NEREL: A Russian Dataset with Nested Named Entities, Relations and Events

Natalia Loukachevitch[1], Ekaterina Artemova[2,3], Tatiana Batura[1,4],
Pavel Braslavski[2,5], Ilia Denisov[1], Vladimir Ivanov[6],
Suresh Manandhar[9], Alexander Pugachev[2], and Elena Tutubalina[2,7,8]

[1]Lomonosov Moscow State University, Russia
[2]HSE University, Russia
[3]Huawei Noah's Ark lab, Russia
[4]Novosibirsk State University, Russia
[5]Ural Federal University, Russia
[6]Innopolis University, Russia
[7]Kazan Federal University, Russia
[8]Sber AI, Russia
[9]Wiseyak, United States

Abstract

In this paper, we present NEREL, a Russian dataset for named entity recognition and relation extraction. NEREL is significantly larger than existing Russian datasets: to date it contains 56K annotated named entities and 39K annotated relations. Its important difference from previous datasets is annotation of nested named entities, as well as relations within nested entities and at the discourse level. NEREL can facilitate development of novel models that can extract relations between nested named entities, as well as relations on both sentence and document levels. NEREL also contains the annotation of events involving named entities and their roles in the events. The NEREL collection is available via https://github.com/nerel-ds/NEREL.

1 Introduction

Knowledge bases (KBs) encompass a large amount of structured information about real-world entities and their relationships, which is useful in many tasks: information retrieval, automatic text summarization, question answering, conversational and recommender systems (Liu et al., 2020; Han et al., 2020; Huang et al., 2020). Even the largest knowledge bases are inherently incomplete, but their manual development is time-consuming and expensive. Automatic population of knowledge bases from large text collections is usually broken down into named entity (NE) recognition, relation extraction (RE), and linking entities to a knowledge base. In turn, training and evaluating models addressing these problems require large and high-quality annotated resources. Currently, most of the available

resources of this kind are in English.

In this paper, we present NEREL (**N**amed **E**ntities and **REL**ations), a new Russian dataset with annotated named entities and relations. In developing the annotation schema, we aimed to accommodate recent advances in information extraction methods and datasets. In particular, nested named entities and relations within named entities are annotated in NEREL. Both of these provide a richer and more complete annotation compared with a flat annotation scheme. Current datasets with nested named entities (Ringland et al., 2019; Benikova et al., 2014) are not annotated with relations. Therefore, most state-of-the-art relation extraction models (Joshi et al., 2020; Alt et al., 2019) do not work with relations between nested and overlapping entities. NEREL aims to address these deficiencies with the addition of nested named entities and relations within nested entities.

Secondly, NEREL relations are annotated across sentence boundaries at the discourse level allowing for more realistic information extraction experiments. Figure 1 illustrates annotation of nested entities, relations between overlapping entities, as well as cross-sentence relations on a sample NEREL sentence.

Finally, NEREL provides annotation for factual events (such as meetings, negotiations, incidents, etc.) involving named entities and their roles in the events. Future versions of the dataset can easily expand the current inventory of entities and relations.

NEREL is the largest dataset for Russian annotated with named entities and relations. NEREL features 29 entity and 49 relation types. At the time of writing the dataset contains 56K entities and
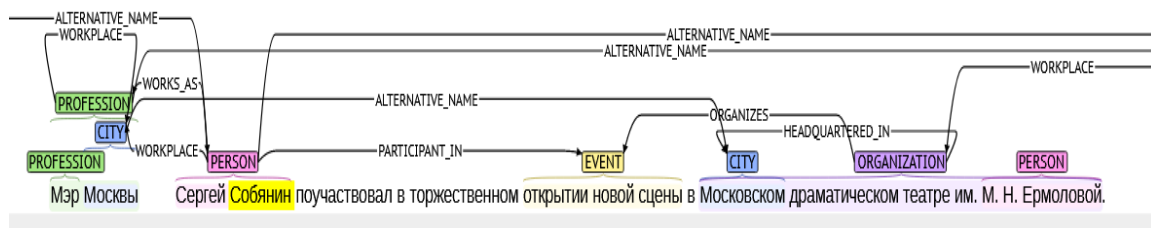
Figure 1: Annotation of the sentence *Moscow Mayor Sergei Sobyanin took part in the grand opening of the new stage of Moscow Ermolova theater* includes nested named entities: *Mayor of Moscow*, *Moscow*, *Mayor*; *Moscow Ermolova Theater*, *Moscow*, *Ermolova*. The intra-entity relations are as follows: *Moscow* is a workplace for *Mayor of Moscow*; *Moscow Ermolova Theater* is headquartered in *Moscow*. *Grand opening of the new stage* is annotated as an event. One can also see ALTERNATIVE_NAME relations linking *Moscow* and *Moscow$_{adj}$* within the sentence and *Moscow Mayor*, *Sergei Sobyanin* with other mentions in neighboring sentences.

39K relations annotated in 900+ Russian Wikinews documents.

In the rest of the paper, we describe the principles behind dataset building process. We also report dataset statistics and provide baseline results for several models. These results indicate that there is a room for improvements. The NEREL collection is freely available.

2 Related Work

Table 1 summarizes most important datasets in the context of NEREL development and provides references to their descriptions.

2.1 Datasets for NER

Most widely used English datasets for named entity recognition in general domain are CoNLL03 and OntoNotes. CoNLL03 is annotated with four basic NE types – persons (PER), organizations (ORG), locations (LOC), and other named entities (MISC), while OntoNotes comprises annotation of 19 NE types, including numeric and temporal ones. Both datasets feature only flat NE annotations.

There are several datasets with annotated nested named entities, see Table 1. NNE is the largest corpus of this kind, both in terms of entity types and annotated NE mentions. NNE provides detailed lexical components such as first and last person's names, units (e.g. *tons*), multipliers (e.g. *billion*), etc. These result in six levels of nestedness in the dataset.

The NoSta-D collection of German Wikipedia articles and online newspapers is annotated with nested named entities of four main classes. Each class can appear in a nominal (proper noun) form, as a part of a token, or as a derivative (adjective) such as "österreichischen" (Austrian). The Digitoday corpus for Finnish is annotated with six types

of named entities (organization, location, person, product, event, and date). It permits nested entities with the restriction that an internal entity cannot be of the same class as its top-level entity. For example, *Microsoft Research* is annotated as a flat entity, without additional annotation of the *Microsoft* entity. Both NoSta-D and Digitoday datasets allow at most two levels of nesting within entities.

Amongst NER datasets in Russian, RURED (Gordeev et al., 2020) provides the largest number of distinct entities with 28 entity types in the RURED dataset of economic news texts. RURED annotation scheme of named entities mainly follows the OntoNotes guidelines with addition of extra named entities (currency, group, family, country, city, etc). Currently, FactRuEval (Starostin et al., 2016) is the only Russian dataset annotated with nested named entities with at most 2 levels of nesting. In FactRuEval, person mentions (PER) can be subdivided into first/last names, patronymics, and nicknames; while organizations and locations – into their description/type and names (e.g. *[[Microsoft]$_{NAME}$ [Corporation]$_{TYPE}$]$_{ORG}$*).

2.2 Datasets for Relation Extraction

One of the largest datasets for relation extraction is the TACRED dataset (Zhang et al., 2017). Relation annotations within TACRED are constructed by querying PER and ORG entities; the returned sentences are annotated by crowd workers (GPE entities are also annotated, the others are treated as values/strings). The dataset consists of 106k sentences with entity mention pairs. Each sentence is labeled with one of 41 person- or organization-oriented relation types, or with a NO_RELATION tag (Table 1 cites the number of "positive" cases). Alt et al. (2020) found that more than 50% of the examples of the TACRED corpus need to be re-

	Dataset	Lang	#NE inst. (Types)	Max Depth	#Rel inst. (Types)
1	CoNLL03 (Tjong Kim Sang and De Meulder, 2003)	en	34.5K (4)	1	–
	Ontonotes (Hovy et al., 2006)	en	104K (19)	1	–
2	ACE2005 (Walker et al., 2006)	en	30K (7)	6	8.3K(6)
	NNE (Ringland et al., 2019)	en	279K (114)	6	–
	No-Sta-D (Benikova et al., 2014)	de	41K (12)	2	–
	Digitoday (Ruokolainen et al., 2019)	fi	19K (6)	2	–
	DAN+ (Plank et al., 2020)	da	6.4K (4)	2	–
3	TACRED (Zhang et al., 2017)	en	(3)	1	22.8K (42)
	DocRED (Yao et al., 2019)	en	132K (6)	1	56K (96)
4	Gareev (Gareev et al., 2013)	ru	44K (2)	1	–
	Collection3 (Mozharova and Loukachevitch, 2016)	ru	26.4K(3)	1	–
	FactRuEval (Starostin et al., 2016)	ru	12K (3)	2	1K (4)
	BSNLP (Piskorski et al., 2019)	ru	9K (5)	1	–
	RuREBUS (Ivanin et al., 2020)	ru	121K (5)	1	14.6K (8)
	RURED (Gordeev et al., 2020)	ru	22.6K (28)	1	5.3K(34)
	NEREL (ours)	ru	56K (29)	6	39K (49)

Table 1: NEREL and its counterparts. 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 presents Russian datasets for information extraction.

labeled to improve the performance of baselines models. RURED (Gordeev et al., 2020) is a Russian language dataset that is similar to TACRED. Several relations for events are added such as the date, place, and participants of an event. The resulting scheme contains 34 relations. The annotation of relations is mainly within sentences.

RuREBus corpus (Ivanin et al., 2020) consists of strategic planning documents issued by the Ministry of Economic Development of the Russian Federation. The data is annotated with eight specialized relations.

DocRED (Yao et al., 2019) is another dataset that is annotated with both named entities and relations. The dataset includes 96 frequent relation types from Wikidata, the relations are annotated at the document level with significant proportion of relations (40.7%) is across sentence boundaries.

FactRuEval (Starostin et al., 2016) is a Russian language dataset that includes about 1,000 annotated document-level relations of four types (OWNERSHIP, OCCUPATION, MEETING, and DEAL).

2.3 Datasets with Annotated Events

Existing NER datasets usually contain annotations of named events such as hurricanes, battles, wars, or sports events (Hovy et al., 2006; Ringland et al., 2019). For knowledge graph population tasks, it is useful to extract information about significant entity-oriented factual events such as funerals, weddings, or concerts (Rospocher et al., 2016). However, such an approach significantly complicates the annotation. In previous specialized event annotation efforts, an event is defined as an "explicit occurrence involving participants" (Song et al., 2015; Bies et al., 2016; Mitamura et al., 2015). Annotators had to tag an event trigger (word or phrase) consisting of the smallest extent of text expressing the occurrence of an event. Mitamura et al. (2015) presented annotation of so-called "event nuggets" that can be discontinuous, for example *found guilty* in the sentence *The court found him guilty*. Additionally, events are annotated with special tags indicating whether or not an event occurred. For example, ACTUAL tag is used when an event actually happened at a particular place and time.

According to ACE and Light ERE guidelines (Linguistic Data Consortium, 2014; Walker et al., 2006), only events of particular types are annotated. The event categories can be: LIFE, BUSINESS, CONFLICT, JUSTICE, and others (Song et al., 2015; Bies et al., 2016; Mitamura et al., 2015). In ACE and ERE datasets (Aguilar et al., 2014) anno-

tated events can be provided with arguments, e.g. CRIME_ARG or SENTENCE_ARG roles for JUSTICE events. There are also universal event attributes, e.g. PLACE and TIME.

TAC-KBP (Aguilar et al., 2014; McNamee et al., 2010) and TACRED (Zhang et al., 2017) annotations contain event-related slots, e.g CHARGES. Such relations can be established between entities, even if the corresponding event is not mentioned explicitly.

3 Dataset Annotation

3.1 Text Selection

The NEREL corpus consists primarily of Russian Wikinews articles.[1] Wikinews publishes news stories under Creative Commons License (CC BY 2.5) allowing reuse of the published materials. An additional advantage of Wikinews as a document source is that a subset of entities mentioned in the news are linked to corresponding Wikipedia pages making it useful for linking of annotated NEs to Wikidata.

To select a subset of Wikinews articles for annotation, we first applied NER trained on RURED (Gordeev et al., 2020) to the whole Wikinews collection. We focused on articles with high density of automatically detected NEs, paying special attention to NEs associated with persons (e.g. PERSON, AGE). Articles about persons are important for further relation extraction and provide opportunity for cross-lingual methods using existing datasets (Walker et al., 2006; Zhang et al., 2017). The extracted articles were inspected manually to balance topics and remove inappropriate documents. Finally, 900+ articles were selected for annotation. At the last step of the selection, we retained texts in the size range 1–5 Kb: very short texts provide little context for annotation, while long documents are usually non-coherent (e.g. lists of movies or events).

3.2 Named Entity Annotation

To define a list of entity types for NEREL, we started with entities in English OntoNotes (Hovy et al., 2006) and RURED (Gordeev et al., 2020) datasets. Additionally, we considered entity types available in Stanford named entity recognizer (Finkel et al., 2005) and TACRED slots such as CRIME and PENALTY. AWARD and DISEASE were added because of their significant frequency in the

gathered collection and importance for personal life.

We followed the following main principles for annotating named entities:

- The entity annotation schema should be easily amenable for further entity linking.
- Annotation of internal entities varies depending on the named entity type. For example, we do not label numbers within numerical entities such as DATE or MONEY, because such annotations are not essential for relation extraction and entity linking.
- We annotate nested named entities and named entities consisting of two disjoint spans, but not intersecting named entities. For example, *deputy chairman* is not annotated within the span *Deputy Chairman of the State Duma Committee* is annotated as *[Deputy [Chairman of the [[State Duma]_{ORG} Committee]_{ORG}] _{PROFESSION}]]_{PROFESSION}*, because it intersects with the longer named entity.
- Adjectives derived from annotated named entities are also annotated with the same tag. Adjectives occur often as internal entities. Figure 1 shows an adjective *moskovskii* (derived from *Moscow*), indicating the theater's location.

Currently, there are 29 entity types in NEREL dataset. Further we describe main groups and specific features of entity annotation.

Basic entity types comprise PERSON, ORGANIZATION, LOCATION, FACILITY, GEOPOLITICAL ENTITIES. The latter are subdivided into COUNTRY, STATE_OR_PROVINCE, CITY, and DISTRICT. We also singled out FAMILY entity to have possibility to describe relations between families and their members.

Temporal and numerical entities include NUMBER, ORDINAL, DATE, TIME, PERCENT, MONEY, AGE. and AGE entity is usually not annotated separately from DATE entities (Hovy et al., 2006; Finkel et al., 2005), but it has their own relations, and therefore it was singled out.

PROFESSION entity denotes jobs, positions in various organizations, and professional titles. This entity type is significant for extracting relationships of specific persons (Zhang et al., 2017; Gordeev et al., 2020; Starostin et al., 2016). Both capitalized and lowercased PROFESSIONS are annotated in NEREL in contrast to other works (Ringland

---

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

et al., 2019; Hovy et al., 2006). PROFESSION entity is one of the most frequent entities in NEREL. It can have a quite complicated nested structure, in particular, longest profession spans include corresponding workplace organization, which allows for a better description of the person's position (Figure 1).

Physical object group of entities includes: WORK_OF_ART, PRODUCT, and AWARD entities. In contrast to OntoNotes, we introduced a special AWARD entity type because the structure and relations of AWARD entities are quite different from WORK_OF_ART entities, and information about awarding is quite frequent in person-oriented texts.

In flat named entity annotations, different guidelines can be used for PRODUCT entities annotation. For example, in OntoNotes (Hovy et al., 2006), manufacturer and product should be annotated separately as ORG+PRODUCT. The same approach is accepted in the Russian Collection3 (Mozharova and Loukachevitch, 2016). In BSNLP-2019 (Piskorski et al., 2019) the manufacturer name should be included into a longer product name. In the NEREL dataset, the PRODUCT entity is annotated as a long span, that can include manufacturer and number subentities.

NORP entities – nationalities, religious, or political groups – are usually capitalized in English but lowercase in Russian. In NEREL, NATIONALITY entity comprises mainly the following expressions: (i) nouns denoting country citizens such as *ukrainec* (*Ukrainian* as a noun); (ii) adjectives corresponding to nations in contexts different from authority-related, for example *russkii pisatel* (Russian writer). The same adjectives are annotated as COUNTRY entity in the context of authorities or the origin of organizations. This decision accounts for most frequent relations in both contexts.

Legal entities (LAW, CRIME, and PENALTY) are significant in person-oriented texts for extraction of relations (Zhang et al., 2017) in the contemporary news flow, however such entities are usually not annotated in named entity datasets. For example, the TACRED dataset contains annotations of the CHARGE relation only. What is more, LAW and CRIME entities can be quite long and specific; they are built from names of organizations, persons, countries, etc. PENALTY entities often contain period of the penalty or monetary values of fine imposed.

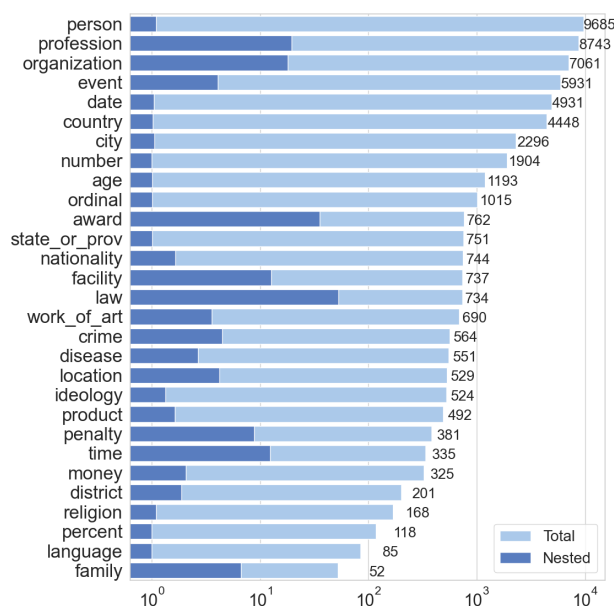Entity type frequencies are presented in Figure 2.

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

As can be seen from the statistics, all but two entity types have at least 100 annotated examples. Manual annotation of named entities and relations was performed by a single annotator, controlled by a moderator. To estimate agreement, 15 documents with about 800 entities were labelled by a moderator (the gold standard) and an annotator. We observed a F_1 measure of 92.95 of the annotator's annotation relative to the gold standard, confirming a high level of agreement. 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 being a kind of pre-professional title).

NEREL annotations do not contain low-level units as for example the NNE dataset (Ringland et al., 2019) featuring e.g. even measurement units as separate entities. Annotation of such units is not challenging because it can be performed using closed word sets. Complex NEREL entities are factual EVENTS similar to event-nuggets (Mitamura et al., 2015) and include PROFESSIONS. These entity types are extremely useful for further relation extraction.

3.3 Events Annotation

As was mentioned, we annotate both named events (traditionally annotated named sports events, ex-

hibitions, hurricanes, battles, wars, revolutions) (Hovy et al., 2006; Ringland et al., 2019) and non-named entity-oriented events significant in the news domain. Annotation of non-named events is most similar to event nugget annotation (Mitamura et al., 2015). Event nuggets can be single words (nouns or verbs) or phrases (noun phrases, verb phrases, or prepositional phrases). As we annotate entities and relation for knowledge graph population, in the current project mainly factual events, which actually happened at a particular place and time, are labeled. Also we annotate future events with exact dates as if for inclusion in a future schedule.

Main types of annotated factual events are as follows: accidents and deaths (*to crash, to attack*); public actions and ceremonies (*to meet, meeting, summit*); legal actions (*to indict, interrogation, to sentence*); transactions (*to buy, to sell*); appointments and resignations; medical events (*hospitalization, surgical operation*); sports events (*match, final*), etc.

We do not restrict subtypes of entities. We define what we exclude. We exclude from event annotation speech acts and cognitive acts, regular activities, changes of numerical indicators (for example, prices or import value), victories and defeats.

3.4 Relation Annotation

Relation types were initially based on TACRED (Zhang et al., 2017) and Russian RURED (Gordeev et al., 2020) corpora. Further, the list of relations has been corrected and expanded from the NEREL corpus analysis; corresponding Wikidata properties were found, when possible.[2] Names for the relations were selected similar to Wikidata property names. The current set of annotated relation types in the NEREL corpus includes 49 relations.

Relations can be subdivided into *person-oriented*, *organization-oriented*, *event-oriented*, and *synonymous* relations (alternative_name, abbreviation). EVENT-oriented relations comprise of role relations, temporal relations, place_of_event relation, causal relations, and others. Figure 3 shows the distribution of relation frequencies in the NEREL dataset. It can be seen that most relations have at least 50 examples.

---

[2]Some relations can 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 calculating the above mentioned age values.
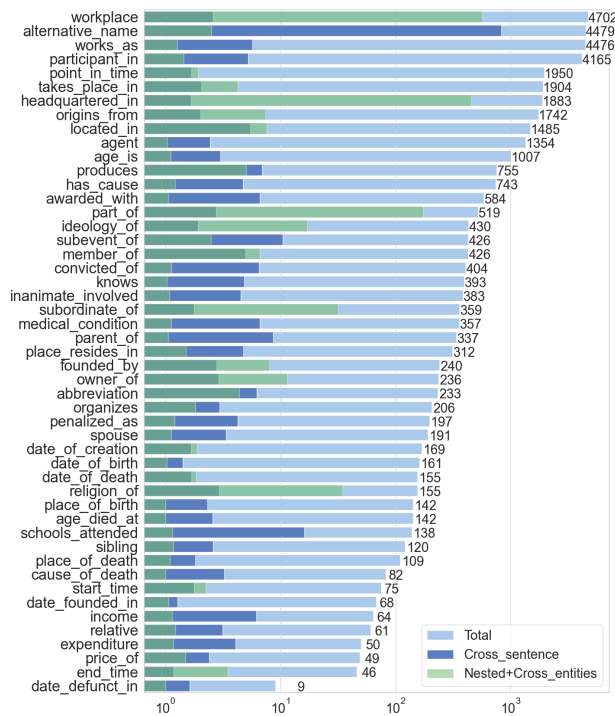

Figure 3: Relation type statistics (log scale). Proportions of cross-sentence relations and relations involving nestedness of entities are shown.

All the annotated relations can be subdivided into cross-sentence relations (24%) and within sentence relations (76%). Annotated cross-sentence relations make it possible to generate document-level relation extraction, which is important for knowledge graph population from texts.

Among relations within a single sentence, we distinguish three types of relations:

- traditional relations between entities, which are located separately, as *Mayor of Moscow* and *Sergei Sobyanin* in Figure 1 – further, external relations (52.38%);
- nested relations, i.e. relations within the longest span of a single nested entity, as *Moscow* within *Mayor of Moscow* in Figure 1 (14.75%);
- relations crossing entity boundaries, i.e. relations between an external named entity and an internal entity within a nested named entity (8.87%) – further cross-entity relations.

The cross-entity relations can be illustrated as follows: in the sentence *Barack Obama is a member of the Democratic party*, external entity *Barack Obama* has the IDEOLOGY_OF relation to entity *Democratic*, which is an internal IDEOLOGY entity in the longer entity *Democratic party*.

Table 2 presents the most frequent types of

Relation	Outer Entity	Inner Entity	Count	%
WORKPLACE	PROFESSION	ORGANIZATION	1,082	19.09
HEADQUARTERED_IN	ORGANIZATION	COUNTRY	846	14.93
WORKPLACE	PROFESSION	COUNTRY	669	11.81
HEADQUARTERED_IN	ORGANIZATION	CITY	333	5.88
PART_OF	ORGANIZATION	ORGANIZATION	281	4.96
HEADQUARTERED_IN	ORGANIZATION	STATE_OR_PROVINCE	125	2.21
SUBORDINATE_OF	PROFESSION	PROFESSION	116	2.05
WORKPLACE	PROFESSION	STATE_OR_PROVINCE	116	2.05
PART_OF	LAW	LAW	111	1.96
IDEOLOGY_OF	ORGANIZATION	IDEOLOGY	100	1.76
ORIGINS_FROM	LAW	COUNTRY	100	1.76

Table 2: The most frequent relation types within nested entities.

nested named entities connected with nested relations. It can be seen that most nested named entities are professions and professional titles, as well as organization names.

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

- if an entity contains an internal entity of the same type (e.g. *President of Russia – President*), all the relations are established with the longer entity. The internal entity helps entity linking if the longer entity is not present in a knowledge base;
- all variants of entity names in a single sentence or neighbour sentences are connected with ALTERNATIVE_NAME or ABBREVIATION relations, other relations are linked to the closest entity mentions among entities' variants;
- cross-sentence relations in neighbouring sentences are annotated with the same detail as in a single sentence;
- relations connecting entities from sentences that are farther than two sentences from each other, should be annotated at least once.

4 Experiments

We exploit multiple deep learning models, which deliver state-of-the-art results for the English data for two tasks, available at NEREL: (i) nested named entity recognition (NER), (ii) relation extraction. To this end, we subdivided NEREL into train, dev, and test sets – 746/94/93 documents, respectively.

4.1 Nested NER

We adopted two publicly available models: Bi-affine (Yu et al., 2020) and Pyramid (Jue et al., 2020) models with default parameters. Addition-

ally 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 answer spans to specialised questions. In our case, the questions are dictionary definitions of the words, corresponding to entity types carefully selected from multiple dictionaries. Word representations used with all models are fastText (fT) embeddings (Mikolov et al., 2018) and pre-trained RuBERT-cased (Kuratov and Arkhipov, 2019). The latter is utilized in the MRC approach, too.

Table 3 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 Bi-affine model is superior to the Pyramid model; (iii) 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 treats a single sentence at a time and is than resource-greedy, the second best solution is still worth consideration.

4.2 Relation Extraction

Recent relation extraction models (Joshi et al., 2020; Alt et al., 2019; Han et al., 2019) 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. Thus we follow the common relation extraction setup and utilize the models to extract relations between longest entity spans (i.e. external relations). To this we adopted three publicly available models: SpanBERT (Joshi et al., 2020), TRE (Alt et al., 2019), and OpenNRE (Han et al., 2019) with default parameters. The TRE

model is build upon the GPT model (Radford et al., 2018). Although initially GPT is trained on English web texts, it still has some limited knowledge of Russian, as Russian tokens are present in its vocabulary. The encoders used with SpanBERT and OpenNRE are multilingual BERT and RuBERT.

Nested relation extraction. We designed a new model, IntModel, aimed at extraction of nested relations within the longest named entity span. As such relations are contained inside a single entity, the whole sentence context can be omitted. To this end, the IntModel classifier inputs the entity features only. IntModel consists of a fully-connected layer with the softmax activation, which inputs fastText embeddings of both entities, trainable embeddings of corresponding entity types, and a binary feature showing whether the two entities are nested.

Table 4 presents the results of relation extraction on the NEREL dataset, grouped with respect to three relation types. The results show that (i) overall, in-sentence relations are much easier to extract than the document-level ones; (ii) the monolingual RuBERT provides better results, when compared to the multilingual version and quasi English GPT; (iii) the OpenNRE model is superior to SpanBERT and copes with all three types of relations, (iv) the simplistic IntModel performs on par with more sophisticated models.

4.3 Discussion

Although the preliminary experiments provide with promising results, there is still some room for improvement. Achieved results are comparable to those, published for English datasets, confirming high quality of the collected dataset. At the same time NEREL annotation schema causes difficulties for the current models: all-together nested named entities, combined with diverse relations, require less straightforward approaches, of which machine reading comprehension is one of the promising directions. Detailed error analysis will help to reveals models' weaknesses and drawbacks.

5 Conclusion

We presented a new Russian dataset NEREL annotated with both nested named entity and relations, which is significantly larger than existing Russian datasets. NEREL dataset has several significant distinctive features, including nested named entities, relations over nested named entities, relations on both sentence and discourse level, and events

Method	P	R	F1
Biaffine, fT	78.84	71.80	75.13
Biaffine, RuBERT	**81.92**	71.54	76.38
Pyramid, fT	72.70	63.01	67.51
Pyramid, RuBERT	77.73	70.97	74.19
MRC	78.70	**80.24**	**79.64**

Table 3: Results of nested NER for NEREL

Rel.	Method	P	R	F1
	In-sentence relations			
External	OpenNRE, mBERT	81.7	81.6	81.7
External	OpenNRE, RuBERT	**85.3**	**84.6**	**84.9**
External	SpanBERT, mBERT	76.8	75.4	76.1
External	SpanBERT, RuBERT	77.4	78.6	78.0
External	TRE	66.4	68.1	67.2
Nested	OpenNRE, mBERT	74.3	77.7	76.0
Nested	OpenNRE, RuBERT	**77.8**	**79.6**	**78.7**
Nested	IntModel	76.3	72.4	74.3
	Document-level relations			
Doc	OpenNRE, mBERT	35.7	51.2	42.1
Doc	OpenNRE, RuBERT	**52.1**	**51.3**	**51.7**

Table 4: Results of relation extraction for NEREL

involving named entities.

NEREL can facilitate development of novel models that address extraction of relations between nested named entities and cross-sentence relation extraction from short texts. NEREL annotation also allows relation extraction experiments on both sentence-level and document-level. Nevertheless, NEREL annotations utilize conventional entity and relations types, enabling cross-lingual transfer experiments.

Our experiments with baseline models for extraction of entities and relations show that there is room for improvement in both. In the nearest future we plan to enrich the dataset by linking the annotated named entities to Wikidata items.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ace, ere, tac-kbp, and framenet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53.

Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. TACRED revisited: A thorough evaluation of the TACRED relation extraction task. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019. Improving relation extraction by pretrained language representations. *arXiv preprint arXiv:1906.03088*.

Darina Benikova, Chris Biemann, and Marc Reznicek. 2014. NoSta-D named entity annotation for German: Guidelines and dataset. In *LREC*, pages 2524–2531.

Ann Bies, Zhiyi Song, Jeremy Getman, Joe Ellis, Justin Mott, Stephanie Strassel, Martha Palmer, Teruko Mitamura, Marjorie Freedman, Heng Ji, et al. 2016. A comparison of event representations in deft. In *Proceedings of the Fourth Workshop on Events*, pages 27–36.

Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. 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)*, pages 363–370.

Rinat Gareev, Maksim Tkachenko, Valery Solovyev, Andrey Simanovsky, and Vladimir Ivanov. 2013. Introducing baselines for russian named entity recognition. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 329–342.

Denis Gordeev, Adis Davletov, Alexey Rey, Galiya Akzhigitova, and Georgiy Geymbukh. 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"]*.

Jiale Han, Bo Cheng, and Xu Wang. 2020. Open domain question answering based on text enhanced knowledge graph with hyperedge infusion. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1475–1481, Online. Association for Computational Linguistics.

Xu Han, Tianyu Gao, Yuan Yao, Demin Ye, Zhiyuan Liu, and Maosong Sun. 2019. OpenNRE: An open and extensible toolkit for neural relation extraction. *EMNLP-IJCNLP 2019*, page 169.

Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, page 57–60, USA. Association for Computational Linguistics.

Luyang Huang, Lingfei Wu, and Lu Wang. 2020. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. *arXiv preprint arXiv:2005.01159*.

Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Wang Jue, Lidan Shou, Ke Chen, and Gang Chen. 2020. Pyramid: A layered model for nested named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5918–5928.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. *arXiv preprint arXiv:1905.07213*.

Xiaoya Li, Jingrong Feng, Yuxian Meng, Qinghong Han, Fei Wu, and Jiwei Li. 2020. A unified MRC framework for named entity recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5849–5859.

Linguistic Data Consortium. 2014. DEFT ERE annotation guidelines: Entities v1.7.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-BERT: enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2901–2908.

Paul McNamee, Hoa Trang Dang, Heather Simpson, Patrick Schone, and Stephanie M Strassel. 2010. An evaluation of technologies for knowledge base population. In *LREC*.

Tomáš Mikolov, Édouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Teruko Mitamura, Yukari Yamakawa, Susan Holm, Zhiyi Song, Ann Bies, Seth Kulick, and Stephanie Strassel. 2015. Event nugget annotation: Processes and issues. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 66–76.

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

Jakub Piskorski, Laska Laskova, Michał Marcińczuk, Lidia Pivovarova, Pavel Přibáň, Josef Steinberger, and Roman Yangarber. 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*, pages 63–74.

Barbara Plank, Kristian Nørgaard Jensen, and Rob van der Goot. 2020. Dan+: Danish nested named entities and lexical normalization. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6649–6662.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Nicky Ringland, Xiang Dai, Ben Hachey, Sarvnaz Karimi, Cecile Paris, and James R Curran. 2019. NNE: A dataset for nested named entity recognition in english newswire. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5176–5181.

Marco Rospocher, Marieke van Erp, Piek Vossen, Antske Fokkens, Itziar Aldabe, German Rigau, Aitor Soroa, Thomas Ploeger, and Tessel Bogaard. 2016. Building event-centric knowledge graphs from news. *Journal of Web Semantics*, 37:132–151.

Teemu Ruokolainen, Pekka Kauppinen, Miikka Silfverberg, and Krister Lindén. 2019. A finnish news corpus for named entity recognition. *Language Resources and Evaluation*, pages 1–26.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 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*, pages 89–98.

Anatoly Starostin, Victor Bocharov, Svetlana Alexeeva, Anastasia Bodrova, Alexander Chuchunkov, Stanislav Dzhumaev, Irina Efimenko, Dmitry Granovsky, Vladimir Khoroshevsky, Irina Krylova, Marina Nikolaeva, Ivan Smurov, and Svetlana Toldova. 2016. Factrueval 2016: Evaluation of named entity recognition and fact extraction systems for russian. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, pages 702–720.

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

Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. 2006. Ace 2005 multilingual training corpus. *Linguistic Data Consortium, Philadelphia*, 57:45.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777.

Juntao Yu, Bernd Bohnet, and Massimo Poesio. 2020. Named entity recognition as dependency parsing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6470–6476.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 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)*, pages 35–45.