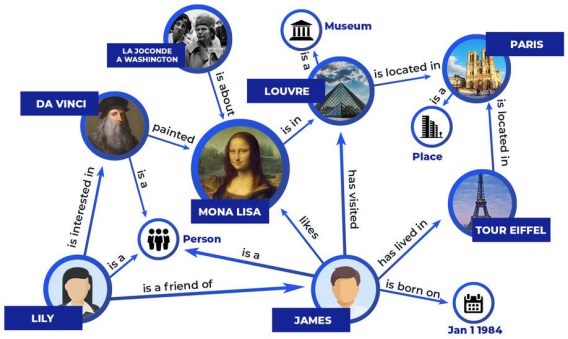


Entity Linking over Nested Named Entities for Russian

Natalia Loukachevitch, Pavel Braslavski, Vladimir Ivanov, Tatiana Batura, Suresh Manandhar, Artem Shelmanov, Elena Tutubalina

Lomonosov Moscow State University, Ural Federal University, HSE University, Novosibirsk State University, Innopolis University, Wiseyak (United States), Kazan Federal University, Sber AI

LREC-2022

Knowledge graph generation from texts

- Knowledge graphs became important instrument in NLP applications in combination with neural networks
  - Available knowledge graphs are not complete
  - The completion of knowledge graph from texts is widely discussed

- Knowledge graph construction from texts
  - Named entity recognition
  - Relation extraction
  - Entity linking

- NEREL – the largest Russian dataset for supervised knowledge graph construction from texts
  - Annotation is based on nested named entities
  - Annotation of named entities should account for next stages of annotation

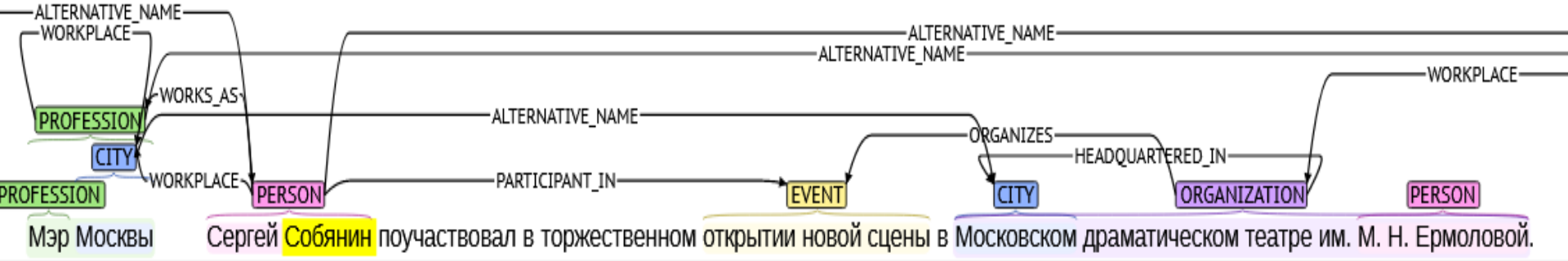
NEREL dataset for the Russian language

- Nested named entities: longer named entities can include shorter named entities
  - Nestedness of entities enables a more accurate and complete description of relations and links to knowledge bases
  - 29 Entity types

- Relation annotated on sentence and document level - 49 relations types
  - Intra entity relations between nested named entities

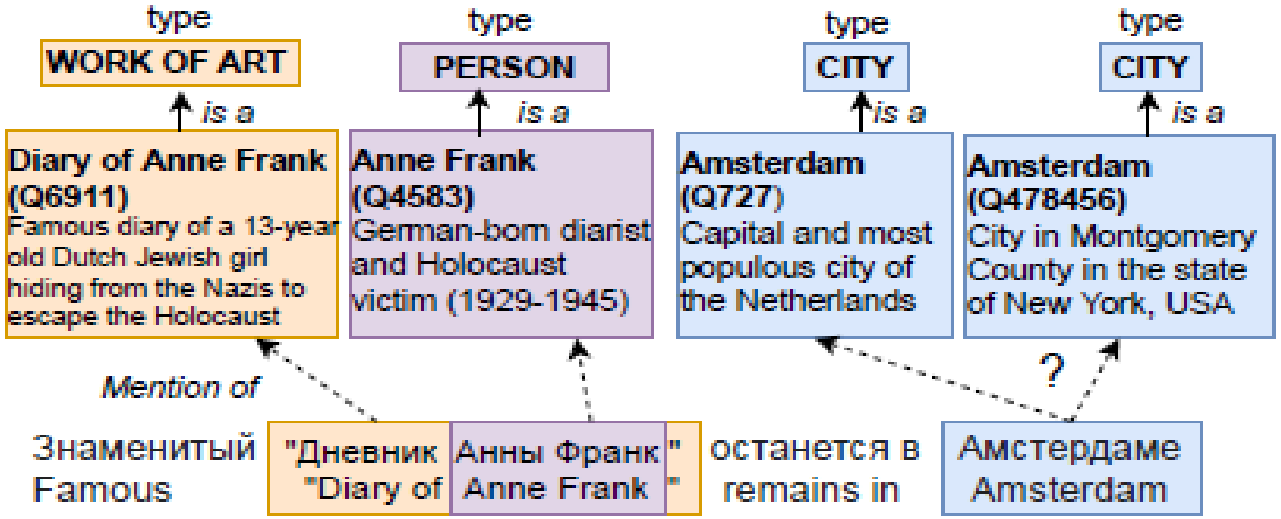
- Wikidata entity links over nested named entities

Example of annotation: nested entities and relations

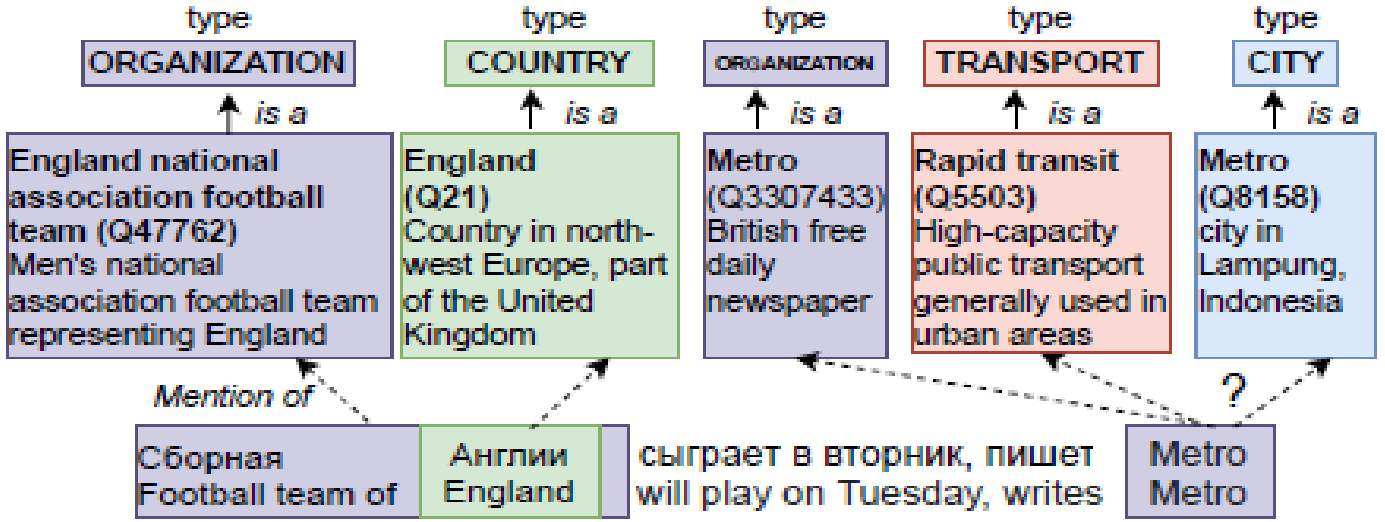


- [[Moscow] [Mayor]] [Sergei Sobyanin] took part in the [grand opening] of the new stage of [[Moscow] [Ermolova] theater]

- Nested named entities:
  - Mayor of Moscow: Moscow, Mayor;
  - Moscow Ermolova Theater: Moscow, Ermolova.
- The intra-entity relations are as follows:
  - Moscow is a workplace for Mayor of Moscow;
  - Moscow Ermolova Theater is headquartered in Moscow.
- Grand opening of the new stage is annotated as an event.


type
WORK OF ART
is a
Diary of Anne Frank (Q6911)
Famous diary of a 13-year old Dutch Jewish girl hiding from the Nazis to escape the Holocaust
type
PERSON
is a
Anne Frank (Q4583)
German-born diarist and Holocaust victim (1929-1945)
type
CITY
is a
Amsterdam (Q727)
Capital and most populous city of the Netherlands
type
CITY
is a
Amsterdam (Q478456)
City in Montgomery County in the state of New York, USA
Mention of
?
Знаменитый "Дневник Анны Франк" останется в Амстердаме
Famous "Diary of Anne Frank" remains in Amsterdam

(a)


type
ORGANIZATION
is a
England national association football team (Q47762)
Men's national association football team representing England
type
COUNTRY
is a
England (Q21)
Country in north-west Europe, part of the United Kingdom
type
ORGANIZATION
is a
Metro (Q3307433)
British free daily newspaper
type
TRANSPORT
is a
Rapid transit (Q5503)
High-capacity public transport generally used in urban areas
type
CITY
is a
Metro (Q8158)
city in Lampung, Indonesia
Mention of
?
Сборная Англии сыграет в вторник, пишет Metro
Football team of England will play on Tuesday, writes Metro

(b)

Dataset collection

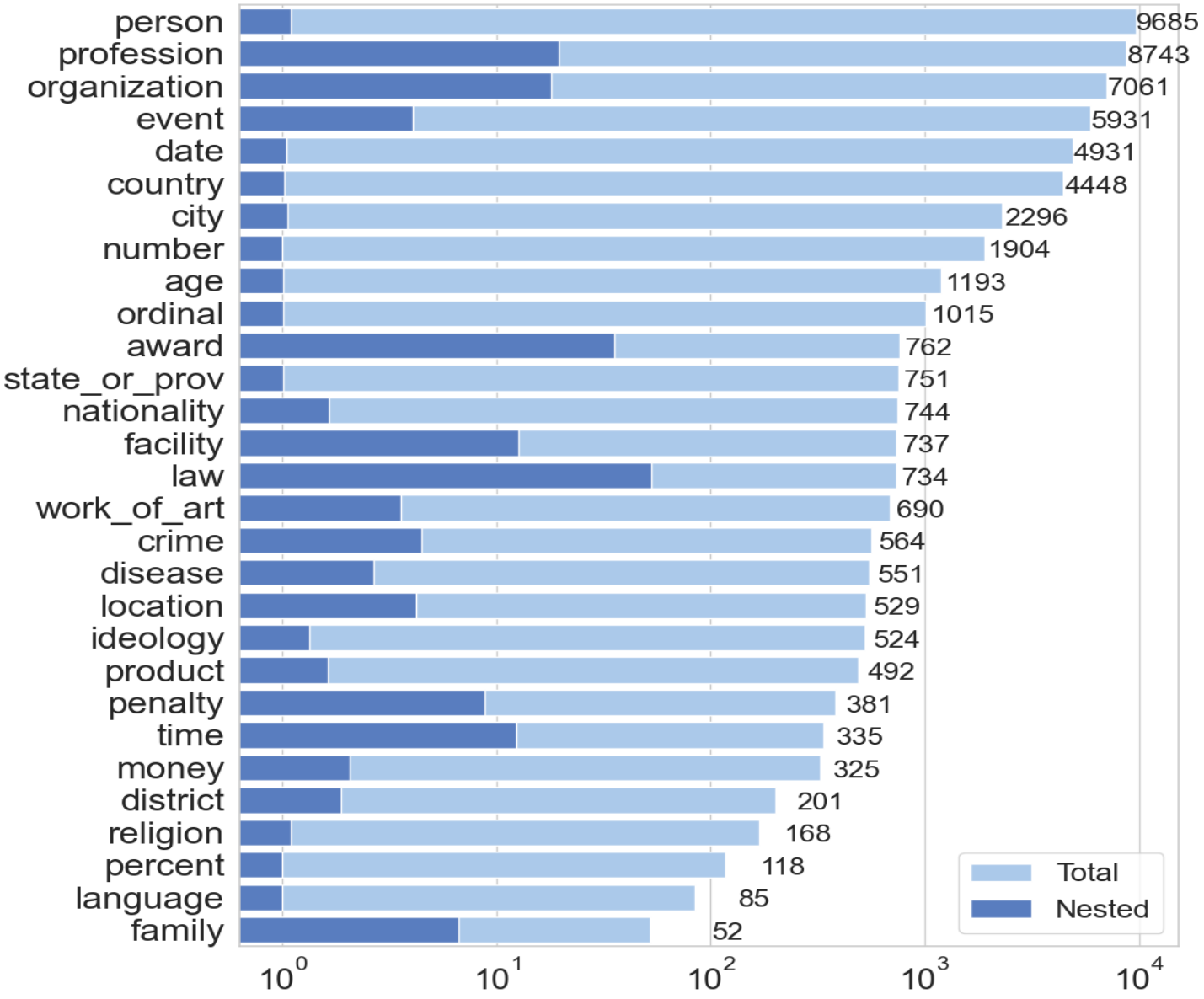
- NEREL dataset consists primarily of Wikinews articles
  - Creative Commons License (CCBY2.5) allowing reuse of the published materials.
  - Wikinews articles are partially linked to Wikipedia articles
  - It is possible to obtain counterparts of some news, written in other languages
- From Wikinews articles we extracted documents containing large number of person-oriented information (age, parents)
  - Models trained on the recent dataset RURED (Gordeev et al., 2020)
- The extracted articles were inspected manually to balance topics and remove inappropriate documents.
  - Articles with size 1-5 Kb were selected for the dataset

	Dataset	Lang	Domain	#NE inst. (Types)	Max Depth	#Rel inst. (Types)	Annot Levels
1	CoNLL03/AIDA [62,23]	en	News	34.5K (4)	1	–	NE/EL
	Ontonotes [24]	en	News/Web	104K (19)	1	–	NE
2	ACE2005 [68,3]	en	News	30K (7)	6	8.3K(6)	NE/RE/EL
	NNE [51]	en	News	279K (114)	6	–	NE
	No-Sta-D [2]	de	News	41K (12)	2	–	NE
	Digitoday [54]	fi	News	19K (6)	2	–	NE
	DAN+ [48]	da	News	6.4K (4)	2	–	NE
3	TACRED [73]	en	News	(3)	1	22.8K (42)	RE
	DocRED [70]	en	Wiki	132K (6)	1	56K (96)	RE
4	DBp. Spotlight [37]	en	News	330	2	–	EL
	AQUAINT [39]	en	News	727	1	–	EL
	TAC-KBP-2010	en	News/Web	1,020	2	–	EL
	VoxEL* [53]	5 lang.	News	204	1	–	EL
	VoxEL** [53]	5 lang.	News	674	2	–	EL
	DWIE [72]	en	News	43,373 (311)	1	16,844 (65)	NE/RE/EL/RF
	SCIERC [36]	en	Science	8,089(6)	1	4,716(9)	TE/RE/RF
5	Gareev [15]	ru	News	44K (2)	1	–	NE
	Collection3 [43]	ru	News	26.4K(3)	1	–	NE
	BSNLP2019 [47]	Slavic	News	9K (5)	1	–	NE/EL
	BSNLP2021 [46]	Slavic	News	9.5K(5)	1	–	NE/EL
	MultiCoNER-2022	11 lang.	News	16.1K(6)	1	–	NE
	FactRuEval [59]	ru	News	12K (3)	2	1K (4)	NE/RE
	RuREBUS [25]	ru	Econ.	121K (5)	1	14.6K (8)	NE/RE
	RURED [17]	ru	Econ.	22.6K (28)	1	5.3K(34)	NE/RE
	WikiOrgs [31]	ru	Wiki	7K (1)	1	7K(2)	RE
	Situations-1000 [64]	ru	News	2.2K (3)	1	(2)	RE
	RuWiki [61]	ru	Wiki	60K	1	–	EL
	RuSERRC [8]	ru	Science	1,337	1	620(6)	TE/RE/EL
	NEREL (ours)	ru	News	56K (29)	6	39K (49)	NE/RE/EL

NEREL Named entity types

- Basic entity types
  - PERSON, ORGANIZATION, LOCATION, FACILITY, GEOPOLITICAL (COUNTRY, STATE OR PROVINCE, CITY, DISTRICT).
- Temporal and numerical entities
  - NUMBER, ORDINAL, DATE, TIME, PERCENT, MONEY, AGE
- Physical object group
  - WORK OF ART, PRODUCT, and AWARD
- NORP entities
  - NATIONALITY, RELIGION, or IDEOLOGY
  - Capitalized in English but not capitalized in Russian
- Legal entities
  - LAW, CRIME,and PENALTY
- Other non-capitalized: PROFESSION, DISEASE

Named entity types statistics



Entity Linking Annotation

- 16 entity types of 29 are linked to Wikidata
  - Numerical and temporal entities are excluded
  - 38 thousand entities should be linked
- Only one entity mention per document should be annotated.
  - Barack Obama, Obama, Obama, Barack Obama
  - Mentions clusters corresponding to the same entity are created
- Mentions are identified according to
  - The same lemma representations
  - Linking via relations
    - ALTERNATIVE_NAME
    - ABBREVIATION

Automatic pre-annotation

- Factors used:
  - Manual entity links from initial Wikinews texts
  - Ranking list of Wikidata titles generated by Elasticsearch retrieval engine
  - Page view statistics to exclude noisy candidates
  - Matching NEREL named entity types to Wikidata general concept
    - CITY – city/town (Q7930989)
    - AWARD – award (Q618779),
    - Wikidata link should correspond to matched superconcept
- If Wikidata link for entity is not found -> special NULL link

Results of automatic pre-annotation

NE type	NE stats		EL stats		Automatic EL		
	#NE	#Nested	#Unique	incl. NULL	L+T+W	L	W
AWARD	767	405	600	186	0.51	0.49	0.39
CITY	2,293	21	1,450	11	0.71	0.65	0.32
COUNTRY	4,444	13	1,991	5	0.75	0.72	0.25
DISTRICT	203	24	156	10	0.66	0.58	0.31
FACILITY	742	285	556	172	0.49	0.45	0.45
LANGUAGE	85	0	70	0	0.77	0.63	0.09
LAW	713	441	584	311	0.29	0.28	0.56
LOCATION	534	121	403	71	0.45	0.40	0.32
NATIONALITY	754	56	532	6	0.32	0.27	0.03
ORGANIZATION	7,066	2,312	4,666	975	0.61	0.58	0.34
PERSON	9,687	103	4,459	908	0.57	0.54	0.43
PRODUCT	492	39	344	27	0.83	0.80	0.25
PROFESSION	8,758	2,873	5,922	1,732	0.54	0.48	0.30
RELIGION	175	3	107	4	0.53	0.48	0.13
STATE_OR_PROVINCE	750	1	473	1	0.81	0.76	0.34
WORK_OF_ART	689	135	544	143	0.55	0.51	0.42
Total	38,152	6,832	22,857	4,562	0.59	0.54	0.34

Manual linking to Wikidata

- Entity linking annotators rely fully on existing annotations of named entities.
  - Some errors can be corrected after agreement with moderator
- If an entity is absent in Wikidata, then it should be linked to NULL, but its internal entities may still have corresponding links.
  - Mayor of Novosibirsk -> NULL link
  - Mayor‘ -> to Q30185, Novosibirsk -> Q883
  - Professions are linked to corresponding professions pages
- Nested named entities allow for annotation of so called iterations of entities
  - [111th [U.S. Congress]$_{ORG}$]$_{ORG}$
  - 111th U.S. Congress -> Q170375, U.S. Congress -> Q11268

Manual linking to Wikidata: adjectives

- In NEREL adjectives are annotated with entity types of corresponding named entities
  - Moscow-> CITY, Moskovskii -> CITY,
  - This provides large coverage for establishing relations

- Adjective annotated as named entities are linked to entities according to corresponding nouns
  - Moscow (Q649) -> Moskovskii (Q649)
  - Such cases are difficult for automatic linking

- Especially difficult for automatic linking nationality-related adjectives (Russian), which can mean in different contexts
  - Nationality, Language, Country

The most frequent nested pairs and their links to Wikidata

Outer NE type	Inner NE type	#Links	# w/o NULLs	# with NULLs		
				outer	inner	both
AWARD	AWARD	186	93	62	8	23
AWARD	PERSON	130	115	15	0	0
LAW	LAW	396	103	207	6	80
LAW	COUNTRY	253	106	147	0	0
ORGANIZATION	ORGANIZATION	1,155	647	365	44	99
ORGANIZATION	COUNTRY	1,046	779	266	0	1
ORGANIZATION	CITY	404	264	139	0	1
ORGANIZATION	PERSON	174	118	55	0	1
ORGANIZATION	STATE_OR_PROVINCE	154	70	84	0	0
PROFESSION	PROFESSION	2,098	1,019	860	25	194
PROFESSION	ORGANIZATION	1,611	329	1004	16	262
PROFESSION	COUNTRY	1,015	664	351	0	0
PROFESSION	CITY	228	81	142	4	1
PROFESSION	STATE_OR_PROVINCE	185	67	116	0	2

State-the art model application
(zero-shot setting)

Entity Type	SapBERT Acc.(top-1)	SapBERT Acc.(top-5)	mGENRE Acc.(+NULLs)
AWARD	0.598	0.750	0.660
CITY	0.281	0.670	0.859
COUNTRY	0.286	0.622	0.911
DISTRICT	0.500	0.833	0.524
FACILITY	0.505	0.667	0.822
LANGUAGE	0.227	0.727	0.667
LAW	0.625	0.750	0.786
LOCATION	0.368	0.632	0.705
NATIONALITY	0.197	0.364	0.231
ORGANIZATION	0.547	0.682	0.754
PERSON	0.552	0.656	0.634
PRODUCT	0.483	0.586	0.900
PROFESSION	0.285	0.468	0.294
RELIGION	0.500	0.688	0.870
STATE_OR_PROVINCE	0.417	0.800	0.946
WORK_OF_ART	0.442	0.687	0.688
Macro-Accuracy	0.426	0.661	0.703
Micro-Accuracy	0.431	0.673	0.637

Conclusion

- We described entity linking annotation within the NEREL dataset, the largest Russian dataset for information
- extraction.
- Entity linking annotation to Wikidata items is provided for 933 documents, 16 entity types, and 38,152 entity mentions.
- The annotation contains a significant share of nested named entities (more than 17%), supporting a broader coverage of linking.
- Currently, NEREL is the only dataset for Russian annotated
- with links to Wikidata entities. It is also the only Russian
- dataset with three levels of annotation..

- The dataset is publicly available
  - https://github.com/nerel-ds/NEREL