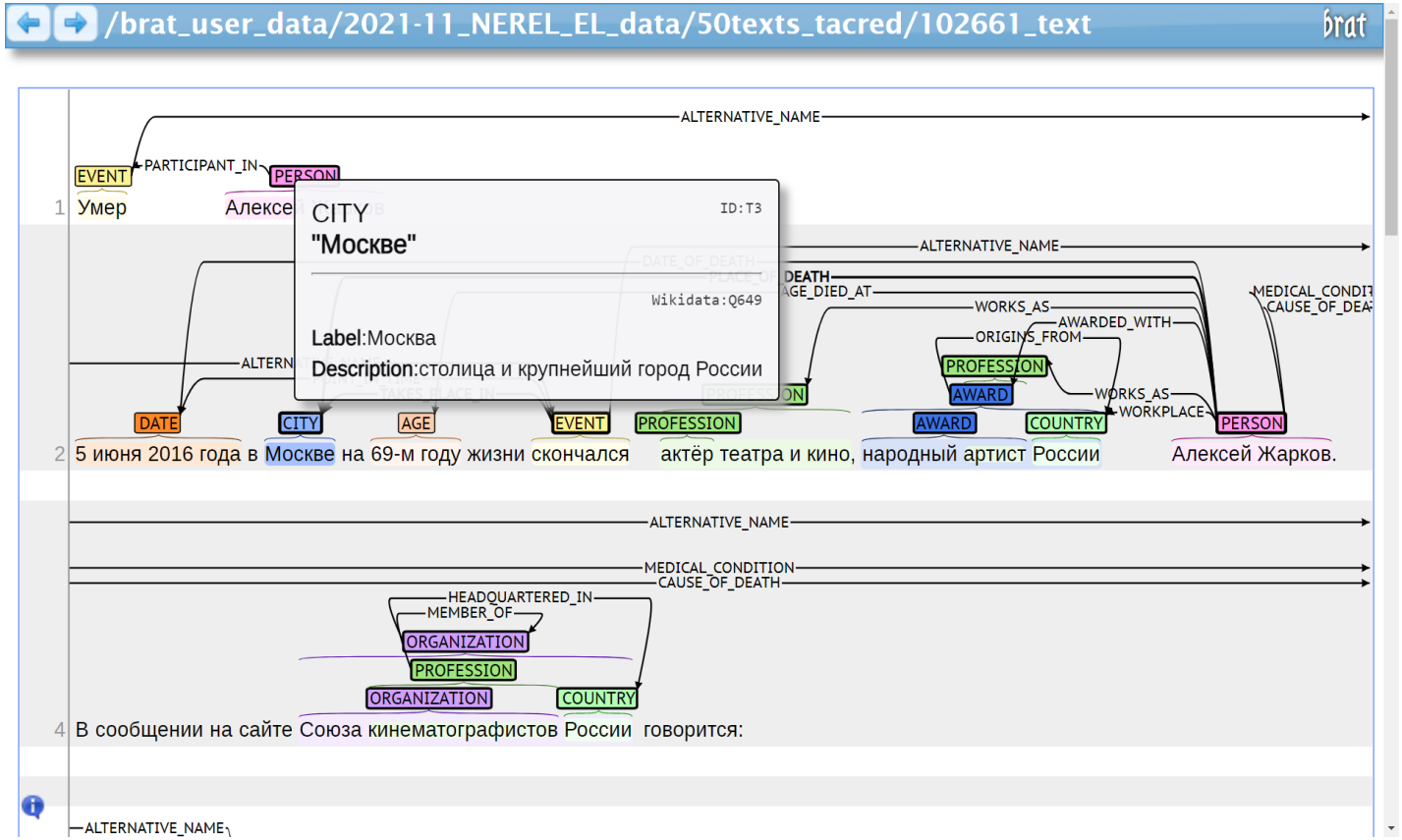
- Содержание
  - Схема или онтология
  - классы, подклассы, экземпляры, типы отношений, аксиомы
  - Большие объемы конкретных сущностей и конкретных отношений (исходно в онтологиях был больший акцент на абстрактные сущности)
- База данных. Часто хранится в виде триплетов: субъект-отношение-объект

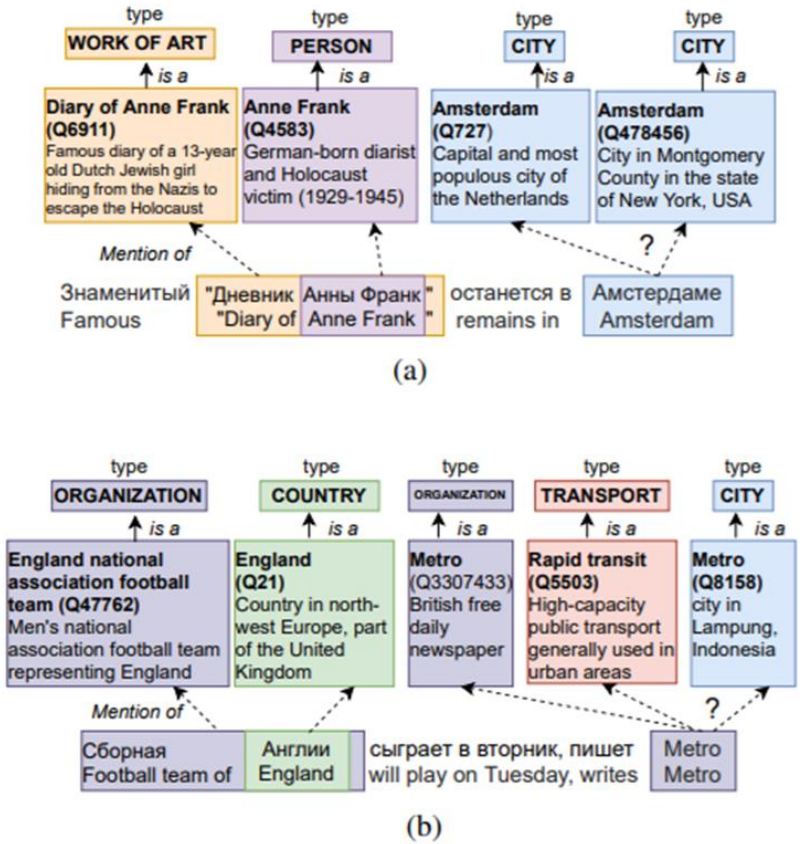
- Граф: можно использовать структуру сети для различного рода задач

instance of	capital of Russia	
	start time	1918
	2 references	
	federal city of Russia	
	start time	1991
	1 reference	
	big city	
	start time	1600s
	0 references	

Фрагмент Викиданных для сущности Москва

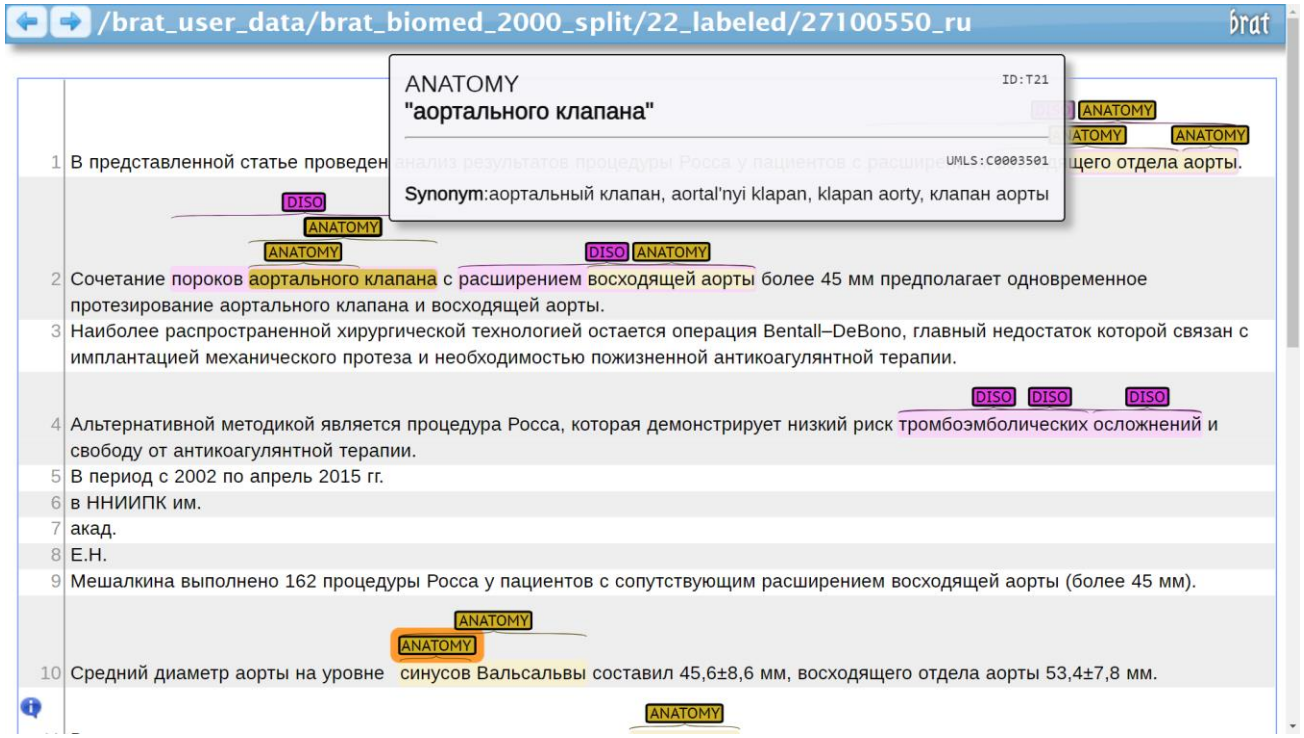
NEREL: все три уровня разметки

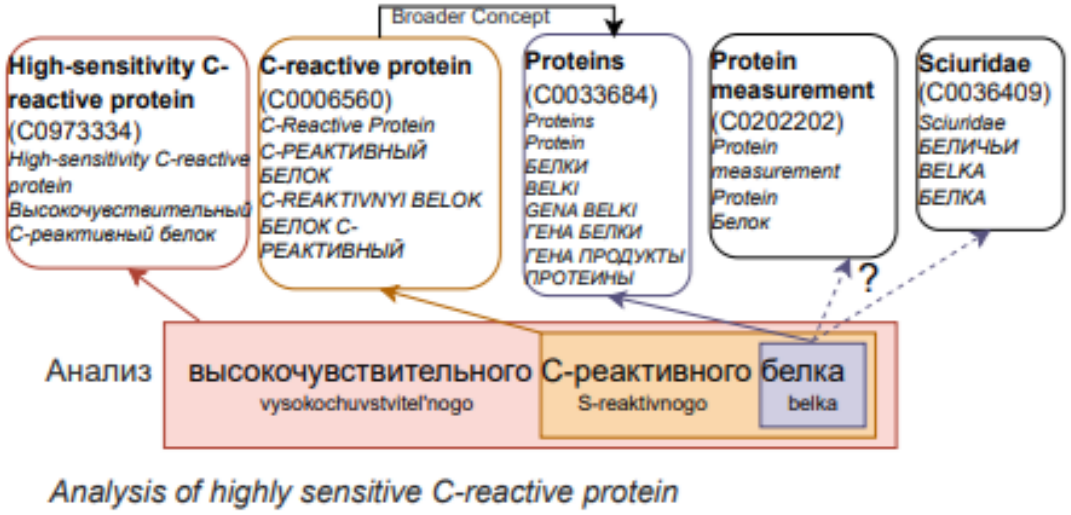




Вложенные сущности позволяют сделать связывание сущностей для всех упомянутых сущностей

Пример связывания медицинских понятий в датасете NEREL-BIO с медицинской базой UMLS



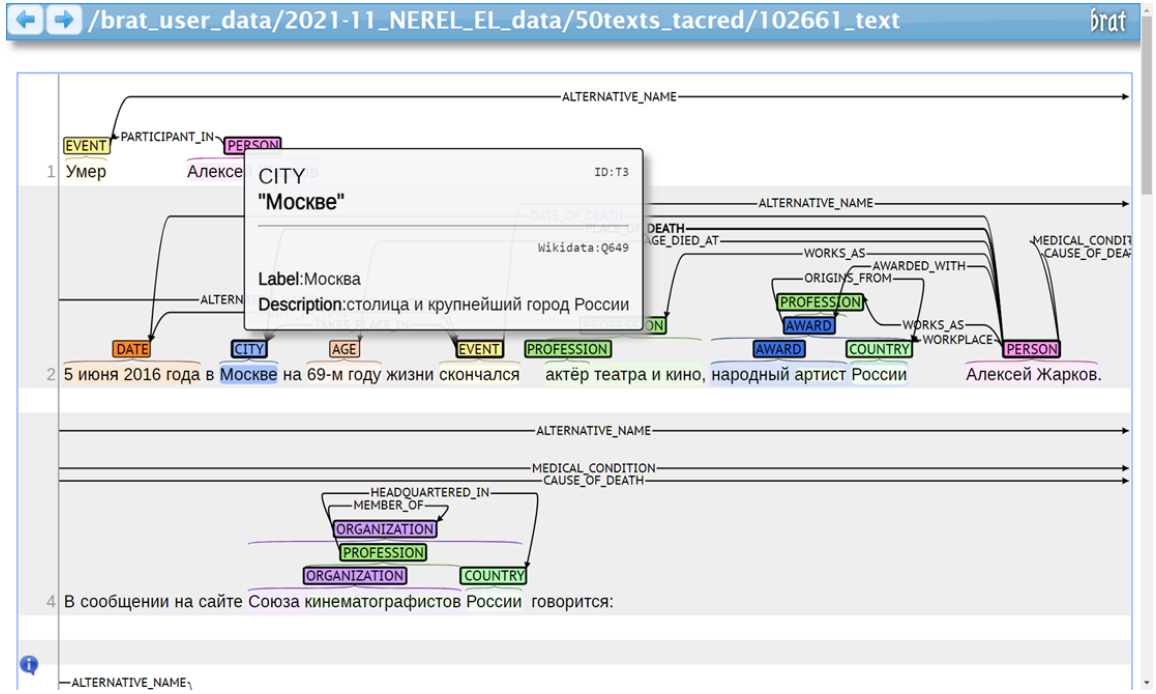


В UMLS не хватает русскоязычных переводов, поэтому в таких случаях автоматическое связывание затруднено

Loukachevitch N., Sakhovskiy A., Tutubalina E. Biomedical Concept Normalization over Nested Entities with Partial UMLS Terminology in Russian //LREC-COLING 2024. – 2024. – С. 2383-2389.

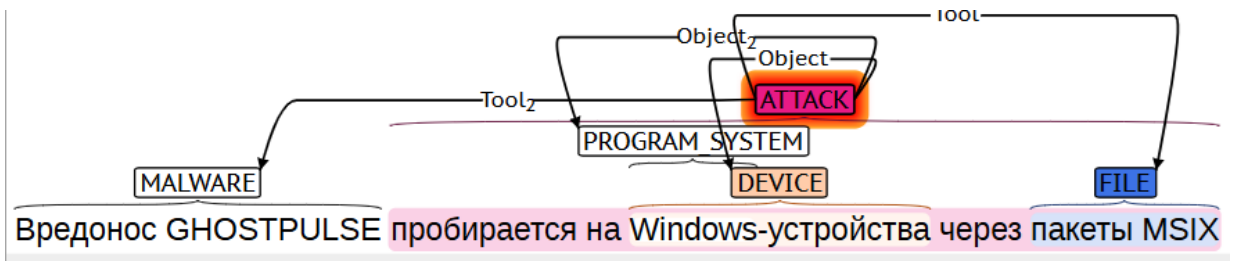
Извлечение событий

- Именованные события
  - *Московский Международный Кинофестиваль*
  - Одна из именованных сущностей
- Неименованные события
  - *Задержать, встретиться, украсть,*
  - Такая разметка важна не сама по себе, а с отношениями-ролями
  - Что именно будем считать событием
  - Что представляет из себя разметка собственно-события – event nugget
    - *Столкнуться с препятствием*



NEREL : разметка главного слова или зависимых слов, если они уточняют смысл происходящего.
Размеченные именованные сущности не включаются

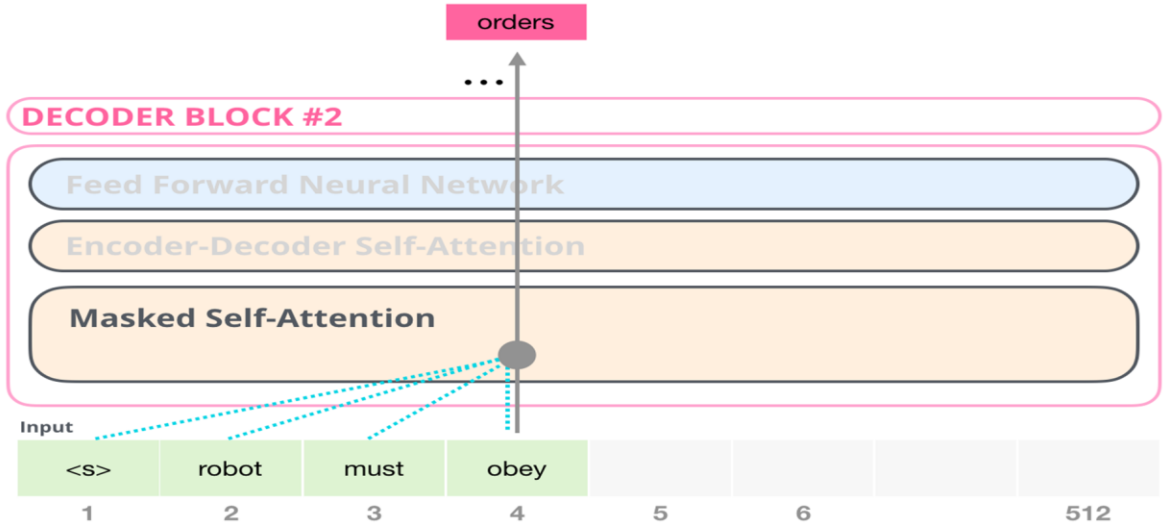
Разметка событий компьютерной безопасности



Разметка события включает не только главное слов, но и участников события, расположенные рядом с ним. После извлечения по названию события сразу понятно, что происходило

Entity	Precision, %	Recall, %	F1, %
ATTACK	69.58	72.55	71.04
DAMAGE	40.13	34.05	36.84
DEVICE	62.56	77.09	69.07
DISCOVER_VULNERABILITY	57.14	16.67	25.81
FILE	51.90	63.56	57.14
GPE	96.90	95.06	95.97
HACKER	90.70	92.58	91.63
HACKER_GROUP	88.55	84.76	86.61
INFOSOURCE	72.00	78.26	75.00
MALWARE	69.51	77.05	73.09
MONEY	80.00	88.89	84.21
ORGANIZATION	77.76	78.57	78.16
PATCH_VULNERABILITY	50.00	50.00	50.00
PERSON	83.33	89.43	86.27
PROGRAM_SYSTEM	60.85	66.82	63.70
PROTECTION	31.43	34.38	32.84
SPECIALIST	88.58	91.51	90.02
TIME	95.00	88.08	91.41
VULNERABILITY	81.12	81.49	81.31
WEBSITE	65.12	69.14	67.07
micro	**72.64**	**76.06**	**74.31**
macro	**70.71**	**71.71**	**70.55**

Большие языковые модели в задачах извлечения информации (GPT и др.)



Model	*Zero-Shot*		*Fine-Tuned*		
	ChatGPT	GPT-3.5	Flair	LUKE	ACE
All	53.2	**53.5**	93.0	93.9	**94.6**
Loc	66.7	**67.1**	94.0	-	-
Per	**87.2**	78.0	97.4	-	-
Org	**51.4**	50.0	91.9	-	-
Misc	4.1	**4.8**	83.0	-	-

Qin C. et al. Is ChatGPT a General-Purpose Natural Language Processing Task Solver? //Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. – 2023. – С. 1339-1384.

Специальное исследование подходов к извлечению информации на основе ChatGPT

- Между методами ChatGPT и SOTA существует значительный разрыв в уровне результатов.
- Чем сложнее задача, тем больше разрыв.
- ChatGPT может сравняться с методами SOTA или превзойти их в нескольких простых случаях.

- использование промптов с несколькими примерами обычно приводит к значительным улучшениям (около 3,0~13,0), но все еще явно отстает от результатов SOTA.

Han R. et al. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors //arXiv preprint arXiv:2305.14450. – 2023.

Промпты для извлечения сущностей: zero-shot и few-shot

- Considering 4 types of named entities including "Organization", "Person", "Location" and "Miscellaneous", recognize all named entities in the given sentence.
- Answer in the format ["entity_type", "entity_name"] without any explanation.
- If no entity exists, then just answer "[]".
- Sentence: "In Home Health said it previously recorded a reserve equal to 16 percent of all revenue related to the community liaison costs."
- Answer

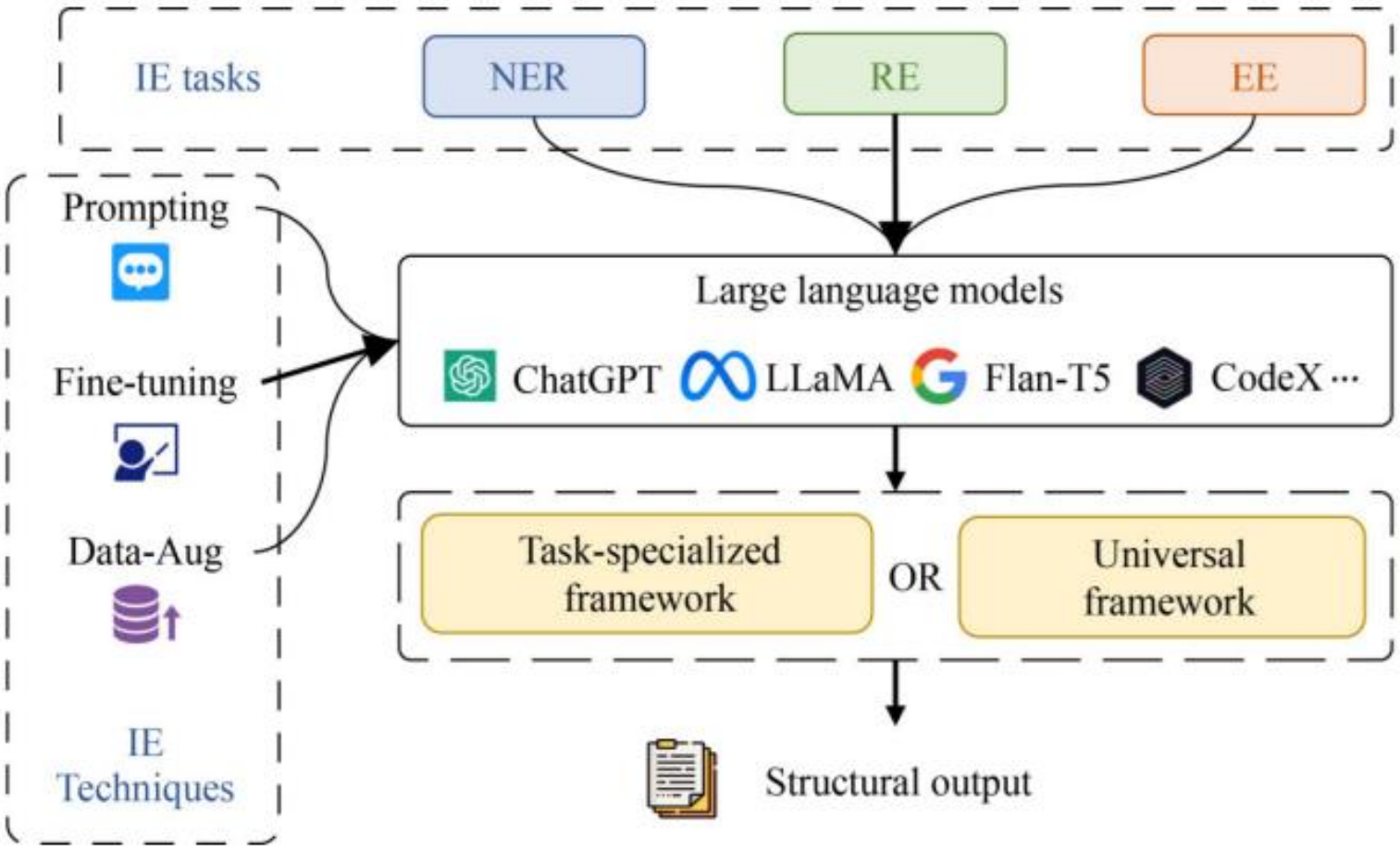
- Considering 4 types of named entities including "Organization", "Person", "Location" and "Miscellaneous", recognize all named entities in the given sentence.
- Answer in the format ["entity_type", "entity_name"] without any explanation. If no entity exists, then just answer "[]".
- Sentence: "The arrangement calls for investors to make additional payments to fund Equitas but also provides them with 3.2 billion stg in compensation to help reduce their prior outstanding liabilities.“
- Answer: ["Organization", "Equitas"]
- Sentence: "Results from the U.S. Open Tennis Championships at the National Tennis Centre on Saturday (prefix number denotes seeding):“
- Answer: ["miscellaneous", "U.S. Open Tennis Championships"], ["location", "National Tennis Centre"] ... (More examples are omitted here.)
- Sentence: "Women’s 3,000 metres individual pursuit qualifying round" Answer: []
- Sentence: "In Home Health said it previously recorded a reserve equal to 16 percent of all revenue related to the community liaison costs." Answer:

Сравнение ChatGPT с лучшими результатами на коллекциях

			Zero-shot		5 примеров
NER-Flat	CoNLL03	94.6(Wang et al., 2021)	65.13	60.10 (3.81)	70.53 (1.44)
	FewNERD	67.1(Ding et al., 2021)	34.28	31.56 (2.44)	36.87 (0.71)
NER-Nested	ACE04	88.5(Yang et al., 2023)	29.55	27.80 (3.10)	38.52 (2.51)
	ACE05-Ent	87.5(Yang et al., 2023)	24.77	23.38 (1.92)	36.17 (1.78)
	GENIA	81.5(Yang et al., 2023)	39.43	38.09 (1.65)	48.82 (1.31)
RE-RC	CoNLL04	-	65.82	59.21 (3.85)	55.32 (4.56)
	NYT-multi	93.5(Zhan et al., 2022)	38.74	30.96 (5.51)	26.88 (2.74)
	TACRED	75.6(Li et al., 2022)	21.58	19.47 (1.49)	27.84 (3.48)
	SemEval2010	91.3(Zhao et al., 2021)	42.32	39.27 (2.20)	39.44 (2.55)
RE-Triplet	CoNLL04	78.8(Lou et al., 2023)	23.04	17.84 (3.43)	24.30 (1.29)
	NYT-multi	86.8(Wang et al., 2023b)	3.79	3.48 (0.24)	12.24 (0.59)
	SemEval2010	73.2(Wang et al., 2023a)	7.65	5.82 (1.29)	12.85 (1.14)

Han R. et al. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors //arXiv preprint arXiv:2305.14450. – 2023.

End-to-end Information extraction (Сквозное извлечение информации)



Xu D. et al. Large language models for generative information extraction: A survey //Frontiers of Computer Science. – 2024. – Т. 18. – №. 6. – С. 186357.

Заключение

- Извлечение информации остается важной задачей автоматической обработки текстов
  - Вложенные именованные сущности
  - Разметка событий
  - Разметка событий в виде вложенных именованных сущностей
  - Три этапа анализа текстов: извлечение сущностей, отношений, связывание сущностей –> пополнение графов знаний

- Методы:
- На основе кодировщиков трансформеров: BERT и др.
  - Контрастивное обучение: Binder
  - Но нужна разметка обучающей коллекции
- На основе порождающих подходов: GPT и др.
  - Попытка обойтись без обучающих коллекций: zero-shot и few-shot подходы
  - Но пока результаты существенно хуже
- Подходы: pipeline vs. end-to-end