Language Resources and Evaluation
https://doi.org/10.1007/s10579-023-09674-z

ORIGINAL PAPER

NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links

Natalia Loukachevitch[1,2] · Ekaterina Artemova[3] · Tatiana Batura[4,5] · Pavel Braslavski[3,6] · Vladimir Ivanov[7] · Suresh Manandhar[8] · Alexander Pugachev[3] · Igor Rozhkov[1] · Artem Shelmanov[1,9,10] · Elena Tutubalina[3,9,11] · Alexey Yandutov[1]

Accepted: 25 May 2023

Abstract
This paper describes NEREL—a Russian news dataset suited for three tasks: nested named entity recognition, relation extraction, and entity linking. Compared to flat entities, nested named entities provide a richer and more complete annotation while also increasing the coverage of relations annotation and entity linking. Relations between nested named entities may cross entity boundaries to connect to shorter entities nested within longer ones, which makes it harder to detect such relations. NEREL is currently the largest Russian dataset annotated with entities and relations: it comprises 29 named entity types and 49 relation types. At the time of writing, the dataset contains 56 K named entities and 39 K relations annotated in 933 person-oriented news articles. NEREL is annotated with relations at three levels: (1) within nested named entities, (2) within sentences, and (3) with relations crossing sentence boundaries. We provide benchmark evaluation of current state-of-the-art methods in all three tasks. The dataset is freely available at https://github.com/nerel-ds/NEREL.

Keywords Named entity recognition · Nested entities · Relation extraction · Nested relations · Entity linking

Mathematics Subject Classification 68T35 · 68T50

1 Introduction

Information extraction (IE) tasks are core to many real-life applications. IE systems rely on the ability of the Natural Language Processing (NLP) models to extract named entities, define relations between entities and link the entities to a structured knowledge base. IE methods are in demand in healthcare, insurance, financial, legal,

Extended author information available on the last page of the article

Published online: 21 September 2023

and many other domains. Although each domain requires domain-specific datasets, general domain datasets provide a valuable resource to evaluate and compare methods.

Research in NLP and IE, in particular, has historically been English-centric. The vast majority of published datasets are collected from English sources. Downstream models tend to be tailored to language phenomena present in English, and annotations adjusted to cultural specifics and biases of the English-speaking community (Shavrina et al., 2020). In this paper, we seek to emphasize the need to annotate novel non-English datasets for IE tasks and to widen the scope of the existing annotated datasets. We present NEREL—a new dataset in Russian collected from a news portal and suited for three tasks: nested named entity recognition (NER), relation extraction (RE), and entity linking (EL).

The Russian language belongs to the group of Slavic languages of the Indo-European language family. Russian is in the top 10 languages by the number of L1 and L2 speakers (estimated by 260 million by Ethnologue).[1] The Russian language has several linguistic features that can affect the performance of NLP models, such as the Cyrillic alphabet, rich inflectional morphology, and quasi-free word order.

This paper describes the main design decisions made when preparing NEREL's annotations: document selection, named entity and relation type selection and definition, and ambiguity resolution for entity linking. NEREL is built upon person-oriented news published in Wikinews.[2] Given the person-oriented nature of the NEREL corpus, named entity and relation types correspond to personal events, ranging from birth, death, and marriage to career changes and participation in various activities. Since person-related texts cover a wide gamut of topics, our annotation scheme comprises a diverse set of entity types and relations. Entities are linked to the Wikidata knowledge base.

Nested named entities, as well as *sentence-* and *document-level* relations are annotated. Nested named entities provide richer and more detailed annotation compared to flat entities (Benikova et al., 2014; Ringland et al., 2019). Nested entities increase the coverage of relations and entity linking. Relations between nested named entities turn out to be particularly challenging to detect when connecting shorter entities nested within longer ones.

Figure 1 shows an example, in which longer entity *Lower House of the Czech Parliament* (Wikidata ID Q320265) includes two other entities *Czech Parliament* (Q2347172) and adjective *Czech* as reference to the *Czech Republic* entity (Q213). Abbreviation *ODS* refers to a Czech political party. This allows for establishing a relation between *ODS* and the embedded *Czech Republic* entity. Also, the access to *Czech Republic* mention would help disambiguate the *ODS* abbreviation.

To the best of our knowledge, NEREL is the first dataset annotated simultaneously with nested entities, relations between those entities and knowledge base links. Such annotations pose challenges to current state-of-the-art relation extraction

[1] https://www.ethnologue.com/.

[2] https://ru.wikinews.org/.

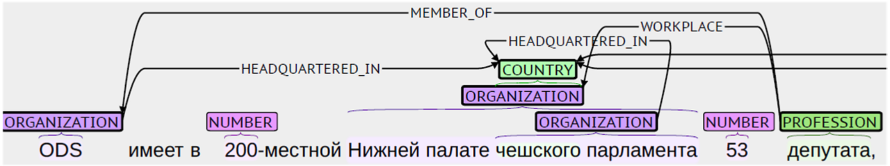

Fig. 1 Annotation of the sentence "ODS has 53 deputies in the 200-seat Lower House of the Czech Parliament" showing nesting as "[Lower House of the [Czech] Parliament]" with each entity having its own Wikidata entry. ODS is a Czech political party. The *headquartered_in* relation connects "ODS " with "Czech"

models that lack support for relations between nested and overlapping entities (Alt et al., 2019; Joshi et al., 2020). NEREL relations may lie inside a sentence or span across sentence boundaries. Entity linking annotations leverage nested named entities, and each nested named entity can be linked to a separate Wikidata entity. NEREL provides a novel testbed for sentence- and document-level relation extraction methods.

At the time of writing, NEREL is the largest Russian dataset annotated with entities and relations compared to the existing Russian datasets (Gareev et al., 2013; Gordeev et al., 2020; Ivanin et al., 2020; Mozharova & Loukachevitch, 2016; Starostin et al., 2016; Trofimov, 2014; Vlasova et al., 2014). NEREL features the largest number of entity and relation types—currently 29 and 49, respectively. Currently, the dataset contains 56K named entities and 39 K relations annotated in 933 person-oriented news articles.

The main contributions of this paper are as follows:

1. We present a novel Russian dataset NEREL annotated with nested named entities, relations between nested entities, and Wikidata entity links.
2. We describe our annotation scheme that can be easily adapted to other languages and genres.
3. NEREL is currently the largest dataset for three IE tasks and the first open entity-linking dataset in Russian;
4. We evaluate current state-of-the-art models on all three tasks.

This paper extends the previous description of NEREL reported in Loukachevitch et al. (2021). In particular:

- we add entity linking annotation over nested named entities;
- we provide benchmark results for entity linking and end-to-end information extraction tasks;
- we provide additional experimental results for named entity recognition and relation extraction.

The remainder of the paper is structured as follows. Section 2 surveys related datasets and their annotation schemes; special attention is paid to Russian datasets. Section 4 overviews the main principles of the dataset annotation. Section 3 describes data collection and annotation in detail. Section 5 contains an experimental evaluation of the dataset and reports several baseline results. Section 6 concludes.

2 Related work

Table 1 summarizes various information extraction datasets that we used as references while creating NEREL. We paid attention to the following choices to design annotation schema: types of named entities, nested NEs, relation types, external KBs for entity linking, and non-English datasets with particular attention to Russian collections. We do not provide an exhaustive overview of datasets (e.g. for NER see a recent survey (Nasar et al., 2021); for entity linking see Sevgili et al. (2020)).

2.1 Datasets for nested NER

CoNLL03 (Tjong & De Meulder, 2003) is a well-known dataset for named entity recognition in the general domain. It provides annotations for four entity types: persons (PER), organizations (ORG), locations (LOC), and other named entities (MISC). OntoNotes (Hovy et al., 2006) provides annotation for 19 named entity types, including numeric (NUMBER, ORDINAL, PERCENT, CARDINAL) and temporal (DATE, TIME) ones. Ontonotes is a genre-diverse dataset: it includes news and magazine articles, broadcast transcripts, web documents, and telephone conversations. However, both datasets only annotate flat named entities.

There are several datasets with annotated nested named entities: ACE-2005 (Walker et al., 2006) and NNE (Ringland et al., 2019) both in English, No-Sta-D (Benikova et al., 2014) in German, Digitoday (Ruokolainen et al., 2019) in Finnish, and FactRuEval2016 (Starostin et al., 2016) in Russian, see Table 1.

The No-Sta-D collection contains Wikipedia articles and online newspapers annotated with four key entity types for German. The dataset introduces unique annotation tags for adjectives (for example, *osterreichischen* is annotated as Location_deriv), considered important for establishing relations. The Digitoday corpus for Finnish is annotated with six named entity types. Nested named entities are annotated with the following restriction: an internal entity cannot be of the same class as its top-level entity (as *Microsoft* entity mentioned within *Microsoft Research* entity)—nestedness, in this case, is considered as redundant information. Systematic nested annotations in Digitoday include a location inside an organization, an organization inside a product, and a person inside an organization. Both NoSta-D and Digitoday datasets include at most two levels of nestedness within entities.

The most extensive corpus annotated with nested named entities is the NNE corpus (Ringland, 2015; Ringland et al., 2019). 114 entity types are annotated. The NNE dataset provides detailed lexical annotation such as first and last person's names, units (*tons*), multipliers (*billion*), and others. There are six levels of nestedness in the dataset.

Table 1 NEREL and its counterparts

	Dataset	Lang	Domain	#NE inst (types)	Max depth	#Rel inst (types)	Annot levels
1	CoNLL03/AIDA Tjong and DeMeulder (2003) and Hoffart et al. (2011)	en	News	34.5 K (4)	1	–	NE/EL
	Ontonotes Hovy et al. (2006)	en	News/Web	104 K (19)	1	–	NE
2	ACE2005 Walker et al. (2006) and Bentivogli et al. (2010)	en	News	30 K (7)	6	8.3 K (6)	NE/RE/EL
	NNE Ringland et al. (2019)	en	News	279 K (114)	6	–	NE
	No-Sta-D Benikova et al. (2014)	de	News	41 K (12)	2	–	NE
	Digitoday Ruokolainen et al. (2019)	fi	News	19 K (6)	2	–	NE
	DAN+ Plank et al. (2020)	da	News	6.4 K (4)	2	–	NE
3	SemEval-2010 Hendrickx et al. (2010)	en	Web	–	1	8.8 K (9)	RE
	TACRED Zhang et al. (2017)	en	News	(3)	1	22.8 K (42)	RE
	DocRED Yao et al. (2019)	en	Wiki	132 K (6)	1	56 K (96)	RE
4	DBp. Spotlight Mendes et al. (2011)	en	News	330	2	–	EL
	AQUAINT Milne and Witten (2008)	en	News	727	1	–	EL
	TAC-KBP-2010	en	News/Web	1020	2	–	EL
	VoxEL(strict) Rosales-Méndez et al. (2018)	5 lang	News	204	1	–	EL
	VoxEL (relaxed) 0 Rosales-Méndez et al. (2018)	5 lang	News	674	2	–	EL
	DWIE Zaporojets et al. (2021)	en	News	43,373 (311)	1	16 K (65)	NE/RE/EL/RF
	SCIERC Luan et al. (2018)	en	Science	8089 (6)	1	4 K (9)	TE/RE/RF

Table 1 (continued)

	Dataset	Lang	Domain	#NE inst (types)	Max depth	#Rel inst (types)	Annot levels
5	Gareev Gareev et al. (2013)	ru	News	44 K (2)	1	–	NE
	Collection3 Mozharova and Loukachevitch (2016)	ru	News	26.4 K (3)	1	–	NE
	BSNLP2019 Piskorski et al. (2019)	Slavic	News	9 K (5)	1	–	NE/EL
	BSNLP2021 Piskorski et al. (2021)	Slavic	News	9.5 K (5)	1	–	NE/EL
	MultiCoNER-2022 Malmasi et al. (2022)	11 lang	News	16.1 K (6)	1	–	NE
	FactRuEval Starostin et al. (2016)	ru	News	12 K (3)	2	1 K (4)	NE/RE
	RuREBUS Ivanin et al. (2020)	ru	Econ	121 K (5)	1	14.6 K (8)	NE/RE
	RURED Gordeev et al. (2020)	ru	Econ	22.6 K (28)	1	5.3 K (34)	NE/RE
	WikiOrgs Kuznetsov et al. (2016)	ru	Wiki	7 K (1)	1	7 K (2)	RE
	Situations-1000 Vlasova et al. (2016)	ru	News	2.2 K (3)	1	(2)	RE
	RuWiki Sysoev and Nikishina (2018)	ru	Wiki	60 K	1	–	EL
	RuSERRC Bruches et al. (2021)	ru	Science	1337	1	620 (6)	TE/RE/EL
	NEREL (ours)	ru	News	56 K (29)	6	39 K (49)	NE/RE/EL

Group 1 includes most known datasets with flat entities without relations annotation. Group 2 comprises datasets with nested named entities without or with a small number of relation types. Group 3 includes most known datasets annotated with relations. Group 4 comprises datasets with knowledge base entity links. Group 5 presents various Russian datasets created for different information extraction tasks

NE named entities, *TE* term extraction, *RE* relation extraction, *EL* entity linking

Several datasets for named entity recognition are available in Russian. These include, dataset developed by Gareev et al. (2013), Persons 1000 and Collection 5 (Mozharova & Loukachevitch, 2016; Vlasova et al., 2014), FactRuEval 2016 (Starostin et al., 2016), the Russian subset of the BSNLP 2019 (Piskorski et al., 2019) and 2021 (Piskorski et al., 2021) Shared Tasks. The BSNLP-2019 Shared Task introduced a new multilingual dataset, annotated with named entities for four Slavic languages: Bulgarian, Czech, Polish, and Russian (Piskorski et al., 2019). Five types of named entities were annotated, including persons, locations, organizations, products, and events. The BSNLP-2021 dataset includes two additional languages, Slovene and Ukrainian, focusing on cross-lingual, document-level extraction of named entities. In addition to Named Entity Mention Detection and Classification includes two additional tasks, Name Lemmatization and Entity Matching, where mentions of the same entity should be assigned the same identifier. A novel dataset for complex Named Entity Recognition with 11 languages (including Russian) published in the MultiCoNER 2022 Shared Task.[3] The dataset focuses on longer spans of named entities but does not include nested named entities.

Among Russian NER datasets, the most number of various entity types is annotated in the recent RURED dataset (Gordeev et al., 2020). The dataset contains 500 news articles about finance and economics. The annotation of named entities is mainly based on the OntoNotes guidelines (Hovy et al., 2006) with additional entities (currency, group, family, age). The GPE entity was subdivided into four categories: *Country*, *Region*, *City*, and *Borough*. Nested named entities were not labeled—only upper-level entities were annotated.

FactRuEval2016 (Starostin et al., 2016) is currently the only Russian dataset annotated with nested named entities. Entities are annotated with at most two levels of nesting. Entity spans comprise person-related spans (names, surnames, patronymics, and nicknames), organization-related spans (descriptors and names), and location-related spans (descriptors and names). Spans are grouped into object mentions. Object types include people, organizations, locations, and the use of location within the organization. Several entities may share common spans, for example, a common descriptor.

To sum up, datasets with annotated nested named entities exist for several languages. The larger number of entity types in annotating nested named entities leads to the larger number of nestedness levels. At the same time, the vast majority of current datasets have limited nestedness or are small in terms of the number of annotations. In particular, the existing Russian dataset with nested named entities (FactRuEval) is not large enough to train usable neural network models.

2.2 Datasets for relation extraction

SemEval-2010 (Hendrickx et al., 2010) is a well known dataset for relation extraction. It provides annotation for 9 different relation types. For each relation, sentences

[3] https://multiconer.github.io.

were automatically collected through pattern-based Web search, based on more than 100 patterns for the relation. Relations (*cause-effect, instrument-agency,* etc.) are treated as mutually exclusive to avoid problems with subjectivity or double annotation. TACRED (Zhang et al., 2017) is currently one of the largest datasets for relation extraction. Relations are annotated at sentence level using crowdsourcing. Each sentence is labeled with one of 41 person- or organization-oriented relation types or the label no_relation for negative instances. Crowd workers were shown the example text, with subject and object mentions highlighted, and asked to select amongst a set of relation label suggestions or assign no_relation. Label suggestions were limited to relations compatible with the subject and object types.

DocRED (Yao et al., 2019) is annotated with both named entities and relations at the document level. A significant proportion of relations (40.7%) is established between entities in different sentences. The dataset includes 96 frequent relation types from Wikidata. Relations are initially annotated using distant-supervision based on Wikidata relations and entities found in documents. Subsequently, annotators reviewed the extracted relations, removed incorrect relation instances, and supplemented omitted ones.

Much less attention is paid to relation extraction than to named entity recognition for the Russian language. Only five datasets are annotated with relations (Ivanin et al., 2020; Kuznetsov et al., 2016; Starostin et al., 2016; Vlasova et al., 2014). The FactRuEval dataset (Starostin et al., 2016) includes annotation of facts, which are relations between multiple named entities. Each fact has a corresponding frame of fields to be filled. Each field has a name and a list of possible types of object that may fill it. The facts are annotated on the document level.

The RURED dataset (Gordeev et al., 2020) relation scheme is inherited from the TACRED relations. It is extended with several new relations, which stand for events, such as the date of an event (*date_take_place_on*), the place of an event (*take_place_in*), participants in an event (*organizes, event_take_place_in*). The annotation of relations lies mainly within sentences. RuREBus corpus (Ivanin et al., 2020) consists of economic documents issued by a state agency. Other existing datasets (Kuznetsov et al., 2016; Vlasova et al., 2014) are much smaller and are not widely used for experiments. Table 1 summarizes characteristics of the Russian datasets available to date.

Currently, there are no other datasets with annotated relations and entity links over nested named entities; see Table 1. Also there are no large datasets with nested named entities and relations between them. Relations are annotated either at sentence-level (i.e., relations do not span over sentence boundaries) or at document-level (i.e., relations cross sentence boundaries). NEREL contains all of these variations.

2.3 Datasets for entity linking

The most widely-used dataset for the evaluation of EL systems is AIDA-CoNLL (Hoffart et al., 2011).[4] It is based on the English CoNLL-2003 corpus (Tjong & De Meulder, 2003) manually annotated with links to YAGO2—a knowledge base automatically built from Wikipedia (Hoffart et al., 2013). AIDA-CoNLL is the most extensive manually annotated dataset. It provides a predefined train/validation/test split making it possible to evaluate systems in a "closed" setting—without using additional datasets for training.

Several EL corpora, including a tri-lingual dataset covering English, Chinese, and Spanish, were released as part of the Entity Discovery and Linking (EDL)/ Knowledge Base Population (KBP) shared tasks at the Text Analysis Conference (TAC) (Ellis et al., 2014; Getman et al., 2017; Ji et al., 2015). The corpora are built from news wire and web forums; the reference knowledge base is derived from Wikipedia infoboxes. Some of the mentions in the TAC KBP corpora do not have a recallable entity in a knowledge base (assigned a "NIL" link). This fact complicates the task requiring that EL systems should also have a mechanism for NIL prediction. TAC KBP datasets (Ellis et al., 2015) are often used for the evaluation of cross-lingual entity linking systems (Zhou et al., 2019). Another multilingual dataset for EL is VoxEL (Rosales-Méndez et al., 2018). It contains 15 manually annotated news articles on politics in 5 different languages. Other widely-used datasets for entity linking are available within the GERBIL platform (Röder et al., 2018): AQUAINT (Milne & Witten, 2008), ACE2004 (Ratinov et al., 2011), DBpedia Spotlight (Mendes et al., 2011), and others.

Besides manual annotation that can produce only relatively small datasets, researchers also leverage automatic labeling for preparing training and validation data. Even though the annotation is imperfect in this case, high-quality annotation can be achieved using techniques such as alignment. The English ClueWeb corpus (collection of web pages) (Gabrilovich et al., 2013) was automatically annotated with entity mentions and links to Freebase (Bollacker et al., 2007). For benchmarking entity linking systems, researchers usually use a subset of this corpus WNED-CWEB and another corpus derived from Wikipedia—WNED-Wiki (Guo & Barbosa, 2018). For the construction of multilingual EL models, Botha et al. (Botha et al., 2020) recently have automatically derived the Mewsli-9 dataset from Wikinews articles. The dataset covers 104 languages and contains links to 20 million WikiData entities.

There are also a few EL datasets for the Russian language. Sysoev and Nikishina presented a Russian entity linking dataset based on Wikipedia. Another Russian dataset for entity linking is RuSERRC (Bruches et al., 2021), which contains abstracts of scientific papers on information technology. The abstracts are labeled with scientific terms, which are linked to Wikidata entities. However, the first dataset is not freely available, and the second one is rather small. Additionally, neither of

[4] AIDA: https://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/ambiverse-nlu/aida.

these datasets are annotated with nested named entities. Table 1 provides comparisons between the above-mentioned datasets.

Most datasets with EL annotation do not support nested entity mentions (e.g., AIDA-CoNLL (Hoffart et al., 2011), ACE2004 (Ratinov et al., 2011), etc.) However, there are a few exceptions. The SemEval 2015 Task 13 (Moro & Navigli, 2015) and DBpedia Spotlight (Mendes et al., 2011) datasets allow for nested entities. VoxEL (Rosales-Méndez et al., 2018) provides two versions of the dataset: strict and relaxed. The strict version contains non-overlapping maximal entity mentions. The relaxed version considers any noun phrase matching a Wikipedia entity as a mention, including overlapping mentions where applicable.

Some datasets are annotated for multiple tasks and involve the joint interaction between them where the information obtained in one task can help to solve another task. Such datasets can be used for research on complex document-level reasoning that goes beyond the local context directly surrounding individual entity mentions. For example, the English DWIE corpus (Zaporojets et al., 2021) combines four annotation sub-tasks: named entity recognition, relation extraction, coreference resolution, and entity linking. The named entities are linked to Wikipedia. Another example is SciERC, a dataset for identification and classification entities, relations, and coreference clusters in scientific articles. The dataset consists of 500 scientific abstracts from the Semantic Scholar Corpus. Coreference links are annotated between identical scientific entities as a cross-sentence relation.

Table 1 presents manually annotated datasets for entity linking. It is important to note that all available Russian named entity or relation annotated datasets are not intended for entity linking or knowledge base population tasks, i.e. annotated entities and relations are not linked to an existing knowledge base and do not account for a knowledge base's ontology.

3 NEREL annotation principles

In designing NEREL annotation guidelines, we aimed to annotate fine grained named entities and relations while at the same time avoiding a long tail of low-frequency entity types or relation types.

We chose the approach of labeling both longer and shorter (internal) named entities (so called nested named entities) because internal entities can also either have a relation with another entity or have a Wikidata link. For example, in phrase *Mayor of Moscow*, we see title *Mayor* and city *Moscow*. All three entities can be linked to Wikidata, internal entity *Moscow* can be required for establishing relations with other entities in a text. Hence, all three entities were annotated.

While annotating common named entities (e.g. PERSON, LOCATION) expressed with nouns and noun groups, we also annotate adjectives derived from named entities, e.g. adjective *Moskovskii* originated from noun *Moscow*. When the corresponding noun is absent in the text, adjectives can be essential for establishing a relation in a sentence or within the whole text (Benikova et al., 2014). Also, an adjective often conveys reference to the same entity in a knowledge base as a noun from which the adjective is derived.

We also annotate nouns as entities that are not proper names when these are needed for relation extraction. These cases include:

- entities, which are written in lowercase in Russian, but are capitalized in English such as RELIGION (*Islam*) or NATIONALITY (*Russian*);
- concrete mentions of authorities, labeled as an ORGANIZATION entity, because they refer to important actors in different relations and events (for example, *the police*);
- concrete entities, which are written in lowercase and include a named entity (*physics department of Lomonosov Moscow State University*);
- entity types, which can be written in lowercase or capitalized, such as PROFESSION, AWARD, EVENT, are annotated regardless their capitalization because of their importance in describing person's relations;
- lowercase written entities, which express significant personal information, such as DISEASE or CRIME.

We annotate so called "news events" as opposed to everyday or regular activity (Thompson et al., 2017; Wang et al., 2022). News events can be subdivided into named events (*French revolution, Hurricane Katrina*) and also non-named events such as accidents, political actions, meetings (*died, born, crash* etc.,). The last event type is close to "event nuggets" as described in Mitamura et al. (2015b). Such no-named events and their relations provide additional connections between mentioned named entities that would otherwise be missed.

NEREL puts stronger emphasis on people and people-related entities as these have a greater significance (with higher frequency) within the NEREL corpus. The corpus consists of general news articles in Russian. For comparison purposes we also annotated a small general news corpus in English. Wikidata (Vrandečić & Krötzsch, 2014) is used as a knowledge graph for entity linking.

Relations are annotated both at the sentence and document level. Existing relations datasets can be subdivided into sentence-level or document-level relation datasets. However, both levels are important for various tasks. Therefore, in NEREL, documents are annotated mainly on the sentence level with cross-sentence links, which provides the possibility for inference of the whole set of entities and relations on the document level. For this, variants of the same entity naming are connected to each other by the *abbreviation* and *alternative_name* relations (further referred as synonymous relations). It should be noted that the usage of such relations is not equivalent to coreference annotation because currently no pronouns are annotated in NEREL. Besides, the synonymous relations should connect entities of the same entity type—coreferential relations can link entities of different types, for example:PERSON and NATIONALITY.

In a sentence, all relations are annotated. If a relation involves an entity located in another sentence, we search for it in a previous sentence to complete the relation. Relations requiring links to an out-of-sentence entity located in longer distances are established using the closest mention of the needed entity. Thus, from the NEREL annotation, three different relation datasets can be obtained: in-sentence relation dataset, two-sentence relation dataset, and document-level relation dataset. The two-sentence relation dataset seems to be easier for extracting relations than the document-level relation extraction and can be used to study relation distribution phenomena in neighbor sentences.

Currently, NEREL is the only dataset with three-level annotation over nested named entities with relations annotated on the sentence level and between sentences with the possibility to derive a document-level entity set and relations. In addition, linkages to the Wikidata make the dataset a unique resource for studies in knowledge extraction and knowledge base construction domains.

4 Data collection and annotation

4.1 Selection of texts and annotation process

The NEREL corpus consists mainly of Russian Wikinews articles published under the Creative Commons Attribution 2.5 License, which allows the reuse of the materials. The additional advantage of Wikinews is partial linking of mentioned entities to Wikipedia pages, which is helpful for further linking annotated NEs to Wikidata. By summer 2020, Russian Wikinews hosted about 23 K articles.

To select articles for annotation, we first applied a NER model trained on RURED data (Gordeev et al., 2020) to the whole Wikinews collection. We focused on articles with a high density of automatically detected NEs, paying particular attention to NEs associated with persons (e.g. PERSON, AGE). We extracted texts in the range of 1–5 kB. Such medium-sized texts are more convenient for annotation: very short texts provide little context, while long documents are usually incoherent lists, e.g., of movies or events. The extracted articles were inspected manually, and 933 articles were finally selected for annotation.

We leveraged the *brat* tool (Stenetorp et al., 2012) for annotation. Three levels of annotation—named entities, relations, Wikidata links—were performed as subsequent independent passes. An annotator performed annotation with moderator control. Due to the complexity of annotation, some guidelines were introduced during the annotation. In such cases, the built-in search mechanism of *brat* was used to implement corrections.

4.2 Named entity annotation

4.2.1 Annotation principles

To define a list of entity types for NEREL annotation, we started with corresponding lists of English OntoNotes (Hovy et al., 2006) and RURED (Gordeev et al., 2020) datasets. Additionally, we considered entity types of the Stanford named entity recognizer (Finkel et al., 2005) and slots of the TACRED corpus such as CRIME

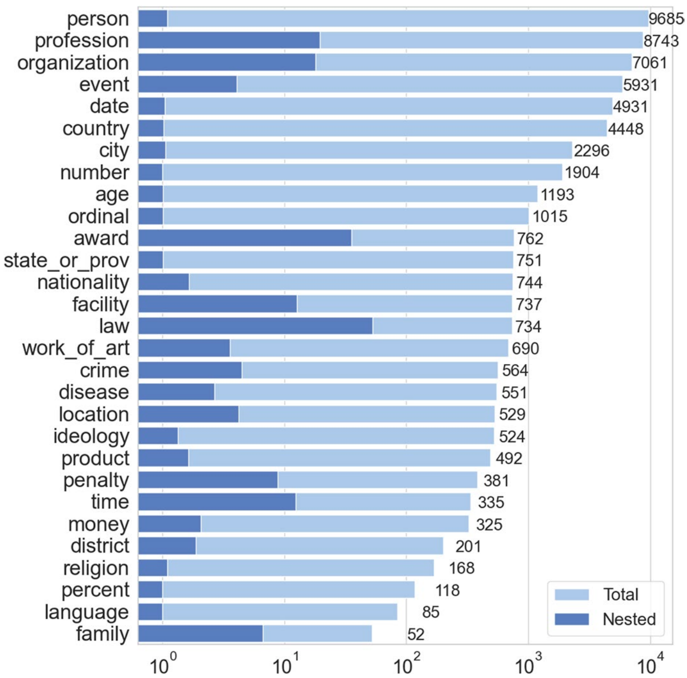

Fig. 2 Entity type statistics (log scale). The proportion of nested named entities is shown

and PENALTY (Zhang et al., 2017). Some entity types (AWARD, DISEASE) were added because of their significant frequency in the gathered collection.

Currently, there are 29 entity types in the NEREL dataset:

- basic entity types: PERSON, ORGANIZATION, LOCATION, FACILITY, geopolitical entities subdivided into COUNTRY, STATE_OR_PROVINCE, CITY, DISTRICT entities;
- 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.

Resulted entity type frequencies are presented in Fig. 2. As can be seen from the statistics, all but two entity types have at least 100 annotated examples.

Regarding the rules of annotation of nested named entities, we tried to choose decisions motivated with necessity of better entity representation for relation annotation and entity linking. Annotation of internal entities varies of specific named entity type. For example, we do not annotate numbers within numerical entities such as date or money, because such annotations are not essential for relation extraction and entity linking tasks. Pronouns are not annotated.

In news texts, geopolitical entities (such as STATE_OR_PROVINCE, CITY, and DISTRICT) are often mentioned in form of adjectives. In NEREL, such adjectives are annotated with the same label as a nominal named entity. COUNTRY adjectives can be annotated with two different tags: COUNTRY and NATIONALITY depending on the context. When the adjective is related to state authorities, headquarters of an organization, positions, and employees working in state bodies (including military units), such adjectives are annotated as COUNTRY. In other cases, when relating to artists, sportspeople, writers, ordinary citizens, adjectives derived from country names are annotated as NATIONALITY. This decision was motivated by relations that are more relevant to each context. For example, in former contexts adjectival forms of countries are more often involved into *located_in* and *headquartered_in* relations.

If compared to previous datasets (Hovy et al., 2006; Weischedel & Brunstein, 2005), all adjectival forms of countries were annotated as a NORP entity (Nationality, Other, Religious, Political) regardless of context, but in these corpora relations were not annotated. In Ringland et al. (2019), the authors use an additional specialised tag NORP: *Nationality* approximately with the same rules of the annotation. In the NoSta-D dataset (Benikova et al., 2014) a special postfix *_deriv* is introduced for adjectival forms of basic entities.

Table 2 shows examples of annotating nested named entities for different entity types and other complicated cases.

For example, the PRODUCT tag is mainly used for annotating product model names, including numbers. In flat named entity annotations, different guidelines can be used for PRODUCT entities annotation. In the OntoNotes (Hovy et al., 2006) dataset, a manufacturer and a product should be annotated separately as ORG+PRODUCT. The same approach is accepted in the Russian Collection3 (Mozharova & Loukachevitch, 2016). In BSNLP-2019 (Piskorski et al., 2019), the manufacturer name should be included in a longer product name. In NEREL, a PRODUCT entity is annotated as a long span, within which the manufacturer and numbers can be marked if necessary.

4.2.2 Annotation process

The labeling of named entities in NEREL started from automatic pre-annotation. A BERT-based named entity recognizer trained on the RURED (Gordeev et al., 2020) dataset was applied to NEREL documents. The recognizer provides a good quality of annotations of basic flat entity types such as PERSON, ORGANIZATION, LOCATION and geopolitical entities. Manual annotation of named entities and relations was performed by a single annotator, controlled by a moderator.

4.2.3 Inter-annotator agreement

We calculated inter-annotator agreement (IAA) through Krippendorff's alpha coefficient (Krippendorff, 2004). We chose Krippendorff's alpha instead of Cohen's kappa due to its inaccuracies noted in several studies (Brandsen et al., 2020; Campillos-Llanos et al., 2021; Jiang et al., 2022). We note that Krippendorff's alpha generalizes several known agreement measures. Additionally, it can handle any number of categories at multiple scales (nominal, ordinal, binary, etc.) containing incomplete or missing data (Checco et al., 2017).

Table 2 Examples of annotating nested named entities of different types

Entity	Example	Annotation
PROFESSION	Governor of California	[Governor of [California]$_{\text{STATE\_OR\_PROVINCE}}$]$_{\text{PROFESSION}}$
	head of Gazprom	[Head of [Gazprom]$_{\text{ORGANIZATION}}$]$_{\text{PROFESSION}}$
ORGANIZATION	Physics department of Lomonosov Moscow State University	[physics department of [[Lomonosov]$_{\text{PERSON}}$ [Moscow]$_{\text{CITY}}$ State University]$_{\text{ORG}}$]$_{\text{ORG}}$
	Russian Government	[[Russian]$_{\text{COUNTRY}}$ government]$_{\text{ORG}}$
NATIONALITY	Citizen of Russia	[Citizen of [Russia]$_{\text{COUNTRY}}$]$_{\text{NATIONALITY}}$
	Russian writer	[Russian]$_{\text{NATIONALITY}}$ [writer]$_{\text{PROFESSION}}$
	Russians	[Russians]$_{\text{NATIONALITY}}$
LAW	Yarovaya law	[[Yarovaya]$_{\text{PERSON}}$ law]$_{\text{LAW}}$
	Article 84 of the Constitution of Kyrgyzstan	[Article [84]$_{\text{ORDINAL}}$ of the [[Constitution]$_{\text{LAW COUNTRY}}$]$_{\text{LAW}}$]$_{\text{LAW}}$
CRIME	Complicity in murder of VGTRK journalists	[[complicity in murder]$_{\text{CRIME}}$ of [[VGTRK]$_{\text{ORG}}$ [journalists]$_{\text{PROFESSION}}$] $_{\text{CRIME}}$
	Armed attack on passers-by	[[Armed attack]$_{\text{CRIME}}$ on passers-by]$_{\text{CRIME}}$
	Attempt to oust Erdogan	[Attempt to oust [Erdogan]$_{\text{PERSON}}$]$_{\text{CRIME}}$
PENALTY	\$129 million fine	[[\$129 million]$_{\text{MONEY}}$ [fine]$_{\text{PENALTY}}$]$_{\text{PENALTY}}$
	Imprisonment for 3 years	[[imprisonment]$_{\text{PENALTY}}$ for [3 years]$_{\text{DATE}}$]$_{\text{PENALTY}}$
PRODUCT	Boeing-737 MAX	[[Boeing]$_{\text{ORG}}$-[737]$_{\text{NUMBER}}$ MAX]$_{\text{PRODUCT}}$
	Apple Watch	[Apple]$_{\text{ORG}}$ Watch]]$_{\text{PRODUCT}}$
AWARD	Merit for the fatherland order	[Merit for the fatherland order]$_{\text{AWARD}}$
	Gold of the olympic games	[Gold of the [olympic games]$_{\text{EVENT}}$]$_{\text{AWARD}}$
	champion of the Olympic Games	[champion of the [Olympic Games]$_{\text{EVENT}}$]$_{\text{AWARD}}$
	Miss Russia-2017	[Miss [Russia]$_{\text{COUNTRY}}$-[2017]$_{\text{DATE}}$]$_{\text{AWARD}}$
	Gold medal	[gold medal]$_{\text{AWARD}}$
EVENT	40th Moscow International film festival	[[40th]$_{\text{ORDINAL}}$ [[Moscow]$_{\text{CITY}}$ International film festival]$_{\text{EVENT}}$]$_{\text{EVENT}}$
	UEFA champions league	[UEFA]$_{\text{ORG}}$ champions league]$_{\text{EVENT}}$
	Sochi-2014	[[Sochi]$_{\text{CITY}}$-[2014]$_{\text{DATE}}$]$_{\text{EVENT}}$

15 documents with 1000 entities were labeled by a moderator (the gold standard) and an annotator. We obtained Krippendorff's alpha of 80.91 demonstrating good quality IAA and reliability of the annotations, as noted by other researchers (Campillos-Llanos et al., 2021; Shabankhani et al., 2020)

Error analysis revealed that the most frequent sources of annotation inconsistencies are as follows: span boundaries of event nuggets, confusing FACILITY and ORGANIZATION entities, confusing EVENT and CRIME entities (such as *murders*) or EVENT and PENALTY entities (such as *arrests*). *Student* role is often annotated as PROFESSION (in spite of being a kind of pre-professional title).

4.3 Event annotation

We annotate two types of expressions as EVENT entities. The first type is named events such as named hurricanes, battles, wars, sports events, as in other NER datasets (Hovy et al., 2006; Ringland et al., 2019) (*40th Moscow International Film Festival, 1966 Tashkent earthquake*). Such EVENT entities usually include geopolitical names (COUNTRY, CITY), LOCATION, ORGANIZATION or DATE entities (Table 2).

The second type of annotated EVENT entities comprises mentions of concrete news events without names such as *kill, arrest, marry* etc. The annotation of this subtype of EVENT entities is similar to annotation in specialized datasets for event annotation (Bies et al., 2016; Mitamura et al., 2015a, 2015b; Song et al., 2015), where an event is defined as an explicit occurrence involving participants. Annotation of non-named events in our dataset is most similar to event nuggets (Mitamura et al., 2015b) defined as the smallest extent of text that expresses the occurrence of an event.

Event nuggets in NEREL can be single words (nouns or verbs) or phrases (noun phrases, verb phrases, or prepositional phrases). A phrase is annotated as an EVENT entity if extra words add significant information to a too general or ambiguous main event word (trigger). Such additional words for noun spans are mainly dependent adjectives or genitive groups (*parliamentary elections, opening of exhibition*). Verb-based EVENT spans may additionally include: direct of indirect objects (*relieved of his post, went free*) and the subject of a sentence. Event nuggets usually should not include other entity types that can be connected with the event entity via available relations. An event nugget can be discontinuous if the main word does not fully convey the sense of the event except within a multiword event phrase, a longer named entity or other extra words are included.

We annotate actual events that occurred at a particular place and time, excluding anticipated or possible events discussed in the text. Also we annotate future events with exact dates. We do not restrict subtypes of EVENT entities in NEREL except for: speech acts and cognitive acts, regular activities, changes of numerical indicators (for example, prices or import values), victories, and defeats.

The analysis of the annotated corpus showed that the most frequent types of annotated non-named events in NEREL are as follows:

- accidents, death of people: *crash, to attack, knife attack*;

- public actions and ceremonies: *demonstration, action of protest, to present, presentation*;
- meetings and gatherings: *to meet, meeting, session, summit*;
- legal actions: *to indict, search, interrogation, to sentence*;
- transactions: *to buy, to sell, selling*;
- appointments and resignations: *to appoint, appointment, to dismiss*;
- medical actions: *hospitalizations, surgical operations*;
- sports events: *match, final, game*.

4.4 Relation types and relation annotation

The list of relation types in NEREL is initially based on the English TACRED corpus (Zhang et al., 2017) and the Russian RURED dataset (Gordeev et al., 2020). Further, the list of relations has been corrected and expanded from the NEREL corpus analysis. When possible, the most similar correlations for all the NEREL relations in Wikidata properties were found.[5] The current set of annotated relation types in the NEREL corpus includes 49 relations.

Relations can be subdivided into several semantic groups according to the most representative entity involved: person-oriented, organization-oriented, event-oriented (e.g. *participant_in*), product-oriented (such as *price_of, part_of, located_in*, etc), and synonymous relations (*alternative_name, abbreviation*). Table 3 shows main groups of relations.

Figure 3 shows the distribution of relation frequencies in the NEREL dataset. It can be seen that least frequent relations have at least 50 examples.

The relations are annotated within and across sentences. *abbreviation* and *alternative name* relations connect different mentions of the same entities located within a single sentence or in different sentences. There are two setups for relation extraction at three levels: sentence-level, and document-level. The document-level dataset is the most important for study of knowledge graph population from texts.

Among all the annotated relations, it is worth considering so-called nested (or internal) relations, established between a longer named entity and its internal named entity. Table 4 presents the most frequent types of nested named entities connected with nested relations. It can be seen that the entities PROFESSION and ORGANIZATION enter most frequently in nested relations as outer entities. We also report the most frequent relations with their nestedness score in Table 5. The last column shows the percentage of relations annotations that are nested relations. Some relations with specific arguments tend to be nested; other relations are never met within a longer entity.

The principles of establishing relations in the NEREL dataset are as follows:

[5] Some relations do not have counterparts in Wikidata properties. For example, AGE and AGE_DIED_AT occur in texts, while Wikidata has only *date of birth (P569)* and *date of death (P570)* that allow calculate the above mentioned age values.

Table 3 Relation groups

Entity	Relation groups	Relations
PERSON	Living relations	*place_of_birth, date_of_birth, place_of_death, date_of_death, place_resides_in, age, age_of_death, medical_condition, cause_of_death, origins_from*
	Family relations	*parent_of, spouse, sibling, relative, origins_from, member_of*
	Education and labour	*schools_attended, workplace, work_as, organizes, subordinate_of, awarded_with, member_of, produces*
	Property relations	*owner_of, expenditure, income, founder_of*
	Mental relations	*knows, ideology_of, religion_of*
	Violation relations	*convicted_of, penalized_as*
	Participance relations	*agent, participant_in*
ORGANIZATION	Existence relations	*date_founded_in, date_defunct_in, founded_by, headquartered_in*
	Between-organization relations	*member_of, part_of*
	Activity relations	*organizes, awarded_with, produces*
	Mental relations	*ideology_of, religion_of*
	Property relations	*owner_of, expenditure, income, founder_of*
	Violation relations	*convicted_of, penalized_as*
	Participance relations	*agent, participant_in*
EVENT	Role relations	*agent, participant_in, inanimate_involved*
	Time and space relations	*point_in_time, start_time, end_time, take_place_in*
	Between-event relations	*subevent_of, has_cause*
Inanimated objects and products		*price_of, part_of, inanimate_involved, date_of_creation, located_in, origins_from*

Entities which can used for reference to PERSON entities such as NATIONALITY or PROFESSION have relation sets similar to PERSON

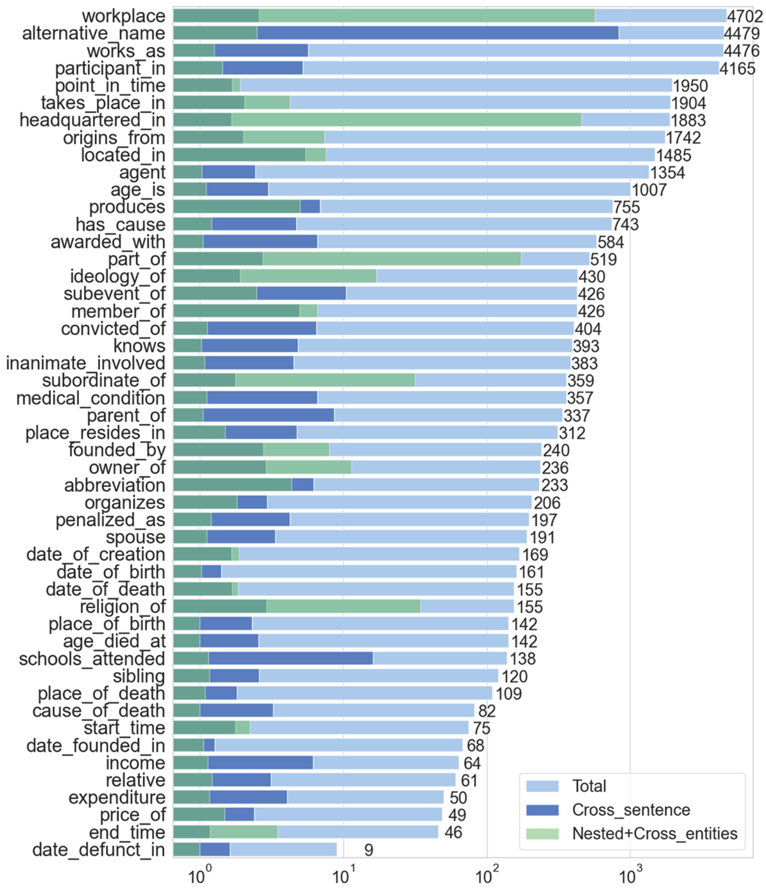

Fig. 3 Relation type statistics (log scale). Proportions of cross-sentence relations and relations involving nestedness of entities are shown

Table 4 The most frequent nested relation types

Relation type	Outer_type	Inner_type	#	% in corpus
WORKPLACE	PROFESSION	ORGANIZATION	672	23.59
HEADQUARTERED_IN	ORGANIZATION	COUNTRY	447	15.69
WORKPLACE	PROFESSION	COUNTRY	372	13.06
PART_OF	ORGANIZATION	ORGANIZATION	151	5.30
HEADQUARTERED_IN	ORGANIZATION	CITY	147	5.16
HEADQUARTERED_IN	ORGANIZATION	STATE_OR_PROVINCE	93	3.26
WORKPLACE	PROFESSION	STATE_OR_PROVINCE	89	3.12
SUBORDINATE_OF	PROFESSION	PROFESSION	77	2.70
ORIGINS_FROM	LAW	COUNTRY	57	2.00
IDEOLOGY_OF	ORGANIZATION	IDEOLOGY_OF	54	1.90

- if a longer and a shorter internal entity of the same type are annotated (*President of Russia—President*), all the relations are established with a longer entity. An internal entity in such cases can help in entity linking if a longer entity does not have a direct counterpart in a knowledge base;
- all variants of naming an entity in a single sentence or neighbour sentences are connected with *alternative_name* or *abbreviation* relations, other relations are linked the closest entity mentions among entities' variants;
- cross-sentences relations in neighbor sentences are annotated with the same detail as in a single sentence;
- relations, which contain entities located in longer distances than two sentences, should be annotated at least once in a text to have a possibility to generate a document-level relation extraction dataset.

4.5 Linking entities to Wikidata

4.5.1 Annotation principles

We use Wikidata (Vrandečić & Krötzsch, 2014) as our target knowledge base (KB). Wikidata is a large open multilingual knowledge base that community members can edit. At the current stage of annotation, we excluded seven numerical types, as well EVENT and IDEOLOGY entities from entity linking, thus ended up with 17 entity types. Although there are entities in Wikidata for numbers, individual years, periods, and dates, linking to these entities does not make much sense—actual values in the KB triples are presented as literals. The EVENT type in NEREL includes named events (e.g. *Monsters of Rock* concert held in Moscow on 28 September 1991 (Q4301946)), as well as events expressed in common nouns (e.g., *funeral* (Q201676)) and verb forms,[6] which greatly complicates the annotation. We will address linking of EVENT entities on the next stages of the resource development.

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 entity is absent in Wikidata, 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

[6] For example, the phrase in bold in the sentence *Natalia accepted Pushkin's proposal, and in April 1830, she became the wife of the famous Russian poet Alexander Pushkin* should be linked to the Wikidata's *marriage* (Q8445).

Table 5 The most frequent relation types with their 'nestedness' score

Relation type	Arg1	Arg2	#	% in corp.	Nestedness (%)
WORKS_AS	PERSON	PROFESSION	2601	13.74	0.08
PARTICIPANT_IN	PERSON	EVENT	1399	7.39	0.57
WORKPLACE	PERSON	ORGANIZATION	969	5.12	0.93
ALTERNATIVE_NAME	PERSON	PERSON	927	4.90	0.00
POINT_IN_TIME	EVENT	DATE	841	4.44	2.62
WORKPLACE	PROFESSION	ORGANIZATION	806	4.26	83.37
HEADQUARTERED_IN	ORGANIZATION	COUNTRY	554	2.93	80.69
AGE_IS	PERSON	AGE	429	2.27	0.00
WORKPLACE	PROFESSION	COUNTRY	426	2.25	87.32
PARTICIPANT_IN	ORGANIZATION	EVENT	372	1.97	3.23

annotation of such entities. The AIDA guidelines prefer to annotate more specific entities (Hoffart et al., 2011). In contrast, the TACKBP annotation guidelines (Ellis, 2012), and later projects (Hamdi et al., 2021) specify that different instantiations 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]$_{ORG}$]$_{ORG}$ linking the entities both to Q170375 [111th U.S. Congress] and Q11268 [U.S. Congress].

NER and EL annotation also suffer from *metonymy*, when an entity is referred to 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 FACILITY and is linked to the Wikidata item *the White House* (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 Wikidata items of the corresponding named entities. For example, the adjective *Moskovskii* derived from *Moscow* is linked to the same item as the initial name: *Moscow* (Q649). The linking of adjectives enhances the coverage of the annotation. At the same time, linking adjectives can be difficult for automatic methods because adjectives are usually not among labels and aliases of Wikidata items. 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 Anti-Doping Agency*), Russian citizens (Q49542, NER type NATIONALITY: *Russian composer*) or Russian language (Q7737, NER type LANGUAGE: *Russian version of magazine*).

We link PROFESSION entities to profession items in Wikidata, not to a person who is currently holding the position because in some texts different candidates to the same post can be discussed. For example, *Mayor of Moscow* is linked to the *Mayor of*

Moscow Wikidata item Q1837906, not to *Sergey Sobyanin* (Q319497), the current Moscow mayor.

4.5.2 Annotation tools

In annotating Wikidata links, we used the normalization system of the *brat* annotation tool (Stenetorp et al., 2012). 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* annotation files. In addition, we retained only one mention per entity in the document based on synonymous relations (see Sect. 4.4). Removing multiple mentions of the same entity within a document decreases the number of named entities linked to Wikidata by 40% (from 38,175 to 22,887 in 933 documents).

We applied an entity linker that was developed for the annotation of a KBQA dataset.[7] The linker builds a search index over a collection of Russian labels and aliases from Wikidata that correspond to around 4 M entities using Elasticsearch. 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 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. Linker's implementation details can be found in the paper (Korablinov & Braslavski, 2020).

Some entity mentions are linked to corresponding Wikipedia pages in original Wikinews articles.[8] For about 15% (3454) of entities to be linked, we could provide 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 *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 the 'local' *brat* knowledge base.

Annotators were presented with documents with highlighted entities. The majority of entities are provided with a candidate Wikidata linkage and its ID, label, and description. Annotators were also able to follow a hyperlink to the Wikidata entity page. To correct an existing linkage or produce a new one, annotators could search the local collection of Wikidata entities using the 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 most convenient for annotators. If no corresponding Wikidata item was found,

7 https://github.com/vladislavneon/kbqa-tools/.

8 Among them, there are many linkages to year or month entities, e.g. *October 2009* → *Q243251*, that are not very helpful for our task.

Table 6 Manual entity linking statistics based on 933 documents: total linkages by type, including NULL

Entity type	#	Incl. NULL	Acc.
AWARD	597	185	0.51
CITY	1434	11	0.71
COUNTRY	1982	5	0.75
DISTRICT	156	10	0.66
FACILITY	553	171	0.49
LANGUAGE	70	0	0.77
LAW	584	311	0.29
LOCATION	403	71	0.45
NATIONALITY	528	6	0.32
ORGANIZATION	4650	969	0.61
PERSON	4430	901	0.57
PRODUCT	343	27	0.83
PROFESSION	5889	1724	0.54
RELIGION	107	4	0.53
STATE_OR_PROVINCE	472	1	0.81
WORK_OF_ART	532	143	0.55
Total	22,730	4539	0.59

The last column reports accuracy of the automatic linkage suggestions against manual annotation

annotators provided the entity mention with a special NULL value. On average, annotators spent an hour processing 100 entity mentions.

Table 6 provides entity linking statistics of almost 23 K entities from 933 documents. The figures give an idea of Wikidata coverage of different entity types. Automatic suggestions greatly facilitate the annotation with an average accuracy of 59%. The lower values, as for example in case of NATIONALITY can be explained by the annotation scheme: the adjective *British* in case of *British actor* must 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 and in contrast to English the cognate words can be quite distant on character level, e.g. *peterburzhskiy—Saint Petersburg*, *rossiyskiy—Rossia*. Note that the Table provides statistics for the 'cleaned' annotation, where only one mention per entity/document is retained. After the Wikidata linking was finished, we restored initial NE/relation annotations and propagated Wikidata linkages to other mentions of the same entity in the corresponding *brat* standoff files.

4.5.3 Annotation statistics

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 5394 pairs, both outer and inner entities are linked to Wikidata. The majority of the remaining

Table 7 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	1155	647	365	44	99
ORGANIZATION	COUNTRY	1046	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	2098	1019	860	25	194
PROFESSION	ORGANIZATION	1611	329	1004	16	262
PROFESSION	COUNTRY	1015	664	351	0	0
PROFESSION	CITY	228	81	142	4	1
PROFESSION	STATE_OR_PROVINCE	185	67	116	0	2

pairs (4454) have NULL-links for the outer entity only; 707 pairs have both entities with NULL-links.

To analyze the NULL-links, we collected statistics on linkages between NEREL and Wikidata for the most frequent pairs of nested entities (Table 7). One can expect to have up to 60% of NULL-links for outer (longer) nested entities. Inner entities inside nested pairs rarely have NULL-links. Geopolitical entities are mainly presented in Wikidata; they appear inside longer 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 (with at least one relation). The remaining 39% of pairs are not connected, which shows a potential of nested entities extraction and linking. The most frequent relation types of connections that present in Wikidata are: COUNTRY (P17), APPLIES TO JURISDICTION (P1001), SUBCLASS OF (P279), PART OF (P361), NAMED AFTER (P138), LOCATED IN THE ADMINISTRATIVE TERRITORIAL ENTITY (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 relations between an outer and inner entities.

In many cases, the absence of a relation between longer and internal entities in Wikidata can be due to insufficient descriptions of Wikidata items. For example, *Mariinsky Theatre Concert hall* item (Q4231897) is not linked to the *Mariinsky Theatre* item (Q207028). Two nested entities from the 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*).

5 Baseline methods

This section presents the evaluation results of off-the-shelf named entity recognition tools and results of baselines for nested named entity recognition, relation extraction, and entity linking trained on NEREL. For experiments, we divided the developed dataset into the train, dev, and test sets—746/94/93 texts, respectively.

5.1 Evaluation of off-the-shelf NER tools

We test three commonly used off-the-shelf NER tools, which support Russian, on the NEREL test set. Natasha[9] is a distilled version of a BERT-based NER model, fine-tuned on data with synthetic annotations. SpaCy[10] and Stanza[11] are equipped with LSTM-based NER models. These tools are distributed as Python standalone packages and can be installed and used without training a model. Note, the scope of such evaluation is limited: the tools support only three entity types (LOCATION, ORGANIZATION, PERSON), while NEREL's entity types are more fine-grained and diverse. We keep two NEREL's entity types (PERSON and ORGANIZATION) intact. We considered, that the tool recognizes the entity correctly, if it assigns LOCATION to entities, annotated in NEREL with location-related types (COUNTRY, CITY, STATE_OR_PROVINCE, FACILITY, LOCATION, DISTRICT). In total, we are able to test the performance of off-the-shelf models with respect to eight entity types. What is more, the tools are aimed at *flat* NER, while NEREL is annotated with nested named entities. The tools disregard the nestedness and extract entities that either contain another entity or are nested in a longer entity. In this evaluation, we do not take the nested annotations into account.

Table 8 shows the results of Natasha, spaCy, and Stanza models tested on NEREL. Natasha outperforms Stanza and spaCy on four out of eight entity types, including most frequent PERSON and ORGANIZATION. The evaluation highlights how challenging NEREL's annotations are. At the same time, the limitations of wide-used NER tools become more evident and emphasize the need for large-scale NER datasets on which the models can be trained.

5.2 Baselines for nested named entity recognition

We used three task-specific state-of-the-art models for nested named entity recognition (NER):

Biaffine model Biaffine model (Yu et al., 2020) scores pairs of start and end tokens to form a named entity.

Pyramid model Pyramid model (Jue et al., 2020) consists of a stack of interconnected layers. Each layer l predicts whether a l-gram is a complete entity mention.

[9] https://github.com/natasha/natasha.

[10] https://spacy.io.

[11] https://stanfordnlp.github.io/stanza/.

Table 8 Performance of NER tools on the NEREL test set, estimated with entity-level precision (P), recall (R) and F1 measure

Type	#	Natasha			spaCy			Stanza		
		P	R	F1	P	R	F1	P	R	F1
PERSON	960	0.91	0.85	0.88	0.89	0.86	0.88	0.93	0.89	**0.91**
ORGANIZATION	673	0.81	0.57	**0.67**	0.78	0.57	0.66	0.71	0.51	0.59
COUNTRY	458	1.00	0.69	**0.82**	1.00	0.70	**0.82**	0.95	0.54	0.69
CITY	240	0.98	0.81	**0.89**	0.98	0.81	0.88	0.99	0.77	0.87
STATE_OR_PROVINCE	112	1.00	0.88	**0.94**	0.98	0.85	0.91	1.00	0.73	0.85
FACILITY	63	0.50	0.26	0.34	0.46	0.22	0.30	0.86	0.53	**0.66**
LOCATION	61	0.68	0.44	0.53	0.55	0.48	0.51	0.59	0.55	**0.57**
DISTRICT	25	1.00	0.80	0.89	0.95	0.83	0.89	1.00	0.84	**0.91**

The best results are highlighted in bold

Second-best Sequence Learning coupled with Decoding (Second Best) model. The model (Shibuya & Hovy, 2020) uses the Conditional Random Field output layer. The model treats the tag sequence for nested entities as the second-best path within the span of their parent entity. In addition, the decoding method for inference extracts entities iteratively from outermost ones to inner ones in an outside-to-inside way.

These models utilize the RuBERT-cased encoder (Kuratov & Arkhipov, 2019) and fastText (fT) embeddings (Mikolov et al., 2018).

Besides, we experimented with the *SpERT* architecture (Eberts & Ulges, 2020) for the joint extraction of entities and relations. In this approach, any token sequence is a candidate for an entity, and any pair of spans can be involved in a relation. This model performs a full search of these hypotheses. We adopted this model with default parameters and used a RuBERT-cased encoder.

Finally, we explored a recently established trend to apply *Machine Reading Comprehension (MRC)* to nested NER (Li et al., 2020). The MRC model treats the NER task as extracting answers to specific questions; each entity type is associated with a specific question. The dataset is converted into triples (question Q, answer A, context C). In our case, the questions are definitions of entity types, carefully selected from multiple dictionaries. The answer is the annotated named entity, and the context is the given sentence. The MRC model is constructed over the BERT model, which obtains the following string as input:

$$\{[CLS], q_1, q_2 ..., q_m, [SEP], t_1, t_2, ...t_n\},$$

where q_i are words in the question, t_i are words in the sentence. The MRC model should extract a continuous span A in the context C:

$$A = \{t_i ..., t_{i+k}, 1 \leq i \leq i + k \leq n\}.$$

Table 9 Performance of NER methods on the NEREL test set

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**

The best results are highlighted in bold

Examples of dictionary definitions are as follows (translated from Russian): "*Age* is the period when someone was alive, or something exists"; "A *city* is a place where many people live, with many houses, shops, businesses."

Table 9 presents the results of nested NER on the NEREL dataset. The results show that (i) contextualized BERT-based models outperform models based on static word representations; (ii) the Biaffine model surpasses the Pyramid model; (iii) the span-based approach of the SpeRT model can identify overlapping or nested entities better than Biaffine and Pyramid models; (iv) the results of MRC approach surpass nested NER models' results, most likely, due to the effective usage of additional external information. However, as the MRC approach needs to loop over each question for each entity candidate and is resource-greedy, the second-best solution is still worth consideration, which is the CRF-based Second Best model in our case.

Table 10 shows the results of the MRC model for all types of named entities in NEREL. The highlighted values can be compared with the results of the off-the-shelf tools (Table 8). The MRC results are much higher for all entity types: PERSON, ORGANIZATION, and general LOCATION integrating geopolitical entities, proper locations and facilities.

5.3 Baselines for relation extraction

Some recent relation extraction models (Alt et al., 2019; Han et al., 2019; Joshi et al., 2020) do not support relations between nested named entities or cross-entity relations. These models are tailored to the common test-beds, such as TACRED and DocRED, which do not possess nested named entities, unlike NEREL.

We adopt the Open NRE classifier (Han et al., 2019) for the task. In our experiments, the pre-trained RuBERT model is used as an encoder.[12] We compared two strategies for pooling: `[CLS]` pooling and entity pooling.

[12] https://huggingface.co/DeepPavlov/rubert-base-cased.

Table 10 Performance of the MRC model for all entity types of NEREL

NE type	P	R	F1
PERSON	0.96	0.98	0.97
COUNTRY	0.96	0.95	0.96
AGE	0.95	0.92	0.94
RELIGION	0.88	0.96	0.92
CITY	0.92	0.90	0.91
DATA	0.91	0.89	0.90
STATE_OR_PROVINCE	0.86	0.90	0.88
NUMBER	0.91	0.83	0.87
ORDINAL	0.89	0.81	0.85
ORGANIZATION	0.85	0.84	0.84
PROFESSION	0.81	0.86	0.84
LAW	0.81	0.85	0.83
WORK_OF_ART	0.93	0.74	0.82
MONEY	0.80	0.81	0.80
PRODUCT	0.75	0.83	0.79
NATIONALITY	0.68	0.94	0.79
TIME	0.77	0.80	0.78
AWARD	0.73	0.80	0.77
DISEASE	0.82	0.71	0.76
IDEOLOGY	0.90	0.61	0.73
FACILITY	0.66	0.74	0.70
DISTRICT	0.68	0.68	0.68
EVENT	0.67	0.63	0.65
PENALTY	0.57	0.76	0.65
LOCATION	0.72	0.54	0.62
LANGUAGE	0.54	0.70	0.61
CRIME	0.50	0.76	0.60
FAMILY	0.67	0.46	0.55
PERCENT	0.50	0.57	0.53

The encoding schema used in OpenNRE entity pooling introduces additional functional tokens for each entity. In the case of relations located inside nested entities, this encoding schema makes the shortest entity occur twice. The input turns into an implausible sentence, affecting the classifier performance. For example, sentence "Moscow State University was found in 1755" with a relation between entities *Moscow State University* and *Moscow* is converted into the format:

"$< E2_{start} >$ `Moscow` $< E2_{end} >$ $< E1_{start} >$ `Moscow State University` $< E1_{end} >$ `was found in 1755.`"

This way, the internal entity `Moscow` is duplicated, which can lead to degradation of the relation extraction results.

We trained the model for 20 epochs with a batch size equal to 64; other hyper-parameters were set to default. Results are presented in Table 11. Table 12 details the best performing entity pooling approach in regard to relation type. Micro- and macro-averaging of F1 provides an aggregated score. For brevity, we present results for the 25 most frequent relations. Aggregated scores are computed for all relations in NEREL.

The results show that while in-sentence relations are easier to detect, cross-sentence relations leave more room for improvement. RuBERT, used as a backbone model, is not tailored to handle long input texts. To this end, contextualized encoders for longer sequences may turn out useful.

Additionally, we trained OpenNRE on in-sentence relations only and obtained much higher results for in-sentence relations and relations between nested entities, which is due to more homogeneous training data (see Table 12).

Finally, OpenNRE performance does not degrade for infrequent relations. For example, the relatively rare relation FOUNDED_BY is detected well in three cases.

When applied to sentence-level relations, the SpanBERT model, equipped with mBERT or RuBERT encoders, obtains close results to in-sentence OpenNRE (Table 12). However, the SpanBERT input format does not allow nested relations. Last but not least, in Loukachevitch et al. (2021) overestimated results for the OpenNRE model were reported due to a performance scoring issue. The current results represent an accurate corrected scores.

5.4 Baselines for entity linking

We evaluate two entity linking baselines: SapBERT (Liu et al., 2021) and mGENRE (Cao et al., 2021).

mGENRE is a sequence-to-sequence multilingual entity linking model. Given an entity mention, mGENRE predicts the entity name 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 up to 75 tokens from left and right contexts of each mention as follows:

```
left context [START] named entity mention [END] right context.
```

In experiments, we use the default mGENRE model without any fine-tuning on NEREL data. mGENRE returns a list of candidate entities along with a certainty score for each mention. Although it does not directly predict NULL links, the certainty scores are used to determine them. When the certainty score for the best candidate is below a threshold θ, then the mention is linked to the NULL entity. In general, setting a universal threshold is difficult, because it is usually domain specific. Therefore, we suggest using the NEREL training set for tuning it. The optimal value for θ obtained on the training set is 0.456.

Table 11 Overall performance of Open NRE with CLS and entity pooling for the 25 most frequent relation types of NEREL

Relation	#	CLS pooling			Entity pooling		
		P	R	F1	P	R	F1
WORKPLACE	439	0.74	0.81	0.77	0.75	0.81	0.78
ALTERNATIVE_NAME	416	0.62	0.56	0.59	0.48	0.65	0.55
PARTICIPANT_IN	416	0.73	0.69	0.71	0.75	0.69	0.72
WORKS_AS	416	0.82	0.86	0.84	0.77	0.87	0.81
TAKES_PLACE_IN	218	0.86	0.88	0.87	0.85	0.89	0.87
POINT_IN_TIME	213	0.82	0.90	0.86	0.88	0.90	0.89
ORIGINS_FROM	208	0.72	0.75	0.74	0.79	0.74	0.77
HEADQUARTERED_IN	194	0.84	0.87	0.85	0.87	0.86	0.86
LOCATED_IN	182	0.80	0.75	0.77	0.76	0.80	0.78
AGENT	116	0.77	0.52	0.62	0.68	0.63	0.65
AGE_IS	99	0.82	0.89	0.85	0.84	0.87	0.86
PRODUCES	87	0.50	0.78	0.61	0.69	0.82	0.75
AWARDED_WITH	81	0.63	0.64	0.64	0.74	0.64	0.69
HAS_CAUSE	71	0.55	0.79	0.65	0.71	0.72	0.71
SUBEVENT_OF	63	0.62	0.32	0.42	0.61	0.48	0.54
PART_OF	63	0.52	0.54	0.53	0.67	0.54	0.60
PLACE_RESIDES_IN	54	0.55	0.31	0.40	0.61	0.41	0.49
INANIMATE_INVOLVED	43	0.63	0.67	0.65	0.79	0.72	0.76
KNOWS	42	0.47	0.36	0.41	0.60	0.50	0.55
FOUNDED_BY	38	0.96	0.68	0.80	0.86	0.79	0.82
MEMBER_OF	36	0.50	0.22	0.31	0.38	0.47	0.42
IDEOLOGY_OF	35	0.67	0.74	0.70	0.76	0.80	0.78
ORGANIZES	33	0.00	0.00	0.00	0.62	0.30	0.41
MEDICAL_CONDITION	33	0.56	0.58	0.57	0.58	0.58	0.58
SUBORDINATE_OF	31	0.56	0.61	0.58	0.64	0.68	0.66
Micro-F1 (all relations)		0.51			0.65		
Macro-F1 (all relations)		0.48			0.63		

The second baseline, SapBERT is a BERT-based model that implements self-alignment of the representation space during pre-training. It produces embeddings of mentions and candidates without taking into account the context. We use one of two publicly available cross-Lingual models (SapBERT-XLMR) and fine-tune it with 'wikititles+muse' pairs of labels as suggested by the SapBERT authors.[13] The length of the input text is restricted to 25 tokens. This limit is spent to encode only a mention or an entity candidate from Wikidata; no tokens from the context are used. After getting embeddings of a mention and candidates, we use cosine similarity

[13] The list of 'wikititles+muse' pairs can be found here: https://github.com/cambridgeltl/sapbert/tree/main/training_data.

Table 12 Performance of OpenNRE for the 25 most frequent relation types in NEREL

Relation	#	F1	#	F1	#	F1
	Between sentences		Inside sentences		Inside entities	
WORKS_AS	66	0.68	348	0.92	2	0.00
PARTICIPANT_IN	89	0.60	305	0.78	22	0.60
WORKPLACE	59	0.61	206	0.80	174	0.87
POINT_IN_TIME	15	0.73	186	0.93	12	0.80
TAKES_PLACE_IN	28	0.74	184	0.92	6	0.80
ALTERNATIVE_NAME	260	0.69	151	0.74	5	0.33
ORIGINS_FROM	18	0.77	136	0.78	54	0.76
LOCATED_IN	38	0.63	135	0.82	9	0.71
AGENT	14	0.00	101	0.65	1	0.00
HAS_CAUSE	18	0.50	52	0.81	1	0.00
PRODUCES	15	0.37	48	0.83	24	0.84
PLACE_RESIDES_IN	10	0.50	42	0.52	2	0.00
HEADQUARTERED_IN	10	0.67	40	0.65	144	0.92
SUBEVENT_OF	28	0.61	32	0.68	3	0.50
INANIMATE_INVOLVED	14	0.78	28	0.90	1	0.00
FOUNDED_BY	8	0.77	25	0.96	5	0.62
IDEOLOGY_OF	2	0.00	23	0.68	10	0.89
ORGANIZES	9	0.20	21	0.50	3	0.50
MEMBER_OF	9	0.17	20	0.55	7	0.57
ABBREVIATION	4	0.13	17	0.40	3	0.00
SUBORDINATE_OF	1	0.00	13	0.46	17	0.83
DATE_OF_CREATION	4	0.33	12	0.71	3	1.00
OWNER_OF	3	0.00	10	0.64	8	0.78
START_TIME	2	0.00	7	0.40	1	0.00
PART_OF	23	0.39	5	0.43	35	0.76
Micro-F1 OpenNRE	0.46		0.71		0.55	
Macro-F1 OpenNRE	0.43		0.70		0.52	
Micro-F1 OpenNRE (in-sent.)	–		**0.80**		**0.77**	
Micro-F1 SpanBERT (mBERT)	–		0.76		–	
Micro-F1 SpanBERT (RuBERT)	–		0.78		–	

F1-scores are provided for all relations in NEREL

The best results are highlighted in bold

between these vectors to find top-k candidates for each mention. If such similarity is below the threshold, then a NULL-link is assigned for the mention.

Table 13 contains the evaluation results. Overall moderate results of the baselines compared to the results reported in recent works for other languages can be explained by the fact that we do not fine-tune the models on the training set. The best performance (accuracy = 0.71) is achieved by the mGENRE model, but top-5 results of the SapBERT model can be considered as competitive. The mGENRE works well on

Table 13 Entity linking accuracy of the pre-trained baselines: mGENRE and SapBERT

Entity type	SapBERT top-1	SapBERT top-5	mGENRE
AWARD	0.504	0.672	0.629
CITY	0.211	0.633	0.841
COUNTRY	0.226	0.516	0.935
DISEASE	0.375	0.464	0.750
DISTRICT	0.520	0.720	0.667
FACILITY	0.641	0.703	0.800
LANGUAGE	0.375	0.500	1.000
LAW	0.532	0.613	0.607
LOCATION	0.279	0.607	0.818
NATIONALITY	0.190	0.333	0.273
ORGANIZATION	0.495	0.639	0.738
PERSON	0.439	0.502	0.671
PRODUCT	0.585	0.811	0.900
PROFESSION	0.559	0.760	0.311
RELIGION	0.417	0.625	0.609
STATE_OR_PROVINCE	0.366	0.795	0.947
WORK_OF_ART	0.516	0.710	0.609
Macro-accuracy	0.425	0.624	0.712
Micro-accuracy	0.437	0.619	0.644

NULL labels are obtained using thresholds fitted on the training set

relatively simple mentions such as COUNTRY, STATE_OR_PROVINCE, PRODUCT, LANGUAGE, and CITY; it has low performance on PROFESSION and NATIONALITY entity types.

6 Conclusion

We presented a new Russian dataset NEREL with three levels of annotation: nested named entity, relations, and links to Wikidata. The dataset is significantly larger than existing Russian datasets. The NEREL dataset has several significant distinctive features, including nested named entities, relations over nested named entities, relations on both sentence and document levels, and events involving named entities. Nevertheless, NEREL annotations utilize conventional entity and relations types, enabling cross-lingual transfer experiments. NEREL will facilitate the construction of derived resources, e.g. for targeted sentiment analysis or discourse analysis because new resources can exploit diverse annotation in NEREL.

NEREL can facilitate the development of novel models that address extraction of relations between nested named entities and cross-sentence relation extraction from texts. NEREL annotation also allows relation extraction experiments on both sentence-level and document-level. NEREL will facilitate the development of two-step (entity and relation extraction) or three-step joint models and also the research on the knowledge graph enrichment applications.

Acknowledgement The work is supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the Russian Academy of Sciences dated November 2, 2021 No. 70-2021-00142.

References

Alt, C., Hübner, M., & Hennig, L. (2019). Improving relation extraction by pre-trained language representations. Preprint at http://arxiv.org/abs/1906.03088

Benikova, D., Biemann, C., & Reznicek, M. (2014). Nosta-d named entity annotation for German: Guidelines and dataset. In *LREC* (pp. 2524–2531).

Bentivogli, L., Forner, P., Giuliano, C., Marchetti, A., Pianta, E., & Tymoshenko, K. (2010). Extending English ace 2005 corpus annotation with ground-truth links to Wikipedia. In *Proceedings of the 2nd workshop on the people's web meets NLP: Collaboratively constructed semantic resources* (pp. 19–27).

Bies, A., Song, Z., Getman, J., Ellis, J., Mott, J., Strassel, S., Palmer, M., Mitamura, T., Freedman, M., Ji, H., & O'gorman, T. (2016). A comparison of event representations in deft. In *Proceedings of the fourth workshop on events* (pp. 27–36).

Bollacker, K., Cook, R., & Tufts, P. (2007). Freebase: A shared database of structured general human knowledge. In *AAAI* (vol. 7, pp. 1962–1963).

Botha, J. A., Shan, Z., & Gillick, D. (2020). Entity linking in 100 languages. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 7833–7845).

Brandsen, A., Verberne, S., Lambers, K., Wansleeben, M., Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., & Goggi, G. (2020). Creating a dataset for named entity recognition in the archaeology domain. In *Conference proceedings LREC 2020* (pp. 4573–4577). The European Language Resources Association.

Bruches, E., Mezentseva, A., & Batura, T. (2021). A system for information extraction from scientific texts in Russian. Preprint at http://arxiv.org/abs/2109.06703

Campillos-Llanos, L., Valverde-Mateos, A., Capllonch-Carrión, A., & Moreno-Sandoval, A. (2021). A clinical trials corpus annotated with umls entities to enhance the access to evidence-based medicine. *BMC Medical Informatics and Decision Making, 21*(1), 1–19.

Cao, N. D., Wu, L., Popat, K., Artetxe, M., Goyal, N., Plekhanov, M., Zettlemoyer, L., Cancedda, N., Riedel, S., & Petroni, F. (2021). Multilingual autoregressive entity linking. Preprint at https://arxiv.org/abs/2103.12528

Checco, A., Roitero, K., Maddalena, E., Mizzaro, S., & Demartini, G. (2017). Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Fifth AAAI conference on human computation and crowdsourcing* (pp. 11–20).

Eberts, M., & Ulges, A. (2020). Span-based joint entity and relation extraction with transformer pre-training. In *ECAI*.

Ellis, J. (2012). *TAC KBP entity selection. KBP 2012 guidelines. Version 1.1. Linguistic Data Consortium (2012)*. Retrieved from https://tac.nist.gov/2012/KBP/task_guidelines/TAC_KBP_Entity_Selection_V1.1.pdf

Ellis, J., Getman, J., Fore, D., Kuster, N., Song, Z., Bies, A., & Strassel, S. M. (2015). Overview of linguistic resources for the tac kbp 2015 evaluations: Methodologies and results. In *TAC*. Retrieved from https://tac.nist.gov/publications/2015/additional.papers/TAC2015.KBP_resources_overview.proceedings.pdf

Ellis, J., Getman, J., & Strassel, S. M. (2014). Overview of linguistic resources for the tac kbp 2014 evaluations: Planning, execution, and results. In *Proceedings of TAC KBP 2014 Workshop, National Institute of Standards and Technology* (pp. 17–18). Retrieved from https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/tackbp-2014-overview.pdf

Finkel, J. R., Grenager, T., & Manning, C. D. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL'05)* (pp. 363–370).

Gabrilovich, E., Ringgaard, M., & Subramanya, A. (2013). *FACC1: Freebase annotation of ClueWeb corpora, version 1 (release date 2013-06-26, format version 1, correction level 0)*. Retrieved from http://lemurproject.org/clueweb09/

Gareev, R., Tkachenko, M., Solovyev, V., Simanovsky, A., & Ivanov, V. (2013). Introducing baselines for Russian named entity recognition. In *International conference on intelligent text processing and computational linguistics* (pp. 329–342).

Getman, J., Ellis, J., Song, Z., Tracey, J., & Strassel, S. M. (2017). Overview of linguistic resources for the tac kbp 2017 evaluations: Methodologies and results. In *TAC*.

Gordeev, D., Davletov, A., Rey, A., Akzhigitova, G., & 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"].*

Guo, Z., & Barbosa, D. (2018). Robust named entity disambiguation with random walks. *Semantic Web, 9*(4), 459–479.

Hamdi, A., Linhares Pontes, E., Boros, E., Nguyen, T. T. H., Hackl, G., Moreno, J. G., & 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* (pp. 2328–2334).

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., & Sun, M. (2019). Opennre: An open and extensible toolkit for neural relation extraction. *EMNLP-IJCNLP, 2019*, 169.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., Pennacchiotti, M., Romano, L., & Szpakowicz, S. (2010). SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th international workshop on semantic evaluation* (pp. 33–38). Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/S10-1006

Hoffart, J., Suchanek, F. M., Berberich, K., & Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence, 194*, 28–61. https://doi.org/10.1016/j.artint.2012.06.001

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., & Weikum, G. (2011). Robust disambiguation of named entities in text. In *Proceedings of the conference on empirical methods in natural language processing, EMNLP '11* (pp. 782–792). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145521

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of the human language technology conference of the NAACL, companion volume: Short papers, NAACL-Short '06* (pp. 57–60). Association for Computational Linguistics.

Ivanin, V., Artemova, E., Batura, T., Ivanov, V., Sarkisyan, V., Tutubalina, E., & 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"].*

Ji, H., Nothman, J., Hachey, B., & Florian, R. (2015). Overview of tac-kbp2015 tri-lingual entity discovery and linking. In *TAC*.

Jiang, H., Hua, Y., Beeferman, D., & Roy, D. (2022). Annotating the tweebank corpus on named entity recognition and building nlp models for social media analysis. In *Conference proceedings LREC 2022* (pp. 7199–7208).

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics, 8*, 64–77.

Jue, W., Shou, L., Chen, K., & Chen, G. (2020). Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5918–5928).

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

Krippendorff, K. (2004). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research, 30*(3), 411–433.

Kuratov, Y., & Arkhipov, M. (2019). Adaptation of deep bidirectional multilingual transformers for Russian language. Preprint at http://arxiv.org/abs/1905.07213

Kuznetsov, A., Braslavski, P., & Ivanov, V. (2016). Family matters: Company relations extraction from wikipedia. In *International conference on knowledge engineering and the semantic web* (pp. 81–92). Springer.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020). A unified MRC framework for named entity recognition. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 5849–5859).

Ling, X., Singh, S., & 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., & Collier, N. (2021). Learning domain-specialised representations for cross-lingual biomedical entity linking. In *Proceedings of ACL-IJCNLP 2021* (pp. 565–574).

Loukachevitch, N., Artemova, E., Batura, T., Braslavski, P., Denisov, I., Ivanov, V., Manandhar, S., Pugachev, A., & 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)* (pp. 876–885).

Luan, Y., He, L., Ostendorf, M., & 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* (pp. 3219–3232).

Malmasi, S., Fang, A., Fetahu, B., Kar, S., & Rokhlenko, O. (2022). MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th international conference on computational linguistics* (pp. 3798–3809).

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

Mikolov, T., Grave, É., Bojanowski, P., Puhrsch, C., & Joulin, A. (2018). Advances in pre-training distributed word representations. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Milne, D., & Witten, I. H. (2008). Learning to link with Wikipedia. In *Proceedings of the 17th ACM conference on information and knowledge management, CIKM '08* (pp. 509–518). ACM. https://doi.org/10.1145/1458082.1458150

Mitamura, T., Liu, Z., & Hovy, E. H. (2015a). Overview of tac kbp 2015 event nugget track. In *TAC*.

Mitamura, T., Yamakawa, Y., Holm, S., Song, Z., Bies, A., Kulick, S., & Strassel, S. (2015b). Event nugget annotation: Processes and issues. In *Proceedings of the the 3rd workshop on EVENTS: Definition, detection, coreference, and representation* (pp. 66–76).

Moro, A., & 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)* (pp. 288–297).

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

Nasar, Z., Jaffry, S. W., & Malik, M. K. (2021). Named entity recognition and relation extraction: State-of-the-art. *ACM Computing Surveys, 54*(1), 1–39.

Piskorski, J., Babych, B., Kancheva, Z., Kanishcheva, O., Lebedeva, M., Marcinczuk, M., Nakov, P., Osenova, P., Pivovarova, L., Pollak, S., Radev, I., Robnik-Sikonja, M., Starko, V., Steinberger, J., & Yangarber, R. (2021). Slav-ner: The 3rd cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages. In *Proceedings of the 8th workshop on Balto-Slavic natural language processing* (pp. 122–133).

Piskorski, J., Laskova, L., Marcińczuk, M., Pivovarova, L., Přibáň, P., Steinberger, J., & Yangarber, R. (2019). The second cross-lingual challenge on recognition, normalization, classification, and linking of named entities across slavic languages. In *Proceedings of the 7th workshop on Balto-Slavic natural language processing* (pp. 63–74).

Plank, B., Jensen, K. N., & van der Goot, R. (2020). Dan+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th international conference on computational linguistics* (pp. 6649–6662).

Ratinov, L., Roth, D., Downey, D., & 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—Volume 1, HLT '11* (pp. 1375–1384). Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002642

Ringland, N. (2015). Structured named entities. Doctor of philosophy Ph.D., School of Information Technologies. Retrieved from http://hdl.handle.net/2123/14558

Ringland, N., Dai, X., Hachey, B., Karimi, S., Paris, C., & 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* (pp. 5176–5181).

Röder, M., Usbeck, R., & Ngomo, A. N. (2018). GERBIL—Benchmarking named entity recognition and linking consistently. *Semantic Web, 9*(5), 605–625. https://doi.org/10.3233/SW-170286

Rosales-Méndez, H., Hogan, A., & Poblete, B. (2018). Voxel: A benchmark dataset for multilingual entity linking. In *International semantic web conference* (pp. 170–186). Springer.

Ruokolainen, T., Kauppinen, P., Silfverberg, M., & Lindén, K. (2019). A finnish news corpus for named entity recognition. In *Language resources and evaluation* (pp. 1–26).

Sevgili, O., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C. (2020). Neural entity linking: A survey of models based on deep learning. Preprint at http://arxiv.org/abs/2006.00575

Shabankhani, B., Charati, J. Y., Shabankhani, K., & Cherati, S. K. (2020). Survey of agreement between raters for nominal data using Krippendorff's alpha. *Archives of Pharmacy Practice, 10*, 160–164.

Shavrina, T., Fenogenova, A., Anton, E., Shevelev, D., Artemova, E., Malykh, V., Mikhailov, V., Tikhonova, M., Chertok, A., & Evlampiev, A. (2020). Russiansuperglue: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 4717–4726).

Shibuya, T., & Hovy, E. (2020). Nested named entity recognition via second-best sequence learning and decoding. *Transactions of the Association for Computational Linguistics, 8*, 605–620.

Song, Z., Bies, A., Strassel, S., Riese, T., Mott, J., Ellis, J., Wright, J., Kulick, S., Ryant, N., & Ma, X. (2015). From light to rich ere: Annotation of entities, relations, and events. In *Proceedings of the the 3rd workshop on EVENTS: Definition, detection, coreference, and representation* (pp. 89–98).

Starostin, A., Bocharov, V., Alexeeva, S., Bodrova, A., Chuchunkov, A., Dzhumaev, S., Efimenko, I., Granovsky, D., Khoroshevsky, V., Krylova, I., Nikolaeva, M., Smurov, I., & 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"]* (pp. 702–720).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012). Brat: a web-based tool for nlp-assisted text annotation. In *Proceedings of the demonstrations at the 13th conference of the European chapter of the Association for Computational Linguistics* (pp. 102–107).

Sysoev, A., & Nikishina, I. (2018). Smart context generation for disambiguation to Wikipedia. In *Conference on artificial intelligence and natural language* (pp. 11–22).

Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2017). Enriching news events with meta-knowledge information. *Language Resources and Evaluation, 51*, 409–438.

Tjong Kim Sang, E.F., & 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* (pp. 142–147). Association for Computational Linguistics.

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

Vlasova, N., Lando, N., Suleymanova, E., & Trofimov, I. (2016). Situations-1000: A tagged corpus for event extraction from texts. In *Computational linguistics and intellectual technologies: Proceedings of the international conference "dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]* (pp. 202–220).

Vlasova, N., Suleymanova, E., & Trofimov, I. (2014). Report on russian corpus for personal name retrieval. In *Proceedings of TEL'2014 conference on computational and cognitive linguistics* (pp. 36–40).

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

Walker, C., Strassel, S., Medero, J., & Maeda, K. (2006). *ACE 2005 multilingual training corpus. LDC2006T06*. Linguistic Data Consortium.

Wang, H. C., Chen, C. C., & Li, T. W. (2022). Automatic content curation of news events. *Multimedia Tools and Applications, 81*(8), 10445–10467.

Weischedel, R., & Brunstein, A. (2005). *Bbn pronoun coreference and entity type corpus*. Philadelphia: Linguistic Data Consortium.

Yao, Y., Ye, D., Li, P., Han, X., Lin, Y., Liu, Z., Liu, Z., Huang, L., Zhou, J., & Sun, M. (2019). Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 764–777).

Yu, J., Bohnet, B., & Poesio, M. (2020). Named entity recognition as dependency parsing. In *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 6470–6476).

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

Zhang, Y., Zhong, V., Chen, D., Angeli, G., & Manning, C. D. (2017). Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP 2017)* (pp. 35–45).

Zhou, S., Rijhwani, S., & Neubig, G. (2019). Towards zero-resource cross-lingual entity linking. In *Proceedings of the 2nd workshop on deep learning approaches for low-resource NLP (DeepLo 2019)* (pp. 243–252). Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-6127

Authors and Affiliations

Natalia Loukachevitch[1,2] · Ekaterina Artemova[3] · Tatiana Batura[4,5] · Pavel Braslavski[3,6] · Vladimir Ivanov[7] · Suresh Manandhar[8] · Alexander Pugachev[3] · Igor Rozhkov[1] · Artem Shelmanov[1,9,10] · Elena Tutubalina[3,9,11] · Alexey Yandutov[1]

✉ Natalia Loukachevitch
louk_nat@mail.ru

1 Lomonosov Moscow State University, Moscow, Russia

2 ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia

3 HSE University, Moscow, Russia

4 Novosibirsk State University, Novosibirsk, Russia

5 A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

6 Ural Federal University, Yekaterinburg, Russia

7 Innopolis University, Innopolis, Russia

8 Madan Bhandari University of Science and Technology, Chitlang, Nepal

9 AIRI, Moscow, Russia

10 Mohamed bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE

11 Sber AI, Moscow, Russia