Entity Linking over Nested Named Entities for Russian

Natalia Loukachevitch[1], Pavel Braslavski[2,3], Vladimir Ivanov[4], Tatiana Batura[5], Suresh Manandhar[6], Artem Shelmanov[7], Elena Tutubalina[7,8]

Lomonosov Moscow State University[1], HSE University[2], Ural Federal University[3], Innopolis University[4], Novosibirsk State University[5], Wiseyak (United States) [6], Sber AI[7], Kazan Federal University[8]

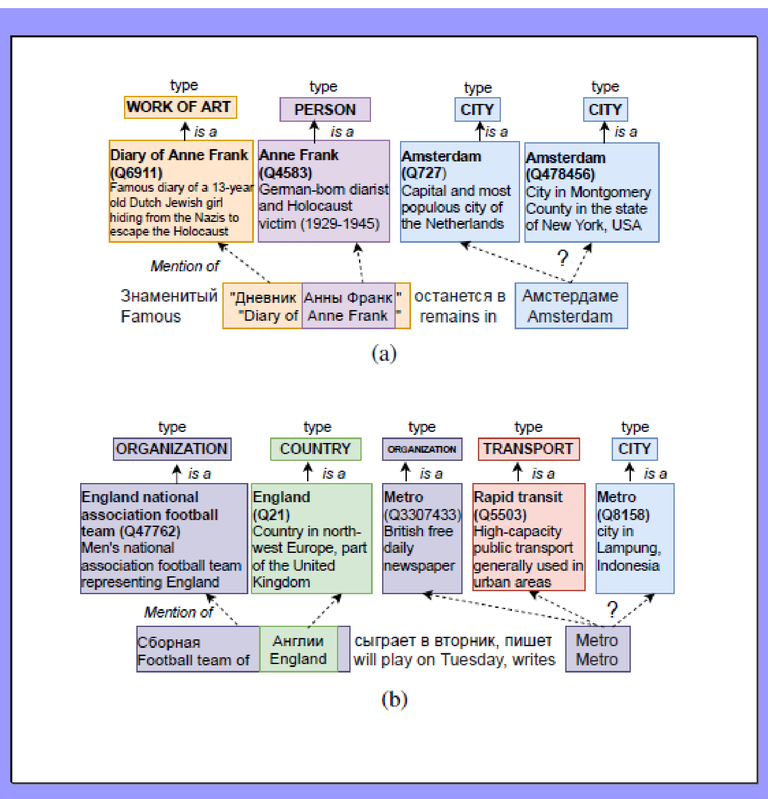
louk_nat@mail.ru, https://github.com/nerel-ds/NEREL

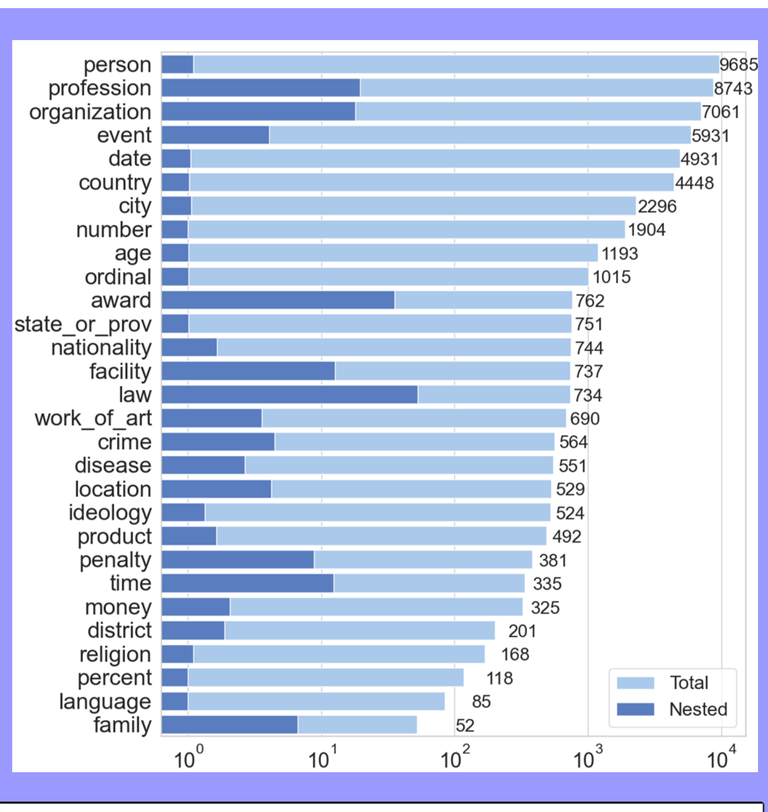
LREC-2022

Knowledge graph generation 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
  - Relations between nested named entities
  - This work is annotation of the third level: annotatng Wikidata links





NEREL dataset

- 933 Wikinews texts
- Nested named entities: longer entities can include shorter entities
  - Nestedness of entities enables a more accurate and complete description of relations and links to knowledge bases
  - 29 Entity types, more than 56K entities
  - 49 relation types
  - Wikidata entity links over nested named entities

Principles of Linking

- 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

Linking of 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

Entity Linking Annotation

- 16 entity types of 29 are linked to Wikidata
  - Numerical and temporal entities are excluded
  - 38 thousand entities are linked
- Only one entity mention per document should be annotated
  - (40% reduction)
  - Barack Obama, Obama, 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

Results of automatic pre-annotation

NE type	NE stats #NE	NE stats #Nested	EL stats #Unique	EL stats incl. NULL	Automatic EL L+T+W	Automatic EL L	Automatic EL 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

Outer NE type	Inner NE type	#Links	# w/o NULLs	# with NULLs outer	# with NULLs inner	# with NULLs 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

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

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%)
- Currently, NEREL is the only Russian dataset with three levels of annotation.