Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4458–4466
Marseille, 20-25 June 2022

Entity Linking over Nested Named Entities for Russian

Natalia Loukachevitch[1], Pavel Braslavski[2,3], Vladimir Ivanov[4], Tatiana Batura[5,6], Suresh Manandhar[9], Artem Shelmanov[7,1], Elena Tutubalina[3,8]
[1]Lomonosov Moscow State University, Moscow, Russia [2]Ural Federal University, Yekaterinburg, Russia
[3]HSE University, Moscow, Russia [4]Innopolis University, Innopolis, Russia
[5]Novosibirsk State University, Novosibirsk, Russia [6]Ershov Institute of Informatics Systems, Novosibirsk, Russia
[7]Artificial Intelligence Research Institute, Moscow, Russia [8]Sber AI, Moscow, Russia
[9]Madan Bhandari University of Science and Technology Development Board, Nepal
louk_nat@mail.ru, pbras@yandex.ru, nomemm@gmail.com, tatiana.v.batura@gmail.com,
suresh.manandhar@mbustb.edu.np, shelmanov@airi.net, tutubalinaev@gmail.com

Abstract

In this paper, we describe entity linking annotation over nested named entities in the recently released Russian NEREL dataset for information extraction. The NEREL collection (Loukachevitch et al., 2021) is currently the largest Russian dataset annotated with entities and relations. The paper describes the main design principles behind NEREL's entity linking annotation, provides its statistics, and reports evaluation results for several entity linking baselines. To date, 38,152 entity mentions in 933 documents are linked to Wikidata. The NEREL dataset is publicly available: `https://github.com/nerel-ds/NEREL`.

Keywords: information extraction, entity linking, nested named entities, Russian language

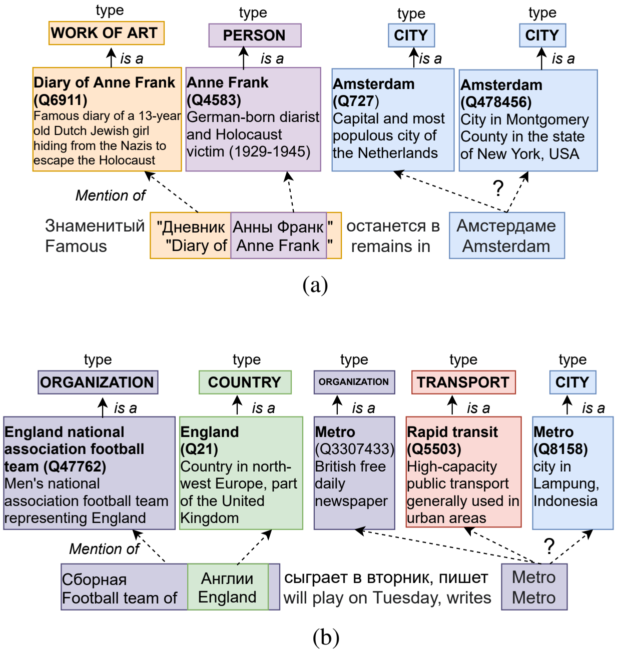
1. Introduction

Entity linking (EL) is a popular NLP task, where a system needs to link a named entity to a concept in a knowledge base such as Wikidata (Vrandečić and Krötzsch, 2014), DBpedia (Auer et al., 2007), or Freebase (Bollacker et al., 2008). Entity linking is a crucial step in automatic knowledge base construction, it helps disambiguate entity mentions in text documents and enrich them with external information (Sevgili et al., 2020; Oliveira et al., 2021).

Entity linking in the general domain is mainly implemented for named entities. Current studies in named entity extraction comprise creation of datasets with so called nested named entities, when an entity can be annotated within another named entity (Benikova et al., 2014; Ringland et al., 2019). In contrast to corpora with "flat" named entities, nested named entity datasets provide possibilities for fine-grained relations between entities and links to knowledge bases. For example, *Lomonosov Moscow State University* (Wikidata ID: Q13164) contains further entities – *Moscow* (the capital of Russia – Q649) and *Mikhail Lomonosov* (the founder of the Moscow State University – Q58720).

In the example in Figure 1a, "Famous "Diary of Anne Frank" remains in Amsterdam", the entities *Diary of Anne Frank*, *Anne Frank*, and *Amsterdam* help disambiguate each other. As Figure 1b shows, the word *Metro* in the sentence "Football team of England will play on Tuesday, writes Metro" refers to a UK newspaper and can be resolved correctly by considering internal entity *England*.

In this paper, we present entity linking annotation over nested named entities in the recently developed Russian dataset NEREL. The NEREL collection (Loukachevitch et al., 2021) is currently the largest Russian dataset annotated with entities and relations, when compared to the existing Russian datasets (Gordeev et al., 2020; Trofimov, 2014; Mozharova and Loukachevitch, 2016; Gareev et al., 2013; Starostin et al., 2016; Vlasova et al., 2014; Ivanin et al., 2020). The dataset includes 933 news texts with annotation of 29 entity and 49 relation types. We linked 16 entity types of NEREL to the Wikidata knowledge base. As a result, we obtain a dataset with three levels of annotation, the first such a dataset for Russian.


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
Famous
"Дневник
"Diary of
Анны Франк
Anne Frank
"
"
останется в
remains in
Амстердаме
Amsterdam
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
Football team of
Англии
England
сыграет в вторник, пишет
will play on Tuesday, writes
Metro
Metro

(a)

(b)

Figure 1: Examples of nested entities and their linking.

2. Related Work

Currently, there are several datasets with multiple levels of annotation that allow solving several tasks simultaneously, either independently or within a multi-task learning setup, where information obtained in one task can help solve an-

Dataset	Lang	Domain	Docs	Mentions	Overlap	Levels
AIDA-CoNLL (Hoffart et al., 2013)	En	News	1,393	34,956	No	NE/EL
DBpedia Spotlight (Mendes et al., 2011)	En	News	10	330	Overlap	EL
TAC-KBP-2010	En	News/Web	1,013	1,020	Overlap	EL
VoxEL*(Rosales-Méndez et al., 2018)	De/En/It/Sp/Fr	News	15	204	No	EL
VoxEL**(Rosales-Méndez et al., 2018)	De/En/It/Sp/Fr	News	15	674	Overlap	EL
DWIE (Zaporojets et al., 2021)	En	News	802	28,482	No	NE/RE/EL/RF
SCIERC (Luan et al., 2018)	En	Science	500	8,089	No	TE/RE/RF
RuWiki (Sysoev and Nikishina, 2018)	Ru	Wikipedia	4,024	60K	No	EL
RuSERRC (Bruches et al., 2021)	Ru	Science	1,680	1,337	No	TE/RE/EL
NEREL (ours)	Ru	News	933	38,152	Nested	NE/RE/EL

Table 1: Datasets with manual entity linking annotation. Levels of annotation include: TE – term recognition, NE – named entity annotation, RE – relation extraction, EL – entity linking, RF – coreference annotation. VoxEL* – a strict version of the VoxEL dataset without overlapping entities, VoxEL** – a relaxed version of VoxEL with overlapping entities.

other task. Table 1 summarizes the characteristics of entity linking datasets. The DWIE corpus (Zaporojets et al., 2021) consists of 802 news articles in English collected from the *Deutsche Welle* news outlet. It combines four annotation sub-tasks: named entity recognition, relation extraction, coreference resolution, and entity linking. DWIE features 311 entity and 65 relation types; named entities are linked to Wikipedia. Another example is SciERC (Luan et al., 2018) – a dataset used to create a system for identification and classification of entities, relations, and coreference clusters in scientific articles. The dataset consists of 500 scientific abstracts from the Semantic Scholar Corpus (Lo et al., 2020). The annotation scheme includes six types specific for scientific domain: Task, Method, Metric, Material, Other-ScientificTerm, and Generic.

The most widely-used dataset for evaluation of entity linking is AIDA-CoNLL (Yosef et al., 2011). It is based on the CoNLL-2003 dataset (Tjong Kim Sang and De Meulder, 2003) manually annotated with links to YAGO2 – a knowledge base automatically built from Wikipedia (Hoffart et al., 2013). Several datasets, including a tri-lingual dataset covering English, Chinese, and Spanish, were released within Entity Discovery and Linking (EDL)/Knowledge Base Population (KBP) shared tasks at the Text Analysis Conference (TAC) (Ellis et al., 2018; Ellis et al., 2017; Ellis et al., 2016). The Mewsli-9 dataset (Botha et al., 2020)is produced fully automatically and contains about 290K entity mentions from 59K Wikinews articles in nine languages linked to Wikidata items. Although the fully automatic approach to dataset creation is very attractive, as our analysis shows, the Wikinews annotations are quite sparse and biased, see Section 5.

Entity linking annotation guidelines for corpora implement different approaches regarding nested named entities. The SemEval 2015 Task 13 (Moro and Navigli, 2015) and DBpedia Spotlight (Mendes et al., 2011) datasets allow for nested entities, while ACE2004 (Ratinov et al., 2011) and AIDA/CoNLL do not support nested mentions.

There are two variants of the VoxEL dataset (Rosales-Méndez et al., 2018): a strict version and a relaxed one. The strict version contains non-overlapping maximal entity mentions (persons, organizations, or locations). The relaxed version considers any noun phrase matching a Wikipedia entity as a mention, including overlapping mentions where applicable. For example, in the sentence "The European Central Bank released new inflation figures today" the strict version would only include *European Central Bank*, while the relaxed version would also include *Central Bank* and *inflation*.

There is an entity linking dataset automatically constructed from Russian Wikipedia (Sysoev and Nikishina, 2018): the collection contains 2,968 Wikipedia articles as a train set (16.3 entity mentions per document) and 1,056 articles as a test set (21.3 entity mentions per document). Another Russian dataset for entity linking is RuSERRC (Bruches et al., 2021), which contains abstracts of 1,680 scientific papers on information technology. In total, 3,386 terms are labeled, 1,337 of which are linked to Wikidata entities. However, the first dataset is not freely available, and the second one is rather small. Additionally, neither of these datasets are annotated with nested named entities. Table 1 provides a comparison of the above-mentioned datasets. Recently, (Nesterov et al., 2022) have released the Russian-language RuCCon dataset related to the clinical domain with annotations of diseases and symptoms linked to the concepts of the UMLS metathesaurus.

3. Nested Named Entity Annotation

To date, the NEREL corpus consists of 933 news articles originating mainly from Russian Wikinews, which discuss persons, their relations, and events (Loukachevitch et al., 2021). Figure 2 shows an example of nested named entities and relations in NEREL.

When designing annotation guidelines for the dataset, we attempted to provide detailed labeling of named entities and relations, at the same time avoiding a long tail of low-frequency entities and relations. We followed three principles.

First, we annotated nested named entities that are shorter named entities within longer ones. A nested named entity can have its own relations and be linked to Wikidata. For example, the phrase *Mayor of Moscow*, being an entity itself, consists of two entities: a title (*Mayor*) and a city (*Moscow*). All three can be linked to Wikidata. The nested entity *Moscow* might be required for establishing relations

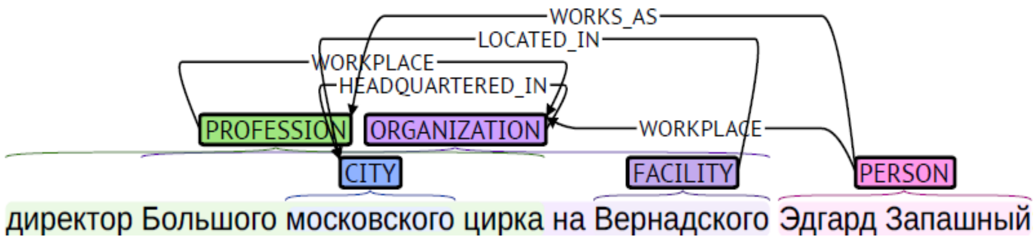


Figure 2: An example of NEREL annotation: nested named entities, relations and links to Wikidata. In the fragment "Director of Moscow State Circus at Vernadskogo avenue Edgard Zapashnyy", the longest entity *Director of Moscow State Circus at Vernadskogo avenue* includes internal entity *Moscow State Circus at Vernadskogo avenue*, which in turn contains entities *Moscow* and *Vernadskogo avenue*. After entity linking annotation, the largest entity will be assigned to NULL link because it is absent in Wikidata. The entity *Moscow State Circus at Vernadskogo avenue* is linked to the Wikidata entity Q154409. This, in turn, includes two additional entities: *Moscow* (Q649) and *Vernadskogo avenue* (Q4381087).

with other entities in a text. Therefore, we decided to annotate nested named entities in the NEREL dataset.

Second, in addition to annotating traditional named entities that are mainly expressed by nouns and noun phrases, we also annotated adjectives such as *British* or *Russian* derived from named entities. An adjective can be essential for establishing a relation in a sentence or in the whole text due to absence or long distance to the corresponding noun (Benikova et al., 2014). Moreover, an adjective often conveys a reference to the same entity in a knowledge base as the noun, from which the adjective is derived.

Third, we annotated entities beyond the traditional named entity typeset in the general domain: crimes, penalties, diseases, professions, and professional titles are annotated independently of their capitalization. This was done due to the significance of corresponding relations in person-related texts. Such entities also provide additional links to the knowledge graph.

Currently, there are 29 entity types in the NEREL dataset. The entity types can be grouped in the following way:

- basic entity types (PERSON, ORGANIZATION, LOCATION, FACILITY, geopolitical entities subdivided into several subgroups);
- numerical entities (NUMBER, ORDINAL, DATE, TIME, PERCENT, MONEY, AGE);
- NORP entities (NATIONALITY, RELIGION, IDEOLOGY) and LANGUAGE;
- law-related entities (LAW, CRIME, PENALTY);
- work-related entities (PROFESSION, WORK_OF_ART, PRODUCT, AWARD) and DISEASE;
- EVENT.

In the first phase of the entity linking annotation, we focused on 16 entity types. We excluded seven numerical types, as well as EVENT, IDEOLOGY, and DISEASE entities; statistics of the remained subset in the NEREL dataset is presented in Table 2. As can be seen from the Table, all but one (LANGUAGE) entity type have at least 100 annotated examples.

4. Entity Linking Annotation Guidelines

Since we establish links to Wikidata for both named entities and instances of general concepts such as CRIME, PENALTY, or LANGUAGE, our entity linking can be viewed as a subtype of general named entity linking (Ling et al., 2015). Entity linking annotators rely fully on existing annotations of named entities. If an annotator encounters an alleged problem with named entity annotations, they report the problem to a moderator and get permission to fix it.

If an entity is absent in Wikidata, then it should be linked to NULL, but its internal entities may still have corresponding links. For example, the entity *Mayor of Novosibirsk* is absent in Wikidata, but *Mayor*' and *Novosibirsk* entities have links to Q30185 and Q883 Wikidata items, respectively.

Nested named entities expose the annotation process to the so-called "iteration problem" – annotating different iterations of the same organization, such as *111th U.S. Congress* and the *112th U.S. Congress*. There exist several approaches to the annotation of such entities. The AIDA guidelines prefer to annotate more specific entities. In contrast, the TACKBP annotation guidelines (Ellis, 2012) and later projects (Hamdi et al., 2021) specify that different iterations of the same entity should not be considered as distinct entities. Both approaches can be problematic. In the former case, a specific iteration of an organization can be missing in the KB; in the latter case, it can be inferred that all congressmen work in the same organization, which distorts extracted relations. We annotate the iteration in the following way: [111th [U.S. Congress]$_{\text{ORG}}$]$_{\text{ORG}}$ using both links Q170375 [111th U.S. Congress] and Q11268 [U.S. Congress].

NER and EL annotation also suffers from *metonymy*, which is a reference to an entity using a semantically related word. In NEREL, in case of official residences (e.g., *the White House*, *the Kremlin*, *Downing Street*) we distinguish between facility vs. administration contexts. For example, *the White house* as a residence is annotated as a FACILITY and is linked to the Wikidata *the White House* item (Q35525). In organizational contexts, *the White House* is annotated as ORGANIZATION and linked to the Wikidata item *Executive Office of the President of the United States* (Q1355327).

Adjectives derived from proper names and annotated as named entities, are linked to the Wikidata items of the corresponding named entities. For example, the adjective

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

Table 2: Statistics of the named entity (NE stats) and entity linking (EL stats) annotation steps in 933 NEREL documents. Note that EL statistics correspond to 'cleaned' documents, where only one mention per entity is retained. The last three columns report the accuracy of the automatic linkage suggestions against manual annotation: 1) combination of the original Wikinews annotation and the linker's output with entity type verification (L+T+W, this variant was presented to the annotators), 2) top-1 linker candidate (L), and 3) original Wikinews annotation (W). In all automatic annotation variants, an empty suggestion was treated as NULL.

Moskovskii derived from *Moscow* is linked to the same item as the initial name: *Moscow* (Q649). Linking of adjectives increases coverage of the annotation. Adjectives derived from nations and nationalities are especially difficult for manual annotation and automatic linking because of their ambiguity. For example, the adjective *russkii* (*Russian*) in different contexts can mean the *Russian Federation* (Q159, NER type COUNTRY), Russian citizens (Q49542, NER type NATIONALITY) or Russian language (Q7737, NER type LANGUAGE).

We linked PROFESSION entities to profession items in Wikidata, not to a person who is currently holding this post. For example, *Mayor of Moscow* mention is linked to the *Mayor of Moscow* Wikidata item Q1837906, not to *Sergey Sobyanin* (Q319497), the current Moscow mayor. This is because specific persons can change their posts in the appointment and resignation contexts.

5. Annotation Process Details

We use Wikidata as our target knowledge base (KB). Wikidata is a large open multilingual KB that is being actively developed by the community. To date, there are about 7.5M Wikidata items with Russian labels.[1]

5.1. Preliminary Automatic Annotation

We used the BRAT annotation tool (Stenetorp et al., 2012), in particular – its normalization feature for entity linking. As a preparation step, we removed markup of the entities that are not intended for linking to Wikidata, as well as relations, from the BRAT standoff annotation files. In addition, we removed all but one mention per entity in the document based on *alternative_name* and *abbreviation* relations annotated in NEREL (Loukachevitch et al., 2021). Removing multiple mentions of the same entity within a document decreased the number of named entities to be linked to Wikidata by 40% (from 38,152 to 22,857 in 933 documents).

We applied an entity linker that was originally developed for the annotation of a question answering dataset.[2] The linker builds a search index over a collection of Russian labels and aliases that correspond to around 4M Wikidata entities using Elasticsearch.[3] The linker converts an input string into a series of phrase and fuzzy search queries, aggregates the search results, and returns a ranked list of candidate entities. The final ranking is performed based on a combination of Elasticsearch matching scores and page view statistics of the corresponding Wikipedia articles. Adding the latter parameter turned out to be very efficient to downrank noisy candidates. The linker implementation details can be found in (Korablinov and Braslavski, 2020).

In the original Wikinews articles, some entity mentions are linked to corresponding Wikipedia pages.[4] For about 15% (3,454) of entities to be linked, we could provide

[1] `https://www.wikidata.org/wiki/User:Mr._Ibrahem/Language_statistics_for_items`

[2] `https://github.com/vladislavneon/kbqa-tools/`

[3] `https://www.elastic.co/`

[4] Among them, there are many links to year or month entities, e.g. *October 2009 → Q243251* that are not very helpful in the context of our task.

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

Table 3: The most frequent nested pairs and their links to Wikidata

Wikidata IDs inferred from the original Wikipedia links. For the rest of the mentions, we took up to three Wikidata entity candidates with non-zero Wikipedia page views returned by the linker. We also associated each entity type with a generic Wikidata concept, e.g. CITY – *city/town* (Q7930989), AWARD – *award* (Q618779), etc. We kept only candidates that are connected with the corresponding superconcepts by a path of the *instance of* (P31) or *subclass of* (P279) properties. The best candidate, if any, was provided as a suggestion for subsequent manual annotation, while the remaining candidates formed a 'local' BRAT knowledge base.

Annotators were presented with documents with highlighted entities. The majority of entities are provided with candidate Wikidata linkages along with their IDs, labels, and descriptions. Annotators were also able to follow a hyperlink to a Wikidata entity page. To correct an existing linkage or produce a new one, annotators could search the local collection of Wikidata entities using a built-in BRAT search interface based on substring matching. Alternatively, they were instructed to use the Wikidata search box or search Wikipedia through a major search engine like Google or Yandex. In the latter case, a Wikidata ID can be easily obtained by following the *Wikidata item* link from the navigation panel of a Wikipedia page. This way appeared to be the easiest and most natural for annotators. If no corresponding Wikidata item was found, the annotators provided the entity mention with a special NULL value. On average, annotators spent 1 hour processing 100 entity mentions.

5.2. Annotation Results and Evaluation of Automatic Pre-Annotation

Table 2 provides a breakdown of entity linking annotations corresponding to almost 23K entities in 933 documents (see section *EL stats*). As can be seen from the Table, the share of NULL labels (i.e. mentions with missing Wikidata entries) is quite uneven across entity types. Thus, more than a half of LAW entities do not have corresponding entries in Wikidata. For AWARD, FACILITY, PROFESSION and WORK_OF_ART types the share of missing Wikipedia items is around 30%. ORGANIZATION, PERSON, LOCATION form the next group with the share of missing knowledge base entries about 20%. "Conventional" entity types such as geopolitical entities (GPE) and nationalities/religious/political groups (NORP) are well presented in the knowledge base.

We also assessed the quality and coverage of automatic linking suggestions against manual annotation, see *Automatic EL* section in Table 2. As can be seen from the Table, automatic suggestions (**L+T+W** variant) greatly facilitate the annotation with an average accuracy of 59%. The coverage of original Wikinews links is rather low: it delivers about one third of correct annotations, but the majority of them are NULL labels. Accounting for entity types improves the accuracy of automatic suggestions by around 5% (cf. **L+T+W** and **L** columns).[5] The lowest quality of automatic suggestions is observed for LAW entities that are scarcely presented in Wikidata: automatic linker produces false candidates based on textual similarity in the majority of cases. This is the only case, when original Wikinews markup (with missing links treated as NULL) outperforms automatic suggestions. A low quality of NATIONALITY link suggestions can be explained by the annotation scheme. For example, the adjective *British* in the case of *British actor* is to be linked to *Britons* (Q842438) according to the annotation guidelines, which is a hard task for a surface matching linker. The same holds for CITY and COUNTRY, where adjectives are often linked to items with nouns labels. In contrast to, e.g. English, Russian cognate words can be quite distant on the character level, e.g. *peterburzhskiy – Saint Petersburg*, *rossiyskiy – Russia*.

Note that entity linking statistics in the table correspond to the 'cleaned' annotations, where only one mention per en-

[5] However, in some cases, it can discard correct linkages. For example, a mention of *Heracles* was correctly linked to *Q122248* but was rejected by the type check since the entity belongs to *demigod (Q466470)*, not *human (Q5)* that was set as a parent class for PERSON.

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

Table 4: Baseline results of entity linking derived from SapBERT and mGENRE models

tity/document is retained (cf. #NE and #Unique columns). After the Wikidata linking was finished, we restored initial NE/relation annotations and propagated Wikidata linkages to other mentions of the same entities in the corresponding BRAT standoff files.

Finally, we provide the distribution of NE types in linked nested named entities. Overall, the dataset contains 10,710 pairs of nested named entities, in 5,394 pairs, both entities (outer and inner) are linked to Wikidata. The majority of the remaining pairs (4,454) have NULL-links for the outer entity only; 707 pairs have both entities with NULL-links.

In Table 3, we provide statistics on linkages to Wikidata for the most frequent pairs of nested entities. It can be seen that geopolitical entities are mainly presented in Wikidata; they are nested inside entities of various types: AWARD, LAW, ORGANIZATIONS, PROFESSION, etc. Internal entities of the PERSON type are mainly well-known; they have corresponding Wikidata items.

In addition, we analyzed connections between nested entities and found that only 61% of nested named entity pairs are actually connected in Wikidata in a single hop (the remaining 39% of pairs have no connection). The most frequent types of such connections are: COUNTRY (P17), APPLIES TO JURISDICTION (P1001), SUBCLASS OF (P279), PART OF (P361), NAMED AFTER (P138), LOCATED IN ... (P131), INSTANCE OF (P31), HEADQUARTERS LOCATION (P159), ORGANIZATION DIRECTED BY THE OFFICE OR POSITION (P2389) and FOUNDED BY (P112). Thus, these nested relations describe different aspects of relation between an outer entity and Locations/Persons/Organizations that are mentioned as an inner entity.

We checked that in many cases the absence of a relation between a longer and internal entities in Wikidata can be due too short descriptions of Wikidata items. For example, *Mariinsky Theatre Concert hall* item (Q4231897) is not linked to the *Mariinsky Theatre* item (Q207028). Also, we noted that two nested entities from NEREL dataset can be connected by up to six different types of properties in Wikidata (e.g., the Q42274:*Google Earth* and the Q95:*Google*).

6. Entity Linking Baselines

As baselines for entity linking, we evaluated two models: (i) a sequence-to-sequence system for the Multilingual Entity Linking (mGENRE) (De Cao et al., 2021) and (ii) a model based on the SapBERT (Liu et al., 2021) that implements self-alignment of representation space during pretraining. The models appear to be state-of-the-art, and can be applied to texts in Russian as they have capabilities to process texts in several languages. Note, that both models were used in the zero-shot evaluation mode without finetuning them on the NEREL data. However, we use the NEREL train set to optimise a threshold for NULL linkage prediction.

The mGENRE model uses Wikidata as the KB while exploiting Wikipedia hyperlinks as the source of supervision during training. Given a mention, mGENRE predicts the name of the entity in an autoregressive fashion (token-by-token). To evaluate the model, we transformed the mentions of entities from the NEREL dataset into the mGENRE input format (with “[START]” and “[END]” tags around each mention), keeping a few tokens from the context of each mention. The mGENRE model was used for inference without any fine-tuning on the NEREL data, which can explain quite moderate performance of the baseline (Table 4).

As mentioned above, in the NEREL dataset, links between entities and Wikidata can have a NULL tag, meaning that there is no corresponding entity in the Wikidata. However, mGENRE outputs a list of candidate entities along with a certainty score for each mention. To deal with NULL linkages, we applied the following approach: fitting a certainty-based threshold for decision about the NULL-linkage. Therefore, the mGENRE either will return a Wikidata entity, or the NULL linkage. The threshold is fitted on the train part of the NEREL dataset and applied to the test subset. The train, dev, and test sets contain 746, 94, and 93 documents, respectively. The last column of the Table 4 shows results of the mGENRE model that correspond to the optimal threshold value (0.51).

Table 4 also contains the results of another baseline based on the SapBERT model (Liu et al., 2021). This model was originally proposed for biomedical entity linking exploiting concepts from the Unified Medical Language System (UMLS) as the source of supervision. Therefore, we trained the model on 'wikititles+muse' pairs of labels to enrich with general-domain knowledge. The training was done according to the SapBERT authors' recommendations[6]. Note, that SapBERT returns vector representations for entity mentions which can be compared to vectors of target Wikidata entities using a similarity. Here, again a threshold for the similarity value can be applied to predict NULL linkages. This threshold was fitted on the NEREL train set to keep the same proportion of NULLs. One can see significant difference between top-1 and top-5 accuracy of SapBERT, which means that top-5 candidates contain correct links, but the ranking is not perfect (probably, due to the lack of context). Using the same motivation, we explored tow options for the target entity sets: (i) with the whole list of Wikidata entities and their labels (more than 6 million QIDs) and (ii) with a cleaned[7] version, which has around 2.9 million QIDs. The former (bigger version) led to much worse results (0.3 accuracy), presumably, due to noisy entity labels. Therefore, in the Table 4 for SapBERT we provide results of linking to the latter (shorter) list of Wikidata entities.

7. Conclusion

In this paper, 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. NEREL can be used for developing and testing entity linking models applied to nested named entities and also for creating end-to-end models accounting for nested entities.

8. Acknowledgements

The project is supported by the Russian Science Foundation, grant # 20-11-20166.

[6] The list can be found here: https://github.com/cambridgeltl/sapbert/tree/main/training_data

[7] Here, by 'cleaning' we mean keeping only Russian labels and dropping from a label any text inside parentheses, which can provide a noisy signal to the entity linking model (e.g. the label 'Andacollo_(Chile)' was substituted with the label 'Andacollo'); labels without Russian characters were dropped.

9. Bibliographical References

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Benikova, D., Biemann, C., and Reznicek, M. (2014). NoSta-D Named Entity Annotation for German: Guidelines and Dataset. In *LREC*, pages 2524–2531.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., and Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Botha, J. A., Shan, Z., and Gillick, D. (2020). Entity Linking in 100 Languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7833–7845.

Bruches, E., Mezentseva, A., and Batura, T. (2021). A system for information extraction from scientific texts in russian. *arXiv preprint arXiv:2109.06703*.

De Cao, N., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., and Petroni, F. (2021). Multilingual autoregressive entity linking. In *arXiv pre-print 2103.12528*.

Ellis, J. (2012). *TAC KBP Entity Selection. KBP 2012 Guidelines. Version 1.1.* Linguistic Data Consortium.

Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., and Ivanov, V. (2013). Introducing baselines for russian named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 329–342.

Gordeev, D., Davletov, A., Rey, A., Akzhigitova, G., and Geymbukh, G. (2020). Relation extraction dataset for the russian language. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*.

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., and Doucet, A. (2021). A multilingual dataset for named entity recognition, entity linking and stance detection in historical newspapers. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2328–2334.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Ivanin, V., Artemova, E., Batura, T., Ivanov, V., Sarkisyan, V., Tutubalina, E., and Smurov, I. (2020). RuREBus-2020 Shared Task: Russian Relation Extraction for Business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

Korablinov, V. and Braslavski, P. (2020). RuBQ: A Russian dataset for question answering over Wikidata. In *ISWC*, pages 97–110.

Ling, X., Singh, S., and Weld, D. S. (2015). Design challenges for entity linking. *Transactions of the Association for Computational Linguistics*, 3:315–328.

Liu, F., Vulić, I., Korhonen, A., and Collier, N. (2021). Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021*, pages 565–574, August.

Lo, K., Wang, L. L., Neumann, M., Kinney, R., and Weld, D. (2020). S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983. Association for Computational Linguistics.

Loukachevitch, N., Artemova, E., Batura, T., Braslavski, P., Denisov, I., Ivanov, V., Manandhar, S., Pugachev, A., and Tutubalina, E. (2021). NEREL: A Russian dataset with nested named entities, relations and events. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 876–885.

Luan, Y., He, L., Ostendorf, M., and Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232.

Mendes, P. N., Jakob, M., García-Silva, A., and Bizer, C. (2011). Dbpedia spotlight: shedding light on the web of documents. In *Proceedings of the 7th international conference on semantic systems*, pages 1–8.

Moro, A. and Navigli, R. (2015). Semeval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 288–297.

Mozharova, V. and Loukachevitch, N. (2016). Two-stage approach in russian named entity recognition. In *International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT)*, pages 1–6.

Nesterov, A., Zubkova, G., Miftahutdinov, Z., Kokh, V., Tutubalina, E., Shelmanov, A., Alekseev, A. M., Avetisian, M., Chertok, A., and Nikolenko, S. (2022). RuCCoN: Clinical concept normalization in Russian. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Oliveira, I. L., Fileto, R., Speck, R., Garcia, L. P., Moussallem, D., and Lehmann, J. (2021). Towards holistic entity linking: Survey and directions. *Information Systems*, 95:101624.

Ratinov, L., Roth, D., Downey, D., and Anderson, M. (2011). Local and global algorithms for disambiguation to wikipedia. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1375–1384.

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., and Curran, J. R. (2019). Nne: A dataset for nested named entity recognition in english newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181.

Rosales-Méndez, H., Hogan, A., and Poblete, B. (2018). Voxel: a benchmark dataset for multilingual entity linking. In *International Semantic Web Conference*, pages 170–186. Springer.

Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., and Biemann, C. (2020). Neural entity linking: A survey of models based on deep learning. *arXiv preprint arXiv:2006.00575*.

Starostin, A., Bocharov, V., Alexeeva, S., Bodrova, A., Chuchunkov, A., Dzhumaev, S., Efimenko, I., Granovsky, D., Khoroshevsky, V., Krylova, I., Nikolaeva, M., Smurov, I., and Toldova, S. (2016). FactRuEval 2016: Evaluation of named entity recognition and fact extraction systems for Russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, pages 702–720.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., and Tsujii, J. (2012). BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *EACL (demonstrations)*, pages 102–107.

Sysoev, A. and Nikishina, I. (2018). Smart context generation for disambiguation to Wikipedia. In *Conference on Artificial Intelligence and Natural Language*, pages 11–22.

Tjong Kim Sang, E. F. and De Meulder, F. (2003). Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, page 142–147, USA. Association for Computational Linguistics.

Trofimov, I. (2014). Identification of personal names in news texts on collections persons-1000/1111-f (in russian). *Proceedings of RCDL-2014*, pages 217–221.

Vlasova, N., Suleymanova, E., and Trofimov, I. (2014). Report on Russian corpus for personal name retrieval. In *Proceedings of TEL'2014 Conference on Computational and Cognitive Linguistics*, pages 36–40.

Vrandečić, D. and Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85.

Yosef, M. A., Hoffart, J., Bordino, I., Spaniol, M., and Weikum, G. (2011). Aida: An online tool for accurate disambiguation of named entities in text and tables. *Proceedings of the VLDB Endowment*, 4(12):1450–1453.

Zaporojets, K., Deleu, J., Develder, C., and Demeester, T. (2021). DWIE: An entity-centric dataset for multi-task document-level information extraction. *Information Processing & Management*, 58(4):102563.

10. Language Resource References

Ellis, J. and Getman, J. and Strassel, S. (2016). *TAC KBP Spanish Cross-lingual Entity Linking - Comprehensive Training and Evaluation Data 2012-2014.* TAC Project, Linguistic Data Consortium, LDC2016T26, 1.0, ISLRN 546-386-811-027-3.

Ellis, J. and Getman, J. and Strassel, S. (2017). *TAC KBP Chinese Cross-lingual Entity Linking - Comprehensive*

Training and Evaluation Data 2011-2014. TAC Project, Linguistic Data Consortium, LDC2017T17, 1.0, ISLRN 464-261-620-634-2.

Ellis, J. and Getman, J. and Strassel, S. (2018). *TAC KBP English Entity Linking - Comprehensive Training and Evaluation Data 2009-2013*. TAC Project, Linguistic Data Consortium, LDC2018T16, 1.0, ISLRN 287-583-243-614-4.