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

Natalia Loukachevitch¹, Ekaterina Artemova²,³, Tatiana Batura¹,⁴, Pavel Braslavski²,⁵, Ilia Denisov¹, Vladimir Ivanov⁶, Suresh Manandhar⁹, Alexander Pugachev², and Elena Tutubalina²,⁷,⁸

¹Lomonosov Moscow State University, Russia  
²HSE University, Russia  
³Huawei Noah’s Ark lab, Russia  
⁴Novosibirsk State University, Russia  
⁵Ural Federal University, Russia  
⁶Innopolis University, Russia  
⁷Kazan Federal University, Russia  
⁸Sber AI, Russia  
⁹Wiseyak, 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 (Named Entities and RELations), 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
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 “österrechischen” (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-
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. 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 its 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 et al. [2018]).
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.

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. 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.

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
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. Additionally, 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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biaffine, fT</td>
<td>78.84</td>
<td>71.80</td>
<td>75.13</td>
</tr>
<tr>
<td>Biaffine, RuBERT</td>
<td><strong>81.92</strong></td>
<td>71.54</td>
<td>76.38</td>
</tr>
<tr>
<td>Pyramid, fT</td>
<td>72.70</td>
<td>63.01</td>
<td>67.51</td>
</tr>
<tr>
<td>Pyramid, RuBERT</td>
<td>77.73</td>
<td>70.97</td>
<td>74.19</td>
</tr>
<tr>
<td>MRC</td>
<td>78.70</td>
<td><strong>80.24</strong></td>
<td><strong>79.64</strong></td>
</tr>
</tbody>
</table>

Table 3: Results of nested NER for NEREL

<table>
<thead>
<tr>
<th>Rel. Method</th>
<th>P</th>
<th>R</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>In-sentence relations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>External</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OpenNRE, mBERT</td>
<td>81.7</td>
<td>81.6</td>
<td>81.7</td>
</tr>
<tr>
<td>OpenNRE, RuBERT</td>
<td><strong>85.3</strong></td>
<td><strong>84.6</strong></td>
<td><strong>84.9</strong></td>
</tr>
<tr>
<td>SpanBERT, mBERT</td>
<td>76.8</td>
<td>75.4</td>
<td>76.1</td>
</tr>
<tr>
<td>SpanBERT, RuBERT</td>
<td>77.4</td>
<td>78.6</td>
<td>78.0</td>
</tr>
<tr>
<td>TRE</td>
<td>66.4</td>
<td>68.1</td>
<td>67.2</td>
</tr>
<tr>
<td>Nested</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OpenNRE, mBERT</td>
<td>74.3</td>
<td>77.7</td>
<td>76.0</td>
</tr>
<tr>
<td>OpenNRE, RuBERT</td>
<td><strong>77.8</strong></td>
<td><strong>79.6</strong></td>
<td><strong>78.7</strong></td>
</tr>
<tr>
<td>IntModel</td>
<td>76.3</td>
<td>72.4</td>
<td>74.3</td>
</tr>
</tbody>
</table>

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.

Acknowledgments

The project is supported by the Russian Science Foundation, grant # 20-11-20166. The experiments were partially carried out on computational resources of HPC facilities at HSE University. We are grateful to Alexey Yandutov and Igor Rozhkov for providing results of their experiments in named entity recognition and relation extraction.
References


